ENST Bretagne

Groupe des écoles
des télécommunications

Applied Stochastic Models and Data Analysis

Brest, France
May, 17–20, 2005

ISBN 2-908849-15-1

Editors: Jacques Janssen and Philippe Lenca

www.enst-bretagne.fr

ENST Bretagne
Technopôle Brest-Iroise – CS 83818
29238 Brest Cedex 3 – France

Dépôt Légal Mai 2005 - GLAT ISBN : 2-908849-15-1

ASMDA 2005 Conference Committees

Chairmen of ASMDA 2005

J.P. Barthélemy, ENST Bretagne, France
N. Limnios, Université de Technologie de Compiègne, France
G. Saporta, CNAM, Paris, France

Program Committee

Chairmen

J. Janssen, EURIA (Brest) and CESIAF (Charleroi), Belgium
P. Lenca, ENST Bretagne, France

Members

N. Balakrishnan, McMaster University, Canada
M. Bardos, Banque de France, France
E. Cinlar, Princeton University, Princeton, USA
J. Glaz, University of Connecticut, Stors, USA
A. Guénoche, Université de Marseille, France
F. Guillet, Université de Nantes, France
G. Govaert, Université de Technologie de Compiègne, France
C. Huber, Université Paris V, France
D. Ho, IBM, Paris, France
S. Lallich, University Lyon 2, France
L. Lebart, CNRS, ENST, Paris, France
N. Limnios, Université de Technologie de Compiègne, France
R. Manca, University La Sapiensa, Roma, Italy
M. Mesbah, Université Paris VI, France
M. Nikulin, Université Bordeaux II, France
J. Teghem, Faculté Polytechnique de Mons, Belgium
M.L. Ting Lee, Harvard University, USA
S.M. Ross, University of Southern California, USA
G. Saporta, CNAM, Paris, France
C. Skiadas, Technical University of Creta, Chania, Greece
L. Simar, Institute of Statistics, Louvain-la-Neuve, Belgium
M. J. Valderrama, Universidad de Granada, Spain
D. Zighed, Université Lyon 2, France

Additionnal reviewers

A. M. Aguilera
J. Azé
V. Barbu
P. Bertrand
G. Biau
P. Y. Boëlle
P. Boët
J. P. Brans
F. Brucker
H. Cancela
P. Cardaliaguet
L. Carvalho
T. Chonavel
G. Coppin
T. Dhorne
T. N. Do
V. Girardin
I. Kojadinovic
V. S. Korolyuk
B. Leclerc
S. Moga
F. A. Ocaña
B. Ouhbi
D. Pastor
R. Pérez-Ocón
F. Poulet
M. Scarsini
B. Vaillant
S. Vaton
J. F. Zagury

Organisation Committee

Chairman

P. Bertrand, ENST Bretagne, France

Members

F. Brucker, ENST Bretagne, France
T. Chonavel, ENST Bretagne, France
G. Coppin, ENST Bretagne, France
G. Le Gall, ENST Bretagne, France
N. L'Hopital, ENST Bretagne, France
S. Moga, ENST Bretagne, France
D. Pastor, ENST Bretagne, France
P. Tanguy, ENST Bretagne, France
B. Vaillant, ENST Bretagne, France

Special thanks

Last, we would like to thank Benoît Vaillant for supervising the editing of these proceedings, Alexandre Sztykgold, Safaa Jarifi, François Legras, Frédéric Cadier, Ghislain Putois and Thomas Lebras for their precious help in this difficult task.

Sponsors

The organizers want to thank the following institutions for their support:

École Nationale des Télécommunications de Bretagne - Brest
Groupe des Écoles des Télécommunications
Laboratoire TAMCIC
International Association for Statistical Computing
Société Francophone de Classification
Association Extraction et Gestion des Connaissances
Euro-Institut d'Actuariat Jean Dieudonné
Laboratoire de Mathématiques de Brest, UMR 6205
Société de Mathématiques Appliquées & Industrielles
Laboratoire de Mathématiques de Brest
Société Française de Statistiques
Société Française de Recherche Opérationnelle et d'Aide à la Décision
SAS
IBM
ISTHMA
SPAD
Cabinet Moeglin
Région Bretagne
Brest Métropole Océane
Conseil Général du Finistère
Ville de Brest

Preface

ASMDA meetings are devoted to serve as the interface between Stochastic Modelling and Data Analysis and their applications particularly in Economy, Business, Finance and Insurance, Management, Production and Telecommunications, Biology, ... To be successful, it is quite necessary that the scientific community approving our objective has regular meetings to measure the involved methods and techniques and also the results in solving real life problems.

This XIth International Symposium on Applied Stochastic Models and Data Analysis (Brest, France, May 17–20), also called now ASMDA 2005, continues our cycle of meetings beginning in 1981 and unfortunately interrupted for 4 years:

Belgium	(Brussels)	1981, 1983, 1985.
France	(Nancy)	1988.
Spain	(Grenada)	1991.
Greece	(Chania, Creta)	1993.
Ireland	(Dublin)	1995.
Italy	(Anacapri, Napoli)	1997.
Portugal	(Lisbon)	1999.
France	(Compiègne)	2001.

This ASMDA 2005 meeting is also the first meeting chaired by N. Limnios and G. Saporta as new co-chairmen of ASMDA, and this time with J-P. Barthélemy as local chairman.

As usual, the papers presented in this Symposium cover a large variety of fields both theoretical and applied. These ASMDA 2005 proceedings include three kinds of papers or abstracts: Keynote speaker papers, Invited session papers and Contributed papers. In these proceedings, the papers are sorted by topic disscussed.

Let us also mention that all the papers submitted were reviewed and we can say that the presented papers are of quality.

We thank all the authors of the papers for their collaboration and also all the members –and non members!– of the Scientific Committee who reviewed all the submitted papers.

We also thank warmly the members of the local Organisation Committee taking in charge all the practical organisation of the ASMDA 2005 meeting.

Lastly, we also thank the different sponsors for making this Symposium in good conditions for the participants.

Brest, May 2005.

Jacques Janssen,
Philippe Lenca,
Chairmen of the ASMDA 2005 Scientific Committee.

Contents

Part I. Keynote Speakers

Part II. Analysis of Textual Data

Part III. Bioinformatics and Statistics

Part IV. Knowledge Management and Data Mining

Part V. Clustering

Part VI. Discriminant analysis and learning

Part VII. Data Analysis

Part VIII. Mathematical statistics

Part IX. Finance and Insurance

Part X. Health

Part XI. Markov Processes

Part XII. Queues and transportation

Part XIII. Reliability and Survival Analysis

Part XIV. Spatial Processes

Part XV. Time Series

Part XVI. Fuzzy Approach

Part XVII. Internet Modelling

Part XVIII. Posters

Part I

Keynote Speakers

Weighted Cramér-von Mises-type Statistics

Paul Deheuvels

LSTA, Université Paris VI
175 rue du Chevaleret
F 75013 Paris, France
(e-mail: pd@ccr.jussieu.fr)

Abstract. We consider quadratic functionals of the multivariate uniform empirical process. Making use of Karhunen-Loève expansions of the corresponding limiting Gaussian processes, we obtain the asymptotic distributions of these statistics under the assumption of independent marginals. Our results have direct applications to tests of goodness of fit and tests of independence by Cramér-von Mises-type statistics.

AMS 2000 classification: 60F05, 60F15, 60G15, 62G30.
Keywords: Cramér-von Mises tests, tests of goodness of fit, tests of independence, weak laws, empirical processes, Karhunen-Lo‘eve decompositions, Gaussian processes, Bessel functions.

1 Introduction and Premiminaries.

1.1 Introduction.

In this paper, we survey some recent results ([14, 15, 13]) related to quadratic functionals of the form

$$\int_0^1 \cdots \int_0^1 t_1^{2\beta_1} \ldots t_d^{2\beta_d} \alpha_{n,0}^2(t_1, \ldots, t_d) dt_1 \ldots dt_d, \tag{1}$$

where $\alpha_{n,0}$ is an appropriate version of the uniform empirical process on $[0, 1]^d$ (see (36) in the sequel for explicit definitions). We first establish conditions on the $\beta_1, \ldots, \beta_d$, under which the statistic in (1) converges to a quadratic functional of a Gaussian process, of the form

$$\int_0^1 \cdots \int_0^1 t_1^{2\beta_1} \ldots t_d^{2\beta_d} \mathbf{B}_0^2(t_1, \ldots, t_d) dt_1 \ldots dt_d, \tag{2}$$

with $\mathbf{B}_0$ denoting a tied-down Brownian bridge. Second, we will characterize the distribution of the random variable in (2), through a Karhunen-Loève expansion of the corresponding weighted Gaussian process.

This problem has been initiated by Cramér [10] (see, e.g., Nikitin [26], Scott [32] and the references therein). In higher dimensions, we refer to Blum, Kiefer and Rosenblatt [6], Cotterill and Csörgő [8, 9], Deheuvels [13], Dugué [17, 18, 19], Hoeffding [20], Kiefer [24], Martynov [27], and Smirnov [33, 35, 34]. Quadratic functionals of Gaussian processes have been studied by Biane and Yor [5], Donati-Martin and Yor [16], Pitman and Yor [28, 29, 30, 31], and Yor [37, 38]. The results of Deheuvels and Martynov [14], and Deheuvels, Peccati and Yor [15], Deheuvels [13], give the core of the present survey paper. The theory of Bessel functions plays here an essential role and we refer to Bowman [7] and Watson [36] for details.

In §1.2 and 1.3, we give some preliminaries. We describe the univariate case in §2.1 and the multivariate case, with $d \geq 2$, in §2.2.

1.2 Some Preliminaries on Gaussian Processes.

Let $\{X(\mathbf{t}) : \mathbf{t} \in [0,1]^d\}$ denote a centered Gaussian process, with $d \geq 1$. We set $\mathbf{s} = (s_1, \dots, s_d) \in \mathbb{R}^d$ and $\mathbf{t} = (t_1, \dots, t_d) \in \mathbb{R}^d$, and set

$$R(\mathbf{s},\mathbf{t}) = \mathbb{E}\big(X(\mathbf{s})X(\mathbf{t})\big) \quad \text{for} \quad \mathbf{s},\mathbf{t} \in [0,1]^d. \tag{3}$$

We will are concerned with the *quadratic functional*

$$\int_{[0,1]^d} X^2(\mathbf{t})d\mathbf{t}, \tag{4}$$

where $d\mathbf{t}$ is the Lebesgue measure. We will work under the assumption that

$$0 < \mathbb{E}\Big(\int_{[0,1]^d} X^2(\mathbf{t})d\mathbf{t}\Big) = \int_{[0,1]^d} R(\mathbf{t},\mathbf{t})d\mathbf{t} < \infty. \tag{5}$$

The condition (5) entails that, almost surely, $X(\cdot) \in L^2\big([0,1]\big)$ belongs to the class of *Hilbert space valued centered Gaussian processes* (see, e.g., §10 in Lifshits [25]). By the Cauchy-Schwarz inequality, for each $\mathbf{s},\mathbf{t} \in [0,1]^d$,

$$R(\mathbf{s},\mathbf{t})^2 = \mathbb{E}\big(X(\mathbf{s})X(\mathbf{t})\big)^2 \leq \mathbb{E}\big(X(\mathbf{s})^2\big)\mathbb{E}\big(X(\mathbf{t})^2\big) = R(\mathbf{s},\mathbf{s})R(\mathbf{t},\mathbf{t}).$$

When combining this last inequality with (5), we obtain that

$$\|R\|_{L^2}^2 := \int_{[0,1]^d}\int_{[0,1]^d} R(\mathbf{s},\mathbf{t})^2 d\mathbf{s}d\mathbf{t} \leq \Big\{\int_{[0,1]^d} R(\mathbf{t},\mathbf{t})d\mathbf{t}\Big\}^2 < \infty, \tag{6}$$

so that $R \in L^2\big([0,1]^d \times [0,1]^d\big)$. Under (6), the Fredholm transformation $y(\cdot) \in L^2\big([0,1]^d\big) \rightarrow \widetilde{y}(\cdot)$, defined by

$$\widetilde{y}(\mathbf{t}) = \int_{[0,1]^d} R(\mathbf{s},\mathbf{t})y(\mathbf{s})d\mathbf{s} \quad \text{for} \quad \mathbf{t} \in [0,1]^d, \tag{7}$$

is a continuous linear mapping of $L^2\big([0,1]^d\big)$ onto itself. The condition (6) also implies the existence of a *convergent orthonormal sequence* [c.o.n.s.],

$\{\lambda_k, e_k(\cdot) : 1 \leq k < K\}$ with the following properties. $\{\lambda_k : 1 \leq k < K\}$ are positive constants and $K \in \{2, \ldots, \infty\}$ a possibly infinite index, with

$$\lambda_1 \geq \ldots \geq \lambda_k \geq \ldots > 0. \tag{8}$$

The $\{e_k(\cdot) : 1 \leq k < K\}$ are orthonormal in $L^2\big([0,1]\big)$, and fulfill

$$\int_{[0,1]^d} e_k(\mathbf{t}) e_\ell(\mathbf{t}) d\mathbf{t} = \begin{cases} 1 & \text{if} \quad k = \ell, \\ 0 & \text{if} \quad k \neq \ell. \end{cases}$$

The function R may be decomposed into the series

$$R(\mathbf{s}, \mathbf{t}) = \sum_{1 \leq k < K} \lambda_k e_k(\mathbf{s}) e_k(\mathbf{t}), \tag{9}$$

convergent in $L^2\big([0,1]^d\big)$. This entails that

$$\|R\|_{L^2} = \int_{[0,1]^d} \int_{[0,1]^d} R(\mathbf{s}, \mathbf{t})^2 d\mathbf{s} d\mathbf{t} = \sum_{1 \leq k < K} \lambda_k^2 < \infty. \tag{10}$$

The λ_k (resp. $e_k(\cdot)$) are the eigenvalues (resp. eigenfunctions) of the Fredholm operator (7), and fulfill the relations, for each $1 \leq k < K$,

$$\widetilde{e}_k(\mathbf{t}) = \int_{[0,1]^d} R(\mathbf{s}, \mathbf{t}) e_k(\mathbf{s}) d\mathbf{s} = \lambda_k e_k(\mathbf{t}). \tag{11}$$

The *Karhunen-Loève* [KL] decomposition of $X(\cdot)$, (see, e.g., Kac and Siegert [23, 22], Kac [21], Ash and Gardner [4], and Adler [2]) decomposes $X(\cdot)$ into

$$X(\mathbf{t}) = \sum_{1 \leq k < K} Y_k \sqrt{\lambda_k}\, e_k(\mathbf{t}), \tag{12}$$

where $\{Y_k : 1 \leq k < K\}$ are independent and identically distributed [i.i.d.] normal $N(0,1)$ random variables. Under (5), the series in (12) is convergent in mean square, since this condition is equivalent to

$$0 < \mathbb{E}\Big(\int_{[0,1]^d} X^2(\mathbf{t}) d\mathbf{t}\Big) = \sum_{1 \leq k < K} \lambda_k < \infty. \tag{13}$$

This, in turn, readily implies that, as $k \uparrow K$ with $k < K$,

$$\mathbb{E}\Big(\int_{[0,1]^d} \Big\{X(\mathbf{t}) - \sum_{m=1}^{k} Y_m \sqrt{\lambda_m}\, e_m(\mathbf{t})\Big\}^2 d\mathbf{t}\Big) = \sum_{m > k} \lambda_k \to 0.$$

The condition (5)–(13) is strictly stronger than (10). It implies that the quadratic functional (4) can be decomposed into the sum of the series

$$\int_{[0,1]^d} X^2(\mathbf{t}) d\mathbf{t} = \sum_{1 \leq k < K} \lambda_k Y_k^2. \tag{14}$$

The latter is almost surely convergent *if and only if* (5) holds. Therefore, we will assume, from now on, that this condition is satisfied.

1.3 A General Convergence Theorem.

With $R(\cdot,\cdot)$ as in (3), we consider independent replicæ $\xi_1(\cdot), \xi(2), \ldots$ of a general stochastic process $\xi(\cdot)$, fulfilling $(H.1–2–3)$ below.

$(H.1)$ $\xi(\cdot) \in L^2\big([0,1]^d\big)$;

$(H.2)$ $\mathbb{E}\big(\xi(\mathbf{t})\big) = 0$ for all $\mathbf{t} \in [0,1]^d$;

$(H.3)$ $\mathbb{E}\big(\xi(\mathbf{s})\xi(\mathbf{t})\big) = R(\mathbf{s},\mathbf{t})$ for all $\mathbf{s},\mathbf{t} \in [0,1]^d$.

Under $(H.1–2–3)$ (see, e.g., Ex. 14, p. 205 in Araujo and Giné [3]), as $n \to \infty$, the convergence in distribution

$$\zeta_n(\cdot) := n^{-1/2} \sum_{i=1}^{n} \xi_i(\cdot) \ \xrightarrow{d}\ X(\cdot), \tag{15}$$

holds *if and only if* (5)–(13)) is satisfied, namely, when

$$\int_{[0,1]^d} \mathbb{E}\big(\xi^2(\mathbf{t})\big) d\mathbf{t} = \int_{[0,1]^d} R(\mathbf{t},\mathbf{t}) d\mathbf{t} < \infty.$$

We have therefore the following theorem.

Theorem 1 *Under* (5) *and* $(H.1–2–3)$*, we have, as* $n \to \infty$*, the convergence in distribution*

$$\int_{[0,1]^d} \zeta_n^2(\boldsymbol{t}) d\boldsymbol{t} \ \xrightarrow{d}\ \sum_{1 \leq k < K} \lambda_k Y_k^2. \tag{16}$$

Proof. Under (5) (or equivalently (13)), it follows from (15) that

$$\int_{[0,1]^d} \zeta_n^2(\mathbf{t}) d\mathbf{t} \ \xrightarrow{d}\ \int_{[0,1]^d} X^2(\mathbf{t}) d\mathbf{t},$$

which, in turn, reduces (16) to a direct consequence of (15).□

Below, we provide some useful statistical applications of Theorem 1.

2 Weighted Empirical Processes.

2.1 The Univariate Case ($d = 1$).

Let $U_1, U_2, \ldots$ be i.i.d. uniform $[0,1]$ random variables. For $n \geq 1$, set

$$F_n(t) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}_{\{U_n \leq t\}}, \tag{17}$$

for the empirical distribution function [df] based upon $U_1, \ldots, U_n$, and let

$$\alpha_n(t) = n^{1/2}\big\{F_n(t) - t\big\} \quad \text{for} \quad t \in [0,1], \tag{18}$$

denote the *uniform empirical process.* Fix $\beta \in \mathbb{R}$, and set, for $n \geq 1$,

$$\xi_n(t) = t^{\beta}\big\{\mathbb{1}_{\{U_n \leq t\}} - t\big\} \quad \text{for} \quad t \in [0,1]. \tag{19}$$

We let $t^0 = 1$ for all $t \in \mathbb{R}$, when $\beta = 0$. In agreement with (15), (18), (19), and the notation of §1.3, we may write

$$\zeta_n(t) = n^{-1/2} \sum_{i=1}^{n} \xi_i(t) = t^{\beta}\alpha_n(t) \quad \text{for} \quad t \in [0,1]. \tag{20}$$

The assumptions (H.1–2–3) in §1.3 are fulfilled with R defined by

$$R(s,t) = s^{\beta}t^{\beta}\big\{s \wedge t - st\big\} \quad \text{for} \quad s,t \in [0,1]. \tag{21}$$

For this choice of R, (5)–(13) hold if and only if

$$\int_0^1 t^{2\beta}\big\{t(1-t)\big\}dt < \infty, \tag{22}$$

which is equivalent to $\beta > -1$. Now, since $s \wedge t - st$ is the covariance function of a standard Brownian bridge $\{B(t) : t \in [0,1]\}$, the kernel R in (21) is nothing else but the covariance function of the weighted Brownian bridge

$$X(t) = t^{\beta}B(t) \quad \text{for} \quad t \in (0,1]. \tag{23}$$

Deheuvels and Martynov [14] have given the KL decomposition of $X(\cdot)$ in (23) when $\beta \neq -1 \Leftrightarrow \nu = 1/(2(1+\beta)) > 0$. For $\nu \in \mathbb{R}$, we the first Bessel function (see, e.g.,§9.1.69 in Abaramowitz and Stegun [1]) is

$$J_\nu(x) = (\tfrac{1}{2}x)^{\nu} \sum_{k=0}^{\infty} \frac{(-\frac{1}{4}x^2)^k}{\Gamma(\nu+k+1)\Gamma(k+1)}\ . \tag{24}$$

Whenever $\nu > -1$, the positive zeros of J_ν are isolated and form an infinite increasing sequence $\{z_{\nu,k} : k \geq 1\}$, such that (see, e.g., Watson [36])

$$0 < z_{\nu,1} < z_{\nu,2} < \ldots, \tag{25}$$

and, as $k \to \infty$,

$$z_{\nu,k} = \big\{k + \tfrac{1}{2}(\nu - \tfrac{1}{2})\big\} + o(1). \tag{26}$$

Given this notation, Theorem 1.4 in [14] asserts that, whnever $\beta > -1$, the KL representation of $X(t) = t^{\beta}B(t)$ is given by

$$X(\mathbf{t}) = t^{\beta}B(t) = \sum_{k=1}^{\infty} Y_k \sqrt{\lambda_k}\ e_k(\mathbf{t}), \tag{27}$$

where $\{Y_k : k \geq 1\}$ are i.i.d. normal $N(0,1)$ random variables,

$$\lambda_k = \Big\{\frac{2\nu}{z_{\nu,k}}\Big\}^2, \quad k = 1, 2, \ldots, \tag{28}$$

and

$$e_k(t) = t^{\frac{1}{2\nu}-\frac{1}{2}}\Big\{\frac{J_\nu(z_{\nu,k}t^{\frac{1}{2\nu}})}{\sqrt{\nu}\ J_{\nu-1}(z_{\nu,k})}\Big\} \quad \text{for} \quad 0 < t \leq 1. \tag{29}$$

Refer to Deheuvels and Martynov [14] for details. We get the theorem:

Theorem 2 *For any $\beta > -1$, setting $\nu = 1/(2(1+\beta))$, we have, as $n \to \infty$, the convergence in distribution*

$$\int_0^1 t^{2\beta}\alpha_n^2(t)dt \ \xrightarrow{d}\ \int_0^1 t^{2\beta}B^2(t)dt = \sum_{k=1}^{\infty}\Big\{\frac{2\nu}{z_{\nu,k}}\Big\}^2 Y_k^2, \tag{30}$$

where $\{Y_k : k \geq 1\}$ is an i.i.d. sequence of normal $N(0,1)$ random variables.

Proof. In view of (28)–(29), it is a direct consequence of Theorem 1.□

2.2 The Multivariate Case ($d \geq 2$).

We now let $d \geq 2$. When $\mathbf{s} = (s_1, \ldots, s_d) \in \mathbb{R}^d$ and $\mathbf{t} = (t_1, \ldots, t_d) \in \mathbb{R}^d$, we denote by $\mathbf{s} \leq \mathbf{t}$ the fact that $s_j \leq t_j$ for $j = 1, \ldots, k$, and set, accordingly,

$$\mathbf{s} \wedge \mathbf{t} = \big(s_1 \wedge t_1, \ldots, s_d \wedge t_d\big).$$

Letting $\mathbf{U} = (U(1), \ldots, U(d)) \in [0,1]^d$ be uniformly distributed on $[0,1]^d$, we let $\mathbf{U}_n = (U_n(1), \ldots, U_n(d)) \in [0,1]^d$, $n = 1, 2, \ldots$ be i.i.d. replicæ of $\mathbf{U}$. For each $n \geq 1$, the empirical df based upon $\mathbf{U}_1, \ldots, \mathbf{U}_n$ is denoted by

$$F_n(\mathbf{t}) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}_{\{\mathbf{U}_i \leq \mathbf{t}\}}, \tag{31}$$

We denote by

$$F(\mathbf{t}) = \mathbb{P}\big(\mathbf{U} \leq \mathbf{t}\big) = \prod_{j=1}^{d} t_j, \tag{32}$$

the (*exact*) distribution function of $\mathbf{U}$, and set

$$\alpha_n(\mathbf{t}) = n^{1/2}\big(F_n(\mathbf{t}) - F(\mathbf{t})\big) \quad \text{for} \quad \mathbf{t} \in [0,1]^d, \tag{33}$$

for the corresponding uniform empirical process. Making use of §1.3, we obtain that the following convergence in distribution holds. As $n \to \infty$,

$$\alpha_n(\cdot) \ \xrightarrow{d}\ \mathbf{B}(\cdot), \tag{34}$$

where $\{\mathbf{B}(\mathbf{t}) : \mathbf{t} \in [0,1]^d\}$ is a standard multivariate Brownian bridge. Namely, $\mathbf{B}(\cdot)$ is a centered Gaussian process, with covariance function

$$\begin{aligned}\mathbb{E}\big(\mathbf{B}(\mathbf{s})\mathbf{B}(\mathbf{t})\big) &= \mathbb{E}\big(\alpha_n(\mathbf{s})\alpha_n(\mathbf{t})\big)\\ &= \prod_{j=1}^{d}\big\{s_j \wedge t_j\big\} - \prod_{j=1}^{d}\big\{s_j t_j\big\}. \qquad (35)\end{aligned}$$

The KL decomposition of $\mathbf{B}(\cdot)$, with covariance function as in (35), is not known explicitly for $d \geq 2$. A more tractable *tied-down* empirical process $\alpha_{n,0}(\cdot)$ is as follows. Set

$$\begin{aligned}\alpha_{n,0}(\mathbf{t}) = \alpha_n(\mathbf{t}) &- \sum_{1\leq j\leq d} t_j\ \alpha_n(t_1,\ldots,t_{j-1},1,t_{j+1},\ldots,t_d)\\ &+ \sum_{1\leq j<\ell\leq d}^{d} t_j t_\ell\ \alpha_n(t_1,\ldots,t_{j-1},1,t_{j+1},\ldots,t_{\ell-1},1,t_{\ell+1},\ldots,t_d)\\ &+\ldots+(1)^d t_1\ldots t_d\ \alpha_n(1,\ldots,1). \qquad (36)\end{aligned}$$

In (36), $\alpha_n(1,\ldots,1)=0$, but this term is stated for convenience. In view of §1.3, we obtain the following convergence in distribution. As $n\to\infty$,

$$\alpha_{n,0}(\cdot)\ \overset{d}{\to}\ \mathbf{B}_0(\cdot), \qquad (37)$$

where $\{\mathbf{B}_0(\mathbf{t}) : \mathbf{t} \in [0,1]^d\}$ is a tied-down multivariate Brownian bridge. Namely, $\mathbf{B}_0(\cdot)$ is a centered Gaussian process, with covariance function

$$\mathbb{E}\big(\mathbf{B}_0(\mathbf{s})\mathbf{B}_0(\mathbf{t})\big) = \prod_{j=1}^{d}\big\{s_j\wedge t_j - s_j t_j\big\}. \qquad (38)$$

We have the following easy consequence of the results of Deheuvels and Martynov [14] (see also Deheuvels, Peccati and Yor [15]).

Theorem 3 *Let $\beta_1,\ldots,\beta_d$ be constants such that $\beta_j>-1$ for $j=1,\ldots,d$. Set $\nu_j = 1/(2(1+\beta_j))>0$ for $j=1,\ldots,d$. Then, the Karhunen-Loève decomposition of the centered Gaussian process*

$$X(\boldsymbol{t}) = t_1^{\beta_1}\ldots t_d^{\beta_d}\mathbf{B}_0(\boldsymbol{t}) \quad \textit{for} \quad \boldsymbol{t}\in(0,1]^d, \qquad (39)$$

is given by

$$X(\boldsymbol{t}) = \sum_{k_1=1}^{\infty}\cdots\sum_{k_d=1}^{\infty}\sqrt{\lambda_{k_1,\ldots,k_d}}\,Y_{k_1,\ldots,k_d}\,e_{k_1,\ldots,k_d}(\boldsymbol{t}), \qquad (40)$$

where

$$\lambda_{k_1,\dots,k_d} = \prod_{j=1}^{d} \Big\{\frac{2\nu_j}{z_{\nu_j,k_j}}\Big\}^2 =: \prod_{j=1}^{d} \mathcal{L}(\nu_j,k_j), \tag{41}$$

and

$$\begin{aligned} e_{k_1,\dots,k_d}(\boldsymbol{t}) &= \prod_{j=1}^{d} \Bigg[t_j^{\frac{1}{2\nu_j}-\frac{1}{2}} \Bigg\{ \frac{J_{\nu_j}(z_{\nu_j,k}\, t_j^{\frac{1}{2\nu_j}})}{\sqrt{\nu_j}\, J_{\nu_j-1}(z_{\nu_j,k})} \Bigg\} \Bigg] \\ &=: \prod_{j=1}^{d} \mathcal{E}(\nu_j,t_j). \end{aligned} \tag{42}$$

Proof. By (38) the covariance function of $X(\mathbf{t})$ in (39) is given by

$$R(\mathbf{s},\mathbf{t}) = \prod_{j=1}^{d} s_j^{\beta_j} t_j^{\beta_j} \{s_j \wedge t_j - s_j t_j\} =: \prod_{j=1}^{d} \mathcal{R}(s_j,t_j). \tag{43}$$

Therefore, via (28)–(29), $\lambda_{k_1,\dots,k_d}$ is an eigenvalue of the Fredholm operator (7) pertaining to $e_{k_1,\dots,k_d}(\cdot)$. To conclude, we show that *all* eigenvalues are so obtained. For this, we combine (10) with (43), to write that

$$\begin{aligned} &\int_{[0,1]^d}\int_{[0,1]^d} R(\mathbf{s},\mathbf{t})^2 d\mathbf{s} d\mathbf{t} = \prod_{j_1=1}^{\infty} \cdots \prod_{j_d=1}^{\infty} \int_0^1 \int_0^1 \mathcal{R}(s_j,t_j)^2 ds_j dt_j \\ &= \prod_{j_1=1}^{\infty} \cdots \prod_{j_d=1}^{\infty} \Big\{ \sum_{k_j=1}^{\infty} \mathcal{L}(\nu_j,k_j)^2 \Big\} = \sum_{k_1=1}^{\infty} \cdots \sum_{k_d=1}^{\infty} \lambda_{k_1,\dots,k_d}^2. \end{aligned}$$

This shows that there is no other remaining eigenvalue of (7).□

The next theorem is an easy consequence of the preceding results.

Theorem 4 *Let $\beta_1,\dots,\beta_d$ be constants such that $\beta_j > -1$ for $j = 1,\dots,d$. Set $\nu_j = 1/(2(1+\beta_j)) > 0$ for $j = 1,\dots,d$. Then, we have, as $n \to \infty$,*

$$\begin{aligned} &\int_{[0,1]^d} t_1^{2\beta_1} \dots t_d^{2\beta_d} \alpha_{n,0}^2(\boldsymbol{t}) d\boldsymbol{t} \overset{d}{\to} \int_{[0,1]^d} t_1^{2\beta_1} \dots t_d^{2\beta_d} \mathbf{B}_0^2(\boldsymbol{t}) d\boldsymbol{t} \\ &= \sum_{k_1=1}^{\infty} \cdots \sum_{k_d=1}^{\infty} \Big\{ \prod_{j=1}^{d} \Big\{ \frac{2\nu_j}{z_{\nu_j,k_j}} \Big\}^2 \Big\} Y_{k_1,\dots,k_d}^2, \end{aligned} \tag{44}$$

where $\{Y_{k_1,\dots,k_d} : k_1 \geq 1,\dots,k_d \geq 1\}$ is an i.i.d. array of normal $N(0,1)$ random variables.

The limiting distribution in Theorem 4 coincides with that of the Blum-Kiefer-Rosenblatt statistic (see, e.g., [6]), when $d = 2$ and $\beta_1 = \ldots = \beta_d = 0$.

Conclusion. For $d \geq 2$, the eigenvalues $\lambda_{k_1,\ldots,k_d}$ in the KL decomposition (41)–(42) are multiple. This renders the numerical computation of the limit distribution of the test statistic in (44) more delicate than in the univariate case. This problem will be investigated elsewhere.

References

1.Abramowitz, M. and Stegun, I. A. (1965). *Handbook of Mathematical Integrals.* Dover, New York.

2.Adler, R. J. (1990). *An Introduction to Continuity, Extrema and Related Topics for General Gaussian Processes.* IMS Lecture Notes-Monograph Series **12**. Institute of Mathematical Statistics, Hayward, California.

3.Araujo, A. and Giné, E. (1980). *The Central Limit Theorem for Real and Banach Valued Random Variables.* Wiley, New York.

4.Ash, R. B. and Gardner, M. F. (1975). *Topics in Stochastic Processes.* Academic Press, New York.

5.Biane, P. and Yor, M. (1987). Variations sur une formule de P. Lévy. *Ann. Inst. Henri Poincaré.* **23** 359-377.

6.Blum, J. R., Kiefer, J. and Rosenblatt, M. (1961). Distribution-free tests of independence based on the sample distribution function. *Ann. Math. Statist.* **35** 138–149.

7.Bowman, F. (1958). *Introduction to Bessel Functions.* Dover, new York.

8.Cotterill, D. S. and Csörgő, M. (1982). On the limiting distribution and critical values for multivariate Cramér-von Mises statistic. *Ann. Math. Statist.* **10** 233–244.

9.Cotterill, D. S. and Csörgő, M. (1985). On the limiting distribution and critical values for the Hoeffding, Blum, Kiefer, Rosenblatt independence criterion. *Statist. Decisions.* **3** 1-48.

10.Cramér, H. (1938). Sur un nouveau théorème limite de la théorie des probabilités. *Actualités Scientifiques et Industrielles.* **736**. Hermann, Paris.

11.Csörgő, M. (1979). Strong approximation of the Hoeffding, Blum, Kiefer, Rosenblatt multivariate empirical process. *J. Multivariate. Anal.* **9** 84–100.

12.Deheuvels, P. (1981). An asymptotic decomposition for multivariate distribution-free tests of independence. *J. Mult. Anal.* **11** 102–113.

13.Deheuvels, P. (2005). Weighted multivariate Cramér-von Mises-type statistis. *Afrika Statistica.* (to appear).

14.Deheuvels, P. and Martynov, G. (2003). Karhunen-Loève expansions for weighted Wiener processes and Brownian bridges via Bessel functions. *Progress in Probability.* **55** 57–93, Birkhäuser, Basel.

15.Deheuvels, P., Peccati, G. and Yor, M. (2004). On quadratic functionals of the Brownian sheet and related processes. *Prépublication n*° 910. *Laboratoire de Probabilités & Modèles Aléatoires.* 40p. Paris.

16.Donati-Martin, C. and Yor, M. (1991). Fubini's theorem for double Wiener integrals and the variance of the Brownian path. *Ann. Inst. Henri Poincaré.* **27** 181-200.

17.Dugué, D. (1967). Fonctions caractéristiques d'intégrales browniennes. *Rev. Roum. Math. Pures Appl.* **12** 1207–1215.

18.Dugué, D. (1969). Characteristic functions of random variables connected with Brownian motion, and the von Mises multidimensional ω_n^2. *Multivariate Analysis* **2**, P. R. Krishnaiah ed., 289–301. Academic Press, New York.

19.Dugué, D. (1975). Sur des tests d'indépendance indépendants de la loi. *C. R. Acad. Sci. Paris Ser. A.* **281** 1103–1104.

20.Hoeffding, W. (1948). A nonparametric test of independence. *Ann. Math. Statist.* **19** 554–557.

21.Kac, M. (1951). On some connections between probability theory and differential and integral equations. *Proc.Second Berkeley Sympos. Math. Statist. Probab.* 180–215.

22.Kac, M. and Siegert, A. J. F. (1947). On the theory of noise in radio receivers with square law detectors. *J. Appl. Physics.* **18** 383–397.

23.Kac, M. and Siegert, A. J. F. (1947). An explicit representation of a stationary Gaussian process. *Ann. Math. Statist.* **18** 438–442.

24.Kiefer, J. (1959). K-sample analogues of the Kolmogorov-Smirnov and Crmér-V. Mises tests. *Ann. Math. Statist.* **30** 420–447.

25.Lifshits, M. A. (1995). *Gaussian Random Functions.* Kluwer, Dordrecht.

26.Nikitin, Y. (1995). *Asymptotic Efficiency of Nonparametric Tests.* Cambridge University Press.

27.Martynov, G. V. (1992). Statistical tests based on empirical processes and related questions. *J. Soviet. Math.* **61** 2195–2275.

28.Pitman, J. W. and Yor, M. (1981). Bessel processes and infinitely divisible laws. In: *Stochastic Integrals*, D. Williams, edit. Lecture Notes in Mathematics **1123** 285–370. Springer, New York.

29.Pitman, J. W. and Yor, M. (1982). A decomposition of Bessel bridges. *Z. Wahrscheinlichkeit. verw. Gebiete.* **59** 425–457.

30.Pitman, J. W. and Yor, M. (1986). Some divergent integrals of Brownian motion. *Analytic and Geometric Stochastics.* D. Kendall, edit., Suppl. to *Adv. Appl. Probab.* 109–116.

31.Pitman, J. and Yor, M. (1998). The law of the maximum of a Bessel bridge. Prépublication n°467, Laboratoire de Probabilités, Université Paris VI.

32.Scott, W. F. (1999). A wieghted Cramér-von Mises statistic, with some applications to clinical trials. *Commun. Statist. Theor. Methods.* **28** 3001-3008.

33.Smirnov, N. V. (1936). Sur la distribution de ω^2. *C. R. Acad. Sci. Paris.* **202** 449–452.

34.Smirnov, N. V. (1937). On the distribution of the ω^2 criterion. *Rec. Math. (Mat. Sbornik).* **6** 3–26.

35.Smirnov, N. V. (1948). Table for estimating the goodness of fit of empirical distributions. *Ann. Math. Statist.* **19** 279–281.

36.Watson, G. N. (1952). *A Treatise on the Theory of Bessel Functions.* Cambridge University Press, Cambridge.

37.Yor, M. (1992). *Some Aspects of Brownian Motion, Part I: Some Special Functionals.* Lectures in Math., ETH Zürich, Birkhäuser, Basel.

38.Yor, M. (1997). *Some Aspects of Brownian Motion, Part II: Some Recent Martingale Problems.* Lectures in Math., ETH Zürich, Birkhäuser, Basel.

The Entire Regularization Path for the Support Vector Machine

Trevor Hastie[1], Saharon Rosset[2], Robert Tibshirani[1], and Ji Zhu[3]

[1] Department of Statistics
Stanford University
Stanford, CA 94305, USA
email: {hastie,tibs}@stanford.edu
[2] IBM Watson Research Center
P.O. Box 218
Yorktown Heights, N.Y. 10598
email: srosset@us.ibm.com
[3] Ji Zhu
Department of Statistics
University of Michigan
Ann Arbor, MI 48109-1092
email: jizhu@umich.edu

Abstract. In this paper we argue that the choice of the SVM cost parameter can be critical. We then derive an algorithm that can fit the entire path of SVM solutions for every value of the cost parameter, with essentially the same computational cost as fitting one SVM model.

Keywords: Support Vector Machine, Regularization, Coefficient Path.

1 Introduction

We have a set of n training pairs x_i, y_i, where $x_i \in \mathbb{R}^p$ is a p-vector of real valued predictors (attributes) for the ith observation, $y_i \in \{-1, +1\}$ codes its binary response. The standard criterion for fitting the linear SVM)[Boser *et al.*, 1992, Cortes and Vapnik, 1995, Schölkopf and Smola, 2001] is

$$\min_{\beta_0,\beta} \frac{1}{2}||\beta||^2 + C\sum_{i=1}^{n} \xi_i, \tag{1}$$
$$\text{subject to, for each } i\text{: } y_i(\beta_0 + x_i^T\beta) \geq 1 - \xi_i.$$

Here the ξ_i are non-negative slack variables that allow points to be on the wrong side of their “soft margin” ($f(x) = \pm 1$), as well as the decision boundary, and C is a cost parameter that controls the amount of overlap. If the data are separable, then for sufficiently large C the solution achieves the maximal margin separator; if not, the solution achieves the minimum overlap solution with largest margin.

Alternatively, we can formulate the problem using a *(hinge) Loss + Penalty* criterion [Wahba *et al.*, 2000, Hastie *et al.*, 2001]:

$$\min_{\beta_0,\beta} \sum_{i=1}^{n} [1 - y_i(\beta_0 + \beta^T x_i)]_+ + \frac{\lambda}{2}||\beta||^2. \tag{2}$$

The regularization parameter λ in (2) corresponds to $1/C$, with C in (1).

This latter formulation emphasizes the role of regularization. In many situations we have sufficient variables (e.g. gene expression arrays) to guarantee separation. We may nevertheless avoid the maximum margin separator ($\lambda \downarrow 0$), which is governed by observations on the boundary, in favor of a more regularized solution involving more observations.

The nonlinear *kernel* SVMs can be represented in this form as well. With kernel K and $f(x) = \beta_0 + \sum_{i=1}^{n} \theta_i K(x, x_i)$, we solve [Hastie *et al.*, 2001]

$$\min_{\beta_0,\theta} \sum_{i=1}^{n} [1 - y_i(\beta_0 + \sum_{j=1}^{n} \theta_i K(x_i, x_j))] + \frac{\lambda}{2} \sum_{j=1}^{n} \sum_{j'=1}^{n} \theta_j \theta_{j'} K(x_j, x'_j). \tag{3}$$

Often the regularization parameter C (or λ) is regarded as a genuine “nuisance”. Software packages, such as the widely used SVM^{light} [Joachims, 1999], provide default settings for C.

To illustrate the effect of regularization, we generated data from a pair of mixture densities, described in detail in [Hastie *et al.*, 2001]. We used an SVM with a radial kernel $K(x, x') = \exp(-\gamma||x - x'||^2)$. Figure 1 shows the test error as a function of C for these data, using four different values for γ. Here we see a dramatic range in the correct choice for C (or $\lambda = 1/C$). When $\gamma = 5$, the most regularized model is called for; when $\gamma = 0.1$, the least regularized.

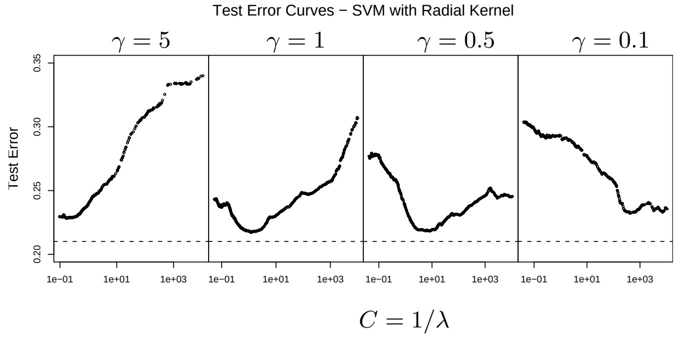


Fig. 1. *Test error curves for the mixture example, using four different values for the radial kernel parameter* γ.

One of the reasons that investigators avoid extensive exploration of C is the computational cost involved. In this paper we develop an algorithm which fits the *entire path* of SVM solutions $[\beta_0(C), \beta(C)]$, for all possible values of C, with essentially the computational cost of fitting a single model for a particular value of C. Our algorithm exploits the fact that the Lagrange multipliers implicit in (1) are piecewise-linear in C. This also means that the coefficients $\hat{\beta}(C)$ are also piecewise-linear in C. This is true for all SVM models, both linear and nonlinear kernel-based SVMs.

2 Problem Setup

We use a criterion equivalent to (1), implementing the formulation in (2):

$$\min_{\beta,\beta_0} \sum_{i=1}^{n} \xi_i + \frac{\lambda}{2}\beta^T\beta \text{ subject to } 1 - y_i f(x_i) \leq \xi_i;\ \xi_i \geq 0;\ f(x) = \beta_0 + \beta^T x. \quad (4)$$

Initially we consider only linear SVMs to get the intuitive flavor of our procedure; we then generalize to kernel SVMs.

We construct the Lagrange primal function

$$L_P : \sum_{i=1}^{n} \xi_i + \frac{\lambda}{2}\beta^T\beta + \sum_{i=1}^{n} \alpha_i(1 - y_i f(x_i) - \xi_i) - \sum_{i=1}^{n} \gamma_i \xi_i \quad (5)$$

and set the derivatives to zero. This gives

$$\frac{\partial}{\partial \beta} : \ \beta = \frac{1}{\lambda} \sum_{i=1}^{n} \alpha_i y_i x_i \quad (6)$$

$$\frac{\partial}{\partial \beta_0} : \ \sum_{i=1}^{n} y_i \alpha_i = 0, \quad (7)$$

along with the KKT conditions

$$\alpha_i(1 - y_i f(x_i) - \xi_i) = 0 \quad (8)$$

$$\gamma_i \xi_i = 0 \quad (9)$$

$$1 - \alpha_i - \gamma_i = 0 \quad (10)$$

We see that $0 \leq \alpha_i \leq 1$, with $\alpha_i = 1$ when $\xi_i > 0$ (which is when $y_i f(x_i) < 1$). Also when $y_i f(x_i) > 1$, $\xi_i = 0$ since no cost is incurred, and $\alpha_i = 0$. When $y_i f(x_i) = 1$, α_i can lie between 0 and 1.

The *usual* Lagrange multipliers associated with the solution to (1) are $\alpha'_i = \alpha_i/\lambda = C\alpha_i$. We prefer our formulation here since our $\alpha_i \in [0, 1]$, and this simplifies the definition of the paths we define.

We wish to find the entire solution path for all values of $\lambda \geq 0$. Our basic idea is as follows. We start with λ large and decrease it toward zero, keeping track of all the events that occur along the way. As λ decreases,

$||\beta||$ increases, and hence the width of the margin decreases. As this width decreases, points move from being inside to outside their margins. Their corresponding α_i change from $\alpha_i = 1$ when they are inside their margin $(y_i f(x_i) < 1)$ to $\alpha_i = 0$ when they are outside their margin $(y_i f(x_i) > 1)$. By continuity, points must linger on the margin $(y_i f(x_i) = 1)$ while their α_i decrease from 1 to 0. We will see that the $\alpha_i(\lambda)$ trajectories are piecewise-linear in λ, which affords a great computational savings: as long as we can establish the break points, all values in between can be found by simple linear interpolation. Note that points can return to the margin, after having passed through it.

It is easy to show that if the $\alpha_i(\lambda)$ are piecewise linear in λ, then both $\alpha'_i(C) = C\alpha_i(C)$ and $\beta(C)$ are piecewise linear in C. It turns out that $\beta_0(C)$ is also piecewise linear in C.

Our algorithm keeps track of the following sets:

- $\mathcal{M} = \{i : y_i f(x_i) = 1,\ 0 \leq \alpha_i \leq 1\}$, $\mathcal{M}$ for Margin
- $\mathcal{I} = \{i : y_i f(x_i) < 1,\ \alpha_i = 1\}$, $\mathcal{I}$ for Inside the margin
- $\mathcal{O} = \{i : y_i f(x_i) > 1,\ \alpha_i = 0\}$, $\mathcal{O}$ for Outside the margin

3 The Algorithm

Due to space restrictions, we show some details here; the rest can be found in [Hastie *et al.*, 2004].

Initialization

The initial conditions depend on whether the classes are balanced or not $(n_+ = n_-)$. The balanced case is easier. For very large λ, $||\beta||$ is small, and the the margin is very wide, all points are in $\mathcal{O}$, and hence $\alpha_i = 1 \forall i$. From (6) this means the orientation of β is fixed until the α_i change. The margin narrows as λ decreases, but the orientation remains fixed. Because of (7), the narrowing margin must connect with an outermost member of each class simultaneously. These points are easily identified, and this establishes the first event, the first tenants of $\mathcal{M}$, and β_0.

When $n_- \neq n_+$, the setup is more complex. In order to satisfy the constraint (7), a quadratic programming algorithm is needed to obtain the initial configuration. See [Hastie *et al.*, 2004] for details.

Kernels

The development so far has been in the original feature space. It is easy to see that the entire development carries through with "kernels" as well. In this case $f(x) = \beta_0 + g(x)$, and the only change that occurs is that (6) is changed to

$$g(x_i) = \frac{1}{\lambda}\sum_{j=1}^{n} \alpha_j y_j K(x_i, x_j),\ i = 1, \ldots, n, \tag{11}$$

or $\theta_j(\lambda) = \alpha_j y_j/\lambda$ using the notation in (3). Hereafter we will develop our algorithm for this more general kernel case.

The Path

The algorithm hinges on the set of points $\mathcal{M}$ sitting on the margin. We consider $\mathcal{M}$ at the point that an event has occurred:

1. The initial event, which means 2 or more points start in $\mathcal{M}$, with their initial values of $\alpha \in [0, 1]$.
2. A point from $\mathcal{I}$ has just entered $\mathcal{M}$, with its value of α_i initially 1.
3. A point from $\mathcal{O}$ has reentered $\mathcal{M}$, with its value of α_i initially 0.
4. One or more points in $\mathcal{M}$ has left the set, to join either $\mathcal{O}$ or $\mathcal{I}$.

Whichever the case, for continuity reasons this set will stay stable until the next event occurs, since to pass through $\mathcal{M}$, a point's α_i must change from 0 to 1 or vice versa. Since all points in $\mathcal{M}$ have $y_i f(x_i) = 1$, we can establish a path for their α_i.

We use the subscript ℓ to index the sets above immediately after the ℓth event has occurred. Suppose $|\mathcal{M}_\ell| = m$, and let α_i^ℓ, β_0^ℓ and λ_ℓ be the values of these parameters at the point of entry. Likewise f^ℓ is the function at this point. For convenience we define $\alpha_0 = \lambda\beta_0$, and hence $\alpha_0^\ell = \lambda_\ell\beta_0^\ell$.

Since

$$f(x) = \frac{1}{\lambda}\left(\sum_{j=1}^{n} y_j\alpha_j K(x, x_j) + \alpha_0\right), \tag{12}$$

for $\lambda_\ell > \lambda > \lambda_{\ell+1}$ we can write

$$\begin{aligned} f(x) &= \left[f(x) - \frac{\lambda_\ell}{\lambda} f^\ell(x)\right] + \frac{\lambda_\ell}{\lambda} f^\ell(x) \\ &= \frac{1}{\lambda}\left[\sum_{j\in\mathcal{M}_\ell} (\alpha_j - \alpha_j^\ell) y_j K(x, x_j) + (\alpha_0 - \alpha_0^\ell) + \lambda_\ell f^\ell(x)\right]. \end{aligned} \tag{13}$$

The second line follows because all the observations in $\mathcal{I}_\ell$ have their $\alpha_i = 1$, and those in $\mathcal{O}_\ell$ have their $\alpha_i = 0$, for this range of λ. Since each of the m points $x_i \in \mathcal{M}_\ell$ are to stay on the margin, we have that

$$\frac{1}{\lambda}\left[\sum_{j\in\mathcal{M}_\ell} (\alpha_j - \alpha_j^\ell) y_i y_j K(x_i, x_j) + y_i(\alpha_0 - \alpha_0^\ell) + \lambda_\ell\right] = 1,\ \forall i \in \mathcal{M}_\ell. \tag{14}$$

Writing $\delta_j = \alpha_j^\ell - \alpha_j$, from (14) we have

$$\sum_{j \in \mathcal{M}_\ell} \delta_j y_i y_j K(x_i, x_j) + y_i \delta_0 = \lambda_\ell - \lambda, \ \forall i \in \mathcal{M}_\ell. \tag{15}$$

Furthermore, since at all times $\sum_{i=1}^n y_i \alpha_i = 0$, we have that

$$\sum_{j \in \mathcal{M}_\ell} y_j \delta_j = 0. \tag{16}$$

Equations (15) and (16) constitute $m+1$ linear equations in $m+1$ unknowns δ_j, and can be solved. The δ_j and hence α_j will change linearly in λ, until the next event occurs:

$$\alpha_j = \alpha_j^\ell - (\lambda_\ell - \lambda) b_j, \ j \in \{0\} \cup \mathcal{M}_\ell. \tag{17}$$

See [Hastie *et al.*, 2004] for more precise details on solving these equations.

¿From (13) we have

$$f(x) = \frac{\lambda_\ell}{\lambda} \left[f^\ell(x) - h^\ell(x) \right] + h^\ell(x), \tag{18}$$

where

$$h^\ell(x) = \sum_{j \in \mathcal{M}_\ell} y_j b_j K(x, x_j) + b_0 \tag{19}$$

Thus the function itself changes in a piecewise-inverse manner in λ.

Finding $\boldsymbol{\lambda_{\ell+1}}$

The paths continue until one of the following events occur:

1. One of the α_i for $i \in \mathcal{M}_\ell$ reaches a boundary (0 or 1). For each i the value of λ for which this occurs is easily established.
2. One of the points in $\mathcal{I}^\ell$ or $\mathcal{O}^\ell$ attains $y_i f(x_i) = 1$.

By examining these conditions, we can establish the largest $\lambda < \lambda_\ell$ for which an event occurs, and hence establish $\lambda_{\ell+1}$ and update the sets.

Termination

In the separable case, we terminate when $\mathcal{I}$ becomes empty. At this point, all the ξ_i in (4) are zero, and further movement increases the norm of β unnecessarily.

In the non-separable case, λ runs all the way down to zero. For this to happen without f “blowing up” in (18), we must have $f^\ell - h^\ell = 0$, and hence the boundary and margins remain fixed at a point where $\sum_i \xi_i$ is as small as possible, and the margin is as wide as possible subject to this constraint.

3.1 Computational Complexity

At any update event ℓ along the path of our algorithm, the main computational burden is solving the system of equations of size $m_\ell = |\mathcal{M}_\ell|$. While this normally involves $O(m_\ell^3)$ computations, since $\mathcal{M}_{\ell+1}$ differs from $\mathcal{M}_\ell$ by typically one observation, inverse updating can reduce the computations to $O(m_\ell^2)$. The computation of $h^\ell(x_i)$ in (19) requires $O(nm_\ell)$ computations. Beyond that, several checks of cost $O(n)$ are needed to evaluate the next move.

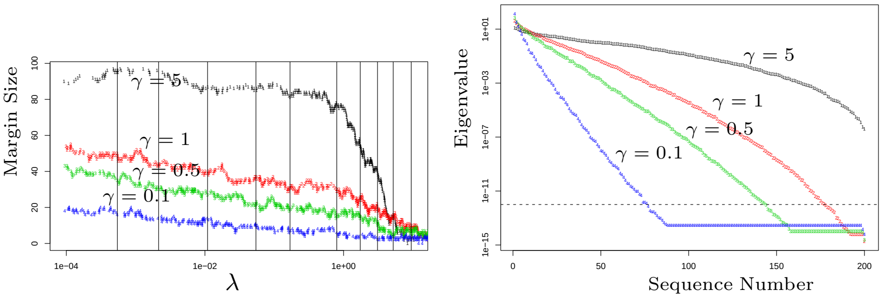


Fig. 2. *[Left] The margin sizes $|\mathcal{M}_\ell|$ as a function of λ, for different values of the radial-kernel parameter γ. The vertical lines show the positions used to compare the times with* `libsvm`*. [Right] The eigenvalues (on the log scale) for the kernel matrices $\mathbf{K}_\gamma$ corresponding to the four values of γ. The larger eigenvalues correspond in this case to smoother eigenfunctions, the small ones to rougher. The rougher eigenfunctions get penalized exponentially more than the smoother ones. For smaller values of γ, the effective dimension of the space is truncated.*

Although we have no hard results, our experience so far suggests that the total number Λ of moves is $O(k \min(n_+, n_-))$, for k around $4-6$; hence typically some small multiple c of n. If the average size of $\mathcal{M}_\ell$ is m, this suggests the total computational burden is $O(cn^2m + nm^2)$, which is similar to that of a single SVM fit.

Our R function `SvmPath` computes all 632 steps in the mixture example ($n_+ = n_- = 100$, radial kernel, $\gamma = 1$) in 1.44(0.02) secs on a Pentium 4, 2Ghz Linux machine; the `svm` function (using the optimized code `libsvm`, from the R library `e1071`) takes 9.28(0.06) seconds to compute the solution at 10 points along the path. Hence it takes our procedure about 50% more time to compute the entire path, than it costs `libsvm` to compute a typical single solution.

4 Mixture simulation continued

The λ_ℓ in Figure 1 are the *entire* collection of change points as described in Section 3. We were at first surprised to discover that not all these sequences achieved zero training errors on the 200 training data points, at their least regularized fit. In fact the minimal training errors, and the corresponding values for γ are summarized in Table 1. It is sometimes argued that the implicit feature space is "infinite dimensional" for this kernel, which suggests that perfect separation is always possible. The last row of the table shows the effective rank of the 200×200 kernel *Gram* matrix $\mathbf{K}$ (which we defined to be the number of singular values greater than 10^{-12}). In general a full rank $\mathbf{K}$ is required to achieve perfect separation. This rank-deficiency of the Gram matrix has been noted by a number of other authors.

γ	5	1	0.5	0.1
Training Errors	0	12	21	33
Effective Rank	200	177	143	76

Table 1. *The number of minimal training errors for different values of the radial kernel scale parameter γ, for the mixture simulation example. Also shown is the effective rank of the 200×200 Gram matrix $\mathbf{K}_\gamma$.*

This emphasizes the fact that not all features in the feature map implied by K are of equal stature; many of them are shrunk way down to zero. Rephrasing, the regularization in (3) penalizes unit-norm features by the inverse of their eigenvalues, which effectively annihilates some, depending on γ. Small γ implies wide, flat kernels, and a suppression of wiggly, "rough" functions.

Writing (3) in matrix form,

$$\min_{\beta_0,\boldsymbol{\theta}} L[\mathbf{y},\mathbf{K}\boldsymbol{\theta}] + \frac{\lambda}{2}\boldsymbol{\theta}^T\mathbf{K}\boldsymbol{\theta}, \tag{20}$$

we reparametrize using the eigen-decomposition of $\mathbf{K} = \mathbf{U}\mathbf{D}\mathbf{U}^T$. Let $\mathbf{K}\boldsymbol{\theta} = \mathbf{U}\boldsymbol{\theta}^*$ where $\boldsymbol{\theta}^* = \mathbf{D}\mathbf{U}^T\theta$. Then (20) becomes

$$\min_{\beta_0,\boldsymbol{\theta}^*} L[\mathbf{y},\mathbf{U}\boldsymbol{\theta}^*] + \frac{\lambda}{2}\boldsymbol{\theta}^{*T}\mathbf{D}^{-1}\boldsymbol{\theta}^*. \tag{21}$$

Now the columns of $\mathbf{U}$ are unit-norm basis functions (in $\mathbb{R}^2$) spanning the column space of $\mathbf{K}$; from (21) we see that those members corresponding to near-zero eigenvalues (the elements of the diagonal matrix $\mathbf{D}$) get heavily penalized and hence ignored. Figure 2 shows the elements of $\mathbf{D}$ for the four values of γ.

5 Discussion

Our work on the SVM path algorithm was inspired by early work on exact path algorithms in other settings. “Least Angle Regression” [Efron *et al.*, 2002] show that the coefficient path for the sequence of “lasso” coefficients is piecewise linear. The lasso uses a quadratic criterion, with an L_1 constraint. In fact, any model with an L_1 constraint and a quadratic, piecewise quadratic, piecewise linear, or mixed quadratic and linear loss function, will have piecewise linear coefficient paths, which can be calculated exactly and efficiently for all values of λ [Rosset and Zhu, 2003]. This includes the L_1 SVM [Zhu *et al.*, 2003].

The SVM model has a quadratic constraint and a piecewise linear (“hinge”) loss function. This leads to a piecewise linear path in the dual space, hence the Lagrange coefficients α_i are piecewise linear.

Of course, quadratic criterion + quadratic constraints also lead to exact path solutions, as in the classic ridge regression case, since a closed form solution is obtained via the SVD.

The general techniques employed in this paper are known as parametric programming in convex optimization. After completing this work, it was brought to our attention that [Pontil and Verri, 1998] reported on the picewise-linear nature of the lagrange multipliers, although they did not develop the path algorithm. [Fine and Scheinberg, 2002, Cauwenberghs and Poggio, 2001] employ techniques similar to ours in incremental learning for SVMs. These authors do not construct exact paths as we do, but rather focus on updating and downdating the solutions as more (or less) data arises. [Diehl and Cauwenberghs, 2003] allow for updating the parameters as well, but again do not construct entire solution paths.

The `SvmPath` has been implemented in the `R` computing environment, and is available from the R website.

Acknowledgements

The authors thank Jerome Friedman for helpful discussions, and Mee-Young Park for assisting with some of the computations. Trevor Hastie was partially supported by grant DMS-0204162 from the National Science Foundation, and grant RO1-EB0011988-08 from the National Institutes of Health.

References

[Boser *et al.*, 1992]B. Boser, I. Guyon, and V. Vapnik. A training algorithm for optimal margin classifiers. In *Proceedings of COLT II*, Philadelphia, PA, 1992.

[Cortes and Vapnik, 1995]C. Cortes and V. Vapnik. Support vector networks. *Machine Learning*, 20:1–25, 1995.

[Schölkopf and Smola, 2001]Bernard Schölkopf and Alex Smola. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond (Adaptive Computation and Machine Learning).* MIT Press, 2001.

[Wahba *et al.*, 2000]G. Wahba, Y. Lin, and H. Zhang. Gacv for support vector machines. In A.J. Smola, P.L. Bartlett, B. Schölkopf, and D. Schuurmans, editors, *Advances in Large Margin Classifiers*, pages 297–311, Cambridge, MA, 2000. MIT Press.

[Hastie *et al.*, 2001]T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning; Data mining, Inference and Prediction.* Springer Verlag, New York, 2001.

[Joachims, 1999]Thorsten Joachims. *Practical Advances in Kernel Methods — Support Vector Learning*, chapter Making large scale SVM learning practical. MIT Press, 1999. see `http://svmlight.joachims.org`.

[Hastie *et al.*, 2004]Trevor Hastie, Saharon Rosset, Robert Tibshirani, and Ji Zhu. The entire regularization path for the support vector machine. *Journal of Machine Learning Research*, (5):1391–1415, 2004.

[Efron *et al.*, 2002]B. Efron, T. Hastie, I. Johnstone, and R.. Tibshirani. Least angle regression. Technical report, Stanford University, 2002.

[Rosset and Zhu, 2003]Saharon Rosset and Ji Zhu. Piecewise linear regularized solution paths. Technical report, Stanford University, 2003. `http://www-stat.stanford.edu/~saharon/papers/piecewise.ps`.

[Zhu *et al.*, 2003]Ji Zhu, Saharon Rosset, Trevor Hastie, and Robert Tibshirani. L1 norm support vector machines. Technical report, Stanford University, 2003.

[Pontil and Verri, 1998]Massimiliano Pontil and Alessandro Verri. Properties of support vector machines. *Neural Comput.*, 10(4):955–974, 1998.

[Fine and Scheinberg, 2002]Shai Fine and Katya Scheinberg. Incas: An incremental active set method for svm. Technical report, IBM Research Labs, Haifa, 2002.

[Cauwenberghs and Poggio, 2001]G. Cauwenberghs and T. Poggio. Incremental and decremental support vector machine learning. In *Advances in Neural Information Processing Systems (NIPS*2000)*, volume 13. MIT Press, Cambridge, MA, 2001.

[Diehl and Cauwenberghs, 2003]Christopher Diehl and Gert Cauwenberghs. Svm incremental learning, adaptation and optimization. In *Proceedings of the 2003 International Joint Conference on Neural Networks*, pages 2685–2690, 2003. Special series on Incremental Learning.

Biological Aggregation at the Interface Between Theory and Practice

William H. E. Day

Port Maitland, Nova Scotia B0W 2V0, Canada
(e-mail: whday@istar.ca)

Abstract. To understand evolutionary processes better, biologists use aggregation methods to estimate evolutionary relationships; yet properties of the methods are sometimes so imprecisely defined, and their interrelationships so poorly understood, that useful formal results may be difficult to obtain. To address this problem I describe a strategy for modeling aggregation methods and studying their properties. The approach accommodates impossibility results for aggregating rankings, nonhierarchical classifications, hierarchies, and phylogenies. It remains to formulate other relevant models of biological aggregation and to characterize methods for solving biological problems of agreement and synthesis.

Keywords: Aggregation, Agreement, Axiom, Consensus, Impossibility, Synthesis.

> *The axiomatic method is, strictly speaking, nothing but this art of drawing up texts whose formalization is straightforward in principle. As such it is not a new invention; but its systematic use as an instrument of discovery is one of the original features of contemporary mathematics.* — Nicolas Bourbaki [Bourbaki, 1968, p. 8]

1 Aggregation problems in biology

Mathematical models of aggregation have long been used in systematic or evolutionary biology [Day and McMorris, 2003]. Given a sequence of trees that estimate phylogenetic relationships among species, for example, one wants to develop methods to synthesize these trees into a single large phylogenetic supertree [Steel *et al.*, 2000, Wilkinson *et al.*, 2004]. If estimating supertrees is an exemplar of biological aggregation, the following questions pertain.

What is a supertree? Most biologists understand biological supertrees and their use to estimate evolutionary history, while mathematicians wish to know no more about supertrees than is necessary to construct appropriate models. Here I assume that supertrees and other relevant objects are defined so that their essential features are expressible by sets of elementary structures.

What is a supertree rule? I describe an abstract framework in which aggregation can be modeled and concepts investigated. Given a profile (sequence) of objects: an agreement rule returns an object having only features in common agreement among the profile's objects, a consensus rule returns an object best representing the profile's objects, and a synthesis rule returns a composite of the profile's objects.

What biologically relevant properties should supertree rules exhibit? Properties of aggregation constrain the formal model so as to improve its capability to approximate a biological process [Wilkinson *et al.*, 2004]. Here I ignore issues of practicality and computational complexity since analyses of time and space resources are best left to computer scientists. I am particularly interested in axioms that if satisfied by an aggregation rule may increase one's confidence in the biological relevance of that rule's results.

Can supertree rules exhibit particular sets of desirable properties? What properties do known supertree rules exhibit? Since little has yet been done to answer such questions, here I simply mention some biologically interesting impossibility results and some open problems concerning the axiomatics of biological aggregation rules.

2 Aggregation models

For 30 years researchers have striven to develop consensus rules for biological applications. Although inappropriate for investigating agreement or synthesis, consensus rules are a useful point of reference. There is a set of voters. Each voter votes by specifying an object. A consensus rule C accepts a profile of objects and returns a unique consensus object that in some sense best represents the profile. A simple model requires that C be a function $C : \mathcal{X}^k \longrightarrow \mathcal{X}$, where $\mathcal{X}$ is a set of *objects* such as those in table 1 and $\mathcal{X}^k$ is the set of all profiles (k-tuples) of $\mathcal{X}$. C is further specified by a set K of k *indices* to name the voters, a set S of n *labels* or species names with which to describe objects, *encoding* functions to represent objects in meaningful ways, and *reduction* functions to reveal the structure of objects. Since the concepts

$\mathcal{X}$	is the set of all ...
$\mathcal{E}$	Nonhierarchical classifications or partitions of S, each being a set of nonempty classes or subsets of S that are pairwise disjoint and that include every element of S.
$\mathcal{O}$	Rankings of S, each being a partition of S the classes of which are linearly ordered from most to least preferred.
$\mathcal{H}$	Rooted trees, each with n leaves, such that the root vertex has degree at least 2, every other interior vertex has degree at least 3, and every leaf is labeled with a distinct element of S.
$\mathcal{P}$	Unrooted trees, each with n leaves, such that no vertex has degree 2 and every leaf is labeled with a distinct element of S.

Table 1. Objects defined in terms of S, $n = |S| > 0$

of object, index, label, encoding, and reduction appear naturally in models of agreement, consensus, or synthesis, the incremental strategy in table 2 can be used to study them.

1. Begin with the basic concepts of object, index, and label.	
2. Design an aggregation model.	Specify axioms and use them to prove things.
3. Add a concept of encoding.	Specify axioms and use them to prove things.
4. Add a concept of reduction.	Specify axioms and use them to prove things.
5. Add other relevant concepts.	Specify axioms and use them to prove things.
6. Repeat steps 2–5 for related aggregation models.	

Table 2. Strategy to investigate aggregation

To specify models let $K = \{1, \ldots, k\}$, $S = \{s_1, \ldots, s_n\}$, and for every $X \subseteq S$ let $\mathcal{X}_X$ be a set of objects defined in terms of each and every label of X. For every $X \subseteq S$ let $\mathcal{X}_{[X]} = \bigcup_{Y \subseteq X} \mathcal{X}_Y$ where $\mathcal{X}_X \subseteq \mathcal{X}_{[X]}$. An object of $\mathcal{X}_X$ has the label set X, but an object of $\mathcal{X}_{[X]}$ may have as its label set any subset of X; thus $\mathcal{H}_S$ is the set of hierarchies having exactly n leaf labels and $\mathcal{H}_{[S]}$ is the set of hierarchies having at most n leaf labels. $\mathcal{X}$, K, S then yield

$$C : \mathcal{X}_S^k \longrightarrow \mathcal{X}_{[S]}, \quad \text{a model of agreement,} \tag{1}$$

$$C : \mathcal{X}_S^k \longrightarrow \mathcal{X}_S, \quad \text{a model of consensus, and} \tag{2}$$

$$C : \mathcal{X}_{[S]}^k \longrightarrow \mathcal{X}_S, \quad \text{a model of synthesis.} \tag{3}$$

The essence of consensus is that profile objects and consensus result have the same label set S. Agreement (1) is more general than consensus (2) since, although the domains are identical, an agreement result's label set may be a proper subset of S. Synthesis (3) is more general than consensus (2) since, although the codomains are identical, the label set of any synthesis profile object may be a proper subset of S.

Models (1)–(3) can be modified into rules that accept profiles of varying lengths or return more than one aggregated object. Let $\mathcal{X}^* = \bigcup_{k \geq 1} \mathcal{X}^k$ be the set of all profiles of finite positive length and call any aggregation rule with domain $\mathcal{X}^*$ a *complete* rule. Let $2^{\mathcal{X}} \setminus \{\emptyset\}$ be the set of all nonempty subsets of $\mathcal{X}$ and call any aggregation rule with codomain $2^{\mathcal{X}} \setminus \{\emptyset\}$ a *multiaggregation* rule. Thus a complete multisynthesis rule is modeled by a function

$$C : \mathcal{X}_{[S]}^* \longrightarrow 2^{\mathcal{X}_S} \setminus \{\emptyset\}. \tag{4}$$

3 Aggregation axioms

The axioms in table 3 address issues of impartiality (whether rules favor one label or index more than another), delegation of authority (whether determining outcomes resides with proper subsets of indices), optimality (how rules

Line	Axiom	Concept	Reference
(5)	*S-Ntr*: Neutrality of Labels	Label	[Steel *et al.*, 2000]
(6)	*Sym*: Symmetry of Indices	Index	[Steel *et al.*, 2000]
(7)	*Prj*: Projection	"	[Barthélemy *et al.*, 1991]
(8)	*Dct*: Weak Dictatorship	Encoding	[Arrow, 1963]
(9)	*Olg*: Oligarchy	"	[Mirkin, 1975]
(10)	*PO*: Pareto Optimality	"	[Arrow, 1963]
(11)	*SO*: Strong Optimality	"	[Steel *et al.*, 2000]
(12)	*RC*: Reduction Consistency	Reduction	[Wilkinson *et al.*, 2004]
(13)	*Ind*: Independence	"	[Arrow, 1963]
(14)	*Dsp*: Display	"	[Steel *et al.*, 2000]
(15)	*Agr*: Agreement	"	[Day and McMorris, 2003]

Table 3. Axioms and their related concepts

behave in the presence of object agreement), contexture (how rules respond to changes in structure or composition), and resolvability (how rules preserve relationships between objects).

To motivate axioms I give informal prose descriptions, but to specify axioms I define them using the logical symbols for negation ($\neg$), conjunction ($\wedge$), disjunction ($\vee$), implication ($\Longrightarrow$), equivalence ($\Longleftrightarrow$), and universal ($\forall$) and existential ($\exists$) quantification. Since axioms may apply to more than one model, in their definitions I assume as little as possible about the model's form: unless stated otherwise let it be a function $C : \mathcal{X}^k \longrightarrow \mathcal{Y}$ for $\mathcal{X}, \mathcal{Y} \subseteq \mathcal{X}_{[S]}$ where $\mathcal{X} \subseteq \mathcal{Y}$ and/or $\mathcal{Y} \subseteq \mathcal{X}$. Such an axiom might be relevant to any of the models (1)–(3).

Let f and g be functions, let x be an element in g's domain, and let $g(x)$ be in f's domain; then I reduce notational clutter by writing fgx instead of $f(g(x))$. Thus $C\sigma P = C(\sigma(P))$ as in (6). To specify objects of $P \in \mathcal{X}^k$ let $P = (T_1, \ldots, T_k)$ as in (7).

3.1 Basic axioms

Three axioms treat objects as atomic and indivisible.

***S-Ntr*: Neutrality of Labels.** Let $\phi : S \longrightarrow S$ be a function that permutes the labels in S. Let $\phi : \mathcal{X} \longrightarrow \mathcal{X}$ permute the labels of an object: for every $T \in \mathcal{X}$, ϕT is the object obtained by using ϕ to permute the labels of T. Let $\phi : \mathcal{X}^k \longrightarrow \mathcal{X}^k$ permute the labels in every object of a profile: $(\forall P \in \mathcal{X}^k)(\phi P = (\phi T_1, \ldots, \phi T_k))$. Although three functions are named ϕ, context shows which ϕ pertains. *Motivation:* If a profile P is described by a data matrix in which each row represents a label then the aggregation of P should be insensitive to the relative order of P's rows (labels). Put another way, for every P and every S-permutation ϕ, the relabeling by ϕ of the aggregation of P should equal the aggregation of the profile in which P's objects are relabeled by ϕ. *Axiom:*

$$(\forall P \in \mathcal{X}^k)(\forall S\text{-permutations } \phi)(C\phi P = \phi CP) \tag{5}$$

***Sym*: Symmetry of Indices.** Let $\sigma : K \longrightarrow K$ be a function that permutes the indices in $K = \{1, \ldots, k\}$. Now σ can permute objects in a profile by permuting the indices of the objects in that profile, i.e., let $\sigma : \mathcal{X}^k \longrightarrow \mathcal{X}^k$ be a function such that $(\forall P \in \mathcal{X}^k)(\sigma P = (T_{\sigma 1}, \ldots, T_{\sigma k}))$. Although two functions are named σ, context shows which σ pertains. *Motivation:* If a profile P is described by a data matrix in which each column represents an object then the aggregation of P should be insensitive to the relative order of P's columns (objects). Put another way, for every P and every K-permutation σ, the aggregation of P should equal the aggregation of the profile in which the positions of P's objects are permuted by σ. *Axiom:*

$$(\forall P \in \mathcal{X}^k)(\forall K\text{-permutations } \sigma)(CP = C\sigma P) \tag{6}$$

***Prj*: Projection (Strong Dictatorship).** *Motivation:* In nontrivial oligarchies and dictatorships the power to control aggregation is shared unequally by voters. In a strong dictatorship, for some index j and every profile P, the aggregation of P is the j^{th} object of P. *Axiom:*

$$(\exists j \in K)(\forall P \in \mathcal{X}^k)(CP = T_j) \tag{7}$$

Thus if P is a point in a k-dimensional space then C projects P onto a single dimension.

3.2 Axioms using object encodings

Whereas in section 3.1 objects were atomic and indivisible, now let every object $T \in \mathcal{X}$ be a set of elementary structures that are defined using the labels of S. Specifically let E_S be a complete set of elementary structures defined using the labels of S, and let r denote an encoding by which every $T \in \mathcal{X}$ is a well-defined subset of E_S. The encodings in table 4 may be familiar to biologists; the axioms in sections 3.2 and 3.3 assume that such an encoding has been applied.

$\mathcal{X}$	r	Using r, $T \in \mathcal{X}$ is a	Reference
$\mathcal{O}$	w	weak order	[Arrow, 1963]
$\mathcal{E}$	e	equivalence relation	[Mirkin, 1975]
$\mathcal{H}$	c	set of clusters	[Margush and McMorris, 1981]
$\mathcal{H}$	t	set of triads	[Colonius and Schulze, 1981]
$\mathcal{H}$	n	set of nestings	[Adams III, 1986]
$\mathcal{P}$	s	set of splits	[Buneman, 1971]
$\mathcal{P}$	q	set of quartets	[Colonius and Schulze, 1981]

Table 4. Encodings (r) to represent objects as sets of elementary structures

***Dct*: Weak Dictatorship.** *Motivation:* In a weak dictatorship, for some index j and every profile P, the aggregation of P contains as a subset the j^{th} object of P. *Axiom:*

$$(\exists j \in K)(\forall P \in \mathcal{X}^k)(T_j \subseteq CP) \tag{8}$$

***Olg*: Oligarchy.** *Motivation:* Oligarchy extends the strong dictatorial concept to forms of aggregation in which ruling power is shared by a set of individuals: for some index set V and every profile P, the aggregation of P is the set intersection of the objects of P that are specified by V. *Axiom:*

$$(\exists V \subseteq K)(\forall P \in \mathcal{X}^k)(\cap_{j \in V} T_j = CP) \tag{9}$$

An oligarchy of one individual is a strong dictator; an oligarchy of k individuals is a form of rule by unanimity.

***PO*: Pareto Optimality.** *Motivation:* Proposals may require for adoption the unanimous support of a society's members. For every profile P the aggregation of P should include those elementary structures (i.e., proposals) that are in every object of P (i.e., are supported by every member). *Axiom:*

$$(\forall P \in \mathcal{X}^k)(\cap_{i \in K} T_i \subseteq CP) \tag{10}$$

***SO*: Strong Optimality.** *Motivation:* Instead of requiring unanimous support, proposals may be adopted if they are unopposed by conflicting proposals. With hierarchies represented by sets of triads (see figure 1), for every profile P and every three labels x, y, z, if $xy|z$ is in some object of P but neither $xz|y$ nor $yz|x$ are in P's objects, then $xy|z$ should be in the aggregation of P. *Axiom:*

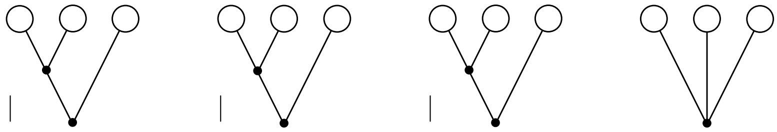


Fig. 1. Triads for representing hierarchies.

$$\begin{aligned}
&(\forall P \in \mathcal{H}^k)(\forall x, y, z \in S)(\\
&\quad [(\exists j \in K)(xy|z \in T_j) \wedge (\forall i \in K)(xz|y \notin T_i \wedge yz|x \notin T_i)] \\
&\quad \implies xy|z \in CP)
\end{aligned} \tag{11}$$

With phylogenies represented by sets of quartets the axiom becomes

$$\begin{aligned}
&(\forall P \in \mathcal{P}^k)(\forall w, x, y, z \in S)(\\
&\quad [(\exists j \in K)(wx|yz \in T_j) \wedge (\forall i \in K)(wy|xz \notin T_i \wedge wz|xy \notin T_i)] \\
&\quad \implies wx|yz \in CP)
\end{aligned}$$

In such cases SO is stronger than PO in the sense that $SO \implies PO$.

3.3 Axioms using object encodings and reductions

Let an encoding (as in table 4) be applied so that every object in $\mathcal{X}$ is represented by a set of elementary structures. Like an X-ray machine, reduction penetrates the surfaces of such objects to reveal hidden structure. For every $X \subseteq S$, let the function $\xi_X : \mathcal{X}_{[S]} \longrightarrow \mathcal{X}_{[X]}$ reduce objects on subsets of S to objects on subsets of X: for every $T \in \mathcal{X}_{[S]}$, $\xi_X T$ is the object obtained by suppressing in T the structure associated with $S \setminus X$. Thus if T were a graph G with vertex set S then $\xi_X T$ might be the subgraph of G that is induced by X. Also let $\xi_X : \mathcal{X}^k_{[S]} \longrightarrow \mathcal{X}^k_{[X]}$ reduce profiles rather than single objects: for every $X \subseteq S$ then $(\forall P \in \mathcal{X}^k_{[S]})(\xi_X P = (\xi_X T_1, \ldots, \xi_X T_k))$. Although two functions are named ξ_X, context shows which ξ_X pertains.

***RC*: Reduction Consistency.** *Motivation:* The order in which reduction and aggregation functions are applied ought not to matter: for every profile P and subset X of labels, the aggregation of the reduction of P to X by ξ should equal the reduction to X by ξ of the aggregation of P. *Axiom:*

$$(\forall P \in \mathcal{X}^k_{[S]})(\forall X \subseteq S)(C\xi_X P = \xi_X CP) \tag{12}$$

***Ind*: Independence.** Profiles $P, P' \in \mathcal{X}^k_{[S]}$ are called equal, i.e., $P = P'$, if and only if $(\forall i \in K)(T_i = T'_i)$. *Motivation:* For all profiles P and P' and every subset X, if P and P' are equal when reduced to X by ξ then the aggregations of P and P' must be equal when reduced to X by ξ. *Axiom:*

$$(\forall P, P' \in \mathcal{X}^k_{[S]})(\forall X \subseteq S)(\xi_X P = \xi_X P' \Longrightarrow \xi_X CP = \xi_X CP') \tag{13}$$

Ind, which Arrow [Arrow, 1963] called independence of irrelevant alternatives, imposes on aggregation rules a form of context insensitivity. *Ind* is weaker than reduction consistency in the sense that *RC* $\Longrightarrow$ *Ind*. Some researchers have confounded *Ind* with *RC*, a result perhaps unsurprising since [Arrow, 1963, pp. 26–27] motivated his definition of *Ind* with examples of both *RC* and *Ind* ([McLean, 1995, p. 108]).

***Dsp*: Display.** An object T is said to resolve an object T' if T' can be obtained from T by a sequence of simplifying elementary transformations. For partitions an elementary transformation forms the union of two classes of the previous partition; for rankings those two classes must be adjacent in the previous linear order. For hierarchies or phylogenies an elementary transformation contracts an interior edge by identifying its endpoints and deleting the resulting loop. For every $T, T' \in \mathcal{X}_{[S]}$, T is said to display T' if, for some $X \subseteq S$, $\xi_X T = T'$ or $\xi_X T$ resolves T'. An object also can display a profile: for every object T and profile P, T is said to display P if T displays T_i for every $i \in K$. *Motivation:* For every profile P if some object displays P then the aggregation of P should display P. *Axiom:*

$$(\forall P \in \mathcal{X}^k_{[S]})[(\exists T \in \mathcal{X}_{[S]})(T \text{ displays } P) \Longrightarrow (CP \text{ displays } P)] \tag{14}$$

***Agr*: Agreement.** For every $P \in \mathcal{X}^k_{[S]}$ let $D(P)$ be the set of all nontrivial objects (i.e., those having nontrivial elementary structures) that are displayed by every $T_i \in P$. *Motivation:* For every profile P if some nontrivial object is displayed by P then the aggregation of P should be nontrivial and should be displayed by P. *Axiom:*

$$(\forall P \in \mathcal{X}^k_{[S]})(D(P) \neq \emptyset \Longrightarrow CP \in D(P)) \tag{15}$$

4 Problems at the interface

Researchers have used the axiomatic approach to prove impossibility results (table 5) for models (1)–(3) of aggregation, a result being called impossible if an undesirable property, e.g., *Dct* or *Olg*, follows from desirable properties, e.g., *Ind* and *PO*. Consequently the following questions may be relevant when assessing the efficacy of such models for biological aggregation.

Model	$\mathcal{X}$	r	Impossibility Result	Reference
consensus	$\mathcal{O}$	w	$\mathit{Ind} \wedge \mathit{PO} \Longrightarrow \mathit{Dct}$	[Arrow, 1963, p. 97]
consensus	$\mathcal{E}$	e	$\mathit{Ind} \wedge \mathit{PO} \Longleftrightarrow \mathit{Olg}$	[Mirkin, 1975, p. 446]
consensus	$\mathcal{H}$	c	$\mathit{Ind} \wedge \mathit{PO} \Longleftrightarrow \mathit{Prj}$	[Barthélemy *et al.*, 1992, p. 63]
agr,con,syn	$\mathcal{H}$	t	$\mathit{SO} \Longleftrightarrow \neg \mathit{SO}$	[Steel *et al.*, 2000, p. 367]
consensus	$\mathcal{P}$	q	$\mathit{Ind} \wedge \mathit{PO} \Longleftrightarrow \mathit{Prj}$	[McMorris and Powers, 1993, p. 54]
consensus	$\mathcal{P}$	q	$\mathit{S\text{-}Ntr} \wedge \mathit{PO} \Longrightarrow \neg \mathit{Sym}$	[Steel *et al.*, 2000, p. 366]
synthesis	$\mathcal{P}$	q	$\mathit{S\text{-}Ntr} \wedge \mathit{Dsp} \Longrightarrow \neg \mathit{Sym}$	[Steel *et al.*, 2000, p. 364]
agreement	$\mathcal{P}$	q	$\mathit{S\text{-}Ntr} \Longrightarrow \neg \mathit{Agr}$	[Day and McMorris, 2003, p. 108]

Table 5. Impossibility results for aggregation models (1)–(3), the representation of objects in $\mathcal{X}$ being determined by the encoding r. For many other such results see [Day and McMorris, 2003] and references therein.

Are we using the right axioms? [Wilkinson *et al.*, 2004] argue that elusive properties of input trees involving tree size, tree shape, or the location or size of conflicting structures may adversely bias methods to build supertrees. How should such properties be included in formal studies of aggregation models? Even devising adequate definitions of such properties may be problematic. Would some particular encoding provide a natural setting in which such properties could be investigated? Since the strategy in table 2 is simplistic, using it to guide the analysis of such models may be ineffective or infeasible.

Engaging but specific problems exist. How strong is S-neutrality? For agreement, consensus, or synthesis rules on phylogenies, characterize those rules that satisfy *S-Ntr*. Since independence (*Ind*) imposes a strong concept of context insensitivity, could it be replaced by biologically useful concepts of context sensitivity?

Are we solving the right problems? Impossibility results encourage mathematicians to explore the boundary areas between feasible and infeasible aggregation rules. Biologists might be more excited by axiomatic characterizations of actual or ideal aggregation rules for biologically relevant objects.

Are we using the right models? Since much is known about complete multiconsensus median rules [McMorris *et al.*, 2000, McMorris *et al.*, 2003], do such axiomatic results generalize to the complete multisynthesis model (4)? Do the concepts of agreement, consensus, synthesis, multiaggregation, and completeness yield useful aggregation models for biological applications? Although an extensive literature on biologically relevant consensus rules exists [Day and McMorris, 2003], axiomatic investigations of agreement and synthesis rules are just beginning [Steel *et al.*, 2000] and show great promise.

Have we the right perspective? If objects are complex structures, one can exploit that complexity to study the interrelationships among objects; but if objects are taken to be atomic and indivisible, one must use object interrelationships to study the basic properties of sets of objects. Would it be useful to investigate agreement or synthesis models from an order theoretic perspective, as was done for consensus models by [Monjardet, 1990] and [Leclerc and Monjardet, 1995]?

For some readers this paper may have little of biological interest since its biological relevance emerges only by specifying undefined terms, e.g., object, and open-ended concepts, e.g., encoding or reduction. If then the axioms or models prove to be inappropriate for analyzing biological problems, perhaps biologists and mathematicians would collaborate to refine the approach.

Acknowledgements. My views have been influenced by [Steel *et al.*, 2000], who investigate supertrees axiomatically and whose clarity of exposition is exemplary, and by [Wilkinson *et al.*, 2004], who examine biologically desirable properties of synthesis rules for supertrees. I am indebted to Robert C. Powers for criticizing preliminary drafts of this paper.

References

[Adams III, 1986]E. N. Adams III. n-Trees as nestings: Complexity, similarity, and consensus. *Journal of Classification*, 3(2):299–317, 1986.

[Arrow, 1963]K. J. Arrow. *Social Choice and Individual Values.* Number 12 in Cowles Foundation for Research in Economics at Yale University: Monographs. Wiley, New York, second edition, 1963. Reprinted by Yale University Press (New Haven) in 1978.

[Barthélemy *et al.*, 1991]J. P. Barthélemy, F. R. McMorris, and R. C. Powers. Independence conditions for consensus n-trees revisited. *Applied Mathematics Letters*, 4(5):43–46, 1991.

[Barthélemy *et al.*, 1992]J. P. Barthélemy, F. R. McMorris, and R. C. Powers. Dictatorial consensus functions on n-trees. *Mathematical Social Sciences*, 25(1):59–64, December 1992.

[Bourbaki, 1968]N. Bourbaki. *Theory of Sets.* Number 1 in Elements of Mathematics. Addison-Wesley, Reading, Massachusetts, 1968.

[Buneman, 1971]P. Buneman. The recovery of trees from measures of dissimilarity. In F. R. Hodson, D. G. Kendall, and P. Tautu, editors, *Mathematics in the Archaeological and Historical Sciences*, pages 387–395. Edinburgh University Press, Edinburgh, 1971.

[Colonius and Schulze, 1981]H. Colonius and H. H. Schulze. Tree structures for proximity data. *British Journal of Mathematical & Statistical Psychology*, 34(2):167–180, 1981.

[Day and McMorris, 2003]W. H. E. Day and F. R. McMorris. *Axiomatic Consensus Theory in Group Choice and Biomathematics.* Number 29 in SIAM Frontiers in Applied Mathematics. Society for Industrial and Applied Mathematics, Philadelphia, 2003.

[Leclerc and Monjardet, 1995]B. Leclerc and B. Monjardet. Latticial theory of consensus. In W. A. Barnett, H. Moulin, M. Salles, and N. J. Schofield, editors, *Social Choice, Welfare, and Ethics: Proceedings of the Eighth International Symposium in Economic Theory and Econometrics*, International Symposia in Economic Theory and Econometrics, chapter 6, pages 145–160. Cambridge University Press, Cambridge, 1995.

[Margush and McMorris, 1981]T. Margush and F. R. McMorris. Consensus n-trees. *Bulletin of Mathematical Biology*, 43(2):239–244, 1981.

[McLean, 1995]I. McLean. Independence of irrelevant alternatives before Arrow. *Mathematical Social Sciences*, 30(2):107–126, October 1995.

[McMorris and Neumann, 1983]F. R. McMorris and D. A. Neumann. Consensus functions defined on trees. *Mathematical Social Sciences*, 4(2):131–136, April 1983.

[McMorris and Powers, 1993]F. R. McMorris and R. C. Powers. Consensus functions on trees that satisfy an independence axiom. *Discrete Applied Mathematics*, 47(1):47–55, 16 November 1993.

[McMorris *et al.*, 2000]F. R. McMorris, H. M. Mulder, and R. C. Powers. The median function on median graphs and semilattices. *Discrete Applied Mathematics*, 101(1–3):221–230, 15 April 2000.

[McMorris *et al.*, 2003]F. R. McMorris, H. M. Mulder, and R. C. Powers. The median function on distributive semilattices. *Discrete Applied Mathematics*, 127(2):319–324, 2003.

[Mirkin, 1975]B. G. Mirkin. On the problem of reconciling partitions. In H. M. Blalock, A. Aganbegian, F. M. Borodkin, R. Boudon, and V. Capecchi, editors, *Quantitative Sociology: International Perspectives on Mathematical and Statistical Modelling*, Quantitative Studies in Social Relations, chapter 15, pages 441–449. Academic Press, New York, 1975.

[Monjardet, 1990]B. Monjardet. Arrowian characterizations of latticial federation consensus functions. *Mathematical Social Sciences*, 20(1):51–71, August 1990.

[Steel *et al.*, 2000]M. A. Steel, A. W. M. Dress, and S. Böcker. Simple but fundamental limitations on supertree and consensus tree methods. *Systematic Biology*, 49(2):363–368, June 2000.

[Wilkinson *et al.*, 2004]M. Wilkinson, J. L. Thorley, D. Pisani, F. J. Lapointe, and J. O. McInerney. Some desiderata for liberal supertrees. In O. R. P. Bininda-Emonds, editor, *Phylogenetic Supertrees: Combining Information to Reveal the Tree of Life*, number 4 in Computational Biology, chapter 10, pages 227–246. Kluwer Academic Publishers, Boston, Massachusetts, 2004.

GARCH Options in Incomplete Markets

Giovanni Barone-Adesi[1], Robert Engle[2], and Loriano Mancini[1]

[1] Institute of Finance, University of Lugano,
Via Buffi 13, CH-6900 Lugano Switzerland
Tel: +41 (0)91 912 47 53, Fax: +41 91 912 46 47
(e-mail: BaroneG@lu.unisi.ch, Loriano.Mancini@lu.unisi.ch)

[2] Dept. of Finance, Leonard Stern School of Business, New York University
(e-mail: REngle@stern.nyu.edu)

Abstract. We propose a new method to compute option prices based on GARCH models. In an incomplete market framework, we allow for the volatility of asset return to differ from the volatility of the pricing process and obtain adequate pricing results. We investigate the pricing performance of this approach over short and long time horizons by calibrating theoretical option prices under the Asymmetric GARCH model on S&P 500 market option prices. A new simplified scheme for delta hedging is proposed.

Keywords: GARCH models.

Introduction

There is a general consensus that asset returns exhibit variances that change through time. GARCH models are a popular choice to model these changing variances. However the success of GARCH in modelling return variance hardly extends to option pricing. Models by [Duan, 1995], [Heston, 1993] and [Heston and Nandi, 2000] impose that the conditional volatility of the risk-neutral and the objective distributions be the same. Total variance, (the expectation of the integral of return variance up to option maturity), is then the expected value under the GARCH process. Empirical tests by [Chernov and Ghysels, 2000], (see also references therein), find that the above models do not price options well and their hedging performance is worse than Black-Scholes calibrated at the implied volatility of each option.

A common feature of all the tests to date is the assumption that the volatility of asset return is equal to the volatility of the pricing process. In other words, a risk neutral investor prices the option as if the distribution of its return had a different drift but unchanged volatility. This is certainly a tribute to the pervasive intellectual influence of the [Black and Scholes, 1973] model on option pricing. However, Black and Scholes derived the above property under very special assumptions, (perfect complete markets, continuous time and price processes). Changing volatility in real markets makes the perfect replication argument of Black-Scholes invalid. Markets are then incomplete in the sense that perfect replication of contingent claims using only the underlying asset and a riskless bond is impossible. Of course markets become complete if a sufficient, (possibly infinite), number of contingent claims are available. In this case a well-defined pricing density exists.

In the markets we consider the volatility of the pricing process is different from the volatility of the asset process. This occurs because investors will set state prices to reflect their aggregate preferences. The pricing distribution will then be different from the return distribution. It is possible then to calibrate the pricing process directly on option prices. Although this may appear to be a purely fitting exercise, involving no constraint beyond the absence of arbitrage, verification of the stability of the pricing process over time and across maturities imposes substantial parameter restrictions. Economic theory may impose further restrictions from investors' preferences for aggregate wealth in different states.

[Carr *et al.*, 2003] propose a similar set-up for Lévy processes. They use a jump process in continuous time. We propose to use discrete time and a continuous distribution for prices. Moreover we use GARCH models to drive stochastic volatility.

[Heston and Nandi, 2000] derived a quasi-analytical pricing formula for European options assuming a parametric linear risk premium, Gaussian innovations and the same GARCH parameters for the pricing and the asset process. In our pricing model we relax their assumptions. We allow for different volatility processes and time-varying, nonparametric risk premia—set by aggregate investors' risk preferences. We use not only Monte Carlo simulation, but also filtered GARCH innovations.

Our method is different from [Duan, 1996], where a GARCH model is calibrated to the FTSE 100 index options assuming Gaussian innovations and the locally risk neutral valuation relationship, which implies that the conditional variance returns are equal under the objective and the risk neutral measures. [Engle and Mustafa, 1992] proposed a similar method to calibrate a GARCH model to S&P 500 index options in order to investigate the persistence of volatility shocks.

The final target is the identification of a pricing process for options that provides an adequate pricing performance. A surprising result concerns hedging performance. Hedging performance, contrary to what is commonly sought in the stochastic volatility literature, cannot be significantly better than the performance of the Black-Scholes model calibrated at the implied volatility for each option. This result stems from the fact that deltas, (hedge ratios), for Black-Scholes can be derived applying directly the (first degree) homogeneity of option prices with respect to asset and strike prices, without using the Black-Scholes formulas. Therefore, hedge ratios from Black-Scholes calibrated at the implied volatility are the "correct" hedge ratios unless a very strong departure from "local homogeneity" occurs. This is not the case for the continuous, almost linear volatility smiles commonly found. In practice, for regular calls and puts, this is the case only for the asset price being equal to the strike price one instant before maturity. In summary, although it may be argued that calibrating Black-Scholes at each implied volatility does not give a model of option pricing, the hedging performance of this common procedure is almost unbeatable. [Barone-Adesi and Elliott, 2004] further investigate the computation of the hedge ratios under similar assumptions.

Our tests use closing prices of European options on the S&P 500 Index over several months. After estimating a GARCH model from earlier S&P 500 index data we search in a neighborhood of this model for the best pricing performance. Care is taken to prevent that our results be driven by microstructure effects in illiquid options.

The structure of the paper is the following. Section 1 presents option and state prices under GARCH models when the pricing process is driven by simulated, Gaussian innovations. Section 2 investigates the pricing performance of the proposed method when the pricing process is driven by filtered, estimated GARCH innovations. Section 3 discusses hedging results and Section 4 concludes.

1 Option and State Prices under the GARCH Model

Consider a discrete-time economy. Let S_t denote the closing price of the S&P 500 index at day t and y_t the daily log-return, $y_t := \ln(S_t/S_{t-1})$. Suppose that under the objective or historical measure $\mathbb{P}$, y_t follows an Asymmetric GARCH(1,1) model; see [Glosten *et al.*, 1993],

$$\begin{aligned} y_t &= \mu + \varepsilon_t, \\ \sigma_t^2 &= \omega + \alpha\varepsilon_{t-1}^2 + \beta\sigma_{t-1}^2 + \gamma I_{t-1}\varepsilon_{t-1}^2, \end{aligned} \tag{1}$$

where $\omega,\ \alpha,\ \beta > 0$, $\alpha + \beta + \gamma/2 < 1$, μ determines the constant return (continuously compounded) of S_t, $\varepsilon_t = \sigma_t\, z_t$, $z_t \sim i.i.d.(0,1)$ and $I_{t-1} = 1$, when $\varepsilon_{t-1} < 0$ and $I_{t-1} = 0$, otherwise. The parameter $\gamma > 0$ accounts for the "leverage effect", that is the stronger impact of "bad news" ($\varepsilon_{t-1} < 0$) rather than "good news" ($\varepsilon_{t-1} \geq 0$) on the conditional variance σ_t^2.

The representative agent in the economy is an expected utility maximizer and the utility function is time-separable and additive. At time $t = 0$, the following Euler equation from the standard expected utility maximization argument gives the price of a contingent T-claim ψ_T,

$$\begin{aligned} \psi_0 &= E_{\mathbb{P}}[\psi_T\, U'(C_T)/U'(C_0)|\mathcal{F}_0] = E_{\mathbb{P}}[\psi_T\, Y_{0,T}|\mathcal{F}_0] \\ &= E_{\mathbb{Q}}[\psi_T\, e^{-rT}|\mathcal{F}_0], \end{aligned}$$

where $E_{\mathbb{G}}[\cdot]$ denotes the expectation under the measure $\mathbb{G}$, r is the risk-free rate, $U'(C_t)$ is the marginal utility of consumption at time t and $\mathcal{F}_t$ is the information set available up to and including time t. The state price density per unit probability process Y is defined by $Y_{t,T} := e^{-r(T-t)}L_t$ and

$$L_t = \frac{d\,\mathbb{Q}_t}{d\,\mathbb{P}_t} = \frac{q\,dS}{p\,dS} = \frac{q}{p},$$

where $\mathbb{Q}$ is the risk neutral measure absolutely continuous with respect to $\mathbb{P}$, the subindex t denotes the restriction to $\mathcal{F}_t$, q and p (time subscripts are omitted) are the corresponding density functions. When the financial market is incomplete, L_t is not unique and is determined by the representative agent's preferences. Intuitively, if $p(\overline{S}_T)$ was a discrete probability, the state price density evaluated at $\overline{S}_T$, $Y_{t,T}(\overline{S}_T)\,p(\overline{S}_T)$, gives at time t the price of \$1 to be received if state $\overline{S}_T$ occurs. The state price per unit probability, $Y_{t,T}(\overline{S}_T)$, is then the market price of a state contingent claim that pays $1/p(\overline{S}_T)$ if state $\overline{S}_T$, which has probability $p(\overline{S}_T)$, occurs. The expected rate of return of such a claim under the physical measure $\mathbb{P}$ is $1/Y_{t,T}(\overline{S}_T) - 1$. As marginal utilities of consumptions decrease when the states of the world "improve", $Y_{t,T}$ is expected to decrease in S_T.

1.1 Monte Carlo Option Prices

Monte Carlo simulation is used to compute the GARCH option prices, because the distribution of temporally aggregated asset returns cannot be derived analytically. We present the computation of a European call option price; other European claims can be priced similarly.

At time $t = 0$ the dollar price of a European call option with strike price $\$K$ and time to maturity T days is computed by simulating log-returns in model (1) under the risk neutral measure $\mathbb{Q}$. Specifically, we draw T independent standard normal random variables $(z_i^\star)_{i=1,\dots,T}$, we simulate (y_i, σ_i^2) in model (1) under the risk neutral parameters $\omega^*, \alpha^*, \beta^*$, γ^*, $\mu = r - d - \sigma_i^2/2$, where r is the risk-free rate and d is the dividend yield on a daily basis, and we compute $S_T^{(n)} = S_0 \exp(\sum_{i=1}^T y_i)$. Then, we compute the discounted call option payoff $C^{(n)} = \exp(-r\,T)\max(0, S_T^{(n)} - K)$. Iterating the procedure N times gives the Monte Carlo estimate for the call option price, $C_{mc}(K,T) := N^{-1}\sum_{n=1}^N C^{(n)}$. To reduce the variance of the Monte Carlo estimates we use the method of antithetic variates; cf., for instance, [Boyle *et al.*, 1997]. Specifically, $C^{(n)} = (C_a^{(n)} + C_b^{(n)})/2$, where $C_a^{(n)}$ is computed using $(z_i^\star)_{i=1,\dots,T}$ and $C_b^{(n)}$ using $(-z_i^\star)_{i=1,\dots,T}$. Each option price C_{mc} is computed simulating $2N$ sample paths for S. In our calibration exercises we set $N = 10{,}000$. To further reduce the variance of the Monte Carlo estimates we calibrate the mean as in the empirical martingale simulation method proposed by [Duan and Simonato, 1998]. Scaling the simulated values $S_T^{(n)}$, $n = 1,\dots,N$, by a multiplicative factor, the method ensures that the risk neutral expectation of the underlying asset is equal to the forward price, i.e. $N^{-1}\sum_{n=1}^N \tilde{S}_T^{(n)} = S_0\exp((r-d)T)$, where $\tilde{S}_T^{(n)} := S_T^{(n)} S_0 \exp((r-d)T)\,(N^{-1}\sum_{n=1}^N S_T^{(n)})^{-1}$. Then, option prices are computed using $\tilde{S}_T^{(n)}$. In our calibration exercises at least 100 simulated paths of the underlying asset end at maturity "in the money" for almost all the deepest out of the money options.

1.2 Calibration of the GARCH Model

The risk neutral parameters of the GARCH model, $\theta^* = (\omega^* \, \alpha^* \, \beta^* \, \gamma^*)$, are determined by calibrating GARCH option prices computed by Monte Carlo simulation on market option prices taken as averages of bid and ask prices at the end of one day.

Specifically, let $P^{mkt}(K, T)$ denote the market price in dollars at time $t = 0$ of a European option with strike price $\$K$ and time to maturity T days. The risk neutral parameters θ^* are determined by minimizing the mean squared error (mse) between model option prices and market prices. The mse is taken over all strikes and maturities,

$$\theta^* := \arg\min_{\theta} \sum_{i=1}^{m} \left(P^{garch}(K_i, T_i; \theta) - P^{mkt}(K_i, T_i) \right)^2 , \qquad (2)$$

where $P^{garch}(K, T; \theta)$ is the theoretical GARCH option price and m is the number of European options considered for the calibration at time $t = 0$.

As an overall measure of the quality of the calibration we compute the average absolute pricing error (ape) with respect to the mean price,

$$\text{ape} := \frac{\sum_{i=1}^{m} \left| P^{garch}(K_i, T_i; \theta^*) - P^{mkt}(K_i, T_i) \right|}{\sum_{i=1}^{m} P^{mkt}(K_i, T_i)} . \qquad (3)$$

1.3 Empirical Results

We calibrate the GARCH model to European options on the S&P 500 index observed on a random date $t :=$ August 29, 2003 and we set $t = 0$. Estimates of σ_0^2 and z_0 are necessary to simulate the risk neutral GARCH volatility and are obtained in the next section.

1.3.1 Estimation of the GARCH Model Percentage daily log-returns, $y_t \times 100$, of the S&P 500 index are computed from December 11, 1987 to August 29, 2003 for a total of 4,100 observations. Model (1) is estimated using the Pseudo Maximum Likelihood (PML) estimator based on the nominal assumption of conditional normal innovations. The parameter estimates are reported in Table 1. The current August 29, 2003 estimates on a daily base of σ_0^2 and z_0 are 0.635 and 0.604, respectively, and will be used as starting values to simulate the risk neutral GARCH volatility in the calibration exercise.

1.3.2 Calibration of the GARCH Model with Gaussian Innovations Initially we calibrate the GARCH model (1) to the closing prices (bid-ask averages) of out of the money European put and call options on the S&P 500 index observed on August 29, 2003. Precisely, we only consider option prices strictly larger than \$0.05—discarding 40 option prices to avoid that our results be driven by microstructure effects in very illiquid options—and maturities $T = 22$, 50, 85, 113 days for a total of $m = 118$ option prices. Strike prices range from \$550 to \$1,250, $r = 0.01127/365$, $d = 0.01634/365$ on a daily basis and $S_0 = \$1{,}008$.

To solve the minimization problem (2) we use the Nelder-Mead simplex direct search method implemented in the Matlab function `fminsearch`. This function does not require the computation of gradients. Starting values for the risk neutral parameters θ^* are the parameter estimates given in Table 1. Calibrated parameters, root mean squared error (rmse) and ape measure for the quality of the calibration are reported in the first row of Table 2. The "leverage effect" in the volatility process under the risk neutral measure $\mathbb{Q}$ ($\gamma^* = 0.288$) is substantially larger than under the objective measure $\mathbb{P}$ ($\gamma = 0.075$). The average pricing error is quite low and equals to 2.54%. Figure 1 shows the pricing performance of the GARCH model which seems to be satisfactory. Figure 2 shows the calibration errors defined as $P^{garch} - P^{mkt}$. Such errors tend to be larger for near at the money options (these options have the largest prices) and for deep out of the money put options.

1.3.3 State Price Density Estimates with Gaussian Innovations

For the maturities $T = 22$, 50, 85, 113 days we compute the state price densities per unit probability of S_T, $Y_{0,T}$, as the discounted ratio of the risk neutral density over the objective density. Under the objective measure $\mathbb{P}$, the asset prices S are simulated assuming the drift $\mu = r + 0.08/365 - \sigma_t^2/2$ in equation (1) and the parameter estimates in Table 1. Under the risk neutral measure $\mathbb{Q}$, $\mu = r - d - \sigma_t^2/2$ and the calibrated GARCH parameters are given in the first row of Table 2. The density functions are estimated by the Matlab function `ksdensity` using the Gaussian kernel and the optimal default bandwidth for estimating Gaussian densities.

Figure 3 shows the estimated risk neutral and objective densities and the corresponding state price densities per unit probability; see also Table 3. As expected the state price densities are quite stable across maturities and monotonic, decreasing in S_T. However, the high values on the left imply very negative expected rate of return for out of the money puts, that appear intuitively "overpriced". As an example, a state price per unit probability of \$6 corresponds to an expected rate of return of $1/6 - 1 = -0.833$ for a simple state contingent claim. State price densities outside the reported values for S_T tend to be unstable, as the density estimates are based on very few observations.

2 GARCH Option Prices with Filtering Historical Simulations

In this section we investigate the pricing performance of the GARCH model when the simulated, Gaussian innovations—used to drive the GARCH process under the risk neutral measure—are replaced by historical, estimated GARCH innovations. We refer to this approach as the Filtering Historical Simulation (FHS) method. [Barone-Adesi *et al.*, 1998] introduced the FHS method to estimate portfolio risk measures.

This procedure is in two steps. Suppose we aim at calibrating the GARCH model on market option prices $P^{mkt}(K_i, T_i)$, $i = 1, \ldots, m$ observed on day $t := 0$. In the first step, the GARCH model is estimated on the historical log-returns of the underlying asset $y_{-n+1}, y_{-n+2}, \ldots, y_0$ up to time $t = 0$. The scaled innovations of the GARCH process $\hat{z}_t = \hat{\varepsilon}_t \, \hat{\sigma}_t^{-1}$, for $t = -n+1, \ldots, 0$, are also estimated.

In the second step, the GARCH model is calibrated to the market option prices by solving the minimization problem (2). The theoretical GARCH option prices, $P^{garch}(K, T; \theta^*)$, are computed by Monte Carlo simulations as in Section 1.1, but the Gaussian innovations are replaced by innovations $\hat{z}_t$'s estimated in the first step, randomly drawn with uniform probabilities. To preserve the negative skewness of the estimated innovations the method of the antithetic variates is not used.

2.1 Calibration of the GARCH model with FHS Innovations

We apply this two steps procedure to the option prices on the S&P 500 observed on a random date July 9, 2003. Specifically, in the first step we estimate the GARCH model (1) on $n = 3{,}800$ historical returns of the S&P 500 index from December 14, 1988 to July 9, 2003 and we estimate the corresponding innovations $\hat{z}$. In the second step, we calibrate the GARCH model to the out of the money put and call options with maturities $T = 10$, 38, 73, 164, 255, 346 days for a total of $m = 151$ option prices; 45 options with bid price lower than \$0.05 are discarded. The PML estimates of model (1) are reported in Table 4. The last panel in Figure 4 shows the estimated scaled innovations, $\hat{z}_t$'s, used to drive the GARCH process under the risk neutral measure. The skewness and the kurtosis of the empirical distribution of $\hat{z}$ are -0.6 and 7.4, respectively. Calibration results are reported in the first row of Table 5 and Figure 5. The average pricing error is 3.5% and the overall pricing performance is quite satisfactory given the wide range of strikes and maturities of the options used for the calibration.

We calibrate the GARCH model using the FHS method also on the same options considered in the calibration for August 29, 2003. The results are reported in the second row of Table 2. Given the limited number of options used in this calibration, the GARCH pricing model with Gaussian innovation has already a very low pricing error. However, using the FHS method both the rmse and the ape measure are reduced by about 10%. The asymmetry parameter γ^* decreases from 0.288 to 0.201 when filtered, estimated innovations rather than Gaussian innovations are used, because of the negative skewness, -0.61, of the filtered innovations.

2.2 State Price Density Estimates with FHS Innovations

The state price densities per unit probability on July 9, 2003, computed similarly as in Section 1.3.3, are shown in Figure 6. Using FHS innovations, the asymmetry parameter γ^* is now very close to γ (cf. Tables 4–5) and state prices per unit probability are still monotone, but much closer to each other. In particular the state prices per unit probability on the left are now in line with the remaining ones. This implies that "excess" out of the money put prices can be explained by the skewness of FHS innovations. The volatility smile—computed using out of the money European put and call options—for 38 days to maturity on this date is reported in Figure 7. Notice that the sample period to estimate the GARCH model (1) starts after the October 1987 crash. Such a large negative return would inflate the variance estimates and this tends to produce non monotone state price densities per unit probability.

The state price densities per unit probability on August 29, 2003 using the FHS method are quite close to those on July 9, 2003 and are omitted.

2.3 Short Run Stability of the GARCH Pricing Model

To investigate the stability of the pricing performance for the GARCH model over a "short" time horizon, i.e. one month, we calibrate the model for several dates from July 9 to August 8, 2003 on out of the money European option prices with maturities less than a year. The calibration results are reported in Table 5. The GARCH parameters tend to change over time, but the pricing performances are quite stable in terms of rmse and ape measures. Moreover, the estimates of the long run level of the risk neutral variance $E_{\mathbb{Q}}[\sigma^2]$ are quite stable and about 1% on a daily base.

To check for the stability of the GARCH parameters we calibrate one GARCH model to the option prices on July 9, 10, 11, and 14, 2003. The initial variances and innovations, σ_0^2's and z_0's, for the dates July 10, 11, 14 are computed updating the corresponding estimates for July 9, i.e. 0.793, -0.667, and using the objective GARCH estimates in Table 4. This procedure ensures that future, not yet available information is not used for the fitting of earlier option prices. The GARCH parameter of the "pooled" calibration are $\omega^*_{pool} = 0.016$, $\alpha^*_{pool} = 0.000$, $\beta^*_{pool} = 0.924$, $\gamma^*_{pool} = 0.121$, which imply a long run level of the risk neutral variance $E_{\mathbb{Q}}[\sigma^2_{pool}] = 0.99$. Table 6 compares the pricing errors—the differences between theoretical and observed option prices—of the pooled calibration with the corresponding errors for the single day calibration given in Table 5. As expected the rmse's for the pool calibration are larger than the corresponding rmse's for the single day calibrations. However, differences are small and the correlation between the two pricing errors is on average 0.92, meaning that the two pricing performances are quite close.

2.4 Long Run Stability of the GARCH Pricing Model and Comparison with CGMYSA Model

To investigate the pricing performance of the GARCH model over a "long" time horizon, i.e. one year, we calibrate the model on out of the money European option prices with maturities between a month and a year for the dates January 12, March 8, May 10, July 12, September 13 and November 8 for the year 2000. For each calibration we use about the last seven years of S&P 500 daily log-returns to implement the FHS method. We also compare the pricing performance of the GARCH model with the CGMYSA model proposed by [Carr *et al.*, 2003] for the dynamic of the underlying asset, which is a mean corrected, exponential Lévy process time changed with a Cox, Ingersoll and Ross process. Average absolute pricing errors are somewhat in favour of the CGMYSA model as this model has nine parameters while the GARCH model has four parameters. The results are reported in Table 7. There is evidence that the GARCH parameters tend to change from month to month, but the pricing performance is quite stable especially in terms of the ape measure. Moreover, the mean and the standard deviation of the ape measures for the GARCH model are 4.07, 1.03 and for the CGMYSA model are 3.91, 1.17, respectively. Hence, the pricing performance of the GARCH model is more stable than the pricing performance of the CGMYSA model, but the last model is superior in terms of average ape measure. [Carr *et al.*, 2003] proposed also more parsimonious (six parameters) models, namely the VGSA and NIGSA models, which are, respectively, finite variation and infinite variation mean corrected, exponential Lévy processes with infinite activity for the underlying asset. For the previous dates, the GARCH model outperforms the VGSA and NIGSA models in five and four out of six cases, respectively.

3 Hedging

Extension to the GARCH setting of the delta hedging, [Engle and Rosenberg, 2002], does not show an improvement on the delta hedging strategy based on the Black-Scholes model calibrated at the implied volatility. To understand why this is the case consider the example presented in Table 8. The three rows in the middle are market option prices from Hull's book. The first row is obtained multiplying the middle row times 0.9 and the last row is obtained multiplying the middle row by 1.1, that is assuming an homogeneous pricing model.

Incremental ratios, that is change in option price over change in stock price, can be computed between the first two and then again the last two rows, i.e. $\Delta_{45} := (5.60 - 2.16)/(49 - 44.1)$ and $\Delta_{55} := (2.64 - 1.00)/(53.9 - 49)$. Taking the average of these two ratios, for the strike price $K = 50$ we obtain an estimate of delta equals to 0.518, which is almost identical to the delta from the Black-Scholes model calibrated at the implied volatility for the middle row, i.e. 0.522—the implied volatility is equal to 0.2 when $r = 0.05$ and $T = 20/52$ years. Hence, the application of first-degree homogeneity to non-homogeneous prices has led to an essentially correct hedge ratio! To understand this paradoxical result consider the sources of errors in the above computations. There is a discretization error and an error due to the volatility smile. In fact, in the absence of a volatility smile, Black-Scholes option prices would be homogeneous functions of the stock and the strike price. The discretization error leads to a discrete delta which is approximately the average of the Black-Scholes deltas computed at the two extremes of each interval and approximated by Δ_{45} and Δ_{55}. Formally, denote by $\Delta(K)$ the delta as a function of the strike price K. For small intervals the delta hedge is approximated by

$$\Delta(50) \approx \frac{\Delta(50) + \overbrace{\Delta'(50)(45-50)}^{>0} + \Delta(50) + \overbrace{\Delta'(50)(55-50)}^{<0}}{2} \approx \frac{\Delta_{45} + \Delta_{55}}{2}.$$

Therefore, the two discrete ratios considered, Δ_{45} and Δ_{55}, are affected by opposite errors up to the first order. Taking their average eliminates these errors. The only error left is due to the smile effect. However, this error is very small if the strike price increment is small relative to the asset price and its volatility. See [Barone-Adesi and Elliott, 2004] for further discussion. The reader may verify this simple result on the options of his choice. It appears therefore that deltas are to a large degree determined by market option prices, independently of the chosen model. Therefore, models alternative to Black-Scholes calibrated at the implied volatility will generally lead to very similar hedge ratios, if they fit well market prices. The only significant deterioration of hedging occurs in the presence of large volatility shocks, which diminish the effectiveness of delta hedging. To observe this compare a day with a modest change in volatility, e.g. $t_2 :=$ July 10, 2003, with a day in which a large negative index return led to a large increase in volatility, e.g. $t_1 :=$ January 24, 2003. Specifically, for the day t_1 we consider out of the money put and call options with maturities equal to 30, 58, 86, 149, 240, 331 days for a total of 160 option prices and for the day t_2 we consider the same options as in Section 2.3. Then, we run the following set of regression for $t+1 = t_1$, t_2

1) $P^{mkt}_{t+1} = \eta_0 + \eta_1 P^{bs}_{t,t+1} + error,$
2) $P^{mkt}_{t+1} = \eta_0 + \eta_1 P^{bs}_{t,t+1} + \eta_2 P^{garch}_{t,t+1} + error,$
3) $P^{mkt}_{t+1} = \eta_0 + \eta_2 P^{garch}_{t,t+1} + error,$

where P^{mkt}_{t+1} are the option prices observed on time $t+1$, $P^{bs}_{t,t+1}$ are the Black-Scholes forecasts of option prices for $t+1$ computed by plugging in the Black-Scholes formula S_{t+1}, r, d at time $t+1$ and the implied volatility observed on time t (i.e. January 23 and July 9, 2003, respectively). $P^{garch}_{t,t+1}$ are the GARCH forecasts obtained using S_{t+1}, the GARCH parameter calibrated at time t and σ_{t+1} updated according to the objective estimates at time t.

The ordinary least square (OLS) estimates of the previous regressions are given in Table 9. In terms of the error variance the Black-Scholes forecasts in regressions 1) are superior to the GARCH forecasts in regressions 3) for both days t_1 and t_2. Moreover, in the regressions 2) the weights η_1 of the Black-Scholes forecasts are larger than the weights η_2 for the GARCH forecasts. This is due to the "initial advantage" of the Black-Scholes forecasts, i.e. the zero pricing error at time t. However, for the day January 24, 2003, from regression 1) to regression 2) the variance of the prediction error is reduced about 60% adding the GARCH forecast as a regressor. Hence, the GARCH model carries on large amount of information on option price dynamics. Moreover, the GARCH model provides a dynamic model for the risk neutral volatility, while the Black-Scholes model does not.

Interestingly, the Black-Scholes forecasts tend to underestimate option prices observed on January 24, 2003 (while the GARCH forecasts tend to overestimate option prices). An explanation is the following. The daily log-return of the S&P 500 for January 24, 2003 is −2.97%, which induces an increase in the volatility of the underlying asset. Such an increase in the volatility can not be detected by the Black-Scholes model with constant implied volatility, but it is reflected in the GARCH forecasts of volatilities and option prices. This effect is stronger in days with large returns. For the day July 10, 2003 the reduction in the variance of the prediction error is only 11%, as the return of the S&P 500 is −1.36% only.

Unfortunately, our GARCH price forecast is conditioned on the current index and it cannot be used to improve significantly delta hedging. Its explanatory power simply indicates that delta hedging is less effective in the presence of large volatility shocks. They are linked to the index return in a nonlinear fashion in the GARCH model.

4 Conclusion

Casting the option pricing problem in incomplete markets allows for more flexibility in the calibration of market prices. Investors' preferences can be inferred comparing the physical and the pricing distributions. Using filtered historical simulation the volatility smile appears to be explained by innovation skewness, with no need of much higher state prices for out of the money puts. Delta hedging does not require a large computational effort under conditions usually found in index option markets, removing a major drawback of simulation-based option pricing. Further refinements of pricing and stability issues are left to future research.

Acknowledgement

Giovanni Barone-Adesi and Loriano Mancini gratefully acknowledge the financial support of the Swiss National Science Foundation (NCCR FINRISK).

References

[Barone-Adesi and Elliott, 2004]Giovanni Barone-Adesi and Robert J. Elliott. Cutting the hedge. Working paper, 2004.

[Barone-Adesi *et al.*, 1998]Giovanni Barone-Adesi, Frederick Bourgoin, and Kostas Giannopoulos. Don't look back. *Risk*, 11:100–103, 1998.

[Black and Scholes, 1973]Fisher Black and Myron Scholes. The valuation of options and corporate liabilities. *Journal of Political Economy*, 81:637–654, 1973.

[Boyle *et al.*, 1997]Phelim Boyle, Mark Broadie, and Paul Glassermann. Monte carlo methods for security pricing. *Journal of Economic Dynamics and Control*, 21:1267–1321, 1997.

[Carr *et al.*, 2003]Peter Carr, Hélyette Geman, Dilip B. Madan, and Marc Yor. Stochastic volatility for lévy processes. *Mathematical Finance*, 13:345–382, 2003.

[Chernov and Ghysels, 2000]Mikhail Chernov and Eric Ghysels. A study towards a unified approach to the joint estimation of objective and risk neutral measures for the purpose of options valuation. *Journal of Financial Economics*, 56:407–458, 2000.

[Duan and Simonato, 1998]Jin-Chuan Duan and Jean-Guy Simonato. Empirical martingale simulation for asset prices. *Management Science*, 44:1218–1233, 1998.

[Duan, 1995]Jin-Chuan Duan. The garch option pricing model. *Mathematical Finance*, 5:13–32, 1995.

[Duan, 1996]Jin-Chuan Duan. Cracking the smile. *Risk*, 9:55–59, 1996.

[Engle and Mustafa, 1992]Robert F. Engle and Chowdhury Mustafa. Implied arch models from options prices. *Journal of Econometrics*, 52:289–311, 1992.

[Engle and Rosenberg, 2002]Robert F. Engle and Joshua V. Rosenberg. Empirical pricing kernels. *Journal of Financial Economics*, 64:341–372, 2002.

[Glosten *et al.*, 1993]Lawrence R. Glosten, Ravi Jagannathan, and David E. Runkle. On the relation between the expected value and the volatility of the nominal excess return on stocks. *Journal of Finance*, 48:1779–1801, 1993.

[Heston and Nandi, 2000]Steven Heston and Saikat Nandi. A closed-form garch option valuation model. *Review of Financial Studies*, 13:585–625, 2000.

[Heston, 1993]Steven Heston. A closed-form solution for options with stochastic volatility, with applications to bond and currency options. *Review of Financial Studies*, 6:327–343, 1993.

Table 1. PML estimates of the GARCH model (1), $y_t \times 100 = \mu + \varepsilon_t$, $\sigma_t^2 = \omega + \alpha\varepsilon_{t-1}^2 + \beta\sigma_{t-1}^2 + \gamma I_{t-1}\varepsilon_{t-1}^2$, $I_{t-1} = 1$ when $\varepsilon_{t-1} < 0$ and $I_{t-1} = 0$ otherwise, $\varepsilon_t = \sigma_t\, z_t$, $z_t \sim i.i.d.(0,1)$, (p-values in parenthesis) for the S&P 500 index daily log-returns y_t in percentage from December 11, 1987 to August 29, 2003.

μ	ω	α	β	γ
0.033	0.009	0.006	0.946	0.075
(0.008)	(0.000)	(0.416)	(0.000)	(0.000)

Table 2. Calibrated parameters of the GARCH model (1), $\sigma_t^2 = \omega^* + \alpha^*\varepsilon_{t-1}^2 + \beta^*\sigma_{t-1}^2 + \gamma^* I_{t-1}\varepsilon_{t-1}^2$, $I_{t-1} = 1$ when $\varepsilon_{t-1} < 0$ and $I_{t-1} = 0$ otherwise, $\varepsilon_t = \sigma_t\, z_t$, $z_t \sim i.i.d.(0,1)$, using Gaussian innovations (first row) and FHS method (second row) on August 29, 2003 out of the money European put and call options ($m = 118$) and time to maturities $T = 22$, 50, 85, 113 days. The root mean squared error (rmse) is in $, the ape measure is defined in equation (3).

	ω^*	α^*	β^*	γ^*	rmse	ape%
Gauss. z	0.037	0.000	0.833	0.288	0.27	2.54
FHS	0.037	0.000	0.870	0.201	0.24	2.29

Table 3. State price densities estimates per unit of probability, $Y_{0,T}$, time to maturities $T = 22$, 50, 85, 113 days for August 29, 2003. $Y_{0,T} := e^{-rT}L_0$ and $L_0 = d\mathbb{Q}_0/d\mathbb{P}_0$, where $\mathbb{Q}$ is the risk neutral measure absolutely continuous with respect to the objective measure $\mathbb{P}$ and the subindex $t = 0$ denotes the restriction to $\mathcal{F}_0$.

S_T	900	1,000	1,100	1,200
$Y_{0,22}$	1.882	1.001	0.437	—
$Y_{0,50}$	1.284	1.011	0.773	0.254
$Y_{0,85}$	1.197	1.003	0.844	0.597
$Y_{0,113}$	1.281	1.028	0.834	0.641

Table 4. PML estimates of the GARCH model (1), $y_t \times 100 = \mu + \varepsilon_t$, $\sigma_t^2 = \omega + \alpha\varepsilon_{t-1}^2 + \beta\sigma_{t-1}^2 + \gamma I_{t-1}\varepsilon_{t-1}^2$, $I_{t-1} = 1$ when $\varepsilon_{t-1} < 0$ and $I_{t-1} = 0$ otherwise, $\varepsilon_t = \sigma_t\, z_t$, $z_t \sim i.i.d.(0,1)$, (p-values in parenthesis) for the S&P 500 index daily log-returns y_t in percentage from December 14, 1988 to July 9, 2003.

μ	ω	α	β	γ
0.033	0.012	0.005	0.936	0.093
(0.008)	(0.000)	(0.547)	(0.000)	(0.000)

Table 5. Calibrated parameters of the GARCH model (1), $\sigma_t^2 = \omega^* + \alpha^* \varepsilon_{t-1}^2 + \beta^* \sigma_{t-1}^2 + \gamma^* I_{t-1} \varepsilon_{t-1}^2$, $I_{t-1} = 1$ when $\varepsilon_{t-1} < 0$ and $I_{t-1} = 0$ otherwise, $\varepsilon_t = \sigma_t \, z_t$, $z_t \sim i.i.d.(0,1)$, under the risk neutral measure $\mathbb{Q}$, using FHS on several days and m out of the money European put and call options. T is the time to maturity in days. The root mean squared error (rmse) is in \$, the ape measure is defined in equation (3).

date	ω^*	α^*	β^*	γ^*	$E_{\mathbb{Q}}[\sigma^2]$	m	$\min(T)$	$\max(T)$	rmse	ape%
Jul 9	0.019	0.000	0.912	0.138	1.00	151	10	346	0.64	3.50
Jul 10	0.008	0.000	0.953	0.078	1.00	148	9	345	0.49	2.75
Jul 11	0.016	0.000	0.921	0.125	0.98	146	8	344	0.64	3.64
Jul 14	0.009	0.000	0.949	0.083	0.96	146	5	341	0.43	2.33
Jul 16	0.011	0.000	0.946	0.086	1.00	141	3	339	0.67	3.59
Jul 21	0.005	0.000	0.964	0.061	0.86	156	26	334	0.94	3.61
Jul 25	0.054	0.000	0.787	0.319	1.03	165	22	330	0.69	4.24
Jul 30	0.010	0.000	0.943	0.092	0.97	161	17	325	0.40	2.26
Aug 1	0.022	0.000	0.912	0.137	1.12	163	15	323	0.59	3.38
Aug 4	0.016	0.000	0.928	0.117	1.21	163	12	320	1.02	5.64
Aug 8	0.017	0.000	0.925	0.119	1.10	159	8	316	0.65	3.69

Table 6. Comparison between pricing errors, i.e. the differences between theoretical and observed option prices, of the calibration pool for July 9, 10, 11, 14, and the single day calibrations. The root mean squared error (rmse) is in \$, corr(err single day, err pool) denotes the correlation between the pricing errors for the single day calibration and the corresponding pricing errors for the pooled calibration.

	Jul 9	Jul 10	Jul 11	Jul 14	average
rmse single day	0.639	0.487	0.636	0.434	0.549
rmse pool	0.725	0.584	0.686	0.481	0.619
corr(err single day, err pool)	0.935	0.877	0.943	0.895	0.915

Table 7. Calibrated parameters of the GARCH model (1), $\sigma_t^2 = \omega^* + \alpha^* \varepsilon_{t-1}^2 + \beta^* \sigma_{t-1}^2 + \gamma^* I_{t-1} \varepsilon_{t-1}^2$, $I_{t-1} = 1$ when $\varepsilon_{t-1} < 0$ and $I_{t-1} = 0$ otherwise, $\varepsilon_t = \sigma_t \, z_t$, $z_t \sim i.i.d.(0,1)$, under the risk neutral measure $\mathbb{Q}$, using FHS on m out of the money European put and call options for the year 2000 and comparison with the CGMYSA model. The root mean squared error (rmse) is in \$, the ape measure is defined in equation (3).

date	ω^*	α^*	β^*	γ^*	$E_{\mathbb{Q}}[\sigma^2]$	m	rmse	ape%	ape% CGMYSA
Jan	0.016	0.000	0.914	0.155	1.80	177	1.62	4.78	3.78
Mar	0.118	0.000	0.635	0.600	1.82	143	1.61	5.13	5.23
May	0.158	0.000	0.526	0.839	2.90	155	1.93	4.74	5.48
Jul	0.006	0.000	0.963	0.065	1.38	159	0.91	2.34	3.26
Sep	0.041	0.000	0.866	0.189	1.04	151	1.08	3.67	2.87
Nov	0.017	0.000	0.903	0.159	0.97	169	1.22	3.74	2.85

Table 8. "Homogeneous hedging of the smile". The three rows in the middle are market option prices form Hull's book. The first row is obtained multiplying the middle row times 0.9 and the last row is obtained multiplying the middle row by 1.1.

Strike price	Asset price	Option price
45	44.1	2.16
45	49	5.60
50	49	2.40
55	49	1.00
55	53.9	2.64

Table 9. OLS regression estimates and variance of forecast errors for time $t+1$, i.e. January 24, 2003 (first panel) and July 10, 2003 (second panel): 1) $P^{mkt}_{t+1} = \eta_0 + \eta_1 P^{bs}_{t,t+1} + error$; 2) $P^{mkt}_{t+1} = \eta_0 + \eta_1 P^{bs}_{t,t+1} + \eta_2 P^{garch}_{t,t+1} + error$, 3) $P^{mkt}_{t+1} = \eta_0 + \eta_2 P^{garch}_{t,t+1} + error$, where P^{mkt}_{t+1} are the option prices observed on time $t+1$, $P^{bs}_{t,t+1}$ are the Black-Scholes forecasts of option prices for $t+1$ computed by plugging in the Black-Scholes formula S_{t+1}, r, d at time $t+1$ and the implied volatility observed on time t (i.e. January 23 and July 9, respectively). $P^{garch}_{t,t+1}$ are the GARCH forecasts obtained using S_{t+1}, the GARCH parameter calibrated at time t and σ_{t+1} updated according to the estimates at time t.

	η_0	η_1	η_2	$Var[error]$
1)	0.823	0.996	—	0.761
2)	−0.037	0.558	0.436	0.316
3)	−1.073	—	0.988	1.035
1)	−0.118	0.997	—	0.188
2)	−0.213	0.293	0.704	0.161
3)	−0.429	—	0.997	0.315

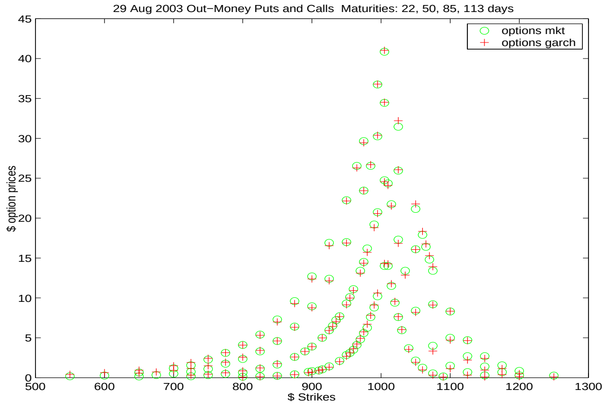


Fig. 1. Monte Carlo calibration results of the GARCH model to $m = 118$ out of the money European put and call option prices observed on August 29, 2003.

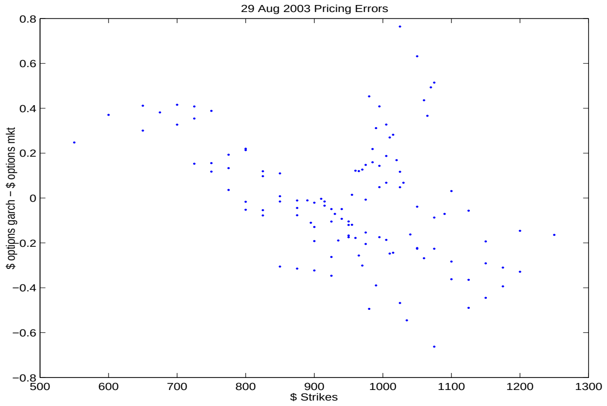


Fig. 2. Pricing errors of the GARCH model for $m = 118$ out of the money European put and call option prices observed on August 29, 2003.

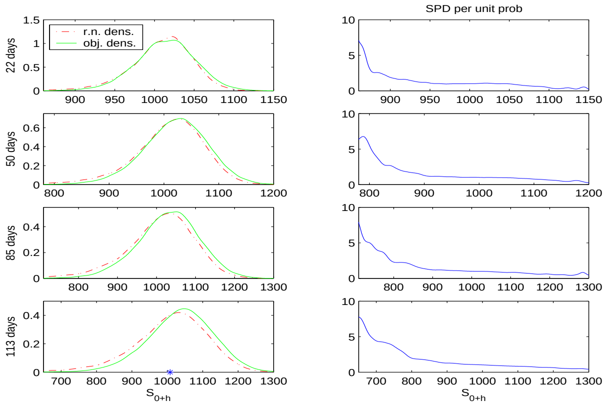


Fig. 3. Risk neutral and objective density estimates (left plots) and state price density estimates per unit of probability (right plots) for August 29, 2003.

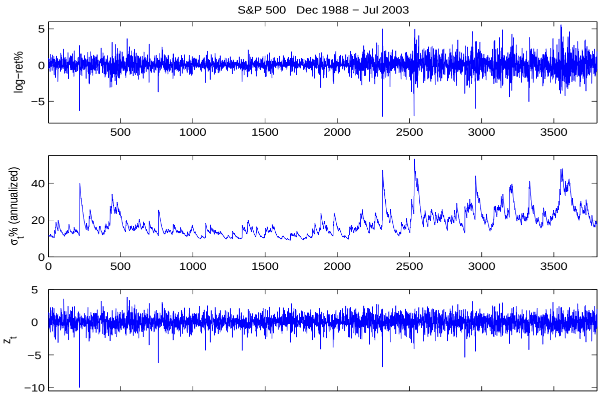


Fig. 4. Daily log-return in percentage of the S&P 500 index from December, 14 1988 to July 9, 2003 (first panel), estimated conditional variances (second panel) and scaled innovations (third panel).

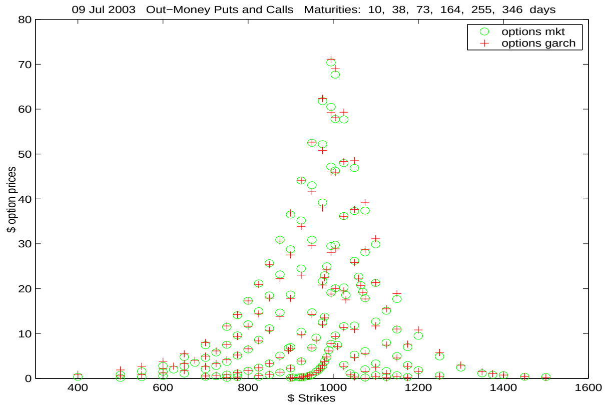


Fig. 5. FHS calibration results of the GARCH model to $m = 151$ out of the money European put and call option prices observed on July 9, 2003.

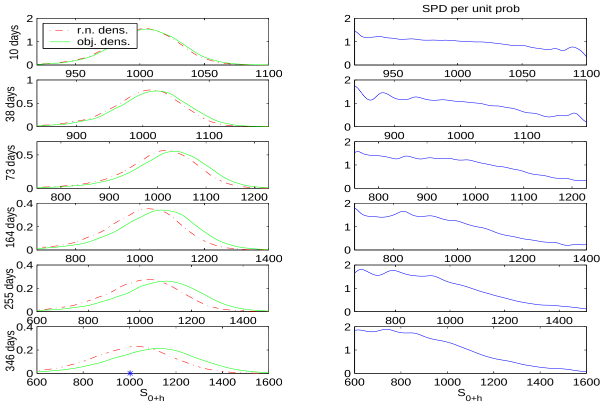


Fig. 6. Risk neutral and objective density estimates (left plots) and state price density estimates per unit of probability (right plots) for July 09, 2003.

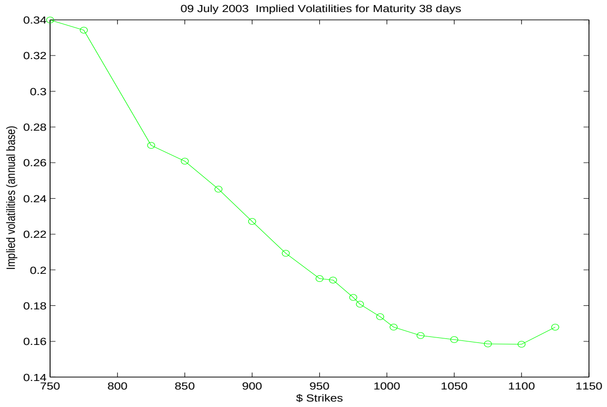


Fig. 7. Implied volatilities observed on July 9, 2003 from out of the money European put and call options with maturity $T = 38$ days.

Part II

Analysis of Textual Data

Visualization of textual data: unfolding the Kohonen maps.

Ludovic Lebart

CNRS - GET - ENST
46 rue Barrault,
75013, Paris, France
(e-mail: ludovic.lebart@enst.fr)

Abstract. The Kohonen self organizing maps (SOM) can be viewed as a visualisation tool that performs a sort of compromise between a high-dimensional set of clusters and the 2-dimensional plane generated by some principal axes techniques. The paper proposes, through Contiguity Analysis, a set of linear projectors providing a representation as close as possible to a SOM map. In so doing, we can assess the locations of points representing the elements via a partial bootstrap procedure.

Keywords: Contiguity analysis, Kohonen maps, SOM, Bootstrap.

1 Introduction

For many users of visualisation tools, the Kohonen self organising maps (SOM) outperform both usual clustering techniques and principal axes techniques (principal components analysis, correspondence analysis, etc.). Indeed, the displays of identifiers of words (or text units) within rectangular or octagonal cells allow for clear and legible printings. The SOM grid, basically non-linear, can then be viewed as a compromise between a high-dimensional set of clusters and the planes generated by any pairs of principal axes. One can regret however the absence of assessment procedures and of valid statistical inference as well. The paper proposes, through Contiguity Analysis (briefly reminded in section 2), a set of linear projectors providing a representation as close as possible to a SOM map (section 3 and 4). An example of application is given in section 5. Via a partial bootstrap procedure, we can now provide these representations with the projection of confidence areas (e.g. ellipses) around the location of words (section 6).

2 Brief reminder about contiguity analysis

Let us consider a set of multivariate observations (n observations described by p variables, leading to a (n, p) matrix $\mathbf{X}$), having an *a priori* graph structure. The n observations are also the n vertices of a symmetric graph $\mathcal{G}$, whose associated (n, n) matrix is $\mathbf{M}$ ($m_{ii'} = 1$ if vertices i and i' are joined by an edge, $m_{ii'} = 0$ otherwise). We denote by $\mathbf{N}$ the (n, n) diagonal matrix having the degree of each vertex i as diagonal element n_i (n_i stands here for n_{ii}). $\mathbf{y}$ is the vector whose i^{th} component is y_i. Note that: $n_i = \sum_{i'} m_{ii'}$. $\mathbf{U}$ designates the square matrix such that $u_{ij} = 1$ for all i and j. y being a random variable taking values on each vertex i of a symmetric graph $\mathcal{G}$ the *local variance* of y, $v^*(y)$, is defined as:

$$v^*(y) = (1/n) \sum (y_i - m_i^*)^2$$

where: $m_i^*= (1/n_i) \sum_{i'} m_{ii'} y_{i'}$. It is the average of the adjacent values of vertex i. Note that if $\mathcal{G}$ is a complete graph (all pairs (i,i') are joined by an edge), $v^*(y)$ is nothing but $v(y)$, the classical empirical variance. When the observations are distributed randomly on the graph, both $v^*(y)$ and $v(y)$ are estimates of the variance of y. The contiguity ratio (analogue to the Geary contiguity ratio [Geary, 1954]), is written: $c^*(y) = v^*(y)/v(y)$. It can be generalized : a) to different distances between vertices in the graph, b) to multivariate observations (both generalizations are dealt with in: [Lebart, 1969]). This section is devoted to the second generalization: multivariate observations having an *a priori* graph structure. The multivariate analogue of the local variance is now the local covariance matrix $\mathbf{V}^*$, given by (using the previously defined notation):

$$\mathbf{V}^* = (1/n)\mathbf{X}'(\mathbf{I} - \mathbf{N}^{-1}\mathbf{M})'(\mathbf{I} - \mathbf{N}^{-1}\mathbf{M})\mathbf{X}$$

The diagonalization of the corresponding local correlation matrix (Local Principal Component Analysis) [Aluja Banet and Lebart, 1984] produces a description of the local correlations that can be compared to the results of a PCA . Comparisons between correlation matrices (local and global) can be done through Procustean Analysis (see: [Gower and Dijksterhuis, 2004]). If the graph is made of k disjoined complete subgraphs, $\mathbf{V}^*$ coincide with the classical *within covariance matrix* used in linear discriminant analysis. If the graph is complete (associated matrix = $\mathbf{U}$ defined above), then $\mathbf{V}^*$ is the classical global covariance matrix $\mathbf{V}$.

Let $\mathbf{u}$ be a vector defining a linear combination $u(i)$ of the p variables for vertex i:

$$u(i) = \sum_j u_j y_{ij} = \mathbf{u}'\mathbf{y}_i$$

The local variance of $u(i)$ is: $v^*(u) = \mathbf{u}'\mathbf{V}^*\mathbf{u}$. The contiguity coefficient of $u(i)$ can be written: $c(u) = \mathbf{u}'\mathbf{V}^*\mathbf{u}/\mathbf{u}'\mathbf{V}\mathbf{u}$. Contiguity Analysis is the search for $\mathbf{u}$ that minimizes $c(u)$. It produces linear functions having the properties of "minimal contiguity". Instead of assigning an observation to a specific class, (case of discriminant analysis) these functions allows one to assign it in a specific part of the graph. Therefore, Contiguity Analysis can be used to discriminate between overlapping classes.

3 SOM maps and associated graphs

The self organizing maps (SOM maps) [Kohonen, 1989] aim at clustering a set of multivariate observations. The obtained clusters are displayed as the vertices of a rectangular (chessboard like) or octagonal graph. The distances between vertices on the graph are supposed to reflect, as much as possible, the distances between clusters in the initial space. Let us summarize the principles of the algorithm:

The size of the graph, and consequently, the number of clusters are chosen *a priori* (for example: a square grid with 5 rows and 5 columns, leading to 25 clusters). The algorithm is similar to the MacQueen algorithm [MacQueen, 1967] in its on-line version, and to the k-means algorithm [Forgy, 1984] in its batch version. Let us consider n points in a p-dimensional space (rows of the (n, p) matrix $\mathbf{X}$). At the outset, to each cluster k is assigned a provisional centre C_k with p components (e.g.: chosen at random). For each step t, the element $i(t)$ is assigned to its nearest provisional centre $C_{k(t)}$. Such centre, together with its neighbours on the grid, is then modified according to the formula: $C_{k(t+1)} = C_{k(t)} + \varepsilon(t)(i(t) - C_{k(t)})$. In this formula, $\varepsilon(t)$ is an adaptation parameter $(0 \leq \varepsilon \leq 1)$ which is a (slowly) decreasing function of t, as those usually involved in stochastic approximation algorithms. The process is reiterated, and eventually stabilizes, but the partition obtained may depend on the initial choice of the centres. In the batch version of the algorithm, the centres are updated only after a complete pass of the data.
Figure 1 represent a stylised symmetric matrix (70, 70) $\mathbf{M}_0$ associated to a partition of n=70 elements in k=8 classes (or clusters). Rows and columns represent the same set of n elements (elements belonging to a same class of the partition form a subset of consecutive rows and columns). The graph consists of 8 cliques. All the cells of the black blocks contains the value 1. All the cells outside these diagonal blocks contains the value 0 . The 8 classes of the previous partition have been obtained through a SOM algorithm from a square 3 x 3 grid (with an empty class).

The left hand side matrix of figure 1 does not take into account the topology of the grid: links between elements do exist only within clusters. In the right hand side of figure 1, two elements i and j are linked ($m_{ij} = 1$) in the graph if they belong either to a same cluster, or to contiguous clusters. Owing to the small size of the SOM grid (figure 2), the diagonal adjacency is not taken into account. (e.g.: elements belonging to cluster 7 are considered as contiguous to those of clusters 4 and 8, but not to the elements of cluster 5). Similarly to matrices $\mathbf{M}_0$ and $\mathbf{M}_1$, a matrix $\mathbf{M}_2$ can be defined, that extends the definition of the edges of the graph to diagonal links. In the simple example of figure 3, the elements of cluster 7, for example, are considered as contiguous to the elements of clusters 4, 8, and 5.

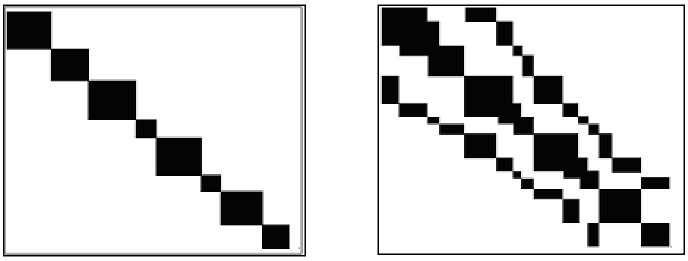
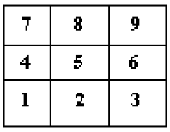


Fig. 1. Stylised incidence matrices $\mathbf{M}_0$ of the graph associated with a simple partition (left), and $\mathbf{M}_1$, relating to a SOM map (right) (all the cells in the white areas contain the value 0 whereas those in the black areas contain the value 1)



7	8	9
4	5	6
1	2	3

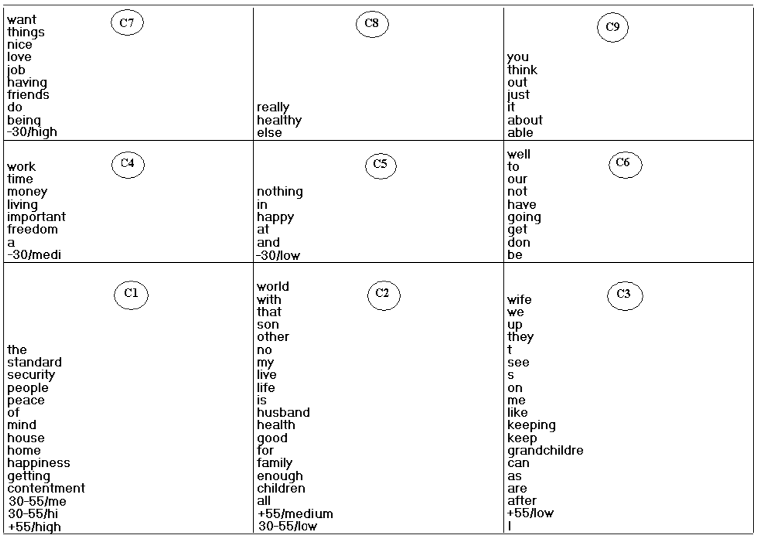
Fig. 2. The *a priori* SOM grid

4 Linear projectors onto the best SOM plane

The matrices $\mathbf{M}_0$, $\mathbf{M}_1$, and $\mathbf{M}_2$ can be easily obtained as a by-product of the SOM algorithm. In the case of contiguity analysis involving the graph G_0 the associated matrix of which is $\mathbf{M}_0$, the local variance coincide with the "within variance", and the result is a classical linear discriminant analysis of Fisher (LDA). In the plane spanned by the two first principal axes, the clusters are optimally located in the sense of the LDA criterion. In the cases of contiguity analysis using the graphs G_1 or G_2 (associated matrices $\mathbf{M}_1$, or $\mathbf{M}_2$), the principal planes strive to reconstitute the positions of the clusters in the SOM map. In the initial p-dimensional space, the SOM map can be schematised by the graph whose vertices are the centroids of the clusters. Those vertices are joined by an edge if the corresponding clusters are contiguous in the grid used in the algorithm. This graph in a high dimensional space will be partially or totally unfolded by the contiguity analysis. The following example will show the different phases of the procedure.

5 An example of application

An open-ended question has been included in a multinational survey conducted in seven countries (Japan, France, Germany, Italy, Nederland, United Kingdom, USA) in the late nineteen eighties [Hayashi *et al.*, 1992]. The respondents were asked : "What is the single most important thing in life for you?" . The illustrative example is limited to the British sample. The counts for the first phase of numeric coding are as follows: Out of 1043 responses, there are 13 669 occurrences (tokens), with 1 413 distinct words (types). When the words appearing at least 25 times are selected, there remain 9815 occurrences of these words, with 88 distinct words. In this example we focus on a partitioning of the sample into 9 categories, obtained by cross-tabulating age (3 categories) with educational level (3 categories). The nine identifiers combine age categories (-30, 30-55, +55) with educational levels (low, medium, high). Note that the SOM map (figure 3) provides a simultaneous representation of words and of categories of respondents. This is due to the fact that the input data are the coordinates provided by a correspondence analysis of the lexical contingency table cross-tabulating the words and the categories. Figure 4 represents the plane spanned by the two first axes of the contiguity analysis using the matrix $\mathbf{M}_1$. We can check that the graph describing the SOM map (the vertices of which C1, C2, ...C9 are the centroids of the elements of the corresponding cells of figure 3), is, in this particular case, a satisfactory representation of the initial map. The pattern of the nine centroids is similar to the original grid exemplified by figure 3. The background of figure 5 is identical to that of figure 4. It contains in addition the convex hulls of the nine clusters C1, C2, ..., C9.. Each of those convex hulls correspond exactly (if we except some double or hidden points) to a cell of figure 3. We note that these convex hulls are relatively well separated. In fact, figure 5 contains much more information than figure 3, since we have now an idea of the shapes and sizes of the clusters, of the



C7 want things nice love job having friends do being -30/high	C8 really healthy else	C9 you think out just it about able
C4 work time money living important freedom a -30/medi	C5 nothing in happy at and -30/low	well C6 to our not have going get don be
C1 the standard security people peace of mind house home happiness getting contentment 30-55/me 30-55/hi +55/high	world C2 with that son other no my live life is husband health good for family enough children all +55/medium 30-55/low	C3 wife we up they t see s on me like keeping keep grandchildre can as are after +55/low I

Fig. 3. A (3 x 3) Kohonen map applied to the words used in the 1043 responses

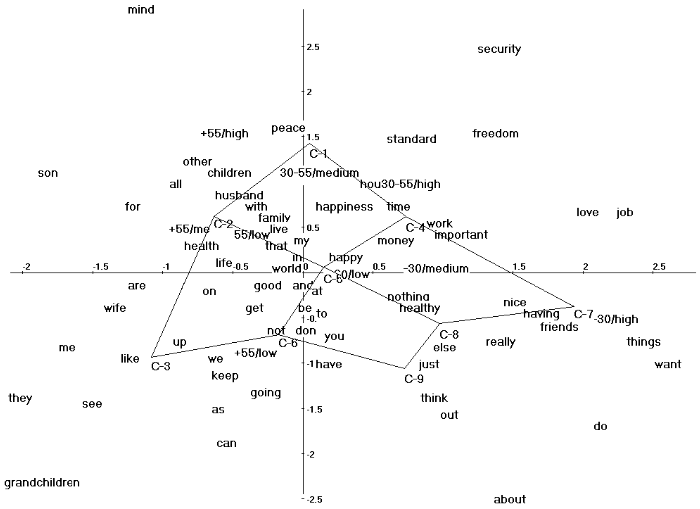


Fig. 4. Principal plane of the contiguity analysis using matrix $\mathbf{M}_1$. The points C1, C2, ...C9 represent the centroids of the 9 clusters derived from the SOM map.

degree to which they overlap. We are now aware of their relative distances, and, another piece of information missing in figure 3, we can observe the configurations of elements within each cluster.

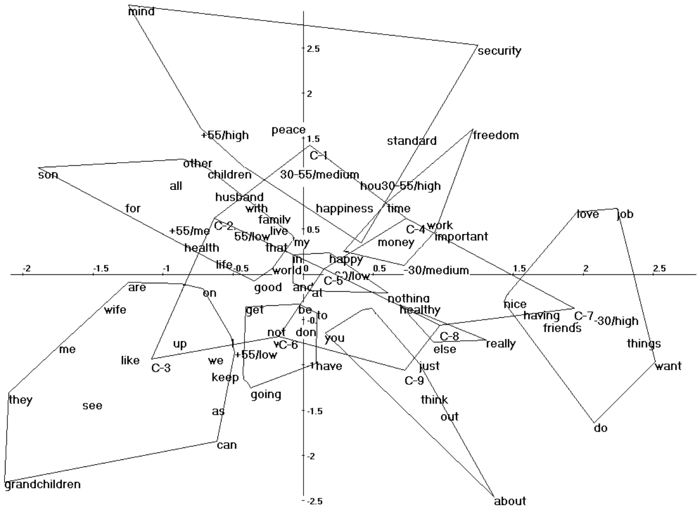


Fig. 5. Principal plane of the contiguity analysis using matrix $\mathbf{M}_1$, with both the centroids of the 9 clusters and their convex hulls

6 Assessing SOM maps through partial bootstrap

We are provided at this stage with a tool allowing us to explore a continuous space. We can take advantage of having a projection onto a plane (and possibly onto a higher dimensional space, although the outputs are much more complicated in that case) to project the bootstrap replicates of the original data set. This can be done in the framework of a partial bootstrap procedure. In the context of principal axes techniques (such as SVD, PCA, and also contiguity analysis), Bootstrap resampling techniques [Efron and Tibshirani, 1993] are used to produce confidence areas on two-dimensional displays. The bootstrap replication scheme allows one to draw confidence ellipses for both active elements (i.e.: elements participating in building principal axes) and supplementary elements (projected a posteriori).

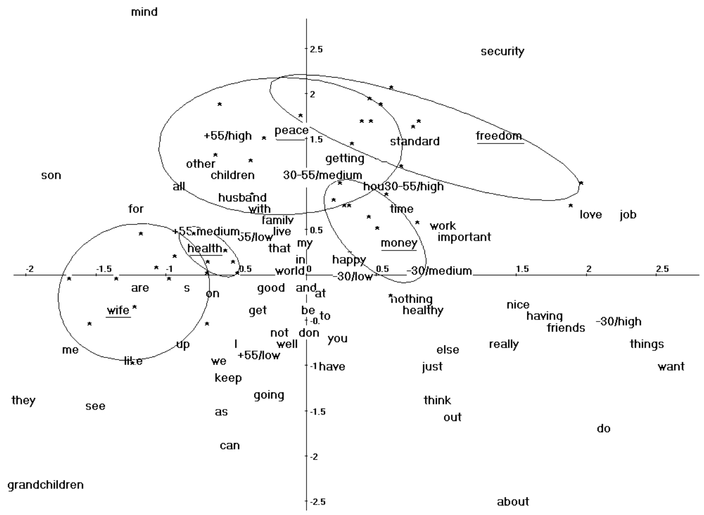


Fig. 6. Bootstrap ellipses of confidence of the 5 words: *freedom, health, money, peace, wife* in the same principal contiguity plane as in figure 4 and 5

In the example of the previous section, the words are the rows of a contingency table. The perturbation of such table under a bootstrap re-sampling procedure leads to new coordinates for the replicated rows. Without recomputing the whole contiguity analysis for each replicated sample (conservative procedure of *total bootstrap*), one can project the replicated rows as supplementary elements on a common reference space, exemplified above by figures 4 and 5. Always on that same space, figure 6 shows a sample of the replicates of five points (small stars visible around the words *freedom, health, money, peace, wife*) and the confidence ellipses that contain approximately 90 % of these replicated points. Such procedures of partial bootstrap [Lebart, 2004] give satisfactory estimates of the relative uncertainty about the location of points. Although the background of figures 5 and 6 are the same, it is preferable, to keep the results legible, to draw the confidence ellipses on a distinct figure. It can be seen for instance that the words *freedom* and *money*, both belonging to cluster C4, have different behaviours with respect to the re-sampling variability. The location of *freedom* is much more fuzzy. That word could belong to some neighbouring clusters as well.

7 Conclusion

We have intended to immerse the SOM maps, obtained through an algorithm often viewed as a black box, into an analytical framework (the linear algebra of contiguity analysis) and into an inferential setting as well (re-sampling techniques of bootstrap). That does not question the undeniable qualities of clarity and readability of the SOM maps. But it may perhaps help to assess their scientific status: like most exploratory tools, they help to rapidly uncover some patterns. However, they should be complemented with statistical procedures whenever deeper interpretation is needed.

References

[Aluja Banet and Lebart, 1984]T. Aluja Banet and L. Lebart. Local and partial principal component analysis and correspondence analysis. In M. Novak T. Havranek, Z. Sidak, editor, *COMPSTAT Proceedings*, pages 113–118, 1984.

[Efron and Tibshirani, 1993]B. Efron and R. J. Tibshirani. *An Introduction to the Bootstrap.* Chapman and Hall, New York, 1993.

[Forgy, 1984]R. Forgy. Cluster analysis of multivariate data : efficiency versus interpretability of classifications. In *Biometric Society Meetings,*, page 768, 1984.

[Geary, 1954]R. C. Geary. The contiguity ratio and statistical mapping. *The Incorporated Statistician*, pages 115–145, 1954.

[Gower and Dijksterhuis, 2004]J. C. Gower and G. B. Dijksterhuis. *Procustes Problem.* Oxford Statistical Science Series, Oxford, 2004.

[Hayashi *et al.*, 1992]C. Hayashi, T. Suzuki, and M. Sasaki. *Data Analysis for Social Comparative Research: International Perspective.* North-Holland, Amsterdam, 1992.

[Kohonen, 1989]T. Kohonen. *Self-Organization and Associative Memory.* Springer Verlag, Berlin, 1989.

[Lebart, 1969]L. Lebart. Analyse statistique de la contiguité. *Publications de l'ISUP*, pages 81–112, 1969.

[Lebart, 2004]L. Lebart. Validation techniques in text mining. In S. Sirmakensis, editor, *Text Mining and its Application*, pages 169–178, 2004.

[MacQueen, 1967]J. B. MacQueen. Some methods for classification and analysis of multivariate observations. *Proc. Symp. Math. Statist. and Probability (5th)*, pages 281–297, 1967.

Efficient Processing of Extra-grammatical Sentences: Comparing and Combining two approaches to Robust Stochastic parsing

Marita Ailomaa[2], Vladimír Kadlec[1], Jean-Cédric Chappelier[2], and Martin Rajman[2]

[1] Faculty of Informatics, Masaryk University
Botanická 68a, 602 00 Brno, Czech Republic
E-mail: xkadlec@fi.muni.cz

[2] Artificial Intelligence Laboratory, Computer Science Department
Swiss Federal Institute of Technology (EPFL)
1015 Lausanne, Switzerland
E-mail: {marita.ailomaa,jean-cedric.chappelier,martin.rajman}@epfl.ch

Abstract. This paper compares two techniques for robust parsing of extra-grammatical natural language that might be of interest in large scale Textual Data Analysis applications. The first one returns a "correct" derivation for any extra-grammatical sentence by generating the finest corresponding most probable optimal maximum coverage. The second one extends the initial grammar by adding relaxed grammar rules in a controlled manner. Both techniques use a stochastic parser that selects a "best" solution among multiple analyses. The techniques were tested on the ATIS and Susanne corpora and experimental results, as well as conclusions on performance comparison, are provided.

Keywords: Robust, Parsing, Coverage.

1 Introduction

Formal grammars are traditionally used in NLP applications to describe well-formed sentences. But in large scale Textual Analysis applications it is not practical to rely exclusively on a formal grammar because of the large fraction of sentences that will receive no analysis. This undergeneration problem has lead to a whole field of research called robust parsing, where the goal is to find domain-independent, efficient parsing techniques that return a correct or usefully "close" analysis for almost all of the input sentences [Carroll and Briscoe, 1996]. Such techniques need to handle not only the problems of undergeneration but also the increased ambiguity which is usually a consequence of the robustification of the parser.

In previous works, a variety of approaches have been proposed to robustly handle natural language. Some techniques are based on modifying the input sentence, for example by removing words that disturb the fluency [Bear *et al.*, 1992, Heeman and Allen, 1994]. More recent approaches are based on selecting the right sequence of partial analyses [Worm and Rupp, 1998, van

Noord *et al.*, 1999]. Minimum Distance Parsing is a third approach based on relaxing the formal grammar, allowing rules to be modified by insertions, deletions and substitutions [Hipp, 1992].

Most of these approaches make the distinction between *ungrammaticality* and *extra-grammaticality.* Ungrammatical sentences might contain errors such as wrong agreement in the case of casual written text like mails, or hesitations and other types of disfluencies in the case of spoken language. On the other hand, extra-grammatical sentences are linguistically correct sentences that are not covered by the grammar.

This paper presents two new approaches that focus on extra-grammatical sentences. The first approach described in section 2 is based on the selection of a most optimal coverage with partial analyses, while the second, presented in section 3, uses controlled grammar rule relaxation. Section 4 describes the comparison of these two approaches and shows that they present differences in behavior when given the same grammar and the same test data.

2 Selecting the most probable optimal maximum coverage

2.1 Concepts

For a given sentence a *coverage*, with respect to an input grammar G, is a sequence of non-overlapping, possibly partial, derivation trees, such that the concatenation of the leaves of these trees corresponds to the whole input sentence (see figure 1).

If there are no unknown words in the input sentence, then at least one trivial coverage is obtained, consisting of the trees that all use only lexical rules (i.e. one rule per tree).

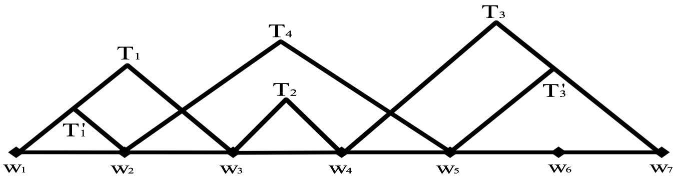


Fig. 1. A coverage $C = (T_1, T_2, T_3)$ consisting of trees T_1, T_2 and T_3. If there are $T_1^{'}$ and $T_3^{'}$, $T_1^{'}$ is a subtree of tree T_1 and $T_3^{'}$ is a subtree of T_3, then we also have coverage $C^{'} = (T_1^{'}, T_4, T_3^{'})$. Conversely $(T_1, T_3^{'})$ and (T_1, T_4, T_3) are not coverages.

A maximum coverage (m-coverage) is a coverage that is maximal with respect to the partial order relation $\leq$, defined as reflexive transitive closure of the subsumed relation $\prec$ (see figure 2). The relation $\prec$ is a relation over coverages such that, for coverages C and $C^{'}$:

$$C^{'} \prec C \text{ iff } \exists i,j,k,\ 1 \le i \le k, 1 \le j \text{ and there exists rule } r \text{ in the grammar } G \text{ such that } C = (T_1, ..., T_i, ..., T_k),\ C^{'} = (T_1, ...T_{i-1}, T_1^{'}, T_2^{'}, ..., T_j^{'}, T_{i+1}, ..., T_k) \text{ and } T_i = r \circ T_1^{'} \circ T_2^{'} ... \circ T_j^{'},$$

i.e. if there exists a sub-sequence of trees in $C^{'}$ that can be connected by rule r and the resulting tree is element of C, the other trees in $C^{'}$ being the same as in C. Notice that the rule r can be an unary rule.

If there is a successful parse (a single derivation tree that covers the whole input sentence) then there are as many m-coverages as full parse trees and every m-coverage contains only one tree.

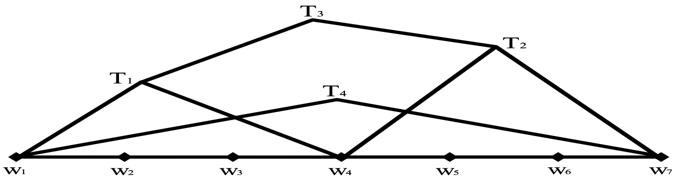


Fig. 2. An example to illustrate a maximum coverage. The coverage $C_1 = (T_3)$ is m-coverage. The coverage $C_2 = (T_1, T_2)$ is not maximum, because $C_2 \le C_1$. There is also another m-coverage $C_3 = (T_4)$. Notice that C_1 and C_3 are not comparable with relation $\le$.

In addition to maximality, we focus on *optimal* m-coverage, where optimality could be defined with respect to different measures. In contrast to maximality, the choice of a measure depends on the concrete application. Several optimality measures could be defined. For instance, the optimality measure can look at the intended structure of trees in a coverage, e.g. it can count the number of nodes in trees. In the presented work, we used the following optimality measure which relates to the average width (number of leaves) of the derivation trees in the coverage. For an m-coverage $C = (T_1, T_2, ...T_k)$ of input sentence $w_1, w_2, ..., w_n$, $n > 1$, we define

$$S_1(C) = \tfrac{1}{n-1}(\tfrac{n}{k} - 1).$$

Notice that $0 \le S_1(C) \le 1$ and $\frac{n}{k}$ is the average width of the derivation trees in the coverage. With this measure, the value of a coverage made exclusively of lexical rules is 0 and the value of a successful parse is 1.

For standard SCFG derivation, the probability of a coverage is defined as the product of the probabilities of the trees it contains. The probability of a coverage could also be viewed as another optimality measure. So the most probable coverages can be found in the same way as optimal m-coverages. But, usually we find all optimal m-coverages (OMC) first (optimal with respect to some other measure than probability) and then the most probable of these is chosen. Notice that both OMC and most probable OMC are not unique.

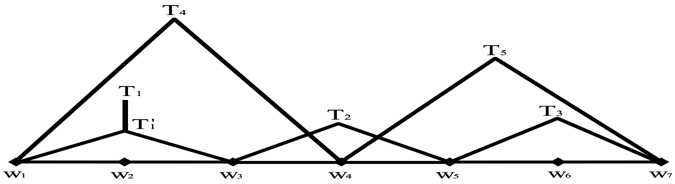


Fig. 3. An example to demonstrate the optimal m-coverage. $C_1 = (T_1, T_2, T_3)$ and $C_2 = (T_4, T_5)$ are m-coverages. The coverage $C_1^{'} = (T_1^{'}, T_2, T_3)$ is not m-coverage. The coverage C_2 is optimal for the measure S_1, $S_1(C_1) < S_1(C_2)$. Notice that the coverages C_1 and C_2 are not comparable with relation $\leq$.

2.2 Algorithm

We use a bottom-up chart parsing algorithm [Chappelier and Rajman, 1998] that produces all possible incomplete parses[1]. The incomplete parses are then combined to find the maximum coverage(s).

The described algorithm finds OMC with respect to the measure S_1 (the average width of the derivation trees in the coverage), but it can be easily adapted to different optimality measures. All operations are applied to a set of Earley's items [Earley, 1970]. In particular, no changes are made during the parsing phase (except some initialization of internal structures for better efficiency of the algorithm).

The Dijkstra's algorithm for shortest path problem in graphs is used to find OMC. The input graph for the Dijkstra's algorithm consists of weighted edges and vertices. The edges are Earley's items and the weight of each edge is 1. The vertices are word positions, thus for n input words we have $n + 1$ vertices. Whenever the Dijkstra's algorithm finds paths with equal length (i.e. identical number of items), we use the probability to select the most probable ones. Notice that, if all the words are known, there exists at least one path from position 0 to n corresponding to the trivial coverage.

The output of the algorithm is a list of Earley's items, which can represent several derivation trees. To get OMC, the most probable tree from each item is selected.

3 Deriving trees with holes

Our second approach to robust parsing is based on the idea that, in the case of a rule-based parser, the parser fails to analyze a given extra-grammatical sentence because one or several rules are missing in the grammar. If a rule-relaxation mechanism is available[2], it can be used to cope with such situ-

[1] Whenever there exists a derivation tree that covers the part of the given input sentence, the algorithm produces that tree

[2] A mechanism that can derive additional rules from the ones present in the grammar

ations. In that case the goal of the robust parser is to derive a full tree where the subtrees corresponding to the used relaxed rules are represented as "holes" (see figure 4).

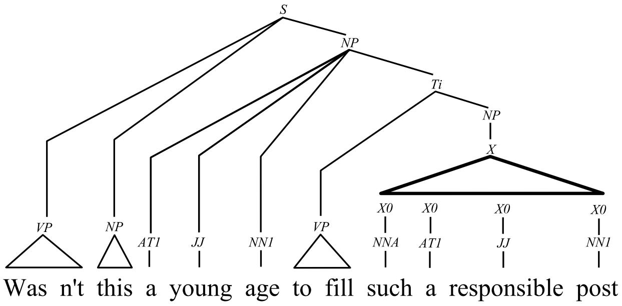


Fig. 4. A tree with a hole representing a missing NP rule $NP \rightarrow NNA\ \ AT1\ \ JJ\ \ NN1$.

We use the principle called Minimum Distance Parsing which has been introduced in earlier robust parsing applications [Hipp, 1992]. This approach relaxes rules in the grammar by inserting, deleting or substituting elements in their right hand side (RHS). Derivation trees are ranked by the number of modifications that have been applied to the grammar rules to achieve a complete analysis. One important drawback is that, in its unconstrained form, the method produces many incorrect derivations and works well only for small grammars [Rosé and Lavie, 2001].

To prevent such incorrect derivations, we make restrictions on how the rules can be relaxed based on observations and linguistic motivations. One such restriction is to only relax grammar rules for which the LHS is frequently represented in the grammar, e.g. NP. Another restriction is to allow only one type of relaxation, namely insertion. The inserted element is hereafter referred to as a filler. A further refinement of the algorithm is to specify what syntactic category a filler is allowed to have when being inserted into a given position in the RHS. To illustrate the ideas, an example is now provided.

Assume that there is a grammar with two NP rules. (The head is indicated with underlined syntactic categories):

$$R1 : NP \rightarrow ADJ\ \ \underline{N}$$
$$R2 : NP \rightarrow POS\ \ \underline{N}$$

According to this grammar "successful brothers" and "your brother" are syntactically correct NPs while "your successful brother" is not. In order to parse the last one, some NP rule needs to be relaxed. We select the second one, R2 (though both are possible candidates). If the filler that needs to be inserted is ADJ (in this case "successful"), then the relaxed NP rule is expressed as:

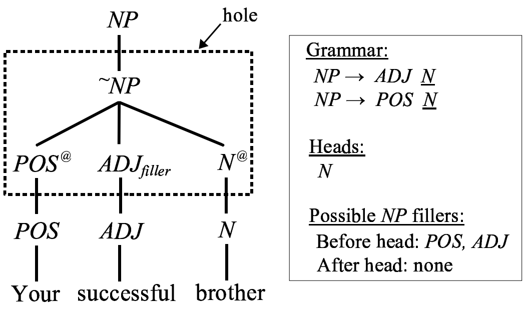


Fig. 5. An example of how a hole is derived by relaxing a rule and inserting a filler.

$$R3 : {}^{\sim}NP \rightarrow POS^{@} \quad ADJ_{filler} \quad N^{@}$$

We use the category ~NP instead of NP to distinguish relaxed rules from initial ones, the "filler" subscripts to identify the fillers in the RHS in the relaxed rule, and the @ to label the original RHS elements. The decision of allowing an insertion of an ADJ as filler is based on whether ADJ is a possible element before the head or not. Since there is a rule in the grammar where an ADJ exists before the head (R1), the insertion is appropriate.

4 Validation

The two robust parsing techniques presented in the previous sections were tested on subsets of two treebanks, ATIS and Susanne. From these treebanks two separate grammars were extracted having different characteristics. Concretely each treebank was divided into a learning set that was used for producing the probabilistic grammar and a test set that was then parsed with the extracted grammar. Around 10% of the sentences in the test set were not covered by the grammar. These sentences represented the real focus of our experiments, as the goal of a robust parser is to process the sentences that the initial grammar fails to describe.

The sentences were first parsed with technique 1 and technique 2 separately and then with a combined approach where the rule-relaxation technique was tried first and only when it failed the most probable OMC was selected. For each sentence the 1-best derivation tree was categorized as good, acceptable or bad, depending on how closely it corresponded to the reference tree in the corpus and how useful the syntactic analysis was for extracting a correct semantic interpretation. The results are presented in table 1. It may be argued that the definition of a "useful" analysis might not be decidable only by observing the syntactic tree. Although we found this to be a quite usable hypothesis during our experiments, some more objective procedure should be defined. In a concrete application, the usefulness might for example be determined by the actions that the system should perform based on the produced syntactic analysis.

	Good (%)	Acceptable (%)	Bad (%)	No analysis (%)
ATIS corpus				
Technique 1	10	60	30	0
Technique 2	24	36	9	31
Technique 1+2	27	58	16	0
Susanne corpus				
Technique 1	16	29	55	0
Technique 2	40	17	33	10
Technique 1+2	41	22	37	0

Table 1. Experimental results. Percentage of good, acceptable and bad analyses with technique 1 (optimal coverage), technique 2 (tree with holes) and with the combined approach.

From the experimental results one can see that, for both grammars, technique 2 is more accurate than technique 1. However, if both good and acceptable results are taken into account, technique 1 behaves better with the ATIS grammar that has relatively few rules, and technique 2 better with Susanne, which is a considerably larger grammar describing a rich variety of syntactic structures.

Regardless of the technique used, the number of bad 1-best analyses that are produced can be explained by the fact that the probabilistically best analysis is not always the linguistically best one. This is a non-trivial problem related to all types of natural language parsing, not only to robust parsers.

An interesting result is that when the sentences are processed sequentially with both techniques, the advantage of each approach is taken into account and the performance is better than when either technique is used alone.

5 Conclusions

In this report we presented and compared two approaches to robust stochastic parsing. First we introduced the optimal maximum coverage framework and associated measures for the optimality of the parser. Then we introduced a rule-relaxation strategy based on the concept of holes, using several linguistically motivated restrictions to control the relaxation of grammar rules.

Experimental results show that a combination of the techniques gives a better performance than each technique alone, because the first one guarantees full coverage while the second has a higher accuracy. The richness of the syntactic structures defined in the initial grammar tends to have some impact on the performance in the second approach but less in the first one. This can be linked to the restrictions that were chosen for the relaxation of the grammar rules. It is possible that different types of restrictions are appropriate for different grammars.

The evaluation of the robust parsing techniques was based on manually checking the derivation trees. An important issue is to integrate the techniques into some target application so that we have more realistic ways of measuring the usefulness of the produced robust analyses.

As a final remark, we would like to point out that this paper has addressed the problem of extra-grammaticality but did not address ungrammaticality, which is also a very important phenomenon in robust parsing, though more relevant in spoken language applications than in textual data analysis.

References

[Bear *et al.*, 1992]John Bear, John Dowding, and Elizabeth Shriberg. Integrating multiple knowledge sources for the detection and correction of repairs in human-computer dialogue. In *Proceedings of the 30th ACL*, pages 56–63, Newark, Delaware, 1992.

[Carroll and Briscoe, 1996]John Carroll and Ted Briscoe. Robust parsing — a brief overview. In John Carroll, editor, *Proceedings of the Workshop on Robust Parsing at the 8th European Summer School in Logic, Language and Information (ESSLLI'96), Report CSRP 435*, pages 1–7, COGS, University of Sussex, 1996.

[Chappelier and Rajman, 1998]J.-C. Chappelier and M. Rajman. A generalized CYK algorithm for parsing stochastic CFG. In *TAPD'98 Workshop*, pages 133–137, Paris, France, 1998.

[Earley, 1970]J. Earley. An efficient context-free parsing algorithm. In *Communications of the ACM*, volume 13, pages 94–102, 1970.

[Heeman and Allen, 1994]Peter A. Heeman and James F. Allen. Detecting and correcting speech repairs. In *Proceedings of the 32th ACL*, pages 295–302, Las Cruces, New Mexico, 1994.

[Hipp, 1992]Dwayne R. Hipp. *Design and development of spoken natural language dialog parsing systems.* PhD thesis, Duke University, 1992.

[Rosé and Lavie, 2001]C. Rosé and A. Lavie. Balancing robustness and efficiency in unification-augmented contextfree parsers for large practical applications. In G. van Noord and J. C. Junqua, editors, *Robustness in Language and Speech Technology.* Kluwer Academic Press, 2001.

[van Noord *et al.*, 1999]Gertjan van Noord, Gosse Bouma, Rob Koeling, and Mark-Jan Nederhof. Robust grammatical analysis for spoken dialogue systems. *Natural Language Engineering*, 5(1):45–93, 1999.

[Worm and Rupp, 1998]Karsten L. Worm and C. J. Rupp. Towards robust understanding of speech by combination of partial analyses. In *Proceedings of the 13th biennial European Conference on Artificial Intelligence (ECAI'98), August 23-28*, pages 190–194, Brighton, UK, 1998.

Clustering units from frequency and nominal variables: definition of a global distance. Application to survey data with closed and open-ended questions

Mónica Bécue[1] and Jerôme Pagès[2]

[1] EIO. Universitat Politècnica de Catalunya
08028 Barcelona - Spain
(e-mail: monica.becue@upc.edu)
[2] Agrocampus Rennes
65 rue de Saint-Brieuc, CS 84215
F-35042 Rennes cedex, France
(e-mail: jerome.pages@agrocampus-rennes.fr)

Abstract. Clustering units from heterogeneous data such as nominal and frequency variables is a relevant challenge. This kind of clustering requires to define a global distance between the units that takes into account the specificity of the data. An important application is clustering the respondents to a questionnaire including both closed and open-ended questions. The main arguments for using a global distance defined through a geometrical and multidimensional approach are exposed and illustrated through an example.
Keywords: Clustering, Heterogeneous data, Multiple factor analysis, Global distance, Clusters description.

1 Introduction

In very different studies, the statistical units are described by both nominal and frequency variables. In ecology, it is common to describe different sites by counting up the occurrences of every species as well as by identifying several soil and climatic attributes. In economic studies, the regions are often characterized by the counts of inhabitants in socioeconomic categories and by nominal attributes. A particular case arises in survey data, when a complex topic is tackled by using closed and open-ended questions: the statistical analysis of the latter starts from counting up the occurrences of the different words in every individual answer.

Each set of nominal variables provides a units × variables subtable. Each frequency variable provides a contingency table units × categories. So, the global table juxtaposes units × nominal variables and contingency subtables, i.e. heterogeneous data.

We want here to cluster the units, using both kinds of data but taking into account their specificity. The starting point consists in defining a global distance, based on a geometrical approach to the data, in such a way that the influence of the different groups is balanced.

Section 2 presents the notations. Section 3 addresses the global distance definition and Section 4 comments the clustering step. The results obtained with an actual example are discussed in Section 5. Finally, Section 6 values the contribution of such a methodology for clustering heterogeneous data.

2 Notation

A set of I statistical units are described by J_c groups of frequency variables (leading to build up J_c contingency tables with dimension $I \times K_j$; K_j here is the number of categories of the variable j) and J_q groups of nominal variables (leading to J_q individuals$\times$indicator variables tables with dimension $I \times K_j$; K_j here is the number of categories of all the nominal variables of the set j).The whole of these J tables ($J = J_c + J_q$) make up a multiple table $I \times K$ ($K = \sum_{j \in J} K_j$).

At the crossing of row i and column k (belonging to table j) we have,

- if j is a contingency table: f_{ikj} the relative frequency, in table j ($j = 1, ..., J_c$), with which row i ($i = 1, ..., I$) is associated to column k ($k = 1, ..., K_j$). ($\sum_{ijk} f_{ikj} = 1$).
- if j is an indicator table : $x_{ik} = 1$ if i belongs to the category k and 0 if not.

We denote: $f_{i.j} = \sum_{k=1}^{K_j} f_{ikj}$ and $f_{.jk} = \sum_i f_{ikj}$ the row and column margins of the contingency table j as subtable of the global table; $f_{i..} = \sum_{kj} f_{ikj}$ the row margin of the table gathering all the J_c contingency tables. The margins of the tables of indicator variables, which are constant, do not intervene in the calculus.

Remark: the row margin of the table gathering all the J_c contingency tables $f_{i..}$; $i = 1, ..., I$ will be used as row weights (and metric in the column space). In the case of tables with notable higher frequencies than others, the former can strongly dominate those weights. As an alternative, it is possible to firstly transform the data into proportions before the concatenation.

3 Definition of a global distance

The first (and fundamental) step for clustering consists in choosing the distance (or the similarity measure) between the units. The case we have to deal with, heterogeneous data with frequency and nominal variables, presents specific problems, close to the problems faced when mixed variables, quantitative and qualitative, are considered. In both cases, we have to choose:

- a distance between units within every group of columns (separate distances)

- an aggregation function of these separate distances in a global distance in such a way that the different groups have a balanced influence.

3.1 Separate distances between units, as induced by every group of columns

For nominal and frequency variables groups, it is usual to consider the χ^2 distance, i.e. the distance between profiles considered, respectively, in correspondence analysis (CA) and in multiple correspondence analysis (MCA).

3.2 Aggregation strategy

It is not a straightforward matter to define an aggregation strategy, even in the case of different groups made up by the same type of variable, that gives a balanced influence to every group of variables.

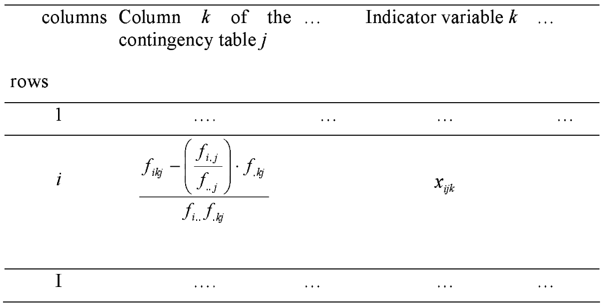
Geometrical approach

[Escofier and Pagès, 1998] propose a geometrical approach in the mixed case, with quantitative and qualitative variables. They consider the structures of the clouds of individuals as induced by every separate distance and propose to re-scale every subcloud in order to have the same greatest axial inertia. For that, a suited principal axes method is performed on every separate table, principal component analysis (PCA) in the case of quantitative variables and multiple correspondence analysis (MCA) in the case of qualitative variables. So, the highest axial inertia λ_1^j is measured, that allows for re-scaling the distances by dividing by λ_1^j the separate distance corresponding to set j. Furthermore, the global distance, as weighted sum of the re-scaled separate distances, is automatically calculated by performing MFA on the juxtaposed table.

Besides, this approach allows for transforming the initial mixed variables into only quantitative variables (that are the principal components) while the genuine distances, as defined from qualitative variables, are conserved in the case of considering all the principal axes. Nevertheless, in some cases, it can be useful to keep only the first principal axes.

Global distance in the case of heterogeneous data

For considering frequency groups, we have already adapted MFA to contingency tables and proposed the multiple factor analysis for contingency tables (MFACT: [Bécue and Pagès, 1999], [Bécue and Pagès, 2004]). With this aim, MFACT transforms the juxtaposed contingency table as in internal CA [Cazes and Moreau, 2000] and adopts the point of view of MFA for balancing the influence of the different tables in the global analysis. Besides, the combination of MFACT with the usual MFA makes possible to deal with contingency tables and indicator tables in a same analysis [Bécue and Pagès, 2001]. The initial global table, multiple table row-wise juxtaposing the contingency and the indicator variables tables, is transformed as shown in Figure 1.



columns rows	Column k of the contingency table j	…	Indicator variable k	…
1	….	…	…	…
i	$\dfrac{f_{ikj} - \left(\dfrac{f_{i.j}}{f_{..j}}\right)\cdot f_{.kj}}{f_{i..}f_{.kj}}$		x_{ijk}	
I	….	…	…	…

Fig. 1. *Multiple table issued from the original table by suited transformations*

Then, a non-normalized weighted PCA is performed using:

- as row weights (and metric in the column space): $f_{i..}; i = 1, ..., I$, where $f_{i..}$ is the mean relative weight of the rows on all the contingency tables that are considered;
- as column weights (and metric in the individuals space) the initial weight of the column divided by λ_1^j, and so $\{(I_k/IJ)/\lambda_1^j;\ k = 1, ..., K_j;\ j = 1, ..., J_q;\ f_{.kj}/\lambda_1^j;\ k = 1, ..., K_j;\ j = J_q{+}1, ..., J_c\}$. I_k (k being an indicator variable) is the number of individuals belonging to category k.

In such a way, MFACT automatically induces the squared distance between rows i and l given by (1):

$$d^2(i,l) = \sum_{j\in J_c} \frac{1}{\lambda_1^j} \sum_{k\in K_j} \frac{1}{f_{.kj}} \left[\left(\frac{f_{ikj}}{f_{i..}} - \frac{f_{lkj}}{f_{l..}}\right) - \frac{f_{.kj}}{f_{..j}}\left(\frac{f_{i.j}}{f_{i..}} - \frac{f_{l.j}}{f_{l..}}\right)\right]^2 + \sum_{j\in J_q} \frac{1}{\lambda_1^j} \sum_{k\in K_j} \frac{I}{K_j I_k} \left[x_{ik} - x_{lk}\right]^2 \qquad (1)$$

The contingency table j brings the contribution to the global distance indicated by the term j of the first block of (1): the deviation between the row profiles i and l is relativized, for each column of table j, by the deviation between the row margins in this table j. The qualitative variable j brings the contribution to the distance indicated by the term j of the second block of (1). Every contribution to the distance is rescaled by $1/\lambda_1^j$, thus balancing the influence of the different groups of variables.

4 Clustering step

Clustering method

For the clustering step, different methods can be used, although a hierarchical clustering, using generalized Ward's criterion, is a suited clustering method when operating from quantitative variables, especially when they are principal components.

Characterization of the clusters and validation of the partition

For every cluster, the significantly over and under represented categories, in the case of the nominal variables, are selected by using a statistical test [Lebart *et al.*, 1998]. A very similar reasoning allows for selecting the significatively frequent words in every cluster. The count m_{iq} of word i in cluster q is compared to the counts that would be obtained with all the samples comprised of $m_{.q}$ occurrences ($m_{.q}$: total length of cluster q) randomly extracted from the whole corpus without replacement [Lebart *et al.*, 1997], [Bécue and Lebart, 2000].

Furthermore, for every cluster, the modal answers are identified. They are actual responses, given by respondents, that are considered as representative according to two different criterions. The first criterion is linked to the frequency of the characteristic words in the answer while the second one is induced by the definition of a distance between the response lexical profile and the cluster lexical profile. It is usual to consider that the most representative answers are those which are selected by both criterions [Lebart *et al.*, 1997].

5 Example: practices and opinions of the children about reading

5.1 Data

The application is extracted from a large study carried out in the outskirts of Barcelona. 895 children studying fifth grade (about 10 or 11 years old) answered a closed questionnaire concerning attitude about reading and had to complete the two following assertions:

1. *Para mí leer es...*(For me, to read means. . .);

2. *Creo que leer es importante porque...*(I believe that reading is important because. . .).

We only keep the 816 children having answered to the active questions. The closed questions concerning the attitudes about reading correspond to the first group (nominal variables) and, respectively, the two open-ended questions make up groups 2 and 3. So, the columns of the first group (indicator variables) are the categories of the closed questions and the columns of the second and third groups correspond to the words used in the corresponding open-ended question, whose frequency is counted up for every child. We only keep the words used at least 8 times by the whole of the respondents.

Additional information is also used, as supplementary, to illustrate the clusters

Table 1. *Closed active questions*

1 . *At school, we read*	(very few, fair enough, a lot)
2 . *At home, we have*	(few, enough, a lot of books)
3 . *I read*	(very little, fair enough, a lot)
4 . *I read*	(very easily, easily, I have difficulties)
5 . *The books given by the teacher*	(I like them, I do not like them)
6 . *I read when*	(I enjoy it, I do scholar work, both)
7 . *I prefer reading*	(silently, aloud, both)
8 . *to read the scholar books*	(I enjoy, I do not enjoy, it depends)

5.2 Results

The three groups of variables contribute, in a balanced way, to the variances of the seven first principal axes, those that are taken into account in the clustering step. So, we can expect that the partition into clusters would depend from the closed questions but also the free answers.

Hierarchical clustering

The hierarchical clustering, from the coordinates on the 7 first principal axes, shares out seven clusters. The between-classes inertia/global inertia ratio is only 43%, showing that the clusters so identified do not correspond to clear frontiers, this could be expected due to the fact that there is a great homogeneity in the socio-economic conditions of the children as well as in their age.

The partition shows that the attitudes and opinions towards reading vary more than the only closed questions was indicating. We find not only bad (clusters 6 and 7) and medium or good readers (clusters 1, 2, 3 and 5) but also nuances that differentiate them. So, the bad readers can be bad students (cluster 7) or fair (cluster 6), the medium or good readers can favor the scholar aspect of reading (cluster 2) or reading as a hobby (cluster 5).

Table 2 presents a detailed description of clusters 1 and 5. These two clusters are clearly different from the closed questions point of view. The children in cluster 1 read only fair enough and with some difficulty while those of cluster 5 read a lot and easily.

Therefore, these clusters do not contain all the children having chosen these items: so, only 56% of the children who declare "I read a lot" are located in cluster 5: the children of this cluster have other characteristics issued from their free answers.

In cluster 5, the children have a rich vocabulary taking into account their youth (10-11 years); for them, reading is not a scholar duty but a hobby. Other children also read a lot but preferably in the scholar context and so they are located in another cluster. This distinction is due to the fact that the open-ended questions are active in the clustering step.

In the cluster 1, the children read only fair enough and with some difficulty. However they are interested by reading, but for scholar reasons. In other clusters, we find children who read fair enough and with some difficulty but without any interest for reading. The characteristics of the children in cluster 1 are also the result of the simultaneous influence of the open-ended and closed questions.

6 Conclusion

To be able to simultaneously take into account open-ended and closed questions for clustering the respondents to a questionnaire is the "natural" wish of the researcher. In such a clustering,

- the closed questions bring a solid framework, whereas they tend to reproduce the *a priori* which underlie the construction of the questionnaire when they are separately analysed;
- the open-ended questions bring their richness, whereas they can disconcert the researcher when they are examined in a separate way.

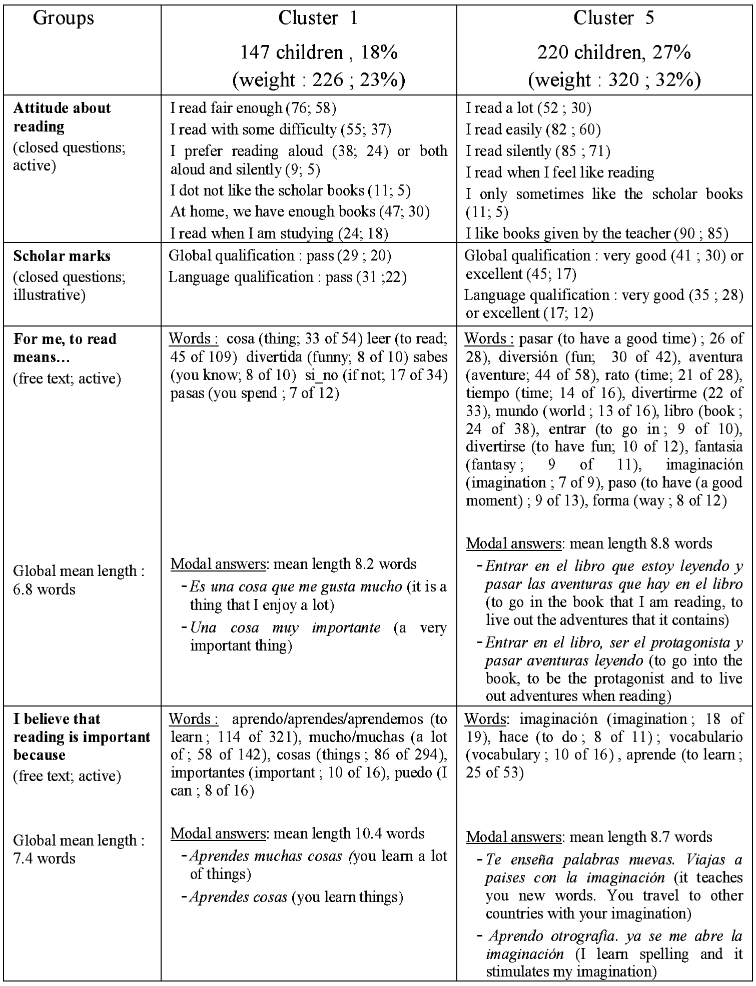
To take into account both points of view seems to allow, at least in this example, for cumulating all their advantages. This approach was little used up to now, indubitably because of the technical problems that are raised. To combine an extension of the MFA, able to deal with contingency tables (MFACT), with the usual MFA, for taking into account the qualitative variables, offers a possibility for such an approach.

References

[Bécue and Lebart, 2000]M. Bécue and L. Lebart. Analyse statistique de réponses ouvertes: application à des enquêtes auprès de lycéens. In J. Moreau, P.A. Doudin, and P. Cazes, editors, *L'analyse des correspondances et les techniques connexes. Approches nouvelles pour l'analyse statistique des données*, pages 59–83, 2000.

[Bécue and Pagès, 1999]M. Bécue and J. Pagès. Intra-sets multiple factor analysis. application to textual data. In H. Bacelar-Nicolau, F. Costa Nicolau, and J. Janssen, editors, *Applied Stochastic Models and Data Analysis*, pages 72–79, 1999.

Table 2. *An excerpt of the description of the clusters*



Groups	Cluster 1 147 children , 18% (weight : 226 ; 23%)	Cluster 5 220 children, 27% (weight : 320 ; 32%)
Attitude about reading (closed questions; active)	I read fair enough (76; 58) I read with some difficulty (55; 37) I prefer reading aloud (38; 24) or both aloud and silently (9; 5) I dot not like the scholar books (11; 5) At home, we have enough books (47; 30) I read when I am studying (24; 18)	I read a lot (52 ; 30) I read easily (82 ; 60) I read silently (85 ; 71) I read when I feel like reading I only sometimes like the scholar books (11; 5) I like books given by the teacher (90 ; 85)
Scholar marks (closed questions; illustrative)	Global qualification : pass (29 ; 20) Language qualification : pass (31 ;22)	Global qualification : very good (41 ; 30) or excellent (45; 17) Language qualification : very good (35 ; 28) or excellent (17; 12)
For me, to read means... (free text; active)	Words : cosa (thing; 33 of 54) leer (to read; 45 of 109) divertida (funny; 8 of 10) sabes (you know; 8 of 10) si_no (if not; 17 of 34) pasas (you spend ; 7 of 12)	Words : pasar (to have a good time) ; 26 of 28), diversión (fun; 30 of 42), aventura (aventure; 44 of 58), rato (time; 21 of 28), tiempo (time; 14 of 16), divertirme (22 of 33), mundo (world ; 13 of 16), libro (book ; 24 of 38), entrar (to go in ; 9 of 10), divertirse (to have fun; 10 of 12), fantasia (fantasy ; 9 of 11), imaginación (imagination ; 7 of 9), paso (to have (a good moment) ; 9 of 13), forma (way ; 8 of 12)
Global mean length : 6.8 words	Modal answers: mean length 8.2 words - *Es una cosa que me gusta mucho* (it is a thing that I enjoy a lot) - *Una cosa muy importante* (a very important thing)	Modal answers: mean length 8.8 words - *Entrar en el libro que estoy leyendo y pasar las aventuras que hay en el libro* (to go in the book that I am reading, to live out the adventures that it contains) - *Entrar en el libro, ser el protagonista y pasar aventuras leyendo* (to go into the book, to be the protagonist and to live out adventures when reading)
I believe that reading is important because (free text; active)	Words : aprendo/aprendes/aprendemos (to learn ; 114 of 321), mucho/muchas (a lot of ; 58 of 142), cosas (things ; 86 of 294), importantes (important ; 10 of 16), puedo (I can ; 8 of 16)	Words: imaginación (imagination ; 18 of 19), hace (to do ; 8 of 11) ; vocabulario (vocabulary ; 10 of 16) , aprende (to learn ; 25 of 53)
Global mean length : 7.4 words	Modal answers: mean length 10.4 words - *Aprendes muchas cosas (*you learn a lot of things) - *Aprendes cosas* (you learn things)	Modal answers: mean length 8.7 words - *Te enseña palabras nuevas. Viajas a paises con la imaginación* (it teaches you new words. You travel to other countries with your imagination) - *Aprendo otrografia. ya se me abre la imaginación* (I learn spelling and it stimulates my imagination)

[Bécue and Pagès, 2001]M. Bécue and J. Pagès. Analyse simulatané de questions ouvertes et de questions fermées. méthodologie, exemple. *Journal de la Société Française de Statistique*, 42(4):91–104, 2001.

[Bécue and Pagès, 2004]M. Bécue and J. Pagès. A principal axes method for comparing contingency tables: Afmtc. *Comput. Statist. Data Anal*, 45(3):481–485, 2004.

[Cazes and Moreau, 2000]P. Cazes and J. Moreau. Analyse des correspondances d'un tableau de contingence dont les lignes et les colonnes sont munies d'une structure de graphe bistochastique. In J. Moreau, P.A. Doudin, and P. Cazes, editors, *L'analyse des correspondances et les techniques connexes. Approches nouvelles pour l'analyse statistique des données*, pages 87–103, 2000.

[Escofier and Pagès, 1998]B. Escofier and J. Pagès. *Analyses Factorielles Simples et Multiples. Objectifs, méthodes et interprétation.* Dunod, Paris, 3 edition, 1998.

[Lebart *et al.*, 1997]L. Lebart, A. Salem, and L. Berry. *Exploring Textual Data.* Kluwer, Dordrecht, 1997.

[Lebart *et al.*, 1998]L. Lebart, A. Morineau, and M. Piron. *Statistique exploratoire multidimensionnelle.* Dunod, Paris, 3 edition, 1998.

Conceptual document indexing using a large scale semantic dictionary providing a concept hierarchy

Martin Rajman, Pierre Andrews, María del Mar Pérez Almenta, and Florian Seydoux

Artificial Intelligence Laboratory, Computer Science Department
Swiss Federal Institute of Technology
CH-1015 Lausanne, Switzerland
(e-mail: Martin.Rajman@epfl.ch, pierre.andrews@cs.york.ac.uk, mariadelmar.perezalmenta@epfl.ch, Florian.Seydoux@epfl.ch)

Abstract. Automatic indexing is one of the important technologies used for Textual Data Analysis applications. Standard document indexing techniques usually identify the most relevant keywords in the documents. This paper presents an alternative approach that aims at performing document indexing by associating concepts with the document to index instead of extracting keywords out of it. The concepts are extracted out of the EDR Electronic Dictionary that provides a concept hierarchy based on hyponym/hypernym relations. An experimental evaluation based on a probabilistic model was performed on a sample of the INSPEC bibliographic database and we present the promising results that were obtained during the evaluation experiments.

Keywords: Document indexing, Large scale semantic dictionary, Concept extraction.

1 Introduction

Keyword extraction is often used for documents indexing. For example, it is a necessary component in almost any Internet search application. Standard keyword extraction techniques usually rely on statistical methods [Zipf, 1932] to identify the important content bearing words to extract. However it has been observed that such extractive techniques are not always efficient, especially in situations where important vocabulary variability is possible.

The aim of this paper is to present a new algorithm that does not extract keywords from the documents, but associates them with concepts representing the contained topics [Rajman *et al.*, 2005]. The use of a concept ontology is necessary for this process. In our work, we use the EDR Electronic Dictionary (developed by the Japan Electronic Dictionary Research Institute [Institute (EDR), 1995]), a semantic database that provides associations between words and all the concepts they can represent, and organizes these concepts in a concept hierarchy based on hyponym/hyperonym relations.

In our approach, the indexing module first divides the documents into topically homogeneous segments. For each of the identified segments, it selects all the concepts in EDR that correspond to all the terms contained in the segment. The conceptual hierarchy is then used to build the minimal sub-hierarchy covering all the selected concepts and this sub-hierarchy is explored to identify a set of concepts that most adequately describes the topic(s) discussed in the document. A "most adequate" set of concepts is defined as a cut in the sub-hierarchy that jointly maximizes specific genericity and informativeness scores.

An experimental evaluation, based on a probabilistic model, was performed on a sample of the INSPEC bibliographic database [INSPEC, 2004] produced by the Institution of Electrical Engineers (IEE). For this purpose, an original evaluation methology was designed, relying on a probabilistic measure of adequacy between the selected concepts and available reference indexings.

The rest of this contribution is organized as follows: in section 2, we describe the EDR semantic database that we use for concept extraction. In section 3, we present the necessary text pre-processing steps that need to be applied for concept extraction to be performed. In section 4, we present the concept extraction algorithm. In section 5, we describe the evaluation framework and the obtained results. Finally, in section 6, we present conclusions and future works.

2 The Data

The EDR Electronic Dictionary [Institute (EDR), 1995] is a set of linguistic resources that can be used for natural language processing. It consists of several parts (dictionaries). For our work, we used the Concept dictionary that provides about 400'000 concepts organized on the basis of hypernym/hyponym relations (See figure 1), and the English word dictionary that provides grammatical and semantic information for each of the dictionary entries. Dictionary entries can be either simple words or compounds.

At the semantic level , the EDR word dictionary provides relations between words and concepts. Notice that, in the case of polysemy, one word can be associated with more than one concept.

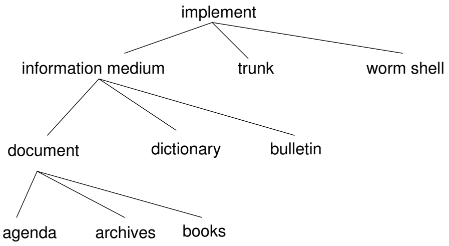


Fig. 1. An example of Concept classification in the EDR Concept dictionary.

3 Pre-processing the texts

Document segmentation The first pre-processing step is document segmentation. Segmentation is necessary because it allows not to have to process simultaneously all the concepts that might be potentially associated with a large document, in which case concept extraction would be computationally inefficient. However, to preserve the quality of the extracted concepts, the used segments must be topically homogeneous. For this purpose, we implemented a simple, well known, Text Tiling technique [Hearst, 1994], where segmentation is based on a measure of proximity between the lexical profiles representing the segments. For the rest of this document, we will consider that the segmentation step has been preformed and the elementary unit for concept association will be the segment, not the document. Once concepts are associated with all the segments corresponding to a document, they are simply merged to produce the set of concepts associated with the document itself.

Tokenization The next pre-processing step is tokenization, which is necessary to decompose the document in distinct tokens that will serve as elementary textual units for the rest of the processing. For this purpose, we used the Natural Language Processing library SLPtoolkit developed at LIA [Chappelier, 2001]. In this library, tokenization is driven by a user defined lexicon, and the resulting tokens can therefore be simple words or compounds. For this purpose, the used lexicon had to be adapted to EDR, so as to contain every possible inflected form of any EDR entry. As EDR does not directly provide these inflected forms, but only the lemmas with inflexion information, we had to write a specific program that exploits the available information to produce the required inflected forms.

Part of Speech Tagging and Lemmatization This pre-processing step consists in identifying, for each token, the lemma and Part Of Speech (POS) category corresponding to its context of use in the document. For our experiments, we used the Brill POS tagger [Brill, 1995].

The output of all the pre-processing steps is the decomposition of each of the identified segments in sequences of lemmas corresponding to EDR entries and associated with the POS category imposed by their context of occurrence. However, because of the polysemy problem already mentioned earlier, this is not sufficient to associate each of the triggered EDR entries with one single concept corresponding to its contextual meaning. Some technique performing semantic disambiguation would be required for that. However, as semantic disambiguation is currently not yet efficiently solved, at least for large scale applications, we decided to keep the ambiguity by triggering all the concepts potentially associated with the (lemma, POS) pairs appearing in the segments. The underlying hypothesis is that some semantic disambiguation will be implicitly performed as a side-effect of the concept selection algorithm. This aspect should however be further investigated.

4 Concept Extraction

The goal of the concept extraction algorithm is to select a set of concepts that most adequately represents the content of the processed document. To do so, we first trigger all the possible concepts that are associated, with the EDR word entries identified in the document. Then, we extract out of the EDR hierarchy the minimal sub-hierarchy that covers all the triggered concepts. This minimal hierarchy, hereafter called the *ancestor closure* (or simply the *closure*), is defined as the part of the EDR conceptual hierarchy that only contains, either the triggered concepts themselves, or any of the concepts dominating them in the hierarchy. Notice that the only constraint imposed on the conceptual hierarchy for the definition of a closure to be valid is that the hierarchy corresponds to a non cyclic directed graph. In such a hierarchy, we call *leaves* (resp. *roots*) all the nodes connected with only incomming (resp. outgoing) links. The EDR hierarchy ideed corresponds to a non cyclic directed graph and, in addition, each of its two distinct parts (the technical concepts and the normal concepts) contains only one single root (hereafter called *the* root).

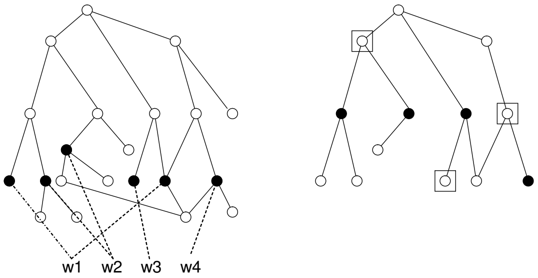


Fig. 2. On the left: links between words and the corresponding triggered concepts. On the right: the corresponding closure and two of its possible cuts (one in black and the other in squares)

Once the closure corresponding to the triggered concepts is produced, the candidates for the possible set of concepts to be considered for representing the content of the document are the different possible *cuts* in the closure.

For any non cyclic directed graph, we define a cut as a minimal set of nodes that dominates all the leaves of the graph. Notice that, by definition, the set of the roots of the graph, as well as the set of its leaves, both correspond a cut.

4.1 Cut Extraction

The idea behind our approach is to extract a cut that optimally represents the processed document. To do so, our algorithm explores the different cuts in the closure, scores them, and select the best one with respect to the used score. As a cut can be seen as a more or less abstract representation of the leaves of the closure, the score of a cut is computed relatively to the covered leaves. In our algorithm, a local score is first computed for the concepts in the cut, and a global score is then derived for the cut from the obtained local scores. Notice also that, as the number of cuts in a closure might be exponential, evaluating the scores of all possible cuts is not realistic for real size closures. A dynamic programming algorithm was therefore developed to avoid intractable computations [Rajman *et al.*, 2005].

In this algorithm, the local score U (the definition of U is given in section 4.2) is computed for each concept c in the cut. This local score measures how much the concept c is representative of the leaves of the closure. The global score of the cut is then computed as the average of U over all concepts in the cut.

4.2 Concept Scoring

The local score U is decomposed into two specific components, genericity and informativeness.

Genericity It is quite intuitive that, in a conceptual hierarchy, a concept is more generic than its subconcepts. At the same time, the higher a concept lays in the hierarchy, the larger is the number of the leaves it covers. Following this, a simple score S_1 was defined to describe the genericity of a concept. We made the assumption that this score should be proportional to the total number $n(c)$ of leaves covered by the concept c. Because of the linearity assumption, the score S_1 of a concept c can therefore be written as:
$S_1(c) = \frac{n(c)-1}{N-1}$, where N is the total number of leaves in the closure.
Informativeness If only genericity would be taken into account, our algorithm would always select the roots of the closure as the optimal cut. Therefore, it is important to also take into account the amount of information preserved about the leaves of the closure by the concepts selected in the cut. To quantify this amount of preserved information, we defined a second score S_2 for which we made the assumption that the score $S_2(c)$ defined for a concept c in a cut should be linearly dependent on the average normalized path length $d(c)$ between the concept c and all the leaves it covers in the closure. Because of the linearity assumption, the score S_2 of a concept c can therefore be written as: $S_2(c) = 1 - d(c)$.
Score Combination As two scores are computed for each concept in the evaluated cut, a combination scheme was necessary to combine S_1 and S_2 into a single score. A weighted geometric mean was chosen:
$U(c) = S_1(c)^{1-a} \times S_2(c)^a$.

The parameter a offers a control over the number of concepts returned by the selection algorithm. If the value of a is close to one, then it will favor the score S_2 over S_1, and the algorithm will extract a cut close to the leaves, whereas a value close to zero will favor S_1 over S_2 and therefore yield more generic concepts in the cut.

5 Evaluation

The evaluation of the Concept Extraction algorithm was made on a sample from the INSPEC Bibliographic database, a bibliographic database about physics, electronics and computing [INSPEC, 2004]. The sample was composed of short abstracts manually annotated with keywords extracted from the abstracts. For the evaluation, a set of 238 abstracts was randomly selected in database, and these abstracts were manually associated with two sets of concepts: the ones corresponding to a simple keyword in the reference annotation, and, the ones corresponding to compound keywords.

In our case, only the concepts of the first kind were considered and all compound keywords were first decomposed into their elementary constituents and then associated with the corresponding concepts.

To measure the similarity between the concept derived from the reference annotation and the ones produced by our algorithm, we used the standard Precision and Recall measures. For any indexed document, *Precision* is the fraction of identified correct concepts in all concepts associated with the document by the algorithm, while *Recall* is the fraction of the identified correct concepts in all concepts associated with the document in the reference annotation. For any set of documents, the quality of the concept association algorithm was then measured by the average Precision and Recall scores over all the documents in the sample.

However, if applied directly, an evaluation based on Precision and Recall scores would be quite inadequate, as it does not take at all into account the hyponym/hypernym relations relating the concepts. For example if a document is indexed by the concept "dog" and the algorithm produces the concept "animal", this should not be considered as a total failure as it would be the case with the standard definition of Precision and Recall. To take this into account, we replaced the binary match between produced and reference concepts by a similarity measure based on the available concept hierarchy. The selected similarity measure was the Leacock-Chodorow similarity [Leacock and Chodorow, 1998] that corresponds to the logarithm of the normalized path length between two concepts. The probabilistic model then used for the evaluation was the following: the normalized version of the concept similarity between a produced concept c_i and a reference concept C_k, denoted by $p(c_i, C_k)$, is interpreted as the probability that the concepts c_i and C_k can match. Then, if $Prod = \{c_1, c_2, ..., c_n\}$ is the set of concepts produced for a document and $Ref = \{C_1, C_2, ..., C_N\}$ is the corresponding set of reference concepts, for each concept c_i (resp. C_k) the probability that it matches the reference set Ref (resp. the produced set $Prod$) is: $p(c_i) = 1 - \prod_{k=1}^{N}(1 - p(c_i, C_k))$ (resp.$p(C_k) = 1 - \prod_{i=1}^{n}(1 - p(c_i, C_k))$), and the expectations for Precision and Recall can therefore be computed as: $E(P) = \frac{1}{n} \times \sum_{i=1}^{n} p(c_i)$ and $E(R) = \frac{1}{N} \times \sum_{i=k}^{N} p(C_k)$.

For the obtained expected values for P and R, the usual *F-measure* can then be computed.

5.1 Results and Interpretation

A first experiment was carried out to select which value of a should be used for the evaluation. Observing the average results obtained for each value of a (see figure 3), one can see that a has a very limited impact on the algorithm performance (the F-measure is quasi constant until a = 0.6). The obtained results therefore seem to indicate that the value of a can be chosen almost arbitrarily between a=0.1 and a=0.7.

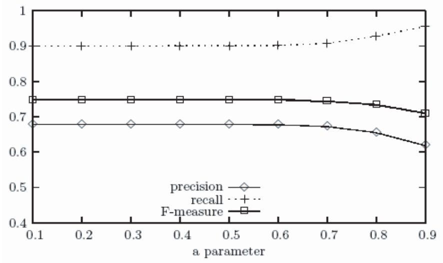
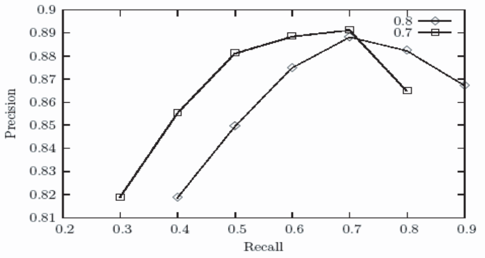


Fig. 3. Comparison of the algorithm results with varying values of a

In a second step, the following procedure was applied to compute the average Precision and Recall: (1) all the probabilities $p(c_i)$ and $p(C_k)$ were computed for each document in the evaluation corpus;(2) the concepts c_i in Prod and C_k in Ref were sorted by decreasing probabilities;(3) for each value Θ in an equi-distributed set of threshold values in [0,1[, an average (Precision, Recall) pair was computed, taking only into account the concepts c for which $p(c) > \Theta$;(4) average values of Precision, Recall and F-Measure were computed over all the produced pairs.



a	Accuracy	Precision	Recall	F-Measure
0.6	0.93746	0.865545	0.763733	0.78103
0.7	0.940054	0.864756	0.779126	0.79017
0.8	0.948855	0.867199	0.837167	0.825773
0.9	0.960364	0.865917	0.911771	0.870373

Fig. 4. averaged(non-interpolated)Precision/Recall curves and the corresponding average result table for two values of the a parameter

The obtained curves shown in figure 4 display an interesting behavior: when Recall increases, Precision first starts to raise and then falls down. This might be explained by the fact that cuts corresponding to higher Recall values contain more concepts and that there is therefore a good chance that these concepts are lower in the hierarchy and have more chances to be close to the concepts in the reference. Then, when the number of produced concepts is too large, its exceeds what is necessary to cover the reference concepts and the added noise therefore entails a drop in Precision. A second interesting observation is that, for a=0.6 and a=0.7, there are no (Precision, Recall) pairs with Recall larger than 0.8. This might be explained by the fact that, for small values of a, there is only a small chance that the extracted cut is specific enough to have a good probability to match all the reference concepts, and therefore makes it hard to reach high values of Recall.

6 Conclusion

Current approaches to automatic document indexing mainly rely, on purely statistical methods, extracting representative keywords out of the documents. The novel approach proposed in this contribution gives the possibility of associating concepts instead of extracting keywords. For that, the construction of the ancestor closure over the segment's concepts is used to choose the best representative set of concepts to describe the document's topics. The novel evaluation method developed to measure the proposed concept extraction algorithm lead to promising results in terms of Precision and Recall, and also gave the opportunity to observe interesting features of the concept association mechanism. It proved that extracting concepts instead of simple keywords can be beneficial and does not require intractable computation.

As far as future works are concerned, more sophisticated methods to solve the ambiguity in concept association related to word polysemy should be investigated. A more general theoretical framework providing some well grounded justification for the scoring scheme should also be worked out.

References

[Brill, 1995]Eric Brill. Transformation-based error-driven learnig and natural language processing: a case study in part-of-speech tagging. *Computational Linguistics*, pages 21(4):543–565, 1995.

[Chappelier, 2001]Jean-Cedric Chappelier. Slptoolkit, http://liawww.epfl.ch/~chaps/SlpTk/, EPFL, 2001.

[Hearst, 1994]Marti Hearst. Multi-paragraph segmentation of expository text. *32nd Annual Meeting of the Association for Computational Linguistics*, 1994.

[INSPEC, 2004]Institution of Electrical Engineers INSPEC. http://www.iee.org/Publish/INSPEC/, United Kingdom, 2004.

[Institute (EDR), 1995]Japan Electronic Dictionary Research Institute (EDR). http://www.iijnet.or.jp/edr, Japan, 1995.

[Leacock and Chodorow, 1998]C. Leacock and M. Chodorow. Wordnet: An electronic lexical database, chapter combining local context and wordnet similarity for word sense identification. *MIT Press*, 1998.

[Rajman *et al.*, 2005]M. Rajman, P. Andrews, M. Pérez Almenta del Mar, and F. Seydoux. Using the edr large scale semantic dictionary: application to conceptual document indexing. *EPFL Technical Report (to appear)*, 2005.

[Zipf, 1932]G.K. Zipf. *Selective Studies and the Principle of Relative Frequency in Language.* Harvard University Press, Cambridge MA, 1932.

Synonym Dictionary Improvement through Markov Clustering and Clustering Stability*

David Gfeller, Jean-Cédric Chappelier, and Paolo De Los Rios

Ecole Polytechnique Fédérale de Lausanne (EPFL)
CH-1015 Lausanne, Switzerland

Abstract. The aim of the work presented here is to clean up a dictionary of synonyms which appeared to be ambiguous, incomplete and inconsistent. The key idea is to use Markov Clustering and Clustering Stability techniques on the network that represents the synonymy relation contained in the dictionary. Each densely connected cluster is considered to correspond to a specific concept, and ambiguous words are identified by the evaluation of the stability of the clustering under noise. This allows to disambiguate polysemic words, introducing missing senses where required, and merge similar senses of a same word if necessary.

Keywords: Complex Network, Markov Clustering Algorithm, Clustering Stability, Community Structure, Synonymy, Word Sense Disambiguation.

Introduction

One aspect of the complexity of text mining comes from the synonymy and the polysemy of the words. The aim of Word Sense Disambiguation (WSD) is precisely to associate a specific sense to every word within context [Besançon *et al.*, 2001, Schütze, 1998]. The determination of the list of possible senses for a given word is a key aspect in this disambiguation. A possible starting point could be a dictionary of synonyms. It happens however that, due to both inherent human errors and errors coming from the automatic (semi-supervised) construction of the dictionary, the synonymy network can contain several mistakes and turn out to be ambiguous, incomplete or inconsistent.

This paper starts with a brief description of the synonymy network derived from the dictionary we used. Then, the clustering algorithm applied to improve it is presented and the general method to identify ambiguities in the clustering is introduced. Finally, the results obtained are discussed.

1 The Synonymy Network

The synonym network we here consider has been built from a French dictionary of synonyms. The synonymy relation is defined between the words in one of their senses and is considered to be symmetric. The resulting network

* This work was financially supported by the EU commission by contracts COSIN (FET Open IST 2001-33555) and DELIS (FET Open 001907).

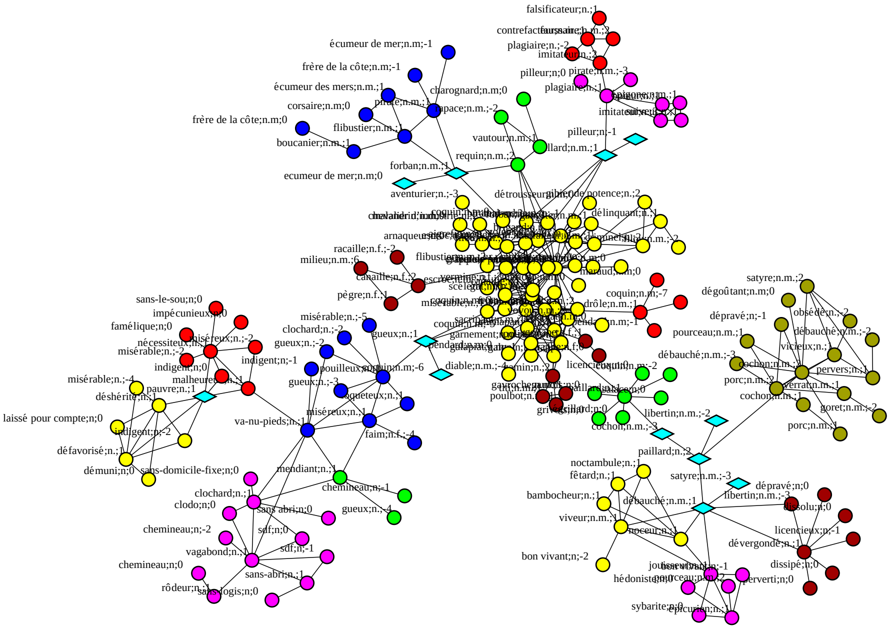


Fig. 1. MCL clustering of the component with 185 elements (different level of grey represent different clusters). Unstable nodes (explained in section 3) are represented with diamonds.

is thus undirected (and unweighted). It is not fully connected but consists of many disconnected components[1], with a power-law size distribution (see fig. 2); the nodes inside a single component representing the same "concept"[2].

The first encountered problem in this dictionary was the existence of much too large synonym components (up to almost 10,000 nodes, see fig. 2). The transitive closure of the synonymy relation connects words which have different senses; e.g. *fêtard* ("merrymaker") and *sans-abri* ("homeless") (see fig. 1). This is due to words which are still ambiguous[3] and relate different "concepts": even if a path exists between two nodes, the slight changes in the senses that could occur at each step along this path may result in a quite different sense between both ends. Moreover these big components clearly show a sub-structure suggesting a partition into smaller clusters.

The second encountered problem was that some words were given too many senses, i.e. senses that actually correspond to the same "concept".

To solve these problems, the Markov clustering algorithm (MCL) [Van Dongen, 2000] was first applied to the synonymy network. The idea is that words with tighter neighborhoods are likely to be less ambiguous than words with

[1] i.e. groups of words which are claimed to be synonyms

[2] We use the word "*concept*" to denote a group of senses that are synonyms.

[3] i.e. the distinction between two of their senses has not been introduced

fuzzier neighborhoods [Sproat and van Santen, 1998], and thus a cluster is likely to correspond to words with a close meaning, and therefore represents one "concept". Then, in order to find ambiguous words, we noise the edges and compare the clusters obtained for several noisy realizations of the network. This provides informations about the network that could not have been extracted with the standard clustering algorithms.

2 Markov Clustering

In this section, MCL, the clustering algorithm we used for splitting the big components into smaller clusters, is briefly described.

Its basis is that "*a random walk on a network that visits a dense cluster will likely not leave it until many of its vertices have been visited*" [Van Dongen, 2000]. The idea is to favor the most probable random walks, increasing the probability of staying in the initial cluster. The algorithm works as follows: (1) consider the adjacency matrix of the network[4] ; (2) normalize each column to one, in order to obtain a stochastic matrix S; (3) square the matrix S; the element $(S^2)_{ij}$ is the probability that a random walk starting at node j ends up at node i after two steps; (4) take the r^{th} power of every element of S^2 (typically $r \approx 1.5 - 2$); this favors the most probable random walks; (5) go back to 2 until convergence.

After several iterations we end up with a matrix stable under MCL. Only a few lines of the matrix have non-zero entries, which give the cluster structure of the network. Note that the parameter r can tune the granularity of the clustering: a small r corresponds to a few big clusters, whereas a big r corresponds to smaller clusters. Comparing the results with different r for some of the components, we chose $r = 1.6$ as a reasonable value.

As an example, the result of MCL on the component of size 185 is displayed in fig. 1. The obtained subdivision into smaller clusters is definitely more meaningful; e.g. *fêtard* is no longer in the same cluster as *sans-abri*.

MCL was applied on the biggest components of the network. We noticed that, as the size of the components becomes smaller than 40, the clustering is often not meaningful anymore, since the components do not show any particular community structure. After clustering the biggest components, the cluster size distribution shown in fig. 2 is obtained. The power-law is still conserved, but the size of the biggest components is much reduced.

3 Unstable Nodes

MCL partitions the network into clusters without overlap, i.e. every node is assigned to a single cluster ("hard-clustering"). However, the resulting

[4] For an undirected and unweighted network, this matrix is symmetric and composed only of zeros and ones.

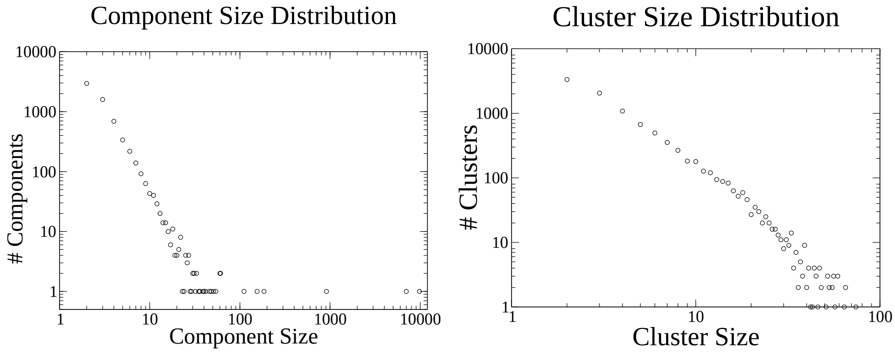


Fig. 2. Left: Distribution of the size of the components in the whole network (log-log scale). Right: Distribution of the size of the clusters after MCL with $r = 1.6$.

clustering is sometimes questionable – both from a topological and a linguistic point of view –, especially for nodes that "lie on the border" between two clusters (see fig. 3).

The problem of finding ambiguities is closely related to the evaluation of the robustness of the clustering. Some attempts, based on particular clustering algorithms, were drawn recently to solve this problem [Wilkinson and Huberman, 2004, Reichardt and Bornholdt, 2004]. We present here a new method based on the introduction of a stochastic noise over the edges of the network, and apply it in the framework of MCL. This consists in adding noise over the non-zero entries of the adjacency matrix[5]. Running MCL with noise several times, some nodes are switching from one cluster to the other (for example node "*reprendre;6*" in fig. 3). This procedure is now detailed.

Let p_{ij} be the probability for the edge between node i and node j of being inside a cluster. After several runs of the clustering algorithm with the noise, a weighted network is obtained where edges with probability 1 are always within a cluster and edges with probability close to 0 connect two different clusters. Edges with a probability smaller than a threshold θ are thus considered as "*external edges*". By removing those edges, one gets a disconnected network[6].

[5] In this study, the noise added over the edges weights (originally equal to 1) is equally distributed in $[-\sigma, \sigma]$, $0 < \sigma < 1$. With σ close to 0, unstable nodes are not detected, while with $\sigma \simeq 1$ the topology of the network changes dramatically. The results were stable for a broad range of values of σ around 0.5. For example in the component displayed in fig. 3, the node "*reprendre;6*" was identified as the only unstable node for $0.35 \leq \sigma \leq 0.8$.

[6] For the choice of the parameter θ, we looked at the distribution of the probabilities p_{ij} over the whole network. As expected, this distribution has a clear maximum in 1, corresponding to edges that are never cut by MCL, preceded by a region corresponding to edges almost never cut. Since for $p_{ij} \leq 0.8$ the distribution is almost flat, we choose $\theta = 0.8$.

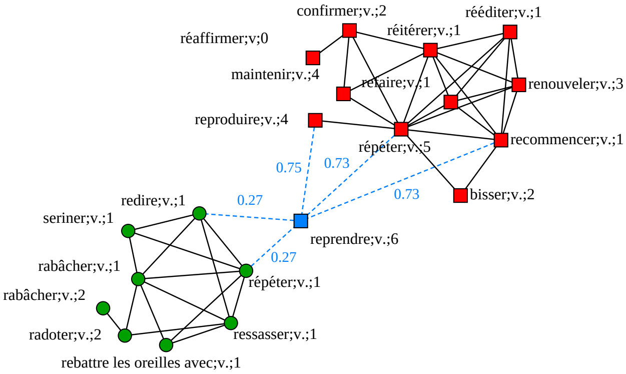


Fig. 3. Small sub-network with one unstable node (“*reprendre;6*”), extracted from a component of 111 nodes. The values over the dashed edges are the probabilities for the edges to be inside a cluster (average over 100 realizations of the clustering with $r = 1.6$, and $\sigma = 0.5$.). Only probabilities smaller than $\theta = 0.8$ are shown. The shape of the nodes indicates the cluster found without noise.

In this section, we use the word “*cluster*” for the clusters obtained without noise, and “*subcomponent*” for the disconnected parts of the network after the removal of the external edges. If the community structure of the network is stable under several repetitions of the clustering with noise, the subcomponents of the disconnected network correspond to the clusters obtained without noise. In the opposite case, a new community structure appears with some similarity with the initial one.

In order to identify which subcomponents correspond to the initial clusters and which are new subcomponents, we introduce the notion of similarity between two sets of nodes. If E_1 (resp. E_2) is the set of clusters (resp. the set of subcomponents), we use the Jaccard index to define the similarity s_{ij} between cluster $C_{1j} \in E_1$ and subcomponent $C_{2i} \in E_2$: $s_{ij} = \frac{|C_{2i} \cap C_{1j}|}{|C_{2i} \cup C_{1j}|}$.

If $C_{1j} = C_{2i}$, $s_{ij} = 1$ and if $C_{1j} \cap C_{2i} = \emptyset$, $s_{ij} = 0$. For every $C_{1j} \in E_1$, we find the component C_{2i} with the maximal similarity and identify it with the cluster C_{1j} (C_{2i} often corresponds to the stable core of the cluster C_{1j}). If there is more than one of such components, none of them is identified with the cluster. In practice, this latter case is extremely rare.

Nodes belonging to subcomponents that have never been identified with any cluster could be defined as *unstable* nodes. However, this definition suffers some drawbacks since it sometimes happens that a big cluster splits into two subcomponents of comparable size. Considering that almost half of the nodes of the cluster are unstable is not realistic and a new cluster should be defined instead. In practice, subcomponents of four nodes or more often correspond to a cluster not detected by the algorithm. We therefore define the unstable nodes as the nodes belonging to subcomponents that have not been identified with a cluster and whose size is smaller than 4.

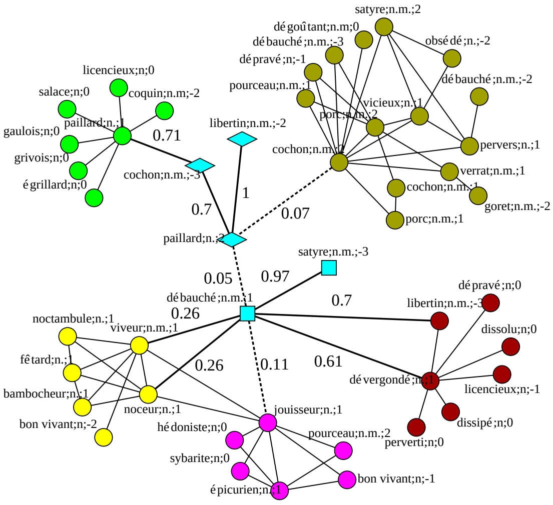


Fig. 4. Zoom over the bottom-right of fig. 1. Five unstable nodes have been found (non-circle nodes). The splitting of them proceeds as follows: removing the edges with $p_{ij} < 1-\theta$ (dashed-line), they are divided into two groups ({*cochon;-3 libertin;-2, paillard;2*} (diamonds) and {*débauché;1, satyre;-3*} (squares). The first group has only one adjacent subcomponent. It is therefore merged to this subcomponent. The second group has two adjacent subcomponents. It is thus duplicated and merged into those two subcomponents.

In the framework of a synonymy network, unstable nodes, which lie on the border of two subcomponents, correspond to polysemic words which have not been clearly identified as such (i.e. one of their senses is not present in the dictionary). We thus decided to split these nodes among their adjacent subcomponents. The adjacent subcomponents are defined as the subcomponents to which the node is connected through at least one edge with a probability higher than a given threshold θ'. Typically we choose $\theta' = 1 - \theta$, where θ was the threshold for defining an edge as external.

If several unstable nodes are connected together, we split them according to the following procedure: first group these nodes keeping only the edges with $p_{ij} > \theta'$; then, for each group, duplicate it and join it to its adjacent subcomponents (see fig. 4).

Finally, the second problem, where the same word appears with different senses in a cluster (e.g. *rabâcher* in fig. 3) has been addressed by simply merging into a single node the nodes that correspond to the same word in a same subcomponent. Indeed, if a node appears twice in a subcomponent, both senses are actually not different, at least not at the level of granularity used. Such a situation occurs 4,642 times in the whole network (the total number of nodes is 50,913). This number is more than four times smaller than before the MCL clustering (21,261 "duplicates").

4 Discussion and Conclusion

The objective evaluation of the work presented here is not easy. How to evaluate whether better results are obtained when no reference to compare to is available[7]? One possible evaluation could be to compare the performances obtained in a targeted WSD experiment using the original and the corrected resource. It is however highly dependent on the targeted application. We thus rather choose to evaluate the method presented here by a subjective (i.e. human centered) validation, achieved by sampling components in the newly obtained network. This evaluation appeared to be very convincing[8].

However, we still wanted to develop objective overall clues from the network to try to objectively grasp the benefits of the method. We, for instance, computed the clustering coefficient of the unstable nodes. The clustering coefficient (C) of a node is the number (N_3) of 3-loops passing through the node, divided by the the maximal possible number of such 3-loops. When the node has degree k, $C = \frac{2\,N_3}{k(k-1)}$. If a node lies in the middle of a densely connected cluster, it quite likely has a high clustering coefficient. If the node lies between clusters, it has a small clustering coefficient[9].

Using MCL, the assumption is made that a community structure is present in the network. Since MCL may also give a partition of random networks without any community structure, it is important to validate this assumption. The introduction of the probability p_{ij} over the edges provides a way to do so. In the case of a random network, the community structure found by the clustering algorithm is expected to be very sensitive to the noise, whereas in the case of a network with a clear community structure, the clusters are quite stable, except for a few unstable nodes. To characterize these two situations, we introduce the *clustering entropy* S_c as a measure of the clustering stability:

$$S_c = -\frac{1}{M} \sum_{(i,j)} \Big(p_{ij} \log_2 p_{ij} + (1 - p_{ij}) \log_2 (1 - p_{ij}) \Big),$$

where M is the total number of edges (i, j) in the network.

Important differences in the clustering entropy between networks with a clear community structure and random networks with no community structure are expected. If the network is totally unstable (i.e. $p_{ij} = \frac{1}{2}$ for all edges), $S_c = 1$, while if the edges are perfectly stable ($p_{ij} = 0$ or 1), $S_c = 0$.

To avoid biasing the results, we compare the components with a randomized version of the same component in which the degree of each node is conserved [Maslov and Sneppen, 2002]. Table 1 shows the comparison for several big components of the network of synonyms. The clustering entropy

[7] had we had one, wouldn't have we developed a method to correct it!

[8] See fig. 1, 2, 4 and 3 for illustrations.

[9] For example, twelve unstable nodes were found in the component displayed in fig. 1. The average clustering coefficient of these nodes is 0.08, while the average clustering coefficient over the whole component is 0.42.

Component Size	S_c(original)	S_c(random)
912	0.25±0.01	0.55±0.01
185	0.19±0.01	0.62±0.02
155	0.27±0.01	0.55±0.03
111	0.21±0.01	0.69±0.02
61	0.20±0.01	0.68±0.04
60	0.19±0.01	0.76±0.04
54	0.21±0.01	0.60±0.07
51	0.21±0.01	0.69±0.05
average	0.21	0.64

Table 1. Comparison for several components. S_c(original) is the clustering entropy of the original components. S_c(random) is the average clustering entropy for 50 randomized versions of the component. We used $r = 1.6$ and $\sigma = 0.5$.

of the randomly rewired components is at least twice bigger than the entropy of the original components. This experimentally shows that the clusters obtained with MCL are not an artifact of the method, but correspond to a real community structure in the network.

Applying the MCL clustering algorithm, the network of synonyms splits into sensible clusters, significantly improving, at least subjectively, the quality of the dictionary. Most of the clusters can be interpreted as groups of synonyms and, at a coarse-grained level of representation, correspond to a general concept of the language. The method introduced to identify nodes which lie between clusters and to check the robustness of the clustering appeared to be fruitful in the splitting of polysemic nodes. We emphasize that this method does not depend of a particular clustering algorithm and can be applied on any complex network.

References

[Besançon *et al.*, 2001]R. Besançon, J.-C. Chappelier, M. Rajman, and A. Rozenknop. Improving text representations through probabilistic integration of synonymy relations. In *Proc. of ASMDA'2001*, pages 200–205, 2001.

[Maslov and Sneppen, 2002]S. Maslov and K. Sneppen. Specificity and stability in topology of protein networks. *Science*, 296:910–913, 2002.

[Reichardt and Bornholdt, 2004]J. Reichardt and S. Bornholdt. Detecting fuzzy community structures in complex networks with a potts model. *Phys. Rev. Lett.*, 93(218701), 2004.

[Schütze, 1998]H. Schütze. Automatic word sense discrimination. *Computational Linguistics*, 24(1):97–123, 1998.

[Sproat and van Santen, 1998]R. Sproat and J. van Santen. Automatic ambiguity detection. In *Proc. of ICSLP'98*, 1998.

[Van Dongen, 2000]S. Van Dongen. *Graph clustering by flow simulation.* PhD thesis, University of Utrecht, 2000.

[Wilkinson and Huberman, 2004]D.M. Wilkinson and B.A Huberman. A method for finding communities of related genes. *PNAS*, 101:5241–5248, 2004.

Evaluation of a Probabilistic Method for Unsupervised Text Clustering

Loïs Rigouste, Olivier Cappé, and François Yvon

Ecole Nationale Supérieure des Télécommunications (GET / CNRS UMR 5141)
46 rue Barrault, 75634 Paris Cedex 13, France
(e-mails: rigouste, cappe, yvon at enst.fr)

Abstract. In this contribution, we investigate the use of a simple probabilistic model for unsupervised document clustering in large collections of texts. The model consists of a mixture of multinomial distributions over the word counts, each component corresponding to a different theme.

The evaluation corpus is a medium size subset of the Reuters news feed, which comes with a manual categorization. The similarity between the clustering produced and this existing categorization is computed in terms of mutual information, and compared to the variations of log-likelihood and perplexity. We analyze the influence of the smoothing parameter, of the size of the vocabulary and of the addition of supervised information.

Our results, which are somewhat more pessimistic than those usually found in the literature, show that it is difficult to reach the quality of the manual categorization when no hint is given at the initialization step. We also show that a side effect of the so-called "curse-of-dimensionality" is that this probabilistic model yields the same results as a simpler, hard clustering algorithm.
Keywords: Text Mining, Unsupervised Clustering, Evaluation.

1 Introduction

Due to the wide availability of huge collections of text documents (news corpora, e-mails, web pages, scientific articles...), unsupervised clustering has emerged as an important text mining task. Several probabilistic models, performing a soft (non-deterministic) clustering of the data, such as Probabilistic Latent Semantic Analysis [Hofmann, 2001] or Latent Dirichlet Allocation [Blei *et al.*, 2002], have been introduced for that purpose. In this contribution, we study the simpler model [Nigam *et al.*, 2000, Clérot *et al.*, 2004] in which the corpus is represented by a mixture of multinomial distributions, each component corresponding to a different "theme". Dirichlet priors are set on the parameters and we use the Expectation-Maximization (EM) algorithm to obtain maximum a posteriori (MAP) estimates of the parameters.

To get a deeper understanding of the potentials of this approach, we consider a reasonably simple corpus, consisting of 5000 Reuters news stories taken from five different categories (as defined by Reuters). After introducing the two measures used for evaluation (perplexity and mutual information between the obtained themes and the Reuters categorization), we investigate the influence of several aspects of the model. An interesting experimental outcome of this study is to show that, due to the high dimensionality of the problem, the model behaves almost like a hard clustering algorithm (with a specific distance measure).

2 The Model

We denote by n_D, n_W and n_T, respectively, the number of documents, the size of the vocabulary and the number of themes (that is, the number of components of the mixture model). Since we use a bag-of-words representation, the corpus is fully determined by the count matrix $C = (C_d(w))_{d=1\ldots n_D, w=1\ldots n_W}$, where the notation C_d is used to refer to the word counts of a specific document d. The multinomial mixture model is such that:

$$\mathrm{P}(C_d; \alpha, \beta) = \sum_{t=1}^{n_T} \alpha_t \frac{l_d!}{\prod_{w=1}^{n_W} C_d(w)!} \prod_{w=1}^{n_W} \beta_{wt}^{C_d(w)} \tag{1}$$

which corresponds to the following probabilistic generative mechanism:

- sample a theme t in $\{1, \ldots, n_T\}$ with probabilities $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_{n_T})$;
- sample l_d (length of document d) words from a multinomial distribution with parameter $(l_d;\, \beta_{1t}, \beta_{2t}, \ldots, \beta_{n_W t})$.

The notation β is used to denote the collection of theme-specific word frequencies. Note that the document length itself is taken as an exogenous variable and its distribution is not accounted for in the model. As all documents are assumed to be independent, the corpus log-likelihood $\mathcal{L}$ is given by $\sum_{d=1}^{n_D} \log \mathrm{P}(C_d; \alpha, \beta)$.

To estimate the model parameters, we use the Expectation-Maximization (EM) algorithm with independent noninformative Dirichlet priors on α (with hyperparameter θ_α) and on the columns $\beta_{\bullet t}$, for $t = 1, \ldots, n_T$ (with hyperparameter θ_β). Denoting the current estimates of the parameters by α' and β' and the latent (unobservable) theme of document d by T_d, it is straightforward to check that each iteration of the EM algorithm updates the parameters according to:

$$\mathrm{P}(T_d = t | C; \alpha', \beta') = \frac{\alpha'_t \prod_{w=1}^{n_W} \beta'^{C_d(w)}_{wt}}{\sum_{t'=1}^{n_T} \alpha'_{t'} \prod_{w=1}^{n_W} \beta'^{C_d(w)}_{wt'}} \tag{2}$$

$$\alpha_t \propto \theta_\alpha - 1 + \sum_{d=1}^{n_D} \mathrm{P}(T_d = t | C; \alpha', \beta') \tag{3}$$

$$\beta_{wt} \propto \theta_\beta - 1 + \sum_{d=1}^{n_D} C_d(w)\, \mathrm{P}(T_d = t | C; \alpha', \beta') \tag{4}$$

where the normalization factors are determined by the constraints $\sum_{t=1}^{n_T} \alpha_t = 1$ and $\sum_{w=1}^{n_W} \beta_{wt} = 1$, for t in $\{1, \ldots, n_T\}$. It turns out that θ_α has little, if any, influence and we set $\theta_\alpha = 1$ in the following. For obvious reasons, we refer to $\theta_\beta - 1$ as the smoothing parameter. We set it to 0.1 to begin with.

3 Evaluation

To evaluate the performance of the model for unsupervised document clustering we use two different measures. The *perplexity*

$$\widehat{\mathcal{P}}^\star = \exp[-\frac{1}{l^\star} \sum_{d=1}^{n_D^\star} \log(\sum_{t=1}^{n_T} \alpha_t \prod_{w=1}^{n_W} \beta_{wt}^{C_d^\star(w)})]$$

quantifies how much the model is able to predict new data, denoted generically by the star superscript. The normalization by the total number of word occurrences $l^\star$ in the test corpus $C^\star$ is conventional and used to allow comparison with simpler models such as the unigram model, which ignores the document level. A second indicator is the *mutual information* between the clustering produced by the model and the Reuters categories, which is more directly related to our ability to accurately cluster the data. It is defined as:

$$\begin{aligned}\widehat{\mathcal{MI}}^\star = \sum_{c=1}^{n_C} \sum_{t=1}^{n_T} (\frac{1}{n_D^\star} \sum_{d=1}^{n_D^\star} \mathrm{P}(\mathit{\Gamma}_c | C_d^\star) \, \mathrm{P}(T_d = t | C_d^\star)) \\ \times \log \frac{n_D^\star \sum_{d=1}^{n_D^\star} \mathrm{P}(\mathit{\Gamma}_c | C_d^\star) p(T_d = t | C_d^\star)}{(\sum_{d=1}^{n_D^\star} \mathrm{P}(\mathit{\Gamma}_c | C_d^\star))(\sum_{d=1}^{n_D^\star} \mathrm{P}(T_d = t | C_d^\star))}\end{aligned}$$

where $\mathrm{P}(\mathit{\Gamma}_c | C_d)$ is the "probability" that document d belongs to category $\mathit{\Gamma}_c$ (usually 0 or 1, as most documents belong to a unique Reuters category) and $\mathrm{P}(T_d = t | C_d)$ is the output of the model (probability that the document d belongs to theme t). The estimated mutual information is then normalized, respectively, by the marginal entropies of the themes and categories. The harmonic average of those scores (between 0 and 1) is referred to as the *(MI) F-Score.*

3.1 Baseline Performance

We selected 5,000 texts from the 2000 Reuters Corpus, from five well-defined categories (arts, sports, health, disasters, employment). All experiments are performed using ten-fold cross-validation (with 10 random splits of the corpus), with 30 iterations of the EM algorithm for each run and with five themes ($n_T = 5$). As will be seen below, initialization of the EM algorithm does play a very important role in obtaining meaningful document clusters. After a bit of experimentation, we found that a good option is to make sure that, initially, all clusters overlap significantly and that none of the theme-dependent word probabilities is too small. The "Dirichlet" initialization thus consists in sampling an initial (fictitious) configuration of posterior probabilities in (2) which is close to equiprobability*.

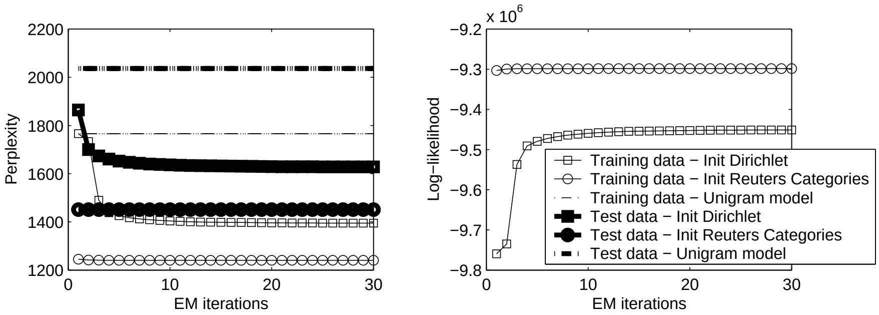


Fig. 1. Evolution of Perplexity and Log-likelihood over EM iterations.

To get an idea about the best achievable performance, we also used the Reuters categories as initialization. We establish a one-to-one mapping between the mixture components and the Reuters categories, setting for every document the initial posterior probability in (2) to 1 for a given theme. Figure 1 displays the corresponding training data likelihood (right) and perplexity as a function of the number of iterations. The first striking observation is that the gap between both initializations is huge. With the "Dirichlet" initialization, we are able to predict the word distribution more accurately than with the unigram model but much worse than with the somewhat ideal initialization. This gap is also patent for the training data log-likelihood. In the following, we report only the values obtained after the last EM iteration, since the variations after the first few iterations are small (note that this phenomenon is particularly marked for the Reuters initialization). Also, we no more report the perplexity on the training data since it conveys the same information as log-likelihood.

The Mutual Information F-Score is similarly oriented with a final value of 0.87 for the Reuters initialization and 0.25 for the "Dirichlet" one. To get an idea of the signification of these numbers, we randomly perturbated a certain amount of the Reuters tags and computed the MI F-Score with the original

* It is not possible to start with exact equiprobability, or, else, it can be seen from the update equations that all word distributions are similar and the clusters never separate from one another. Hence we sample from a Dirichlet distribution with the same parameter for every component. This variable controls the variance of the probabilities sampled. It also has an interesting influence on the results that we do not develop here.

categorization. Proceeding this way, perturbing (respectively) 5%, 15% and 50% of the document labels gives F-Score of 0.9, 0.7 and 0.25. Hence 0.25 corresponds to a rather poor performance. Now we check if this gap between both initializations can be reduced when tuning the smoothing parameter.

3.2 Influence of the Smoothing Parameter

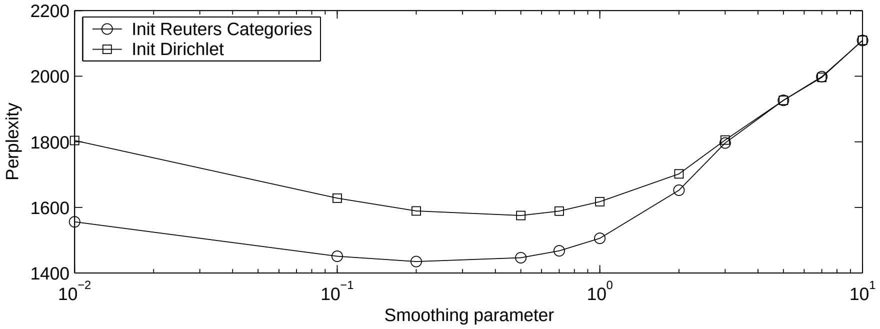


Fig. 2. Perplexity as a function of smoothing.

Figure 2 depicts the influence of the smoothing parameter $\theta_\beta - 1$ in terms of perplexity. For both initializations, the best performances are obtained for smoothing parameters between 0.1 and 2, with an optimum at 0.5. Clearly using some prior information about the fact that word probabilities should not get too small helps to fit the distribution of new data, even for words that are rarely (or even never) seen in association with a given theme.

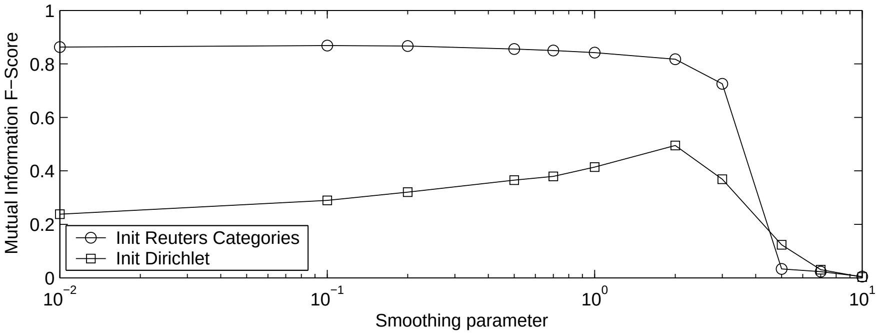


Fig. 3. Evolution of mutual information as a function of smoothing

Figure 3 reveals a slightly different behavior for the MI F-Score. First, except when using very large (5 or more) values of the smoothing parameters, which yields a serious drop in performance, the categorization accuracy is rather insensitive to smoothing for the Reuters initialization. Of more practical interest however is the behavior for the "Dirichlet" initialization, which is roughly consistent with what is observed in Figure 2, except for the fact that the optimum is obtained for higher values of the smoothing parameter (around 2). A possible explanation of this observation that more smoothing improves categorization capabilities (even if it slightly degrades distribution fit) is that the model is so coarse and the data so sparse that only quite frequent words are helpful in categorizing; the other words are essentially misleading, unless properly initialized. This suggests that removing rare words from the vocabulary should improve the classification accuracy.

As an aside, it is interesting to observe, in figure 4, that the variations of the MI F-Score is highly dependent on the initialization and the smoothing parameter. For large (unrealistic) values, the more iterations we conduct, the more inaccurate prior information we give to the model and the worst the performances get. For the initialization "Dirichlet", the optimal value of $\theta_\beta - 1$ (2) clearly corresponds to the higher increasing curve. From 3, the clustering begins to degrade after 5 or 6 iterations.

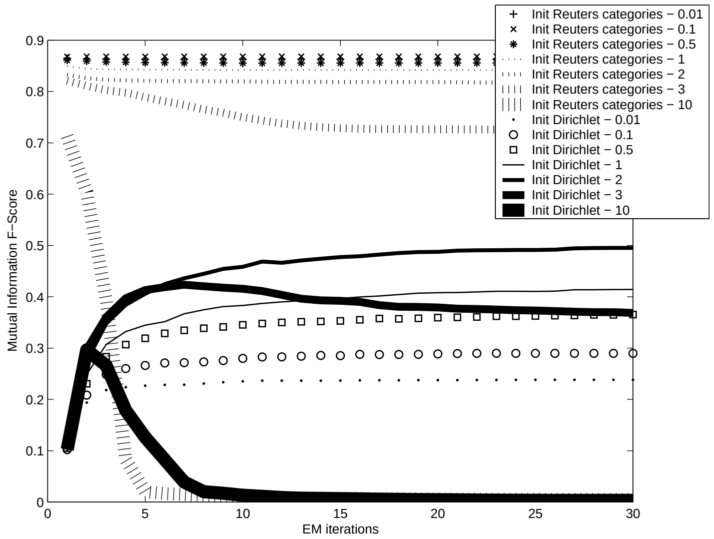


Fig. 4. Evolution of mutual information as a function of EM iterations, with different smoothing values

3.3 Adjusting the Vocabulary Size

A valid question, after having decided to ignore part of the vocabulary, is if we should rather cut rare words (hapax) or frequent words (stop-words). We try both strategies, removing consecutevely tens, hundreds and thousands of terms from the vocabulary. The words discarded are simply not taken into account in the count matrix**.

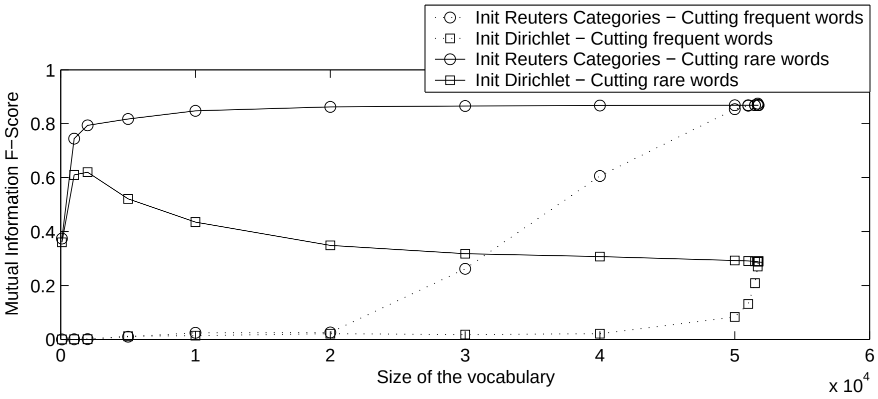


Fig. 5. Evolution of mutual information when removing rare words.

Results in term of perplexity are not helpful, since the size of the vocabulary has an impact on perplexity which is hard to distinguish from the variations due to a possible better fit of the model. The MI F-Score, on the other hand, is meaningful even when the vocabulary sizes are different. The results in Figure 5 suggest that we can substantially improve the performance of the model with the "Dirichlet" initialization, by keeping a very limited number of frequent words (around 2,000). Note that the obtained F-Score is still far from reaching the performance attained with the Reuters initialization. This agrees with our previous observation that even the rarest word may be informative, when properly initialized.

On the other hand, removing frequent words almost always hurts as one can see when reading the dashed curves from right (full vocabulary) to left (all words removed from vocabulary). Only in the case of the "Reuters Categories" initialization, discarding the 50 or 100 most frequent words leads to a slightly better performance but it is hardly visible on the figure. Then the MI F-Score steadily decreases when cutting frequent words. The score is almost 0 with 20,000 rare words, which is not surprising, since, in this case, the vocabulary only consists of words with 1 occurence in the whole corpus and a text is therefore reduced to at most a dozen of terms.

** We do not study here the effect of another common trick: grouping all unknown words under the token "Out Of Vocabulary".

3.4 Adding Supervised Information

Clearly none of the variants discussed so far is susceptible of bridging the gap between the ideal results, obtained using Reuters categories, and the results achievable in practice. To this aim, we consider using a limited number of texts (2, 5, 10, 20 or 50) from each theme to initialize the theme-dependent word frequency parameters. Note that in this case, the EM algorithm is used in "semi-supervised" mode, updating only the posterior probabilities for the texts whose category is truly unknown. In each case and each repetition (we are still using ten-fold cross validation), we repeat the experiment ten times to make up for the chances of picking "unrepresentative" texts.

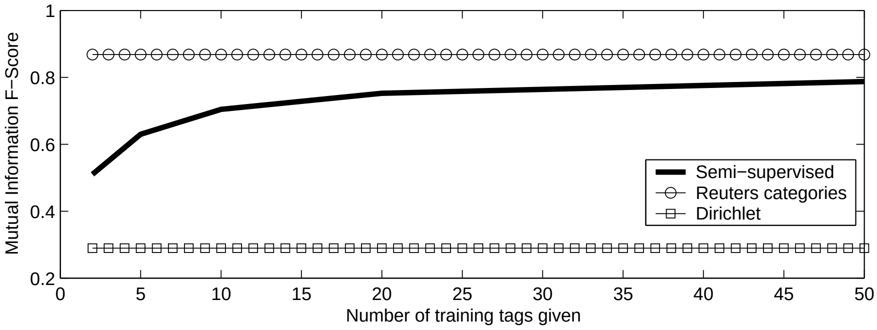


Fig. 6. Evolution of mutual information when using partial category information.

Figure 6 shows that, as expected, results improve with the number of known text tags and that acceptable values are obtained quite fast: with 10 tags per theme (that is 1.1% of the training documents labeled), the obtained F-Score is already about 0.7 (to be compared with 0.8 when 5.5% of the labels are known and 0.9 when all the labels are known).

Figure 7 conveys the same impression and suggests that knowing 20 or 50 labels per category is almost equivalent in terms of perplexity and log-likelihood. Hence, knowing a few percents of the document labels is enough to catch up on word distribution modelling (perplexity) and a few additional percents suffice to obtain very good categorization performance.

3.5 Equivalence with a Non-Probabilistic Algorithm

A surprising fact, when working with this model, is the huge fraction of posterior probabilities (that a document belongs to a given theme) dramatically close to 0 or 1. Indeed, when starting from Reuters categories, the proportion of texts classified in only one given theme (that is, with probability one up to machine precision) is almost 100%. Since we start from the opposite point of "extreme fuzziness", this effect is not as strong with the "Dirichlet"

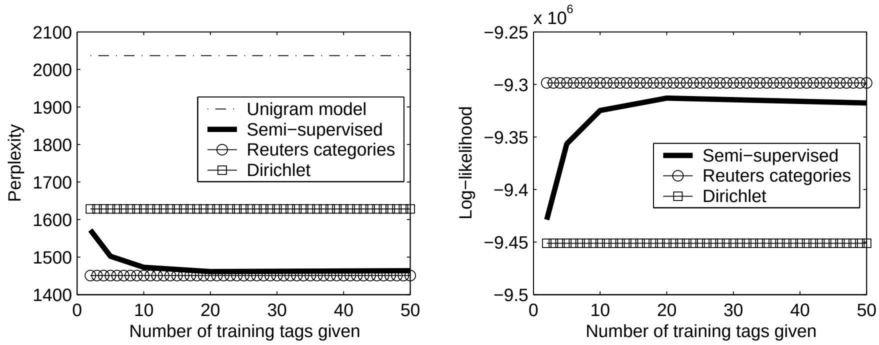


Fig. 7. Evolution of perplexity and log-likelihood when using partial category information.

initialization. Still, after the fifth iteration, more than 90% of the documents are categorized with absolute certainty.

Therefore, we compare the results obtained with an algorithm similar to EM but based on hard clustering. This is in fact a version of K-means, with the following distance between a text $d \in \{1, \ldots, n_D\}$ and theme (or cluster) $t \in \{1, \ldots, n_T\}$:

$$dist(d,t) = \frac{1}{\alpha_t \prod_{w=1}^{n_W} \beta_{wt}^{C_d(w)}}$$

This distance is computed for every document and every theme and each document is assigned to its closest theme. The reestimation of the parameters β_{wt} is done according to (4) where the posterior "probabilities" are always either 0 or 1. α_t simply becomes the proportion of documents in theme t and β_{wt} the ratio of the number of occurrences of w in theme t over the total number of occurrences in documents in theme t.

$$\alpha_t = \frac{1}{n_D} \sum_{d=1}^{n_D} \delta_{\{d \in t\}}$$

$$\beta_{wt} = \frac{\sum_{d \in t} C_d(w)}{\sum_{w=1}^{n_W} \sum_{d \in t} C_d(w)}$$

We applied this algorithm to the same dataset, with the same initialization procedures as above. At the end of each iteration, we compute the Mutual Information F-Score between the fuzzy clustering produced by EM and the hard clustering produced by this version of K-means.

- With the "Reuters Categories" initialization, the Mutual Information F-Score between the clusterings produced is 1 after one iteration.
- With the "Dirichlet" initialization, which is somehow the opposite of a hard clustering, the F-Score between the soft and hard clustering converges very fast to 1 and is greater than 0.99 after five iterations.

In both cases, the different outputs of the fuzzy and hard methods become indiscernible after very few iterations. We believe that this behavior of EM can be partly explained by the large dimensionality of the space of documents[⋆⋆⋆]. This assumption can be verified with experiments on artificially simulated datasets.

4 Conclusion

In this article, we study a mixture model of thematic multinomial distributions for corpus clustering. We show that, even though some parameters have a real influence and actually help reduce the gap, there exists a large difference between the best achievable performance and the ones we are able to obtain without prior supervised information. Eventually, we note that in this case, a fuzzy clustering approach is just uselessly time consuming since we get exactly the same results with a hard clustering version of the algorithm.

In future work, it would be interesting to check if the same conclusions apply to more complicated models such as PLSA and LDA. Besides, we are still looking for ways to improve the performances of the model with the "Dirichlet" initialization, for example using other inference methods.

5 Acknowledgment

This work has been supported by France Télécom, Division R&D, under contract n°42541441.

References

[Blei *et al.*, 2002]David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent Dirichlet allocation. In Thomas G. Dietterich, Suzanna Becker, and Zoubin Ghahramani, editors, *Advances in Neural Information Processing Systems (NIPS)*, volume 14, pages 601–608, Cambridge, MA, 2002. MIT Press.

[Clérot *et al.*, 2004]Fabrice Clérot, Olivier Collin, Olivier Cappé, and Eric Moulines. Le modèle "monomaniaque" : un modèle statistique simple pour l'analyse exploratoire d'un corpus de textes. In *Colloque International sur la Fouille de Texte (CIFT'04)*, La Rochelle, 2004.

[Hofmann, 2001]Thomas Hofmann. Unsupervised learning by probabilistic latent semantic analysis. *Machine Learning Journal*, 42(1):177–196, 2001.

[Nigam *et al.*, 2000]Kamal Nigam, Andrew K. McCallum, Sebastian Thrun, and Tom M. Mitchell. Text classification from labeled and unlabeled documents using EM. *Machine Learning*, 39(2/3):103–134, 2000.

[⋆⋆⋆] The vocabulary contains more than 50,000 words.

Part III

Bioinformatics and Statistics

A Hidden Markov Model applied to the analysis of protein 3D-structures

AC. Camproux[1], F. Guyon[1], R. Gautier[2], J. Laffray[1], and P. Tufféry[1]

[1] Equipe de Bioinformatique Génomique et Moléculaire, INSERM U726,
Université Paris 7, case 7113, 2 place Jussieu, 75251 Paris, France
(e-mail:
camproux@ebgm.jussieu.fr,guyon@ebgm.jussieu.fr,tuffery@ebgm.jussieu.fr)

[2] Institut de Pharmacologie Moleculaire et Cellulaire
UMR 6097 CNRS / UNSA
660 route des Lucioles
06560 Sophia Antipolis
(e-mail: gautier@ipmc.cnrs.f)

Abstract. Understanding and predicting protein structures depends on the complexity and the accuracy of the models used to represent them. We have setup a Hidden Markov Model to optimally compress three dimensional (3D) conformation of protein into a structural alphabet, i.e. a library of exhaustive and representative states (describing short fragments) learnt simultaneously with connection logic. The discretization of protein backbone local conformation as a series of states results in a simplification of protein 3D coordinates into a unique unidimensional (1D) representation. We present some evidence that such approach can constitute a very relevant way to the analysis of protein architecture in particular for protein structure comparison or prediction.

Keywords: Hidden Markov Models, structural alphabet, protein structural organization.

1 Introduction

The recent genome sequencing projects [Waterston *et al.*, 2002] have provided sequence information for large number of proteins. In most cases, an accurate 3D structural knowledge of the proteins is necessary for a detailed functional characterization of these sequences. However, even in the days of high-throughput methods, experimental determination of protein structures by X-ray crystallography or NMR is quite time-consuming. Thus, there is an increasing gap between the number of available protein sequences and experimentally derived protein structures, which makes it even more important to improve the methods for predicting protein 3D structures. The structural biology community has long focused on the very hard task of developing algorithms for solving the ab initio protein folding problem - namely, predicting protein structure from sequence. In its initial phase, the exploration of protein structure consisted in simplifying the 3D structure into secondary structures, included the well-known repetitive and regular zone - the α-helix (30%)

of protein residues and the β-sheet (20%). The remaining elements constitute a category, often considered as variable (50% of the structures). With the increasing of available 3D structures of proteins, many studies [Unger *et al.*, 1989],[Rooman *et al.*, 1990],[de Brevern *et al.*, 2000],
[Micheletti *et al.*, 2000],[Kolodny *et al.*, 2002] have focused on the identification of a more detailed but finite set of generic protein fragments. Despite the fact that such libraries provide an accurate approximation of protein conformation, their identification teaches us little about the way protein structures are organized. They do not consider the rules that govern the assembly process of the local fragments to produce a protein structure. An obvious mean of overcoming such limitations is to consider that the series of representative fragments that can describe protein structures are in fact not independent but governed by a Markovian process. For this purpose we have used Hidden Markov Models (HMM). HMM have been applied in several area of computational biology, for example to model protein families, to construct multiple sequence alignment or to determine protein domain in a query sequence [Krogh *et al.*, 1994],[Durbin *et al.*, 1998],[Bateman *et al.*, 2004]. In this study, we apply HMM to identify a library of representative fragments and their transition process, called Structural Alphabet (SA) or HMM-SA. Such an approach can constitute a very relevant way to the analysis of protein architecture in particular for protein structure comparison or prediction.

2 Materials and Methods

2.1 Datasets and describing three dimensional conformations

The extraction of SA is performed from a collection of 1429 non-redundant protein structures presenting less than 30% sequence identity. The structures are described using the α_Carbons (Figure (a.1)), as series of overlapping fragments of 4 residue length (Figure (a.2)) [Camproux *et al.*, 1999]. Each fragment h is described by a 4-descriptors vector $y(h)$ with the three distances between the non consecutive α_Carbons, i.e. $d_1(h)$= d{C$_{\alpha 1}$(h) - C$_{\alpha 3}$(h)}, d$_2$(h)=d{C$_{\alpha 1}$(h) - C$_{\alpha 4}$(h)}, d$_3$(h)=d{C$_{\alpha 2}$(h) - C$_{\alpha 4}$(h)}, where $C_{1,..,4}$ denotes the 4 residues of fragment h, and the oriented projection P$_4(h)$ of the last alpha-carbon C$_{\alpha 4}$(h) to the plane formed by the three first ones, as shown in Figure (a.3). The collection of 1429 proteins represent a total of 332493 four-residues fragments.

2.2 Identification of the optimal structural alphabet (SA)

Models
Suppose that polypeptidic chains are made up of representative fragments of (R) different types $\{S_1, S_2, ..., S_R\}$. We then assume that there are (R) states of the model. Each state is associated with a multi-normal function

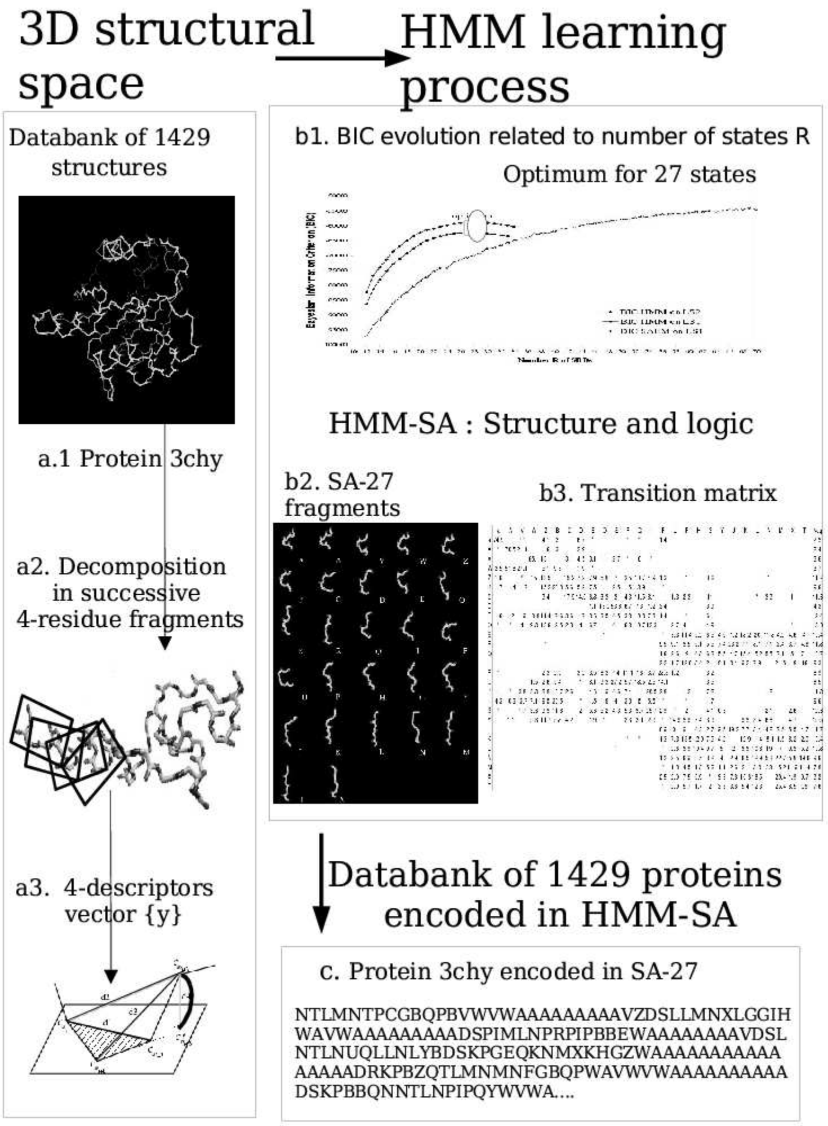

3D structural space
HMM learning process
Databank of 1429 structures
a.1 Protein 3chy
a2. Decomposition in successive 4-residue fragments
a3. 4-descriptors vector {y}
b1. BIC evolution related to number of states R
Optimum for 27 states
Bayesian Information Criterion (BIC)
Number R of states
HMM-SA : Structure and logic
b2. SA-27 fragments
b3. Transition matrix
Databank of 1429 proteins encoded in HMM-SA
c. Protein 3chy encoded in SA-27
NTLMNTPCGBQPBVWVWAAAAAAAAAAVZDSLLMNXLGGIH
WAVWAAAAAAAAAAADSPIMLNPRPIPBBEWAAAAAAAAAVDSL
NTLNUQLLNLYBDSKPGEQKNMXKHGZWAAAAAAAAAAAA
AAAAADRKPBZQTLMNMNFGBQPWAVWVWAAAAAAAAAAA
DSKPBBQNNTLNPIPQYWVWA....

Fig. 1. Encoding of 3D conformation of proteins using HMM-SA with 27 states: right part "3D structural space" represents the polypeptidic chain of protein 3chy (a1) scanned in overlapping windows that encompassed 4 successive-carbons C_α (a2), thereby producing a series of four-residue fragments. Each fragment is described by a vector of four-descriptors (a3). Center part: Figure b1 represents the BIC evolution *versus* the number of states considered, Figures b2 and b3 illustrate the optimal HMM-SA corresponding to both 27 average four-residue fragments associated to 27 states and transition matrix between states. Bottom part represents the corresponding encoded chain 3chy (c1) as a states series.

of parameters θ describing the descriptors and their variability. We consider two types of model to identify a SA corresponding to R states: a process without memory or a process with memory of order 1.
(i) Model without memory *(order 0)*, assuming independence of the R states is identified by training simple finite Mixture Models (MM) of R multi-normal distributions.
(ii) Model with memory *(order 1)* is identified, by training a Hidden Markov Model. Here, the aim is to a learn hidden sequence of states. The succession of underlying states $\{x_1, x_2, ..., x_N\}$ emits the series of vectors $\{y_1, y_2, ..., y_N\}$, describing consecutive overlapping fragments of the proteins, *via* a multi-normal density $b_{S_i}(y)$ of parameters θ_i associated to each state $S_{i,1 \leq i \leq R}$. We assume a common state dependence process for all polypeptidic chains governed by a Markov chain. The evolution of the Markov chain is completely described by:
1) the law $V = V(i)$ of the initial state of each polypeptidic chain, i.e. the probability that a polypeptidic chain starts in each of the R different states
2) the matrix of transition probabilities $\Pi = (\pi_{ii'})_{1 \leq i,i' \leq R}$ between R different states of the Markov chain, where $\pi_{ii'} = P(X_j = S_i' \mid X_{j-1} = S_i)$ is the probability for different proteins to evolve from state S_i to S_i' at any position j. For a given set of proteins and a given number (R) of states, unknown parameters $\lambda = (\Pi, V, \theta)$ of the selected model are estimated with an Expectation and Maximization (EM) algorithm [Baum *et al.*, 1970] applied on the complete likelihood.

Complete likelihood of N four-residue fragments $\{y_1, y_2, ..., y_N\}$ describing a protein of N+3 residue

$$V(y_1, y_2, ..., y_N|\lambda) = \sum_{\{x_1,x_2,...,x_N\}} V(x_1) b_{x_1}(y_1) \prod_{t=1}^{N-1} \pi_{x_t x_{t+1}} b_{x_{t+1}}(y_{t+1}) \quad (1)$$

For practical details on application to protein structures, see
[Camproux *et al.*, 1999].

Encoding proteins using Viterbi algorithm

Our ultimate goal is to reconstruct the unobserved (hidden) states sequence $\{x_1, x_2, ..., x_N\}$ of the polypeptide chains, given the corresponding four-dimensional vectors of descriptors $\{y_1, y_2, ..., y_N\}$, and to provide a classification of successive fragments in R states. For a given 3D conformation and a selected model (fixed number R of states), the corresponding best state sequence among all the possible paths in $\{S_1, \ldots, S_R\}^N$ can be reconstructed by a dynamic programming algorithm based on Markovian process (Viterbi algorithm [Rabiner, 1989]).

Statistical criteria to determine the optimal number of states

Structural alphabets of different size (R), noted SA$-R$ are learnt using HMM and MM by progressively increasing R and compared using Bayesian Information Criterion (BIC, [Schwartz, 1978]).

2.3 Assessing the discretization of protein structures

For a given state, the average C_α Root-Mean-Square deviation (RMSd) between C_α coordinates, that is an euclidean distance, of the fragments to their centroid is used to measure the structural dispersion of each state. To reconstruct the protein 3D structures from their description as a series of states, and to keep some comparison possible, we use the building procedure employed by Kolodny et al. [Kolodny *et al.*, 2002]. Briefly, the fragments are assembled using an iterative concatenation procedure to adjust 3D conformation.

2.4 Quantifying structure similarity

During the HMM-SA encoding of proteins of known structures, the probabilities of substituting one state for another are directly provided by the forward-backward algorithm [Rabiner, 1989]. A lod-score or substitution matrix is derived from these probabilities:

$$S(i,j) = ln[\frac{P(S_i, S_j)}{P(S_j)P(S_j)}] \tag{2}$$

which can be rewritten as

$$S(i,j) = ln[\frac{P(S_i|S_j)}{P(S_j)}] \tag{3}$$

with $P(S_i|S_j)$, the probability of letter S_i substitutes for letter S_j at one position and $P(S_j)$, the probability of state S_j (computed as the proportion of observed letter S_j). This lod-score matrix quantifying similarity between states is shightly modified. The score values S(i,j) get to $-\infty$ when the substitution of state S(i) by S(j) is impossible. All the finite values of S(i,j) are shifted and made positive, and the infinite one are replaced by large negative values.

2.5 Measuring sequence-structure consistency

Amino acid / state dependence can be learnt *a posteriori* from the database of 1429 proteins encoded in HMM-SA and the corresponding amino acid sequences. The specificity of each state in terms of amino acid is assessed using the “relative entropy” [Kullback and Leibler, 1951].

These amino acid sequence / states dependence can be used to quantify the consistency of a candidate 3D structure encoded in HMM-SA and its corresponding amino acids sequence. Emission probabilities of 20 amino acids $a_{j,1\leq j\leq 20}$ from each state $S_{i,1\leq i\leq R}$: $P(a_j|S_i)$ are introduced in the HMM to compute the likelihood of an amino acids sequence $\{a_1, a_2, ..., a_N\}$ corresponding to a structure encoded in states sequence x=$\{x_1, x_2, ..., x_N\}$.

$$V(a_1, a_2, ..., a_N, x|\lambda) = V(x_1)P(a_1|x_1)\prod_{t=1}^{N-1}\pi_{x_tx_{t+1}}P(a_{t+1}|x_{t+1}) \qquad (4)$$

3 Results

3.1 HMM-SA validation

HMM-SA is few dependent on the learning set
We learn SA of increasing sizes using either HMM or MM, and we compare them on the basis of their goodness of fit (Figure (b.1)). The influence of the Markovian process is large, as illustrated by the very different behaviors of the BIC associated with MM_0 or HMM_1. For MM, no BIC optimum is reached until alphabet sizes of 70 whereas for HMM, an optimum is reached for a number of states of 27 (SA-27), larger than that obtained using MM, which means a better fit of the data using HMM. Interestingly, the Markov classification takes advantage of information implicitly contained in the succession of the observations to greatly reduce the number of states, keeping a minimal representativity for each (at least 1.5%). Similar results are obtained using two independent learning sets of 250 proteins with similar BIC curves evolution. The optimum is reached for 27 states in both cases, and we find that the two SA-27 very similar. It follows that, at the optimum, the HMM-derived structural alphabet (HMM-SA) is very weakly dependent on the learning set, which in turn suggests that the learnt model can be considered as representative of all protein structures.

Geometrical and logical description of the structural alphabet
The 27 identified states are denoted as structural letters: [a, A, B,..., Y, Z]. The set of letters, sorted by increasing stretches in figure (b.2) and their transitions constitute the SA The "local fit approximation" is low, as quantified by the average alpha-carbon RMSd to the centroid associated with each state (0.23 ± 0.14 Å). SA-27 shown very reasonable performance (RMSd value less than 1Å) in terms of reconstruction of the whole protein structure accuracy, compared to other recent libraries fragments optimized in a purpose of reconstruction [Micheletti *et al.*, 2000, Kolodny *et al.*, 2002]. Concerning description of logic of protein architecture, 66% of 729 transitions between states have probabilities less than 1% (see Transition matrix between 27 states in b.3), i.e.. We observe the existence of pathways between the states, that obey some precise and unidirectional rules. Looking in detail, we observe that the states associated with close shapes have different logical roles. For instance, the two closest states [A, a] in term of geometry, close to canonical alpha helix, are distinguished by different preferred input and output states.

Moreover, the learning process attempts to optimize the likelihood associated with the entire trajectories of the proteins, resulting in propagation of such long range conditioning to the short range constraints that are learnt. For instance, three major types of alpha-helices categorized as linear, kinked or curved by Kumar and Bansal [Kumar and Bansal, 1998], seem identified in HMM-SA by [AAAAAAAAAA] series, [AAAAVWAAAA] series and [aaaaaaaaaa] series. These results are detailed in [Camproux *et al.*, 2004].

3.2 HMM-SA application: HMM structural alphabet as a general concept to simplify 3D protein structure analysis ?

Discretization of 3D structural space of proteins in SA space
Subsequently, HMM-SA provides some kind of compression from the 3D protein coordinate space into the *1D structural alphabet space*, see Figure 1. We have explored two directions in which this facility could be of interest.

Categorizing structural similarity
The detection and analysis of structural similarities of proteins can provide important insights into their functional mechanisms or relationship and offer the basis of classifications of the protein folds. The global 3D alignment of two proteins is NP-hard [Lathrop, 1994]. Therefore, approximate methods have been proposed to achieve fast similarity searching, based on the direct consideration of protein alpha-carbon coordinates [Gibrat *et al.*, 1996], [Holm and Sander, 1993, Shindyalov and Bourne, 1998]. Using HMM model, the lod-score matrix of similarity between states (Eq(2)) allows to quantify the similarity of protein fragments encoded as different series of states. It is possible to use it with classical methods developed for the amino acid sequences similarity search and thus to reduce 3D searches as a 1-dimensional sequence alignment problem [Guyon *et al.*, 2004]. Although we currently obtain performance poorer than pure 3D methods, this approach can perform fast 3D similarity search such as the extraction of exact words using a suffix tree approach, or the search for fuzzy words and is very promising in a perspective of combining with prediction procedure.

Applying sequence-structure consistency measures
All the states of SA-27 have some significant amino acid sequence specificity compared to the profiles of the collection of 1429 protein fragments ("relative entropy", $p<0.001$). Ab initio prediction is commonly viewed as composed of two problems (1) generating candidate folds, called decoys ; and (2) devising a scoring function that discriminates between near native folds and other non-native folds amongst the decoys [Kolodny *et al.*, 2002]. Concerning point (2), we can use significant dependence between states and sequence (Eq(3)) to evaluate the consistency of a set of decoys encoded in SA-27 with its corresponding amino acids sequence. Preliminary results to discriminate 3D decoys proposed in CASP6 (Critical Assessment of Techniques for Protein

Structure Prediction) show some correlation with RMSd for decoys library and this work is in progress.

4 Discussion and perspectives

In the present study, we have discussed an HMM derived 27 states SA based on a Markov process of order 1. Higher order Markovian dependence could be considered, but at the cost of a much larger number of parameters, which may pose practical computational problems. HMM-SA fits well the previous knowledge related to protein architecture organization and seems able to grab some subtle details of protein organization, while using a reduced number of states. Results on dependence between letters and amino acid sequence confirms that, despite we have learnt SA using only geometric information, we have not over-split sequence information and that all states present some sequence signature. The resulting 1D representation of protein structure can be applied to a large variety of problems recurrent to the field of protein structure analysis and prediction. Here, we have presented some evidence of its relevance for categorizing structural similarity, or measuring some sequence / structure consistency. Work is under progress to enlarge this to fold classification and prediction.

References

[Bateman *et al.*, 2004]A. Bateman, R. Coin, L.and Durbin, R.V. Finn, V. Hollich, S. Griffiths-Jones, A. Khanna, M. Marshall, S. Moxon, E. L. L. Sonnhammer, D. J. Studholme, C. Yeats, and S.R. Eddy. The pfam protein families database. *Nucleic Acids Research*, 32:138–141, 2004.

[Baum *et al.*, 1970]L. E. Baum, T. Petrie, G. Soules, and N. Weiss. A maximization technique occurring in the statistical analysis of probabilistic functions of markov chains. *The Annals of Mathematical Statistics*, 41(1):164–171, 1970.

[Camproux *et al.*, 1999]A. C. Camproux, P. Tuffery, J. P. Chevrolat, J. F. Boisvieux, and S. Hazout. Hidden markov model approach for identifying the modular framework of the protein backbone. *Protein Eng*, 12(12):1063–73, 1999.

[Camproux *et al.*, 2004]A. C. Camproux, R. Gautier, and P. Tuffery. A hidden markov model derived structural alphabet for proteins. *J Mol Biol*, 339(3):591–605, 2004.

[de Brevern *et al.*, 2000]A. G. de Brevern, C. Etchebest, and S. Hazout. Bayesian probabilistic approach for predicting backbone structures in terms of protein blocks. *Proteins*, 41(3):271–87, 2000.

[Durbin *et al.*, 1998]R. Durbin, S.R. Eddy, and G.J. Krogh A.and Mitchison. *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids.* Cambridge University Press, Paris, 1998.

[Gibrat *et al.*, 1996]J. F. Gibrat, T. Madej, and S. H. Bryant. Surprising similarities in structure comparison. *Curr Opin Struct Biol*, 6(3):377–85, 1996.

[Guyon *et al.*, 2004]F. Guyon, A. C. Camproux, J. Hochez, and P. Tuffery. Sa-search: a web tool for protein structure mining based on a structural alphabet. *Nucleic Acids Res*, 32(Web Server issue):W545–8, 2004.

[Holm and Sander, 1993]L. Holm and C. Sander. Protein structure comparison by alignment of distance matrices. *J Mol Biol*, 233(1):123–38, 1993.

[Kolodny *et al.*, 2002]R. Kolodny, P. Koehl, L. Guibas, and M. Levitt. Small libraries of protein fragments model native protein structures accurately. *J Mol Biol*, 323(2):297–307, 2002.

[Krogh *et al.*, 1994]A. Krogh, M. Brown, I.S. Mian, K. Sjolander, and D. Haussler. Hidden markov models in computational biology. applications to protein modeling. *J Mol Biol.*, 235(5):1501–31, 1994.

[Kullback and Leibler, 1951]S. Kullback and R. A. Leibler. On information and sufficiency. *Annals of Mathematics and Statistics*, 22:79–86, 1951.

[Kumar and Bansal, 1998]S. Kumar and M. Bansal. Geometrical and sequence characteristics of alpha-helices in globular proteins. *Biophys J*, 75(4):1935–44, 1998.

[Lathrop, 1994]R. H. Lathrop. The protein threading problem with sequence amino acid interaction preferences is np-complete. *Protein Eng*, 7(9):1059–68, 1994.

[Micheletti *et al.*, 2000]C. Micheletti, F. Seno, and A. Maritan. Recurrent oligomers in proteins: an optimal scheme reconciling accurate and concise backbone representations in automated folding and design studies. *Proteins*, 40(4):662–74, 2000.

[Rabiner, 1989]L. R. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *Proc. of the IEEE*, 77(2):257–285, 1989.

[Rooman *et al.*, 1990]M. J. Rooman, J. Rodriguez, and S. J. Wodak. Automatic definition of recurrent local structure motifs in proteins. *J Mol Biol*, 213(2):327–36, 1990.

[Schwartz, 1978]G. Schwartz. Estimating the dimension of a model. *Annals of statistics*, 6:461–464, 1978.

[Shindyalov and Bourne, 1998]I. N. Shindyalov and P. E. Bourne. Protein structure alignment by incremental combinatorial extension (ce) of the optimal path. *Protein Eng*, 11(9):739–47, 1998.

[Unger *et al.*, 1989]R. Unger, D. Harel, S. Wherland, and J. L. Sussman. A 3d building blocks approach to analyzing and predicting structure of proteins. *Proteins*, 5(4):355–73, 1989.

[Waterston *et al.*, 2002]R. H. Waterston, E. S. Lander, and J. E. Sulston. On the sequencing of the human genome. *Proc Natl Acad Sci U S A*, 99(6):3712–6, 2002.

Classification of Domains with Boosted Blast

Cécile Capponi[1], Gwennaele Fichant[2], Yves Quentin[2], and François Denis[1]

[1] LIF - CNRS, Université de Provence, 39, avenue Joliot Curie
13453 Marseille Cedex 13, France
(e-mail: capponi@cmi.univ-mrs.fr, fdenis@cmi.univ-mrs.fr)

[2] LMGM - IBCG - CNRS, Université Paul Sabatier, 118, route de Narbonne,
31062 Toulouse Cedex, France
(e-mail: fichant@bibcg.biotoul.fr, quentin@ibcg.biotoul.fr)

Abstract. This paper presents the first real experimentations of the boosting techniques applied to Blast for producing a model of functional domains whose amino-acids primary sequences are not conserved during evolution. The BlastBoost algorithm is depicted, and first results are analysed, showing the relevance of our approach.

Keywords: Bioinformatics, Boosting, Sequence similarity, Functional domains.

1 Introduction

The function of a single protein is mainly carried out by a *domain* which is a subsequence of amino-acids within the whole sequence of the protein. During evolution, the sequence of such a domain can be significantly modified while the function is still conserved. One way of predicting the function of a protein is to identify a known domain in the protein, despite sequence modifications such as deletions or substitutions. Domains can be grouped in functional families, themselves subdivided in subfamilies. Our work deals with functional families whose domains are not well conserved during evolution, which means that, given the sequence of a protein, it is hard to predict whether it carries the domain associated with the function. Formally, let F be a functional family, let $P = \{p_1, ..., p_n\}$ a set of annotated proteins which are known to belong or not to F, our problem is to decide whether any new protein p belongs to F. It is a supervised binary classification problem: how to build a rule from the annotated proteins of P in order to determine the class of new unannotated proteins.

In many cases, comparing a new sequence of protein p with some sequences of the family F is enough for predicting whether $p \in F$. Such a similarity search may be achieved by using either an alignment program such as Blast [Altschul *et al.*, 1997] or any model of the family's sequences, for example stochastic and probabilistic models such as Hidden Markov Models. Unfortunately, none of these methods is satisfactory whenever the sequences of the domains of the family are not conserved: the models are hard to build. For example, Membrane Spanning Domains (MSD) play the role of a pore through which a substrate goes in and/or out of the cell. The composition in amino-acids of such a domain depends essentially on the nature of membrane of the considered species, so their sequences are not conserved. As a consequence, using usual alignment programs, or any current probabilistic model, is not satisfactory for retrieving such a domain onto a protein.

One way for identifying such unconserved domains of a family F on a new protein p is to successively (1) specialize the family F into several subfamilies, (2) search significant similarities between p and each known protein of each subfamily, and (3) check back the obtained predictions in order to remove false positives. It is the Iris strategy, as presented in [Quentin *et al.*, 2002] for retrieving proteins that carry a MSD domain. Despite good results, such a strategy is tedious to set up for many reasons. Among others, the subdivision of the whole family into many subfamilies is possible only when the considered domain has been widely experimentally studied. Moreover, since it is still hand-made, the subdivision may suffer from annotation mistakes.

Our proposal is to use a classification technique which avoids the subdivision of the functional family: how to learn a rule for deciding whether a given protein's sequence contains a domain of a given functional family. Usual techniques in supervised classification consist in computing at once one *strong* classification rule, *i.e.* a rule that is both selective (few false positives) and sensitive (few false negatives). Instead our proposal is to progressively learn a sequence of *weak* rules: each weak rule is not efficient on the whole training set albeit better than a random prediction. Each weak rule is learnt from an example that was badly classified using the previous weak rules. A strong rule is eventually computed by combining weak rules weighted by the confidence we have on each. Such a technique is named *boosting* [Schapire, 1990] [Freund, 1995]. Our proposal is to learn those classifiers using Blast, which is an algorithm to produce and evaluate local sequence alignments based on a stochastic model.

2 Boosting blast

In the following of the paper, let X be the instance space and $Y = \{-1, +1\}$ be the label set. A learning algorithm takes as input a training set $S = \{(x_1, y_1) \cdots (x_n, y_n)\}$ where $x_i \in X$ and $y_i \in Y$. The learnt classifier is a function $H : X \mapsto Y$ that predicts the label of any example of X according to a model computed from the training set S. The training error of H is the error rate made on the training set, while the test error of H is an approximation of the real error rate made by H on all the instances of X. The Iris strategy starts from observations which are positive examples described by their sequence of amino-acids, and applies a combination of alignment programs in order to tag any new protein. Our approach is to consider the problem as a classification problem which must be solved by taking advantage of the predictive power of local alignment programs.

Since boosting is a general method for improving the accuracy of any given learning algorithm, we propose to use it in order to improve the significancy of learning algorithms computed over local sequences alignments. We chose to boost Blast as it is the most popular tool for investigating sequence alignments, and because it relies on a stochastic model.

2.1 Boosting: principles and algorithm

Boosting is a machine-learning method which is based on the observation that finding many moderately inaccurate rules of thumb can be a lot easier than finding a single, highly accurate prediction rule. Let M be an algorithm or a method for finding the rules of thumb: let name it a "weak" or "base" learning algorithm. The boosting approach calls M repeatedly, each time feeding it with a different subset of S, more precisely a different distribution over S. Each time M is called, it generates a new weak prediction rule; after many rounds, the boosting algorithm combines these weak rules into a single prediction rule that is proven to be much more accurate than any one of the weak rules when enough data is available [Schapire, 1990]. On each round, the distribution of S is updated in such a way that the weight on the examples misclassified by the preceding weak rule is increased: this forces the base learner M to focus its attention on the "hardest" examples. The final combination of the weak rules is a simple weighted majority vote of their predictions: the weight asssigned to a weak rule should actually account for the confidence one can have on it.

We focus here on the AdaBoost algorithm (*cf.* Algorithm 1 further on, introduced by [Freund and Schapire, 1997]), which is of reference. A complete and easy presentation of practical and theoretical results about boosting and AdaBoost is available in [Schapire, 2002], especially results concerning error bounds.

Algorithm 1 AdaBoost(T), where T is the number of rounds (iterations)

```
Given: $(x_1, y_1), \cdots, (x_n, y_n)$ where $x_i \in X$ and $y_i \in Y = \{-1, +1\}$
Initialize $D_1(i) \leftarrow 1/n, \forall i \in 1..n$ ($D$ is indexed by the indices of the examples)
for all $t = 1, \cdots, T$ do
    Train base learner $M$ using distribution $D_t$
    Get the base classifier $h_t : X \to \{-1, +1\}$
    Compute $\alpha_t \in \mathbb{R}$
    Update:
        $$D_{t+1}(i) \leftarrow \frac{D_t(i) \exp(-\alpha_t y_i h_t(x_i))}{Z_t}$$
    where $Z_t$ is a normalization factor chosen so that $D_{t+1}$ will be a distribution
    of probabilities.
end for
Output the final classifier: $H(x) = \text{sign}\left(\sum_{t=1}^{T} \alpha_t h_t(x)\right)$
```

Let us here comment the Algorithm 1 which considers the simplest case: the range of each h_t is binary. D is a distribution over the sample S which is updated at each iteration t. Let $\epsilon_t = \Pr_{D_t}[h_t(x_i) \neq y_i]$ be the training error of the base classifier h_t. The parameter α_t should measure the importance assigned to h_t: it is usually related to ϵ_t. For binary base classifiers, we typically set $\alpha_t = \frac{1}{2} \ln \left(\frac{1-\epsilon_t}{\epsilon_t} \right)$ as suggested in [Schapire, 2002].

2.2 Aligning sequences with Blast

For finding similarities between protein sequences, some algorithms compare, in a pairwise fashion, a *query* sequence to all the sequences of a specified database. Each comparison is given a *score* reflecting the degree of similarity between the sequences. The similarity is measured and shown by *aligning* two sequences, globally or locally. A global alignment is an optimal alignment that includes all characters from each sequence, whereas a local alignment is an optimal alignment that includes only the most similar local region(s) (e.g. [Smith and Waterman, 1981]).

Among these algorithms, heuristic algorithms such as Blast and Fasta trade reduced accuracy for improved efficiency. Blast [Altschul *et al.*, 1990] is actually a set of sequence comparison algorithms that are used to search sequence databases for optimal local alignments to a query. Blast improves the overall speed of searches while retaining good sensitivity by breaking the query and database sequences into fragments (words), and initially seeking exact matches between fragments. The algorithm then tries to significantly raise the length of each match: the obtained extended fragments are named high-scoring segment pairs (HSPs). Hence, each pairwise sequence alignment is first assigned a raw score S (which accounts for the score of its HSP). Then, if the raw score is over a given threshold, a statistical score is computed in such a way one can discriminate between real and artefactual matches: the expected number of HSPs with score at least S is given by : $E = Kmn \exp^{-\lambda S}$ where K and λ are parameters of the scoring system (gap costs and the matrix of amino-acids substitutions), and m and n are the lengths of the sequences. This score E, named the e-value of an alignment, is the expected number of chance alignments with a score larger than (or equal to) S [Altschul *et al.*, 1997]. The smaller the e-value is, the most significant the alignment is.

Let $\mathcal{A}(x, D, \tau)$ be a formatted result of the `blastp` program (program of the Blast software for aligning proteins sequences) with x as a query, and D as the database: it is a set that contains all the proteins of D aligned with x with an e-value less than τ.

2.3 Boosting Blast

Experimentations of the Iris strategy show that some functional subfamilies of MSD are more difficult than others to be delimited. Consequently, our major intuition was that the boosting principle should help to slim over the covering of the description space of the subfamilies, by focusing on the proteins that are on the boundaries hence improving the whole covering of each subfamily. Indeed, many *false positives* are produced by all the tested recognition methods (HMM, profiles, etc.) while they could help to make more accurate the model of the subfamilies We then supposed that the boosting techniques would help to focus on proteins of the boundary of each subfamily.

The Algorithm 2 depicts the backbone of "boosting blast" for computing a model of one type of functional domains. At each iteration of the boosting algorithm, the obtained weak classifier h_t is a decision tree, built from a Blast alignment of sequences whose query has been randomly selected in the learning sample with respect to the distribution D_t. In order to compute a decision tree of deepness 1 (a stump), a Blast program is launched with query x_t over the set of known protein's sequences X, which leads to the set of proteins aligned with x_t under the given threshold τ of e-value: $\mathcal{A}(x_t, X, \tau)$. The base classifier h_t generated at iteration t is then obvious: for any sequence of protein $x \in X$, if $x \in \mathcal{A}(x_t, X, \tau)$ (*i.e.* x is aligned with x_t) then x is tagged as x_t, otherwise it is classifed like the majority (according to D_t) of the learning examples which are out of $\mathcal{A}(x_t, X, \tau)$. With such a decision tree, we can expect that the training error of each weak classifier is usually less than 0.5 (which is a condition of the boosting technique). Indeed, the Blast model of alignments is usually very predictive locally, so a few errors should be observed for examples aligned with x_t; moreover, since the majority class is chosen for classifying proteins that are not aligned with the x_t, the training error over them is less than 0.5. We expect that boosting would then extend the local predictivity of Blast to a global predictivity.

We actually set up many different kinds of weak classifiers. Two possible variants among others are:

1. *the deepness of the decision trees.* The algorithm 2 considers trees with only one test. The generalization to decision trees with d tests (a comb) is straightforward: if the test j does not lead to a valuable alignment, another example is selected from the same distribution D_t from which a new test $j+1$ is achieved, and so on until d `blastp` have been launched with d different queries of the sample. Such a generalization should raise the number of examples covered by each weak classifier, hence hopefully decrease the training error ϵ_t.
2. *The class of the selected example.* If considering all proteins of a species, the ratio between positive and negative proteins is very low. As a consequence, we chose only negative examples which are known to be close to the positive examples. Then, since both classes may be randomly selected, we authorize to allow different thresholds τ_+ and τ_- for the e-value, whether the query is a positive or a negative example. An important variant is to only authorize the selection of positive examples: in

Algorithm 2 BlastBoost(τ,T)

```
Given: $(x_1, y_1), \cdots, (x_n, y_n)$ where $x_i \in X$ and $y_i \in Y = \{-1, +1\}$
Initialize $D_1(i) \leftarrow 1/n$
for all $t = 1, \cdots, T$ do
    Select $x_{i,t}$ according to the distribution $D_t$
    Compute $\mathcal{A}_t = \mathcal{A}(x_{i,t}, X, \tau)$ with blastp
    Get $h_t : X \rightarrow \{-1, +1\}$ such that $\forall x \in X$:
        $$\text{if } x \in \mathcal{A}_t \text{ then } h_t(x) = y_{i,t} \text{ else } h_t(x) = \text{argmax}_{k \in \{-1,+1\}} \sum_{j, y_j = k, x_j \notin \mathcal{A}_t} D_t(j)$$
    Compute
        $$\epsilon_t = \sum_{i=1..n, h_t(x_i) \neq y_i} D_t(i) \text{ and } \alpha_t = \frac{1}{2} \ln \left( \frac{1 - \epsilon_t}{\epsilon_t} \right)$$
    Update:
        $$D_{t+1}(i) \leftarrow \frac{D_t(i) \exp(-\alpha_t y_i h_t(x_i))}{Z_t}$$
end for
Output the final classifier:
    $$H(x) = \text{sign} \left( \sum_{t=1}^{T} \alpha_t h_t(x) \right)$$
```

such a way, the learning algorithm does not try to learn negative examples, instead it focuses on the boundaries of the family.

3 Experimentation

3.1 Protocols

The performance of the approach has been evaluated on a specific domain found as a component of ABC transporters: the Membran Spanning Domain (MSD). These domains have been chosen because they are poorly conserved in sequence and their identification led to a large number of false positives in previous analysis [Quentin *et al.*, 2002]. We used five genomes to learn the model, and five other genomes to test it. Genomes, both in the learning and the test sets, were chosen according to the phylogeny. These sets are made up of all the positive proteins (i.e. those proteins that carry a MSD domain), and all the negative proteins that are close to the positive ones (i.e. they are known to be aligned with one or more proteins of a MSD subfamily). Each experiment has been launched 10 times with the same parametrization: the reported results of one experiment are actually a mean of the results. These first experiments helped us to investigate the role of parameters (d, τ, etc.).

Ideally, the BlastBoost algorithm should run each Blast against the whole set of known protein sequences. As it would be too long during both the learning and the testing steps, we precomputed with Blastp, and stored, the alignement of each sequence of a species with each sequence of all the species.

As far as we known, no other algorithm seeking similarities between proteins have been integrated within a boosting algorithm. As a consequence, we compare the results of BlastBoost with the Iris strategy which previously subdivided the MSD functional family into 18 subfamilies in order to be able to annotate any new example with a low error test.

3.2 Results and discussion

The figure 1 presents the best results that we obtained by tuning up the set of parameters and alternative algorithms presented section 2.3. The four categories of test share the same $d = 3$; in all categories, positive and negative examples were selected during the learning step.

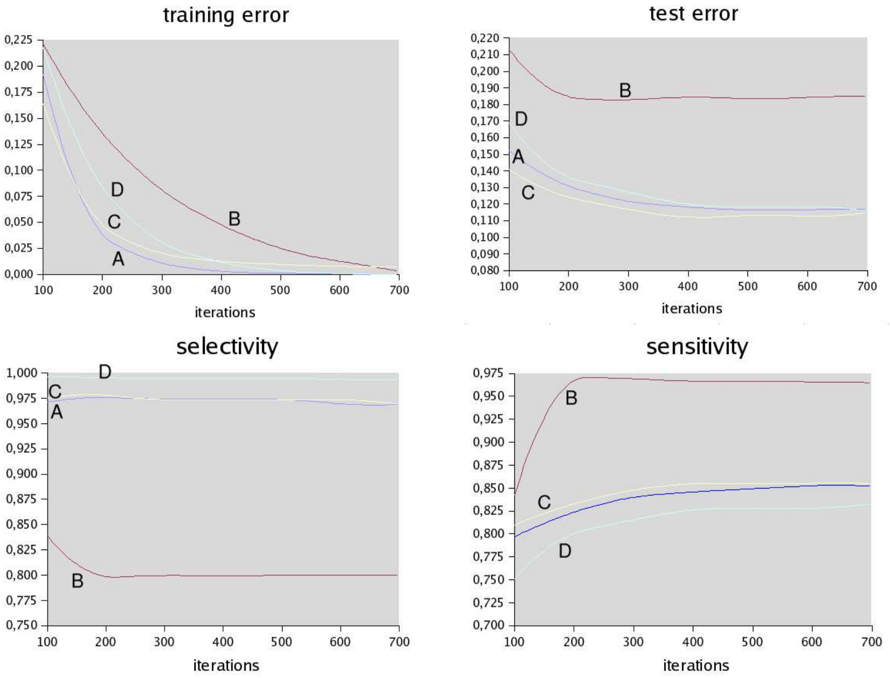


Fig. 1. The number of iterations corresponds to the parameter T in the algorithms. In tests of category A, $\tau_+ = 10^{-2}$ and $\tau_- = 10^{-10}$. In tests of category B, $\tau_+ = 10^{-10}$ and $\tau_- = 10^{-2}$. In tests of category C, $\tau_+ = 10^{-2}$ and $\tau_- = 10^{-2}$. In tests of category D, $\tau_+ = 10^{-5}$ and $\tau_- = 10^{-5}$.

The training (resp. testing) set contains 161 (resp. 172) positive proteins and 73 (resp. 84) negative proteins. The selectivity of our method is almost as good as this of the Iris method applied to MSD, while our sensitivity is better (up to 0.999 with BlastBoost, and 0.946 with Iris). Yet, their method previously subdivides the functional families into 18 subfamilies, so their learning and testing steps are independant from one subfamily to another, therefore more accurate. As a consequence, the results of BlastBoost are good: with comparable results, BlastBoost is efficient on the whole functional family even if the proteins sequences are not conserved. When analyzing the misclassified data, we noticed that the problematic subfamilies identified by Iris (M7 and M9) were better characterized by BlastBoost, while one subfamily has not been circled by BlastBoost whereas it was an "easy" family for Iris (M12). This last point results from an under-representation of M12 members in the learning set (1 out of 161 proteins), leading to false negatives during the tests. So, in order for our result to be statistically significant, we still have to work out the samples so that each subfamily is represented.

We improved our results by increasing the number of genomes in both samples: with eight in each, the test error is less than 0.1 in the categories of test B and C while the selectivity gets perfect in these categories. The accurate study of the produced weak classifiers shows that some sets of aligned sequences have a poor significance in their globality, for example when proteins are dimers. Thus, we think of defining and importing the significancy of an alignment within the boosting model, in addition to the significancy of pairwise alignments (e-value). The InfoBoost algorithm [Aslam, 2000] should be a first step towards the integration of sequences alignements properties within a boosting algorithm, for it pays attention on the quantitative *and* qualitative performance of each weak classifier, which in our case can be measured from the properties of each sequences alignement.

4 Conclusion

We presented a new way for learning a model of unconserved functional families, without dividing them into subfamilies, which is based on amino-acids sequences, by applying the boosting techniques to one performant alignment program, Blast. Our first results are good and promising. We think that several improvments could be carried out, independently from the tuning of the involved parameters. Among other, our first perspective is to integrate in the boosting model, and especially in the rated confidence of weak classifiers, some properties of alignment algorithms such as the density of an alignment, which involves its covering rate of the database and the inner significancy of the e-value's distribution.

References

[Altschul *et al.*, 1990]S.F. Altschul, W. Gish, W. Miller, E.W. Myers, and D.J. Lipman. Basic local alignment search tool. *Journal of Molecular Biology*, 215:403–410, 1990.

[Altschul *et al.*, 1997]S.F. Altschul, T.L. Madden, A.A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D.J. Lipman. Gapped blast and psiblast: a new generation of protein database search programs. *Nucleic Acid Research*, 25:3389–3402, 1997.

[Aslam, 2000]J. A. Aslam. Improving algorithms for boosting. In *13th Conference On Learning Theory, COLT'2000*, pages 200–207, Stanford, CA, USA, 2000. Morgan Kaufmann.

[Freund and Schapire, 1997]Y. Freund and R.E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55(1):119–139, aug 1997.

[Freund, 1995]Y. Freund. Boosting a weak learning algorithm by majority. *Information and Computation*, 121(2):256–285, 1995.

[Quentin *et al.*, 2002]Y. Quentin, J. Chabalier, and G. Fichant. Strategies for the identification, the assembly and the classification of integrated biological systems in completely sequenced genomes. *Computers and Chemistry*, 26:447–457, 2002.

[Schapire, 1990]R.E. Schapire. The strength of weak learnability. *Machine Learning*, 5(2):197–227, 1990.

[Schapire, 2002]R.E. Schapire. The boosting approach to machine learning: an overview. In *MSRI Workshop on Nonlinear Estimation and Classification*, 2002.

[Smith and Waterman, 1981]T.F. Smith and M.S. Waterman. Identification of common molecular subsequences. *Journal of Molecular Biology*, 147:195–197, 1981.

A Segmentation-Clustering problem for the analysis of array CGH data

F. Picard[1], S. Robin[1], E. Lebarbier[1], and J-J. Daudin[1]

Institut National Agronomique Paris-Grignon
UMR INA P-G/ENGREF/INRA MIA 518
16 rue Claude Bernard,75231 Paris cedex 05.
(e-mail: picard@inapg.fr)

Abstract. Microarray-CGH experiments are used to detect and map chromosomal imbalances, by hybridizing targets of genomic DNA from a test and a reference sample to sequences immobilized on a slide. A CGH profile can be viewed as a succession of segments that represent homogeneous regions in the genome whose representative sequences (or BACs) share the same relative copy number on average. Segmentation methods constitute a natural framework for the analysis, but they do not assess a biological status to the detected segments. We propose a new model for this segmentation-clustering problem, combining a segmentation model with a mixture model. We present an hybrid algorithm to estimate the parameters of the model by maximum likelihood. This algorithm is based on dynamic programming and on the EM algorithm. We also propose to adaptively estimate the number of segments when the number of clusters is fixed. An example of our procedure is presented, based on publicly available data sets.

Keywords: Segmentation methods, Mixture Models, Dynamic Programming, EM algorithm, Model Selection.

Introduction

Chromosomal aberrations often occur in solid tumors: tumor suppressor genes may be inactivated by physical deletion, and oncogenes activated via duplication in the genome. The purpose of array-based Comparative Genomic Hybridization (array CGH) is to detect and map chromosomal aberrations, on a genomic scale, in a single experiment. Since chromosomal copy numbers can not be measured directly, two samples of genomic DNA (referred as the reference and the test DNA) are differentially labelled with fluorescent dyes and competitively hybridized to known mapped sequences (referred as BACs) that are immobilized on a slide. Subsequently, the ratio of the intensities of the two fluorochromes is computed and a CGH profile is constituted for each chromosome when the $\log_2$ of fluorescence ratios are ranked and plotted according to the physical position of their corresponding BACs on the genome.

Each profile can be viewed as a succession of 'segments' that represent homogeneous regions in the genome whose BACs share the same relative copy number on average. Array CGH data are normalized with a median set to $\log_2$(ratio)= 0 for regions of no change, segments with positive means represent duplicated regions in the test sample genome, and segments with negative means represent deleted regions. It has to be noted that even if the underlying biological process is discrete (counting of relative copy numbers of DNA sequences), the signal under study is viewed as being continuous, because the quantification is based on fluorescence measurements, and because the possible values for chromosomal copy numbers in the test sample may vary considerably, especially in the case of clinical tumor samples that present mixtures of tissues of different natures.

Segmentation methods seem to be a natural framework to handle the spatial coherence on the genome that is a specificity of array CGH data [Autio *et al.*, 2003, Jong *et al.*, 2003]. These methods provide a partition of the data into segments, each segment being characterized by its mean and variance μ_k and σ_k^2 in the Gaussian case. Nevertheless, even if the data are instrinsically segmented, they are also structured into clusters which have a biological interpretation: we can define a group of deleted segments, a group of unaltered segments, and many groups of amplified segments for instance. This refinement means that the mean and variance of each segment should be restricted to a finite set such that $\mu_k \in \{m_1, \dots, m_P\}$ and $\sigma_k^2 \in \{s_1^2, \dots, s_P^2\}$ if the segments are structured into P clusters.

We propose to handle this segmentation-clustering problem combining a segmentation model and a mixture model to assign a biological status to segments. Section 1 is devoted to the precise definition of such model. In Section 2 we propose an hybrid algorithm combining dynamic programming and the EM algorithm to alternatively estimate the break-point coordinates and the parameters of the mixture. The convergence properties of this algorithm are presented.

Once the parameters of the model have been estimated, a key issue is the estimation of the number of segments and of the number of clusters. We propose to estimate the number of segments when the number of groups is fixed, using a penalized version of the likelihood. We propose to apply the procedure defined by [Lavielle, 2005], that has been successfully applied to array CGH data [Picard *et al.*, 2005]. An example of our method is provided in Section 3, using publicly available data sets.

1 A new model for the segmentation-clustering problem

Let y_t represent the $\log_2$ ratio of the t^{th} BAC on the genome and $y = \{y_1 \dots, y_n\}$ the entire CGH profile constituted by n data points. We suppose that y is the realization of a Gaussian process Y whose mean and variance are affected by $K+1$ abrupt changes at unknown coordinates $T = \{t_0, t_1, \dots, t_K\}$ with the convention $t_0 = 1$ and $t_K = n$. This defines a partition of the data into K segments of length n_k. We write Y as $\{Y^1, \dots, Y^K\}$, where

$Y^k = \{Y_t, t \in I_k\}$, with $I_k = \{t, t \in]t_{k-1}, t_k]\}$. We suppose that the mean and the variance of the process are constant between two break-points and they are noted μ_k and σ_k^2.

More than classical segmentation models, we assume that the mean and variance of the segment Y^k can only take a limited number of values with $\mu_k \in \{m_1, \ldots, m_P\}$, and $\sigma_k^2 \in \{s_1^2, \ldots, s_P^2\}$. In addition to the spatial organization of the data, via the partition T, there exists a secondary structure of the process into P clusters, and we adopt a mixture model approach to handle this problem.

We assume that the partitionned data $\{Y^1, \ldots, Y^K\}$ are structured into P clusters with weights π_p ($\sum_p \pi_p = 1$). We introduce a sequence of independent hidden random variables, $Z^k = \{Z_1^k, \ldots, Z_P^k\}$ such that Z^k is distributed according to a multinomial distribution consisting of one draw on P categories with probabilities $\pi_1, \ldots, \pi_P$. The mixing proportions $\pi_1, \ldots, \pi_P$ then represent the *prior* probability for segment Y^k to belong to the p^{th} component, while the *posterior* probability of membership to the p^{th} component with y^k having been observed is: $\tau_p^k = \Pr\left\{Z_p^k = 1 | Y^k = y^k\right\}$. Contrary to classical mixture models, where the indicator variables provide informations about the labelling of individual data points (which would be Y_t in our case), our model focuses on the belonging of the segments Y^k to different clusters.

We focus on the case where the data are supposed to be drawn from a mixture of Gaussian densities, with parameters $\theta_p = (m_p, s_p^2)$. If we suppose the indepence of individual data points Y_t within a segment, the model can be formulated as follows:

$$Y^k | Z_p^k = 1 \sim \mathcal{N}(m_p \mathbb{1}_{n_k}, s_p^2 I_{n_k}).$$

We note $\psi = \{\pi_1, \ldots, \pi_{P-1}, \theta_1, \ldots, \theta_P\}$ the vector of unknown independent parameters of the mixture, and the log-likelihood of the model is:

$$\log \mathcal{L}_{KP}(T, \psi) = \sum_{k=1}^{K} \log \left\{ \sum_{p=1}^{P} \pi_p f(y^k; \theta_p) \right\}.$$

$f(y^k; \theta_p)$ represents the conditional density of a vector of size n_k. Our purpose is to optimize this likelihood to estimate the parameters of the model using an hybrid algorithm.

2 An hybrid algorithm combining the EM algorithm and Dynamic Programming

The principle of our algorithm is simple: when the break-point coordinates T are known, the EM algorithm is used to estimate the mixture parameters ψ, and once ψ has been estimated, the break-point coordinates are computed using dynamic programming. This algorithm requires the *prior* knowledge of both the number of segments K and the number of populations P. The choice for these components of the model will be discussed in a later section.

2.1 Estimating the break-point coordinates when the mixture parameters are known

When the number of segments K and the parameters of the mixture are known, the problem is to find the best K-dimensional partition of the data according to the log-likelihood $\log \mathcal{L}_{KP}(T,\psi)$. Since the number of of partitions of a set with n elements into K segments is $\mathcal{C}_{n-1}^{K-1}$, and because of the additivity in K of the log-likelihood, we use a dynamic programming approach to reduce the computational load from $\mathcal{O}(n^K)$ to $\mathcal{O}(n^2)$, as suggested by [Auger and Lawrence, 1989].

Let $\hat{C}_{k+1,P}(i,j;\psi)$ be the maximum log-likelihood obtained by the best partition of the data $Y^{ij} = \{Y_i, Y_{i+1}, ..., Y_j\}$ into $k+1$ segments, when the mixture parameters ψ are known. The algorithm starts as follows: for $k=0$ and for $(i,j) \in [1,n]^2$, with $i<j$, calculate:

$$\hat{C}_{1,P}(i,j;\psi) = \log\left\{\sum_{p=1}^{P} \pi_p f(y^{ij};\theta_p)\right\} = \log\left\{\sum_{p=1}^{P} \pi_p \prod_{t=i+1}^{j} f(y_t;\theta_p)\right\}.$$

$\hat{C}_1(i,j;\psi)$ represents the local log-likelihood for segment Y^{ij}. Then the algorithm is run as follows:

$$\forall k \in [1, K_{max}] \quad \hat{C}_{k+1,P}(1,j;\psi) = \max_{h}\left\{\hat{C}_{k,P}(1,h;\psi) + \hat{C}_{1,P}(h+1,j;\psi)\right\}$$

Dynamic programming considers that a partition of the data into $k+1$ segments is a union of a partition into k segments and a set containing 1 segment. More than a reduction in the computational load, this approach provides an exact solution for the global optimum of the likelihood, that will be central for downstream model selection procedures.

2.2 Estimate the mixture model parameters when the break-point coordinates are known

When the break-point coordinates are known, we dispose of a partition of the data into K segments $\{Y^1, \ldots, Y^K\}$. This partition defines the statistical units of a mixture model whose parameters have to be estimated. The purpose is then to maximize the log-likelihood of the model $\log \mathcal{L}_{KP}(T,\psi)$ according to ψ. As it is the case in classical mixture models, the direct optimization of the likelihood is impossible, but can be handled using the EM algorithm in the complete-data framework [Dempster *et al.*, 1977]. Let us define the complete-data log-likelihood:

$$\log \mathcal{L}_{KP}^{c}(T,\psi) = \sum_{k=1}^{K}\sum_{p=1}^{P} z_p^k \log\left\{\pi_p f(y^k;\theta_p)\right\}.$$

The EM algorithm is as follows:

- E-**step**: compute the conditional expectation of the complete-data log-likelihood, given the observed data Y, using the current fit $\psi^{(h)}$ for ψ.

$$Q_{KP}(\psi|\psi^{(h)};T) = \sum_{k=1}^{K}\sum_{p=1}^{P} \tau_p^{k(h)} \log\left\{\pi_p f(y^k;\theta_p)\right\},$$

with

$$\tau_p^{k(h+1)} = \frac{\pi_p^{(h)} f(y^k;\theta_p^{(h)})}{\sum_{\ell=1}^{P} \pi_\ell^{(h)} f(y^k;\theta_\ell^{(h)})}.$$

- M-**step**: The M-step on the $(h+1)^{th}$ iteration requires the global maximization of $Q_{KP}(\psi|\psi^{(h)};T)$ with respect to ψ to give the updated estimate $\psi^{(h+1)}$:

$$\psi^{(h+1)} = \underset{\psi}{\text{Argmax}}\left\{Q_{KP}(\psi|\psi^{(h)};T)\right\}.$$

2.3 Convergence properties of the hybrid algorithm

The proof of the convergence of our algorithm is based on the properties of both dynamic programming and EM. It can be seen that both algorithms are linked through the likelihood they alternatively optimize: the incomplete-data likelihood of the mixture of segments.

Dynamic programming globally optimizes the likelihood with respect to T. At iteration (ℓ) we have:

$$\log \mathcal{L}_{KP}\left(T^{(\ell+1)};\psi^{(\ell)}\right) \geq \log \mathcal{L}_{KP}\left(T^{(\ell)},\psi^{(\ell)}\right).$$

On the other hand, the key convergence property of the EM algorithm is the increase of the incomplete-data log-likelihood at each step [Dempster *et al.*, 1977]:

$$\log \mathcal{L}_{KP}\left(T^{(\ell)},\psi^{(\ell+1)}\right) \geq \log \mathcal{L}_{KP}\left(T^{(\ell)},\psi^{(\ell)}\right).$$

Put together, our algorithm generates a sequence $\left(T^{(\ell)},\psi^{(\ell)}\right)_{\ell\geq 0}$ that increases the incomplete-data log-likelihood such as:

$$\log \mathcal{L}_{KP}\left(T^{(\ell+1)},\psi^{(\ell+1)}\right) \geq \log \mathcal{L}_{KP}\left(T^{(\ell)},\psi^{(\ell)}\right).$$

3 Estimating the number of segments K when the number of clusters P is fixed.

Once the parameters of the model have been estimated (for a fixed K and a fixed P), the next question is the estimation of the number of segments and of the number of clusters. Since the principal objective of biologists is rather the detection of biological events on the genome rather than the clustering of those events into groups, we choose to focus on the estimation of the number of segments when the number of groups is fixed.

The maximum of the log-likelihood $\log \hat{\mathcal{L}}_{KP} = \log \mathcal{L}_{KP}(\hat{T}, \hat{\psi})$ can be viewed as a quality measurement of the fit to the data of the model with K segments. In classical segmentation models, this quantity is maximal when the number of segments equals the number of data points. Nevertheless, as our model also considers the clustered nature of segments, it appears that the quality of fit of the model is not always increasing with the number of segments, as shown in Figure 1. For $P = 2$ the incomplete-data log-likelihood is decreasing for a number of segments $K \geq 12$ for instance. This behavior of the model can be interpreted as follows: since the segmentation-clustering model is under the constraint $P \leq K$, the addition of new segments can lead to contiguous segments affected to the same cluster. This configuration leads to an increase in the number of parameters (one additional break-point) without any gain for the fit of the mixture model. These considerations imply that there will be a number of segments above which the addition of a new segment will not increase the log-likelihood.

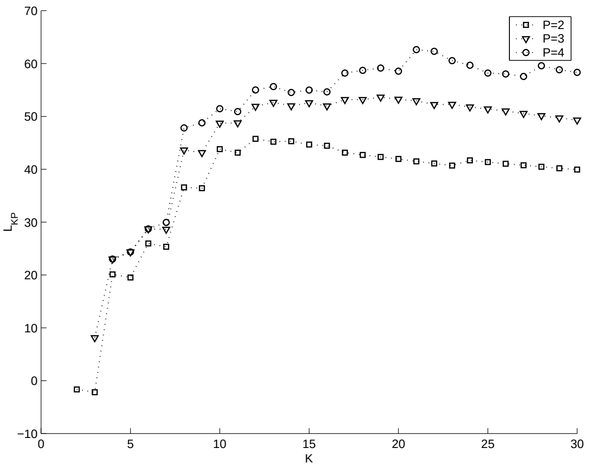


Fig. 1. Evolution of the incomplete-data log-likelihood $\log \hat{\mathcal{L}}_{KP}$ with the number of segments K for different number of clusters ($P = 2, 3, 4$).

A penalized version of the likelihood is used as a trade-off between a good adjustement and a reasonnable number of break-points. The estimated number of segments is such as:

$$\hat{K}_P = \underset{K}{Argmax} \left(\hat{\mathcal{L}}_{KP} - \beta_P pen(K) \right),$$

with $pen(K)$ a penalty function that increases with the number of segments, and β_P a penalty constant. The definition of an appropriate penalty function and constant has lead to theoretical developments in the context of breakpoint detection models. Recently, [Lavielle, 2005] proposed to use an adaptive procedure to estimate the penalty constant, that has been successfully applied to array CGH data [Picard *et al.*, 2005]. The principle of this procedure is to find the number of segments for which the log-likelihood ceases to increase significantly. It is geometrically linked to the finding of the number of segments for which the second derivative of the log-likelihood function is maximal (see [Lavielle, 2005] for further details). A result of our procedure is shown in Figure 2. For a number of clusters $P = 3$, the adpative procedure estimates a number of segments $\hat{K}_3 = 10$. This leads to a profile which presents three types of segments that can be interpreted in terms of biological groups, as shown in Figure 2.

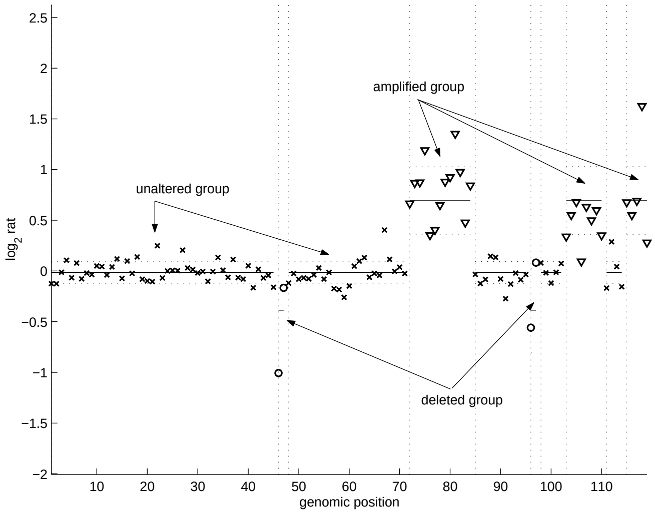


Fig. 2. Result of the segmentation-clustering procedure for a fixed number of clusters $P = 3$ and an estimated number of segments $\hat{K}_3 = 10$. These data concern chromosome 1 of breast cancer cell lines Bt474.

4 Discussion

Microarray CGH currently constitutes the most powerful method to detect gain or loss of genetic material on a genomic scale. We introduced a statistical methodology for the analysis of CGH microarray data, that combines segmentation methods and clustering techniques. It terms of modeling, the discovery of homogeneous regions clustered into groups could have been handled using Hidden Markov Models, as in [Fridlyand *et al.*, 2004]. In those models, the segmented structure of the data is recovered using the *posterior* probability of membership of individual data points into a fixed number of hidden groups, whereas our method focuses on the labelling of segments to hidden groups. Moreover, a property of Hidden Markov Models is that the distance between two 'break-points' is dependent on the probability distribution of the hidden sequence: the within-class sojourn time is geometrically distributed. Our approach is free from those constraints, since break-point coordinates are 'real' parameters of the model that are not randomly distributed.

The definition of this new model leads to unusual statistical considerations: it appears that the statistical units of the mixture model (when the segmentation is known) are segments of different size. Since the partition of the data is random, the individuals of the mixture model themselves are random. This explains the difficulty of the joint estimation of K the number of segments, and P the number of clusters, since classical model selection procedures are based on a compromize between a reasonnable number of parameters to estimate given a fixed number of statistical units. To these extents, this problem of model selection for two components remains an open question.

References

[Auger and Lawrence, 1989]I.E. Auger and C.E. Lawrence. Algorithms for the optimal identification of segments neighborhoods. *Bull. Math. Biol.*, 51:39–54, 1989.

[Autio *et al.*, 2003]R. Autio, S. Hautaniemi, P. Kauraniemi, O. Yli-Harja, J. Astola, M. Wolf, and A. Kallioniemi. CGH-plotter: MATLAB toolbox for cgh-data analysis. *Bioinformatics*, 13:1714–1715, 2003.

[Dempster *et al.*, 1977]A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society Series B*, 39:1–38, 1977.

[Fridlyand *et al.*, 2004]J. Fridlyand, A. Snijders, D. Pinkel, D.G. Albertson, and A.N. Jain. Hidden markov models approach to the analysis of array CGH data. *Journal of Multivariate Analysis*, 90(1):132–1533, 2004.

[Jong *et al.*, 2003]K. Jong, E. Marchiori, A. van der Vaart, B. Ylstra, M. Weiss, and G. Meijer. *Applications of Evolutionary Computing: EvoWorkshops 2003: Proceedings*, volume 2611, chapter chromosomal breakpoint detection in human cancer, pages 54–65. Springer-Verlag Heidelberg, 2003.

[Lavielle, 2005]M. Lavielle. Using penalized contrasts for the change-point problem. *(to appear in) Signal Processing*, 2005.

[Picard *et al.*, 2005]F. Picard, S. Robin, M. Lavielle, C. Vaisse, and J-J. Daudin. A statistical approach for CGH microarray data analysis. *BMC Bioinformatics*, 6:27, 2005.

The operons, a criterion to compare the reliability of transcriptome analysis tools

A.-S. Carpentier, A. Riva, G. Didier, J.-L. Risler, and A. Hénaut

Laboratoire Génome et informatique UMR 8116
Tour Evry2, 523 Place des Terrasses
91034 EVRY, France
(e-mail: carpentier@genopole.cnrs.fr)

Abstract. The number of statistical tools used to analyze transcriptome data is continuously increasing and no one, definitive method has so far emerged. There is a need for comparison and a number of different approaches has been taken to evaluate the effectiveness of the different statistical tools available for microarray analyses. In this paper we describe a simple and efficient protocol to compare the reliability of different statistical tools available for microarray analyses. It exploits the fact that genes within an operon exhibit the same expression patterns. We have compared five statistical tools using *Bacillus subtilis* expression data: ANOVA, PCA, ICA, the *t*-test and the paired *t*-test. Our results show ICA to be the most sensitive and accurate of the tools tested.

Keywords: operon, criterion of comparison, transcriptome, expression analysis.

1 Introduction on microarrays and their analysis

Protein activities are the bases of cell and organism functioning. In order to fit to changes in extern or interne physiological conditions the expression level of some genes and the quantity of the corresponding proteins may vary. As proteins are much harder to analyze than mRNAs, techniques for transcriptome analysis have been more popular up to now. In the last decades a tool has been developed in order to measure the expression levels of many genes (several thousands of genes) at the same time.

As microarrays allow measuring the expression levels of thousands of genes at the same time, this opens the possibility to identify differentially expressed genes [Callow *et al.*, 2000] and to cluster those genes sharing similar expression patterns [Heyer *et al.*, 1999]. This allow the identification of gene functions, regulation and networks.

Different tools have been developed for or adapted to the analysis of the huge amount of data created in microarray experiments. The number of tools is continuously increasing and no one, definitive method has so far emerged. There is a need of comparing the tools, but identifying an unbiased and biologically relevant criterion for the comparison is difficult [He *et al.*, 2003]. A number of different approaches has been taken to compare the effectiveness, or reliability, of the different statistical tools available for microarray analyses:

* Some are based on artificial data to define precisely the specificity and sensitivity of these statistical tools ([Reiner *et al.*, 2003]).

* Others are based on experimental data. The quality of a statistical tool can be measured by the number of differentially expressed genes which it reveals. A statistical parameter like the p-value may be used [Pan, 2002].

* Finally some authors combine two criteria, the number of identified genes and their physiological coherence, based on an a priori knowledge of the biological phenomenon studied [Troyanskaya *et al.*, 2002].

In this paper we try to establish a protocol for the comparison of statistical tools (available for microarray analysis) which is objective, reflects a biological reality and is not bound to one, particular set of experimental conditions. It is based on the expression coherence of genes belonging to the same operon. In bacteria a number of genes are organized in operons, that is to say clusters of contiguous genes transcribed from one promoter. For an operon a single mRNA corresponds to several genes whereas for isolated genes one mRNA corresponds to one gene. It has been shown that the genes within an operon exhibit the same expression patterns [Sabatti *et al.*, 2002].

That is why, a good and reliable statistical tool is one that, when detecting an over- or under-expression for a gene belonging to an operon, also detects this pattern for the other genes belonging to this operon. This criterion, based on the expression coherence of genes belonging to the same operon, therefore reflects a biological property that is not bound to a particular set of experimental conditions.

We have tested this criterion on five statistical tools using *Bacillus subtilis* expression data [Sekowska *et al.*, 2001]: The Analysis of Variance (ANOVA), the Principal Component Analysis (PCA), the Independent Component Analysis (ICA), the t-test and the paired t-test. Note: ANOVA and the t-tests need the a priori definition of factors, which could influence the level of gene expression; ICA and PCA do not need the definition of any factor for their use.

2 Methods

The microarray data used in this study stem from experiments on the sulphur metabolism of *Bacillus subtilis* [Sekowska *et al.*, 2001]. The experiments were carried out using *B. subtilis* gene arrays; each array contained all of *B. subtilis'* genes and one gene is represented by one spot. Each gene spot is represented twice on the array.

The aim of these experiments was to identify the genes differentially expressed when the bacteria are grown with methionine or methyl-thioribose as sulphur source. The experiments followed a fully crossed factorial design with 4 factors (sulphur source, day of experiment, amount of RNA used and duplicate of each spot).

We have used the logarithm (base 10) of these raw data in order to remove much of the proportional relationship between random error and signal intensity. We have normalized the data (mean equal to 0 and variance equal to 1 for each experimental condition).

We have chosen to analyze the expression data for the two experimental factors "sulphur source" and "day of experiment". For ICA and PCA the axes which correspond to these two factors are determined a posteriori. For PCA the factor "day" corresponds to the third axis and the factor "sulphur source" to the fifth. The fourth axis corresponds to an interaction between these two factors.

For each gene, the model used for ANOVA is the following:

$$Y_{ijkl} = \mu + S_i + J_j + C_k + D_l + \epsilon_{ijkl}$$

where Y_{ijkl} is the gene intensity

μ is the mean of the intensities of expression measured for the gene

S_i, J_j, C_k and D_l are, respectively, the effects of sulphur source i, experiment day j, RNA concentration k and duplicate l on the gene intensity

ϵ_{ijkl} is the residual error.

We need to know how the genes of *Bacillus subtilis* are organized into operons. A presumed operon is defined as a group of contiguous genes that are on the same reading strand delimited either by a promoter and a terminator (predicted or not) or a gene, which lies on the other DNA strand. This allowed to find the operons in *Bacillus subtilis* (Subtilist).

To compare statistical tools, one needs to define quantitative criteria that will measure the "tool reliability": sensitivity, accuracy and the detection of false positives need to be evaluated.

The following procedure was applied:

1. The genes are ranked as a function of their expression changes (rank #1 is the most significant).
   In order to compare the five tools under the best possible conditions, the genes are ranked according to the most relevant criterion for each tool, that is to say, the one that gives the most coherent results for the tool:
   * for ANOVA and the *t*-tests, the p-value obtained for each gene;
   * for PCA and ICA, the remoteness from the cloud centre of the projection of the gene on the axis studied.
   We thus obtain for each tool a list of genes, ranked according to a specific criterion. The order of the genes on the lists obtained may differ from each other.
2. "Detected Operons" are identified based on the ranks (one gene of the operon with rank ≤ 20 and another gene with rank ≤ 100).
   It should be noted that a priori the "Detected Operons" may be different for the various tools tested.
3. The Most Significant Interval (MSI) is determined.

In order to facilitate the analysis and comparison of the statistical tools we introduce the Most Significant Interval (MSI). It is calculated for each "Detected Operon" in the following manner:

$$MSI_j = median_j - first_j$$

Where MSI_j is the MSI of "Detected Operon" j
$median_j$ is the median of the rank values of the genes belonging to "Detected Operon" j
$first_j$ is the smallest rank value within "Detected Operon" j

4. False positives are evaluated (MSI≥700).
   The reliability of a statistical tool will also be measured by the absence of false positives. For the definition of false positives we exploit the fact that each gene spot had been duplicated on the microarrays and any difference measured for two spots belonging to the same gene cannot have a biological cause. We ranked the genes according to this "duplicate factor ", as described under point 1 and identified "Detected Operons" as described under point 2. As there is no biological cause for this detection, we find ourselves with false positives. The results of this analysis lead us to conclude that a "Detected Operon" is a false positive when MSI≥700 (see table 1 for details).

Operon name	Operon size	MSI (most significant interval)				
		ANOVA	*t*-test	Paired *t*-test	PCA	ICA
fliLMY cheY fliZPQR flhBAF ylxH cheBAWCD sigD ylxL	19	2385	2242	1613	1193	2499
yonRSTUVX yopAB	8	61	134	124	127	251
hemAXCDBL	6	1360	1547			
ruvAB queA tgt yrbF	5				1005	707

Table 1. Quantification of false positives

[We find ourselves with false positives. One exception is the operon yonRSTUVXyopAB, detected by all four tools, with small MSIs. As we cannot give a biological reason, we suspect that its detection is due to a default on the microarray used in the experiments.]

5. "Relevant Detected Operons" are identified (MSI<700). The definition of "Relevant Detected Operons" follows from the definition of false positives: "Relevant Detected Operons" have an MSI<700.
6. The accuracy of a "Relevant Detected Operon" is evaluated (MSI<150). We define that an operon is detected with good accuracy if its MSI is lower then a given threshold. Our results lead us to state that: Operons detected with good accuracy have an MSI<150.
7. The sensitivity of a tool is evaluated.

The sensitivity of the tools is estimated by comparing the number or "Relevant Detected Operons" identified by each tool.

We have decided to compare the five statistical tools under three experimental conditions biologists are frequently faced with:

* The experimental factor is identified and fully controlled. In the case of the microarray data used in this study, this factor is the sulphur source contained in the growth medium. In one case the sulphur source was methionine, in the other case it was methylthioribose. The five statistical tools were tested on these experimental data. The results obtained are displayed in table 2.

Operon name	Operon	MSI (most significant interval)				
	size	ANOVA	*t*-test	Paired	PCA	ICA
				t-test		
yqiXYZ	3	1	1	4	3	6
argCJBD carAB argF	7	15	28	29	201	56
argGH ytzD	3	1	1	6	6	2
ahpCF	2	46	7	85	11	13
lctEP	2	26			36	8
levDEFG sacC	5	316	220	287		
sunAT yolIJK	5		634			13
ydcPQRST yddABCDEFGHIJ	15				1313	116
ytmIJKLM hisP ytmO ytnIJ ribR hipO ytnM	12			45	92	
flgM yvyG flgKL yviEF csrA hag	8					509
fliLMY cheY fliZPQR flhBAF ylxH cheBAWCD sigD ylxL	19					350
yxbBA yxnB asnH yxaM	5				15	
yvrPONM	4			494		
ycbCD	2			40		
comGABCDEFG yqzE	8			49		
Relevant detected operons		6	6	9	7	9

Table 2. Comparison of the statistical tools when the experimental factor is identified and fully controlled

* The experimental factor is identified but not under control. In this case it was "day". The experiments were carried out twice, on different days. The protocol followed was the same on these two days; however, parameters like "room temperature" were not necessarily the same, thus introducing a factor in the experimental setup that was identified but not under control. The results obtained are displayed in table 3.

* The interaction between experimental factors. The aim of a protocol is to separate completely the different experimental factors. However, the

Operon name	Operon size	MSI (most significant interval)				
		ANOVA	*t*-test	Paired *t*-test	PCA	ICA
comGABCDEFG yqzE	8	16	26	28	6	4
comFABC yvyF	4	339			66	19
cotVWXYZ	5		148		315	417
groESL	2			37		
yvaVWXY	4				53	
yqxM sipW cotN	3				79	
comEABC	3					35
Relevant detected operons		2	2	2	5	4

Table 3. Comparison of the statistical tools when the experimental factor is identified but not under control

expression of certain genes may be under the control of more than one factor. In this case one talks of an "interaction between experimental factors". ANOVA and the *t*-tests are adapted to the analysis of variations due to a single experimental factor; they are not well suited for the study of interactions between factors; they were not tested under this condition. On the other hand, ICA and PCA are well adapted to cope with possible interactions; these interactions are identified because more than one factor plays a major role in the definition of an axis. The results obtained are displayed in table 4.

Operon name	Operon size	MSI	
		PCA	ICA
purMNHD	4	71	57
ybaC rpsJ rplCDWB rpsS rplV rpsC rplP rpmC rpsQ rplNXE rpsNH rplFR rpsE rpmD rplO secY adk map	25	51	56
alsS alsD	2		25
rpsL rpsG fus tufA	4		21
yvaVWXY	4		73
yxbBA yxnB asnH yxaM	5		126
yyaEF rpsF ssb rpsR	5	408	
Relevant detected operons		3	6

Table 4. Comparison of the statistical tools to detect possible interactions between the experimental factors

3 Results and discussion

Microarrays are defined as a tool for analyzing gene expression that consists of a small membrane or glass slide containing samples of many genes arranged in a regular pattern. They are widely used for analyzing the relative transcription level of genes. The number of statistical tools for analyzing the huge amount of data created in the experiments is continuously growing and no-one of these tools has yet emerged as the definitive one.

We have developed a protocol for the comparison of statistical tools applied to the analysis of transcription data. We have applied this method to compare five statistical tools (ANOVA, *t*-test, paired *t*-test, ICA and PCA) under three typical experimental conditions. All five tools were compared under two of these conditions (see tables 2 and 3 for details), whilst only ICA and PCA, which do not need the a priori definition of experimental factors, could be tested under the third condition (see table 4 for details).

Based on our observations, we have defined threshold values to define "Relevant Detected Operons" ($MSI<700$), false positives ($MSI\geq700$) and to define a good accuracy ($MSI<150$); the sensitivity of the tools is estimated by comparing the number of "Relevant Detected Operons" identified by each tool.

	ANOVA	*t*-test	Paired *t*-test	PCA	ICA
Relevant detected operons					
Table 2-4	8	8	11	15	19
Table 2-3	8	8	11	12	13
Accuracy of Detection					
Table 2-4	75%	75%	82%	80%	84%
Table 2-3	75%	75%	82%	83%	77%

Table 5. Overview of the results

[The table sums up the results obtained in this study. The first part of the table relates to the number of "Relevant Detected Operons" identified and thus to the tools' relative sensitivities. "Tables 2 - 4": adding the results from Tables 2, 3 and 4, the total of "Relevant Detected Operons" has been calculated for each tool. The entries for "Tables 2 - 3" have been obtained accordingly. The second part of the tables relates to the tools' accuracies: the percentage of "Relevant Detected Operons" identified with a "good accuracy" (MSI<150) has been calculated for each tool, adding the results from Tables 2, 3 and 4 ("Tables 2 - 4") etc.]

Table 5 sums up the results obtained. Overall, we observe that ANOVA and *t*-test have the lowest sensitivity, whilst ICA is the tool with the highest sensitivity. The same observations can be made regarding the accuracies of the tools. It is interesting to note that even under the two experimental conditions for which ANOVA and the *t*-test were conceived (tables 2 and 3), it performs less well than ICA. The paired *t*-test has a high accuracy but a lower sensitivity than ICA just like PCA. However, each tool may detect operons not identified by the other tools.

The results obtained by testing the five statistical tools show us that ICA has overall the best performance.

In this paper we have set out to describe a simple and efficient protocol to compare the reliability of different statistical tools available for microarray analyses. The criterion used in our method is based on the expression coherence of genes belonging to the same operon. The method is objective, reflects a biological reality and is not bound to one, particular set of experimental conditions. It allows to compare the sensitivity, the accuracy and the detection of false positives of different statistical tools.

Here we have used this method to compare statistical tools applied to the analysis of differential gene expression. However, the above protocol can also be applied without modification to compare the statistical tools developed for other types of transcriptome analyses, like the study of gene co-expression.

References

[Callow *et al.*, 2000]M. J. Callow, S. Dudoit, E. L. Gong, T. P. Speed, and E. M. Rubin. Microarray expression profiling identifies genes with altered expression in hdl-deficient mice. *Genome Res*, pages 2022–9., 2000.

[He *et al.*, 2003]Y. D. He, H. Dai, E. E. Schadt, G. Cavet, S. W. Edwards, S. B. Stepaniants, S. Duenwald, R. Kleinhanz, A. R. Jones, D. D. Shoemaker, and R. B Stoughton. Microarray standard data set and figures of merit for comparing data processing methods and experiment designs. *Bioinformatics*, pages 956–65., 2003.

[Heyer *et al.*, 1999]L. J. Heyer, S. Kruglyak, and S. Yooseph. Exploring expression data: identification and analysis of coexpressed genes. *Genome Res*, pages 1106–15., 1999.

[Pan, 2002]W. Pan. A comparative review of statistical methods for discovering differentially expressed genes in replicated microarray experiments. *Bioinformatic*, 2002.

[Reiner *et al.*, 2003]A. Reiner, D. Yekutieli, and Y. Benjamini. Identifying differentially expressed genes using false discovery rate controlling procedures. *Bioinformatics*, pages 368–75., 2003.

[Sabatti *et al.*, 2002]C. Sabatti, L. Rohlin, M. K. Oh, and J. C. Liao. Co-expression pattern from dna microarray experiments as a tool for operon prediction. *Nucleic Acids Res*, pages 2886–93., 2002.

[Sekowska *et al.*, 2001]A. Sekowska, S. Robin, J. J. Daudin, A. Henaut, and A. Danchin. Extracting biological information from dna arrays: an unexpected link between arginine and methionine metabolism in bacillus subtilis. *Genome Biol*, pages 0019.1–0019.12, 2001.

[Troyanskaya *et al.*, 2002]O. G. Troyanskaya, M. E. Garber, P. O. Brown, D. Botstein, and R. B. Altman. Nonparametric methods for identifying differentially expressed genes in microarray data. *Bioinformatics*, pages 1454–61., 2002.

Statistical and computational methods for haplotype reconstruction

Pierre-Yves Boëlle

Université Pierre et Marie Curie
Assistance Publique - Hôpitaux de Paris - INSERM U707
184 Rue du Faubourg Saint-Antoine
75012 PARIS

1 Introduction and biological background

With currently available genomic methods, it is increasingly easy to obtain detailed information on the genetic code, found in most organisms in the form of DNA strands or *chromosomes.* Due to the nature of DNA, it is not a surprise that the smallest detectable difference between two chromosomes is the "Single Nucleotide Polymorphism" (SNP), corresponding to a change in a base occurring at a given position (or *locus*) in two chromosomes. SNPs happen to be fairly common in the genome (≈ 1 every 100 to 300 base pairs), and have become of primary importance for mapping purposes (i.e. locating a gene within a chromosome) because they provide a very dense set of markers.

In *diploid* organisms, chromosomes are found in homologous pairs. Therefore, the genetic information, or *genotype*, consists of the sequences of both copies. While this information is readily obtained from sequencing, it is not technically feasible, at least with today's high throughput methods, to obtain *phased* information, corresponding to the exact two sequences underlying the genotype. Consider for example a situation where at a first locus, the genotype A T was found (A on one chromosome, T on the other), and on a nearby locus the genotype G C was found. Knowing *phase* amounts to know if one chromosome bore A and G and the other T and C, or alternatively A and C and T and G. When *phase* is known, a genotype may be split in two *haplotypes*: these correspond to a combination of SNPs on the same chromosome. Haplotypes give a more global picture of genetic variation, are more closely related to the notion of allele, and provide more opportunity to detect a dysfunctional version of a gene: it is therefore important to obtain this information, especially to correlate it with *phenotypic* information, corresponding to symptoms or conditions seen in individuals in case of polygenic disease.

In this text, a presentation of statistical haplotype reconstruction is given, and a review of currently used algorithms is presented. We conclude with some considerations regarding inclusion of these haplotypes in the analysis of correlation between haplotypes and phenotypes.

2 Notation

Consider a DNA region where K SNPs have been identified, and sequenced in a sample of individuals. We obtain a $n-$sample of genotypes $G = (g_i)_{1\leq i\leq n}$ taken from (assumed) independent individuals, each consisting of $K\times 2$ values. Each genotype is made of two unobserved haplotypes $H_i = (h_{i,1}, h_{i,2})$ taken in a collection $H = (h_j)_{1\leq j\leq m}$ haplotypes existing in the population. The distribution of probability on haplotype s in the population is $\Theta = (\theta_j)_{1\leq j\leq m}$, with $\sum \Theta_j = 1$
The precise number of haplotypes is generally unknown, but is obviously bounded upwards by 2^{K}. It is generally far less, and probably limited in most cases to a few dozens. Furthermore, in a given sample, random fluctuations may cause the absence of some haplotypes.
A pair of haplotypes is *consistent* with a genotype if the union of the pair sums up to the genotype. Such a pair is called a resolution of the genotype. The covering of a haplotype h is the number of individuals whose genotype may be resolved using h and another haplotype.

3 Methods based on parsimony

These methods, first proposed by Clark, provide a very straightforward approach to haplotype reconstruction(1). First, the set of haplotypes H is set to that of the "unambiguous" haplotypes H_{U} determined from all individuals who have at most one discordant SNP among the K sequenced sites. Some ambiguous subjects may readily be resolved using pairs of haplotypes found in H_{U} . In case of multiple solutions, one is taken at random. Some subjects may be resolved using one haplotype in H $_{\mathrm{U}}$ and another haplotype h ,in this case the latter is added to H. By repeatedly applying the last step with unresolved genotypes , the set of haplotypes is grown to explain the maximum number of genotypes. Limitations of the method include that some genotypes may not be resol ved at all by this procedure; furthermore it is dependant on the order of presentation of the genotypes. Clark advocated repeating the procedure several times to choose the most parsimonious solution.
A more systematic approach was presented recently, using a branch and bound algorithm(2). Instead of adding sequentially haplotypes from randomly chosen genotypes, the set of resolutions consistent with each ambiguous haplotype is first enumerated. Then, starting from a solution (for example take the first resolution of each genotype), all combinations are sequentially explored. When it appears that the explored solution will require more haplotypes than the best current solution, it is discarded at once. When an explored solution requires less haplotypes than the best current, it replaces this latter. A solution with the least possible haplotypes is ultimately recovered. With minor improvements, this approach is able to deal with missing data at some SNPs: it suffices to include as resolutions all pairs of haplotypes consistent with the observed sites.

4 Haplotype reconstruction as perfect phylogeny

One model for describing genetic evolution is known as the *coalescent*. In summary, evolution is described along a tree, starting from a single branch corresponding to a unique ancestral allele, and where each embranchment corresponds to the occurrence of a new haplotype, appearing by mutation from an already existing one. The resulting tree is called a *phylogeny*, where all leaves correspond to existing haplotypes. In practical problems, phylogenies are unknown, but because haplotypes are thought to have occurred by the coalescent, it is tempting to impose that the set of haplotypes used to explain a sample of genotypes should form a phylogeny(3). In this approach, a further hypothesis is that recombination has been rare, whereby new haplotypes as a mixture of already existing ones is neglected.
The set of genotypes is presented as a $n \times K$ matrix, with values 0, 1, 2 corresponding to a "wild" homozygous, "mutated" homozygous, and heterozygous site. The Perfect Phylogeny Haplotype problem is then to find a $2n \times K$ binary matrix M of resolutions, with each row a haplotype, and a phylogeny where each row of M corresponds to a leaf.
An algorithm has been proposed to efficiently find a solution to this problem , when it exists. It rests on the characterization of a matrix M as defining a perfect phylogeny if no submatrix of size $2n \times 2$ may be extracted that contains all rows to exclude possible resolutions. A bound is available for the number of solutions: if K−K0 is the number of sites where heterozygosity has been observed, then there are at most 2^{K0} solutions allowing perfect phylogeny.

5 Maximum likelihood with the EM algorithm

EM algorithm (4)

Under the assumption of random mating, the probability of finding a genotype made of the pair ($h_{.,1}$= h_{j}, $h.,\ _2$=h_{k}) is the product θ_i, θ_j of the individual haplotype frequencies. If the pairs making a genotype g are not observed, it is still possible to write the likelihood of this genotype by summing the probabilities over a ll its resolutions. Therefore, the likelihood is available, and maximum likelihood estimates may be obtained.
It turns out that a solution may be obtained by the *EM* algorithm. Write Θ^t for the distribution of the m g enotypes. A formal EM algorithm is obtained by iterating over equation

$$\theta_g^{t+1} = \frac{E_{\theta^t}(n_g|G)}{2n}$$

until probabilities do not change much. Uncertainty on the frequencies may be obtained from the associated Fisher's information matrix.

Contrary to the two methods described above, the method does not end up with a single possibility for each genotype. On the contrary, the probability of each consistent resolution may be determined and taken into account in further calculations. Like all instances of the *EM* algorithm, convergence may be rather slow, all the more when K increases.

6 Haplotype reconstruction using Bayesian methods

PHASE(5)

To improve on *EM* reconstruction, Bayesian methods have been proposed that incorporate imputation of haplotypes using Gibbs sampling. In this approach, convergence to a stationary distribution of haplotypes may theoretically be obtained.
Starting form an initial set of resolutions $H^{(0)} = (H_i^{(0)})_{1 \leq i \leq n}$ for G, where each $H_i^{(0)}$ corresponds to a pair of haplotypes resolving individual i, the following steps are repeatedly applied to obtain an updated resolution $H^{(\mathrm{t}+1)}$ from the current set $H^{(\mathrm{t})}$:

1. choose an individual i from all ambiguous individuals,
2. sample $H_i^{(t+1)}$ from the law of $H_i^{(t+1)}|G, H_{-i}^{(t)}$, where $H_{-i}^{(t)}$ is the current set of resolutions excluding subject i,
3. set $H^{(t+1)} = (H_i^{(t+1)})_{1 \leq i \leq n}$

The distribution is updated a large number of times, and samples from the distribution on haplotypes is obtained by states of $H^{(t)}$, after an appropriate burn−in period has been discarded, and with suffic ient thinning to avoid correlation in the output.
The only problem left in this approach is the determination of a convenient proposal law for $H_i^{(t+1)}|G, H_{-i}^{(t)}$. Stephens has shown that this law was proportional to $\pi(h_{i,1}|H_{-i})\pi(h_{i,2}|H_{-i}, h_{i,1})$, where $\pi\,(h\,|H\,)$ was the conditional probability of a haplotype h given a set H of previously sampled haplotypes. Fur ther, they proposed, from an analysis of the distribution of haplotypes generated under the coalescent theory in randomly sampled individuals that this conditional probability could be approx imated by a parametric law depending on a mutation rate and mutation matrix that could efficiently be sampled from.
However, when haplotypes are made of a large number of SNPs, it becomes impractical to adopt the above approach. Therefore, instead of updating the whole haplotype for subject i, only a subset of SNPs is updated at a given time, giving a local updating strategy.

HAPLOTYPER(6)

Another take at updating a large number of haplotypes is to explicitly subset the problem, using a “divide and conquer” strategy. In this approach, the set of K SNPs is split in adjacent blocks of moderate length L (≤ 8 for example). Because there are less than 2^{L} haplotypes, it is possible to enumerate all haplotypes in the block, and to sample from their distribution by Gibbs sampling, using a Dirichlet prior for the frequency of haplotypes. Once convergence is met on the separate blocks, ligation may occur: adjacent blocks are united either sequentially or hierarchically. At each ligation, a set of haplotypes for exploration is made by combination of the best B haplotypes of each block. This strategy leads to much improved computational efficiency.

7 Conclusion

Several strategies have been described for haplotype reconstruction from genotypic data. The first are combinatorial, and proceed by a systematic exploration of all resolutions. These methods have two characteristics: they are easily understood, and efficient algorithms have been found to reach a solution when it exists. However, these methods are not cast in a statistical framework, and may give a false sense of certainty when a solution is found. Indeed, statistical uncertainties due to sampling and *ad hoc* simplifications are not taken into account.

The second kind of methods is based on statistical maximum likelihood estimation, either in a frequentist or Bayesian framework. The *EM* approach was until recently the only available approach of this kind. Of practical importance is that it is possible to analyse the association between phenotypes and haplotypes, even if these have not been observed(7).

In fact, it is possible by spreading every observed genotype on the set of compatible haplotypes.

Methods based on more Bayesian sampling, using the Gibbs sampler have emerged as a very efficient alternative, consistently outperforming the previous methods. Software packages have been released that make the approach available to the community. They differ in how much data they can handle in the same run; and also in how missing data is dealt with. Some progress is possible on the algorithms : for example, Stephens recently incorporated the idea of partition/ligation in their approach, leading to much improved performance(8). It is still unknown if perfect sampling could be used in this respect.

Finally, it should be remarked that the presented methods have been evaluated mostly using simulated data. It may now be technically possible to obtain phased information on small samples, which will provide an opportunity to test the methods with real data.

References

[Bafna *et al.*, 2003]V. Bafna, D. Gusfield, G. Lancia, and S. Yooseph. Haplotyping as perfect phylogeny: a direct approach. *J Comput Biol*, pages 323–40, 2003.

[Clark, 1990]AG Clark. Inference of haplotypes from pcr-amplified samples of diploid populations. *Mol Biol Evol*, pages 111–22, 1990.

[Excoffier and Slatkin, 1995]L. Excoffier and M. Slatkin. Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population. *Mol Biol Evol*, pages 921–7, 1995.

[Niu *et al.*, 2002]T. Niu, ZS. Qin, X. Xu, and JS. Liu. Bayesian haplotype inference for multiple linked single-nucleotide polymorphisms. *Am J Hum Genet*, pages 157–69, 2002.

[Schaid *et al.*, 2002]DJ. Schaid, CM. Rowland, DE. Tines, RM. Jacobson, and GA. Poland. Score tests for association between traits and haplotypes when linkage phase is ambiguous. *Am J Hum Genet*, pages 425–34, 2002.

[Stephens and Donnelly, 2003]M. Stephens and P. Donnelly. A comparison of bayesian methods for haplotype reconstruction from population genotype data. *Am J Hum Genet*, pages 1162–9, 2003.

[Stephens *et al.*, 2001]M. Stephens, NJ. Smith, and P. Donnelly. A new statistical method for haplotype reconstruction from population data. *Am J Hum Genet*, pages 978–89, 2001.

[Wang and Xu, 2003]L. Wang and Y. Xu. Haplotype inference by maximum parsimony. *Bioinformatics*, pages 1773–80, 2003.

Modeling and identification of biological networks

Florence d'Alché–Buc[1] and Vincent Schachter[2]

[1] LaMI UMR CNRS 8042 and Epigenomics Programme, Genopole
523 Place des terrasses
91 000 Evry, France
(e-mail: dalche@lami.univ-evry.fr)
[2] Genoscope, Consortium National de Recherche en Génomique
91 000 Evry, France
(e-mail: vs@genoscope.cns.fr)

Abstract. Modeling, reverse-engineering and analysis of macromolecular networks has spurred increasing interest in the computational biology and the biostatistics communities. Biologists need rigorous and flexible tools to describe, infer and study these complex systems. This survey focuses on some of the latest advances on the corresponding direct and reverse modeling approaches.
Keywords: biological networks, gene networks, metabolic networks, machine learning.

1 Introduction

With the availability of complete genome sequences and high-throughput, post-genomics experimental data, the last 5 years have witnessed a growing interest in the study of networks of macromolecular interactions.

During the last few years, modeling efforts have targeted several distinct types of networks at the molecular level : gene regulatory networks, metabolic networks, signal transduction networks or protein-protein interaction networks, not to mention networks of interactions that are not restricted to a cell (intercellular communications) or take place at an altogether different level of detail (immunological networks, ecological networks). Here, we focus exclusively on molecular processes that take place within a cell, and more specifically on two distinct types of cellular mechanisms : transcriptional regulation and metabolism.

A major challenge consists in identifying with reasonable accuracy those complex macromolecular interactions that take place at different levels from genes to metabolites through proteins. Once identified, a network model can be used to simulate the process it represents, or for a variety of analyses, ranging from statistical properties of its topology to predictions of features of its dynamic behavior, or even prediction of cellular phenotypes.

This review focuses on modeling frameworks for biological networks, and on the existing methods to identify models from data within these frameworks. Framework choice and design are influenced both by targeted analyses, and by the need for model identification methods that will yield exploitable results given available experimental data and prior knowledge.

2 Models of biological networks

Why design models of biological networks ? A first motivation is to present a synthetic view of the current state of biological knowledge on a given network, and to structure it in a way that brings to sight relevant properties that might remain hidden without the model, or with a less relevant model. A second motivation is to allow predictions of (properties of) the network's dynamical behavior, one key point being that if these predictions can be compared with experimental results, they should allow either confirmation of the model's accuracy or, better yet, correction of the model.

Recent modeling framework proposals abound (see [de Jong, 02] for a detailed review), resulting in significant advances in the biological network modeling field, but also in a conceptual landscape that seems somewhat cluttered and unstructured. This impression is only superficial, however. The landscape can be simplified by regrouping frameworks that have similar underlying mathematical structure. In addition, models are very often goal-oriented, each framework was originally designed with some analytical aim in mind. In the rest of this section, we review families of formalisms classified according to the types of analyses and predictions for which they are best suited.

As we will see in the section 3, however, such a classification is only one-half of the story : available experimental data and model identification methods can also have a strong influence on the choice or design of a framework. The final modeling choice is often the result of a subtle balancing act between the requirements of model identification and the goals of the intended analyses.

2.1 Gene regulatory networks

Transcriptional regulation is the process by which genes regulate the transcription of other genes. A gene A *directly regulates* a gene B if the protein that is encoded by A is a transcription factor for gene B, ie if it binds to DNA on a specific site near the sequence coding for B, called a *regulatory region* of B, and activates or inhibits its rate of transcription. Regulation can be indirect, e.g. A activates B, which activates C, and *cooperative*, i.e. several genes regulate the same target gene in a non purely additive manner.

Several types of experimental data provide information on the transcriptional regulation process, some of which can be produced at high-throughput, while others still result from targeted, context-dependent assays and therefore can be acquired only for small networks. The main high-throughput technology is DNA chips which measure the concentration of mRNAs (a.k.a the *expression level of genes*) corresponding to all genes (or a large set of genes) in the organism under study, for several time-points or several different *conditions*, i.e. environmental changes, genetic or chemical perturbations of the system [Spellman *et al.*, 98]. These experiments can be seen as providing instantaneous pictures of the state of the regulatory network. Other sources of information include ChIP-chip ([Lee *et al.*, 02]) assays, that detect direct regulatory influences by identifying the binding of a protein to a regulatory region (in other words, a protein-DNA interaction), as well as the identification of sequence segments that are similar to known regulatory sequences, using sequence comparison methods. So far, large-scale expression and protein-DNA datasets have been generated mainly for model organisms, i.e. *Saccharomyces cerevisiae* (common yeast) and to a lesser extent the bacterium *E.coli*, making these the most likely candidates for any large scale reverse modeling effort.

2.1.1 Analyses of network topology : directed graphs The first category of properties of interest in biological networks are those related to their (static) network structure. Such topological analyses are most meaningful when applied to large ('genome-scale') networks, the aim being to identify statistical properties that can be interpreted as 'traces' of underlying biological mechanisms or design principles, related for instance to their dynamics [Shen-Orr *et al.*, 02, Watts and Strogatz, 98] (how the connectivity structure of the biological process reflects its dynamics), to their evolution [Jeong *et al.*, 00, Wagner, 01] (i.e. likely scenarios for the evolution of a network exhibiting the observed property or properties), or to both [Jeong *et al.*, 01, Milo *et al.*, 04].

One should emphasize that those analyses that are motivated by the search for insights into network dynamics focus on network structure mostly because large-scale data on network dynamics is not yet available. They can provide valuable insight insofar as the interpretative leap between static structure and dynamic behavior is performed carefully. Statistical graph properties that have been studied in this context include the distribution of vertex degrees [Jeong *et al.*, 01], the distribution of the clustering coefficient and other notions of density [Newman, 03, Guelzim *et al.*, 02], the distribution of vertex-vertex distances [Ravasz *et al.*, 02], and the distribution of network motifs occurrences [Milo *et al.*, 02].

The framework of choice to study these properties is also the most straightforward one. A gene regulatory network is viewed as a *directed graph* : a pair (V,E) where V is a set of vertices and E a set of directed edges, i.e. pairs (i,j) of vertices, where i is the source vertex and j the target vertex. Vertices of the graph represent genes, edges represent regulatory influences. Note that in some cases, it may be preferable to work with undirected graphs instead, for instance when only the existence of a correlation between the expression levels of two genes is known, but not the causal direction.

This simple model can be enriched by adding information (labels) on vertices or edges : for instance, '+' or '-' labels on edges may indicate positive or negative regulatory influence, the existence of an edge may be specified as conditional on the cell being in a specific global state, or on the source gene (the regulator) being expressed above a given threshold. These latter types of additional information, however, refer implicitly to notions of state and temporal evolution, and thus lead naturally towards qualitative dynamical models.

Finally, it is worth mentioning that enriched graph representations are also at the core of most existing biological pathways databases [Cary *et al.*, 05]. One reason is their simplicity, another one is that basic or complex queries on biological networks often correspond to classical operations on graphs, e.g. the search for paths between genes obeying given conditions.

2.1.2 Analyses of network dynamics : continuous models, discrete models The dynamics of regulatory processes has been the object of intense recent scrutiny. Whereas understanding the detailed dynamics of a regulatory network requires more experimental information than deciphering its static structure, dynamics is obviously one step closer to biological function.

Models can be used to run simulations of the biological system under study, with various choices of values for parameters corresponding either to unknown system characteristics or to environmental conditions. Comparison of simulated dynamics with experimental measurements can help refine the model or provide insight on qualitative properties of the system's dynamical behavior. The latter can also be addressed directly, by reasoning on or identifying properties of the system's behavior instead of simulating it, with the help of theoretical tools that depend on the choice of formalism. Dynamical properties of interest include the identification of steady states or limit cycles, identification of multistable (e.g. switch-like) behavior , identification of oscillatory behavior, characterization of the role of some parts of the network in terms of signal processing (e.g. amplifiers, derivators, logic gates) , and assessment of robustness environmental changes or genetic perturbation (see [Tyson *et al.*, 03, Wolf and Arkin, 03] for detailed review).

The default modeling option to simulate the dynamics of regulatory processes is to write a system of differential equations that govern the evolution of mRNA and protein concentrations. Typically, a gene regulatory network is modeled as a system of rate equations of the form : $\frac{dx_i}{dt} = f_i(\mathbf{x})$, $1 \leq i \leq n$ where $\mathbf{x}= (x_1, \ldots, x_n)$ is the vector of concentrations (of mRNAs, proteins or small molecules) and $f_i : \Re \rightarrow \Re^n$ a function, not necessarily linear. The level of detail and the complexity of these *kinetic models* can be adjusted, through the choice of the rate functions f_i. Typical tradeoffs include :

- using a more or less simplified set of entities and reactions, e.g. choosing whether to take into account mRNA and protein degradation,
- including delays to account for transcription, translation or diffusion time
- using more or less detailed kinetics, i.e. specific forms of f_i

Systems of differential equations as a modeling framework for biological networks presents two major drawbacks. Each equation in the model requires the knowledge of one or several parameter values (thermodynamic constants, rate constants), which is out of the present reach of high-throughput data production techniques. It is thus difficult to instantiate models of large networks directly, and reverse-engineering techniques are limited in how much information they allow to extract from limited datasets. Moreover, deriving meaningful dynamical properties of large differential equations system is a challenge : the f_i being nonlinear, analytical solutions are not known in the general case. So far these systems have been mainly used for numerical simulations within given parameter ranges (realistic or not), possibly complemented by bifurcation analysis, rather than submitted to analytical approaches [de Jong, 02].

These limitations have motivated two main tracks of investigation on alternative modeling frameworks for biological networks : simplified kinetic models on one hand, and discrete[1] models on the other hand.

Simplified continuous frameworks include piecewise-linear differential equations, a special case of rate equations where the response of a gene to regulatory stimuli (the function f_i) is approximated by a step function [de Jong, 02]. Linearity facilitates the analytical treatment of some dynamical properties, such as steady states. Systems of piecewise-linear differential equations can also be analyzed qualitatively by discretizing and recasting them within the framework of *qualitative differential equations*, where variables and their derivatives take qualitative (discrete) values and functions f_i are abstracted into sets of qualitative constraints.

Several discrete modeling frameworks have been proposed, each with a specific tradeoff between the level of detail of its chosen observables and the type of analyses that it enables : boolean networks (see below), generalized logical networks [Thomas *et al.*, 1995](a generalization of boolean networks that increases biological realism by allowing variables to have more than two values and using asynchronous transitions), petri-nets [Matsuno *et al.*, 2000], process-algebra [Regev *et al.*, 2001], rule-based formalisms [Chabrier *et al.*, 04].

[1] Here, we mean that *time* is discretized, leading to frameworks where the dynamics is governed by state transitions between t and t+1. Discretization of expression levels and/or of rate functions are a different path to simplification of either time-continuous or time-discrete frameworks.

2.2 Metabolic networks

Metabolism is the set of processes by which a cell extracts energy and raw material from its environment, and uses both to produce the components (DNA, proteins, lipids...)necessary for its survival and function, and to interact with its environment. Metabolic networks are thus networks of biochemical reactions : each reaction transforms one or several substrates (metabolites, i.e. small organic molecules) into one or several products (metabolites as well).To occur within a cell at a significant rate, a metabolic reaction needs to be catalyzed by an enzyme (a protein with catalytic activity) specific to that reaction.

Much of the classification introduced above for regulatory networks applies to metabolic networks ; indeed, several formalisms and analytical tools have been used on both. One should not be misled by these similarities, however : metabolic networks and regulatory networks represent very distinct, albeit interrelated, biological mechanisms, and this does translate into mathematical differences.

The framework of choice to capture the connectivity structure of a metabolic network is a directed bipartite graph (rather than a simple directed graph) : vertices correspond respectively to metabolites and reactions, edges represent production or consumption of a metabolite by a reaction. Two types of simpler graphs can be extracted from such a bipartite graph : *enzyme graphs*, where an edge between two reaction vertices denotes the fact that a product of the source reaction is a substrate of the target reaction (and can also denote the causal ordering of reactions in metabolic processes), and *metabolite graphs*, where vertices representing metabolites are linked when a reaction consumes one to produce the other. For all three graph types, active areas of research include the definition of biologically meaningful distances and the design of relevant and computable subgraph similarity measures to allow comparative studies.

Metabolic networks dynamics can be expressed as described above, using systems of rate equations and a given approximation for rate functions. Attempts at analytical reasoning have spurred the development of various simplified frameworks, including *Biochemicals Systems Theory* [Savageau, 1991] where production and consumption rates are expressed using a power-law approximation, and *Metabolic Control Analysis* [Westerhoff *et al.*, 1994], which focuses on a first-order approximation of the dynamical system in the neighborhood of steady-state. It is worth noticing that discrete frameworks have seldom been used to model metabolism : metabolite fluxes and concentrations are the key variables of interest here, in contrast with regulatory networks where an on/off discretization of the state of a gene already provides valuable information on the regulatory logic. Another type of abstract, scalable framework have been successfully applied to metabolic modeling : constraint-based modeling.

Constraint-based modeling is a framework dedicated to the modelling of metabolic processes at steady state : a global state of the metabolic network is defined as a distribution of fluxes within the network reactions. It emerged in the 90s as a simplification of kinetic models (mostly in the Schuster and Palsson groups), and was developed to allow tractable modelling of genome-scale metabolic networks [Price *et al.*, 04]. The steady-state hypothesis positions the framework at a level of detail intermediate between description of static network structure and representation of network dynamics. It is designed to represent incomplete information, yet to allow some prediction of metabolic behavior. The focus, rather than being on fully instantiated descriptions of the system's behavior, is on sets of such descriptions, i.e. sets of flux distributions compatible with a set of constraints representing the current knowledge on the structure of the network, on thermodynamic and kinetic parameters, and on input/output relationships of the network with its environment. The solution set can be refined incrementally as new constraints are added, ensuring some robustness in structural analyses and metabolic behaviour predictions with respect to modifications of the model. As this framework has been applied successfully to a variety structural analyses and predictive tasks on large metabolic networks in bacteria and yeast, yielding interesting biological results, efforts are under way to extend it while preserving simplicity and tractability.

3 Model identification: a machine learning problem

Once a formal framework is defined to describe models of biological networks, the question of how to choose parameters arises. Various works have shown that this identification problem can be expressed in the framework of machine learning. Given a family of mathematical models of gene interactions and a set of observations, learning consists here in optimizing the parameters of the model in such way that it captures the observed behavior of the true system. The ability of the instantiated model to be used in prediction is referred as the generalization property. A model is able to generalize if learning ensures a trade-off between a good fit to the data and simplicity of the model. Solving a learning problem leads to three key questions : the representation problem, the optimization problem and the validation problem. The **representation** problem concerns mostly the choice of the formalism in which data and the model are going to be expressed, and the method to encode them into this formalism. Both symbolic and numerical learning leads to an **optimization** problem whose nature is combinatorial (for symbolic learning) and numeric (for statistical learning). Statistical approaches generally use maximum likelihood criteria penalized by a parsimony constraint. Combinatorial approaches are solved using heuristics to ensure a large exploration of the models spaces. At last **validation** is required to identify how one can trust the inferred model. In this area statistical approaches benefits from an important background in statistical validation of estimation methods. The validation question is far from being solved in the context of network reconstruction, however.

Most efforts on biological networks identification fall into this framework, albeit most of them do not address each of the above key issues. They can be divided into two categories: on one hand, the static approaches that neglect temporal aspects and focus on the sole reconstruction of the interaction graph, and on the other hand, the dynamic approaches that aim at modeling the underlying dynamic system providing both structure and dynamics parameters.

3.1 Static approaches

Several static approaches have yielded promising results in reconstructing gene networks. In order to focus on the structure identification problem, they ignore the temporal aspects and search for causality chains among the variables at hand. This point of view is based on the implicit assumption that the underlying dynamical process is at equilibrium, and that no circuit exists among studied genes.

Bayesian networks are undoubtedly the most successful approach to gene networks structure reconstruction.They represent the expression levels of genes as random variables, whose joint probability law has to be identified. This model has two major advantages : it takes into account the inherent stochastic character of biological processes and it is able to cope with noisy data. A graphical display of such models can be obtained by considering directed acyclic graphs whose vertices are genes and whose edges are modeled by conditional probabilities distributions. Choosing discrete or continuous variables, parametric or non parametric forms for the cpd's are the main questions in the representation problem.

Learning bayesian networks consists in estimating the joint probability distribution of the variables using available data. The core issue is to find the decomposition of the joint law in the conditional probability distributions (cpd's) among the relevant variables. This decomposition defines the graph structure. Once the structure of a network is given, the task of learning cpd 's is not difficult. Learning the structure, however, is an NP-hard problem that can only be tackled by heuristics. Several pioneering results in this area have been achieved using a constructive strategy. Reconstruction has been shown to be successful on the yeast cell cycle dataset of [Spellman *et al.*, 98].

Extensions of these results were obtained by integration of prior knowledge into the model. For instance, [E.Segal *et al.*, 01] introduced an enriched formalism, probabilistic relational networks (PRN) that allows to deal with object variables instead of simple discrete or continuous variables. Information about promoters and genomic sequence can be thus be introduced. While information propagation in the net is modified and for this reason learning becomes also more complex, this work opens the door to new formalisms that couple high level descriptions with a probabilistic framework.

Another variations on bayesian networks models is the so-called 'module networks' approach. Module networks introduced in [E.Segal *et al.*, 03] have been proposed as bayesian networks with a special structure, where the variables sharing the same parents are gathered into a so-called module, i.e. a set of genes that appear to be co-regulated in some experimental conditions. Elucidating which are the genes that belong to the same modules and what are the conditions under which these regulations occur can be solved using a Expectation-Maximization(EM) based algorithm that starts from relevant initializations.

Validation can consist in the comparison between the inferred structure and the true structure. The simplest way consists in a comparison between the inferred network with other sources of knowledge. Precision and recall measures, ROC curves have also been proposed to evaluate the power of learning algorithms [Husmeier, 03].

However all these static approaches are not able to discover circuits in a graph interaction. The reason is that without considering time, it is not possible to elucidate feedback interactions that can only be observed an through time.

3.2 Dynamic approaches

Dynamic approaches aim at identifying the dynamics of the system implemented by a biological network while extracting the structure. Only discrete-time models are considered for learning since experimental data come from discrete-point measurements. In the area of genetic networks, the available data take the form of gene expression kinetics measured after some perturbation of the studied organism. While these data are more expensive to generate than static data, a few subsets exist and mainly concerns model organisms such as bacteria or yeast. Modeling dynamics of a network can serve both exploratory and explanatory goals. A long-term goal is of course to exploit these models in simulation and prediction for drug-design and therapeutical targeting. However its should be stresses that this feature has not yet been fully exploited in the existing works.

Dynamic models that have been considered for learning include Boolean networks, artificial recurrent neural networks, dynamic bayesian networks including state-space models. Learning in boolean networks has first been tackled with combinatorial algorithms [Akutsu *et al.*, 1999]and then renewed by using a randomized algorithm. However the best way to reduce complexity of the problem is to reduce the class of boolean functions as proposed in[Gat-Viks *et al.*, 03] with the so-called chain functions. Promising new directions have also been introduced by [Shmulevich *et al.*, 02] with the introduction of Probabilistic Boolean networks and learning algorithms devoted to their reverse engineering.

While boolean networks simplify the description of the system's dynamical behavior, quantitative models have attracted much attention from machine learning community because of the existence of a large set of efficient learning algorithms for numerical data. These models are usually based on the quantization of differential equations. Again the representation issue implies to choose among linear/non linear models and discrete/continuous variables. Keeping the model deterministic allows its implementation as a recurrent artificial network (see for instance [D'haeseleer *et al.*, 00]and [Mjolness *et al.*, 00]) for which learning algorithms such as genetic algorithms and backpropagation through time have been designed .This last feature avoids avoids data-overfitting. This framework can be compared to dynamic Bayesian networks that are inspired from stochastic differential equations [Hoon *et al.*, 03], or can simply be obtained by adding a noise component to the equation [Perrin *et al.*, 03]. These approaches aim at estimating the joint probability of the temporal sequence of network states. The optimization task takes the form of a likelihood maximization problem with a parsimony constraint.

Several dynamic approaches have been applied to different models, first order [Hoon *et al.*, 03], second order models [Perrin *et al.*, 03], and from linear to non linear [Nachman *et al.*, 04] to splines-based models. Validation of dynamical approaches can be performed by measuring the ability of the model to make k-step predictions or to predict the last part of the sequence used for training. However the most difficult point remains the ability of the algorithm to retrieve the structure of the network which can be deduced from the identified parameters.

4 Conclusion and perspectives

We have reviewed modeling formalisms for biological networks and their relationship to down stream analysis and reverse engineering methods. As this field of research matures, it is becoming increaslingly clear that there is no one-size-fits-all solution, but rather a range of frameworks and methods, each with its specific trade-off between abstraction and tractability, the ultimate test being the ability to answer relevant biological questions. Indeed, network models are only starting to become useful tools for biological investigation. Promising research directions include the design of frameworks that allow joint modeling of metabolism and regulation, the refinement of stochastic rule-based frameworks that are meant to capture intrinsic stochasticity in regulatory networks dynamics, the design of dedicated process calculi, and the development of model-checking tools. Another key direction, towards, efficient model inference is the elaboration of formalisms that are able to support high level language of description while managing uncertainty in the data.

References

[Akutsu *et al.*, 1999]T. Akutsu, S. Miyano, and S. Kuhara. Identification of genetic networks from a small number of gene expression patterns under the boolean network model. In *Proc. of PSB*, volume 4, pages 17–28, 1999.

[Cary *et al.*, 05]M. P. Cary, G. D. Bader, and C. Sander. Pathway information for systems biology. *FEBS Lett*, 579(8):1815–20, 05. 0014-5793 Journal Article.

[Chabrier *et al.*, 04]N. Chabrier, M. Chiaverini, V. Danos, F. Fages, and V. Schachter. Modeling and querying biomolecular networks. *Theoretical Computer Science*, 325(1):25–44, 04.

[de Jong, 02]H. de Jong. Modeling and simulation of genetic regulatory systems: a literature review. *J Comput Biol*, 9(1):67–103, 02. 1066-5277 Journal Article Review.

[D'haeseleer *et al.*, 00]P. D'haeseleer, S. Liang, and R. Somogyi. Genetic network inference: from co-expression clustering to reverse engineering. *Bioinformatics*, 16(8):707–726, 00.

[E.Segal *et al.*, 01]E.Segal, T.Taskar andA.Gasch, N.Friedman, and D.Koller. Rich probabilistic models for gene epxression. *Bioinformatics*, 17:S243–252, 01.

[E.Segal *et al.*, 03]E.Segal, D. Pe'er, A. Regev, D. Koller, and N. Friedman. Learning module networks. In *Proc. of UAI*, pages 523–534, 03.

[Gat-Viks *et al.*, 03]I; Gat-Viks, R.Shamir, R.M. Karp, and R. Sharan. Reconstructing chain fucntions in genetic networks. *Bioinformatics*, 19:i108–i117, 03.

[Guelzim *et al.*, 02]N. Guelzim, S. Bottani, P. Bourgine, and F. Kepes. Topological and causal structure of the yeast transcriptional regulatory network. *Nat Genet*, 31(1):60–3, 02.

[Hoon *et al.*, 03]M. De Hoon, S. Imoto, K. Kobayashi, N. Ogasawara, and S. Miyano. Inferring gene regulatory networks form time-ordered gene expression data of bacillus subtilis using differential equations. In *PSB*, pages 17–28, 03.

[Husmeier, 03]Dirk Husmeier. Sensitivity and specificity of inferring genetic regulatory interactions from microarray experiments with dynamic bayesian networks. *Bioinformatics*, 19(17):2271–2282, 03.

[Jeong *et al.*, 00]H. Jeong, B. Tombor, R. Albert, Z.N. Oltvai, and A.-L. Barabasi. The large-scale organization of metabolic networks. *Nature*, 407:651–654, 00.

[Jeong *et al.*, 01]H. Jeong, S. P. Mason, A. L. Barabasi, and Z. N. Oltvai. Lethality and centrality in protein networks. *Nature*, 411(6833):41–2, 01.

[Lee *et al.*, 02]T. I. Lee, N. J. Rinaldi, F. Robert, D. T. Odom, Z. Bar-Joseph, G. K. Gerber, N. M. Hannett, C. T. Harbison, C. M. Thompson, I. Simon, J. Zeitlinger, E. G. Jennings, H. L. Murray, D. B. Gordon, B. Ren, J. J. Wyrick, J. B. Tagne, T. L. Volkert, E. Fraenkel, D. K. Gifford, and R. A. Young. Transcriptional regulatory networks in saccharomyces cerevisiae. *Science*, 298(5594):799–804, 02.

[Matsuno *et al.*, 2000]H. Matsuno, A. Doi, M. Nagasaki, and S. Miyano. Hybrid petri net representation of gene regulatory network. *Pac Symp Biocomput*, pages 341–52, 2000. Journal Article.

[Milo *et al.*, 02]R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, and U. Alon. Network motifs: simple building blocks of complex networks. *Science*, 298(5594):824–7, 02.

[Milo *et al.*, 04]R. Milo, S. Itzkovitz, N. Kashtan, R. Levitt, S. Shen-Orr, I. Ayzenshtat, M. Sheffer, and U. Alon. Superfamilies of evolved and designed networks. *Science*, 303(5663):1538–42, 04.

[Mjolness *et al.*, 00]E. Mjolness, T. Mann, R. Castao, and B. Wold. From coexpression to coregulation: An approach to inferring transcriptional regulation among gene classes from large-scale expression data. In *Advances in Neural Information Processing Systems*, volume 12, pages 928–934. 00.

[Nachman *et al.*, 04]I. Nachman, A. Regev, and N. Friedman. Inferring quantitative models of regulatory networks from expression data. *Bioinformatics*, 20:i248–i258, 04.

[Newman, 03]M. E. Newman. Properties of highly clustered networks. *Phys Rev E Stat Nonlin Soft Matter Phys*, 68(2 Pt 2):026121, 03.

[Perrin *et al.*, 03]B.-E. Perrin, L. Ralaivola, A. Mazurie, S. Bottani, and F. d'Alché Buc. Gene networks inference using dynamic bayesian networks. *Bioinformatics*, 19:i38–i49, September 03.

[Price *et al.*, 04]N. D. Price, J. L. Reed, and B. O. Palsson. Genome-scale models of microbial cells: evaluating the consequences of constraints. *Nat Rev Microbiol*, 2(11):886–97, 04. 1740-1526 Journal Article Review Review, Tutorial.

[Ravasz *et al.*, 02]E. Ravasz, A. L. Somera, D. A. Mongru, Z. N. Oltvai, and A. L. Barabasi. Hierarchical organization of modularity in metabolic networks. *Science*, 297(5586):1551–5, 02.

[Regev *et al.*, 2001]A. Regev, W. Silverman, and E. Shapiro. Representation and simulation of biochemical processes using the pi-calculus process algebra. *Pac Symp Biocomput*, pages 459–70, 2001. Journal Article.

[Savageau, 1991]M. A. Savageau. Biochemical systems theory: operational differences among variant representations and their significance. *J Theor Biol*, 151(4):509–30, 1991. 0022-5193 Journal Article.

[Shen-Orr *et al.*, 02]S. S. Shen-Orr, R. Milo, S. Mangan, and U. Alon. Network motifs in the transcriptional regulation network of escherichia coli. *Nat Genet*, 31(1):64–8, 02.

[Shmulevich *et al.*, 02]I. Shmulevich, E. R. Dougherty, S. Kim, and W. Zhang. Probabilistic boolean networks: a rule-based uncertainty model for gene regulatory networks. *Bioinformatics*, 18(2):261–274, 02.

[Spellman *et al.*, 98]P. T. Spellman, G. Sherlock, M. Q. Zhang, V. R. Iyer, K. Anders, M. B. Eisen, P. O. Brown, D. Botstein, and B. Futcher. Comprehensive identification of cell cycle-regulated genes of the yeast saccharomyces cerevisiae by microarray hybridization. *Mol Biol Cell*, 9(12):3273–97., 98.

[Thomas *et al.*, 1995]R. Thomas, D. Thieffry, and M. Kaufman. Dynamical behaviour of biological regulatory networks–i. biological role of feedback loops and practical use of the concept of the loop-characteristic state. *Bull Math Biol*, 57(2):247–76, 1995. 0092-8240 Journal Article.

[Tyson *et al.*, 03]J. J. Tyson, K. C. Chen, and B. Novak. Sniffers, buzzers, toggles and blinkers: dynamics of regulatory and signaling pathways in the cell. *Curr Opin Cell Biol*, 15(2):221–31, 03. 0955-0674 Journal Article Review Review, Tutorial.

[Wagner, 01]A. Wagner. The yeast protein interaction network evolves rapidly and contains few redundant duplicate genes. *Mol Biol Evol*, 18(7):1283–92, 01.

[Watts and Strogatz, 98]D. J. Watts and S. H. Strogatz. Collective dynamics of 'small-world' networks. *Nature*, 393(6684):440–2, 98.

[Westerhoff *et al.*, 1994]H. V. Westerhoff, J. H. Hofmeyr, and B. N. Kholodenko. Getting to the inside of cells using metabolic control analysis. *Biophys Chem*, 50(3):273–83, 1994. 0301-4622 Journal Article.

[Wolf and Arkin, 03]D. M. Wolf and A. P. Arkin. Motifs, modules and games in bacteria. *Curr Opin Microbiol*, 6(2):125–34, 03.

Hidden Markov Model for protein secondary structure

Juliette Martin, Jean-Francois Gibrat, and Francois Rodolphe

Unité Mathématique Informatique et Génome,
INRA, Domaine de Vilvert,
78350 Jouy-en-Josas Cedex, France
(e-mail: [Juliette.Martin,Jean-Francois.Gibrat,
Francois.Rodolphe]@jouy.inra.fr)

Abstract. We address the problem of protein secondary structure prediction with Hidden Markov Models. A 21-state model is built using biological knowledge and statistical analysis of sequence motifs in regular secondary structures. Sequence family information is integrated *via* the combination of independent predictions of homologous sequences and a weighting scheme. Prediction accuracy with single sequences reaches 65.3% and raises to 72% of correct classification with profile information.

Keywords: α-helix, β-sheet, prediction.

1 Introduction

Proteins are the main actors of living cells. Many cellular constituents are made out of proteins. Almost all enzymes are proteins, cellular pumps and motors are made out of proteins.

The function of a protein strongly depends of its 3D-structure. For instance, enzymes need to have a tight spatial complementarity with their substrates (reaction partners). Thus knowledge of a protein structure gives relevant clues to its function.

Since genome sequencing started, the even widening gap between the number of protein sequences and protein structures available in databases enhances the utility of structure prediction methods. Because of the structure-function relationship, structures are more conserved than sequences during evolution and therefore different sequences can have the same 3D structure.

Structure prediction methods fall into two categories:

- comparative modeling if a related structure is known and can be used to derive a global model,
- *de novo* prediction if there is no related structure available.

We are presently interested in the latter. *De novo* prediction methods often require a first step of local structure prediction: secondary structure prediction in our case. Three canonical classes of secondary structures are considered : α-helices, β-strands and coil, see figure 1.

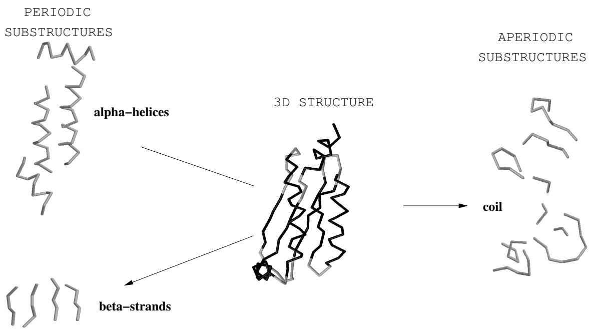


Fig. 1. Secondary structure of proteins. A 3D protein structure (center) can be described in term of secondary structures: α-helices, β-strands (left side) and coil (right side). Only C-α are shown, periodic substructures are indicated in black in the full 3D structure.

α-helices and β-strands are geometrically periodic sub-structures frequently occurring in 3D structures (about 50% of residues in proteins are involved in α-helices and β-strands). Coil denotes all sequence segments which do not fall into one of these two categories.

We use Hidden Markov Models to predict the three classes of secondary structure. The model is built using prior biological knowledge and pattern analysis in protein sequences.

2 Data set

The data set is a subset of 2530 structural domains taken from ASTRAL 1.65 [Brenner *et al.*, 2000], determined by X-ray, with a resolution factor less than 2.25 $\AA$ and less than 25% sequence identity. Secondary structure definition is given by an assignment method developed in our laboratory (manuscript in preparation) or by STRIDE method [Frishman and Argos, 1995]. 489743 residues have a defined secondary structure in our data set. 2024 sequences, randomly selected, are used in a four-fold cross validation procedure: three quarters of these sequences are used for parameter estimation and one quarter is used for the test. The remaining 506 sequences are used as an *independent* test set. This test set is never used to estimate model parameters. The use of an independent test set allows to check that no bias is introduced during the model design when searching for characteristic motifs in secondary structures (see hereafter). The number of residues with a defined secondary structure are 94790, 101521, 99796 and 99031 in the cross validation subsets and 94605 in the independent test set. The secondary structure contents are similar in all the subsets: about 39% of residues in α-helix, 24% in β-strand and 37% in coil with our assignment and 38%/22%/40% with STRIDE assignment.

3 Hidden Markov Models: application to secondary structure

In a Markovian sequence, the character appearing at position t only depends on the k preceding characters, k being the order of the Markov chain. Hence, a Markov chain is fully defined by the set of probabilities of each character given the past of the sequence in a k-long window: the transition matrix. In the hidden Markov model, the transition matrix can change along the sequence. The choice of the transition matrix is governed by another Markovian process, usually called the *hidden process*. Hidden Markov models are thus particularly useful to represent sequence heterogeneity. These models can be used in predictive approaches: some algorithms like the Viterbi algorithm and the forward-backward procedure allow to recover which transition matrix was used along the observed sequence.

In our case, it is known that different structural classes have different sequence specificity. Intuitively we want to use different Markov chains to model different secondary structures. Figure 2 illustrates the HMM-translation of our secondary structure prediction problem.

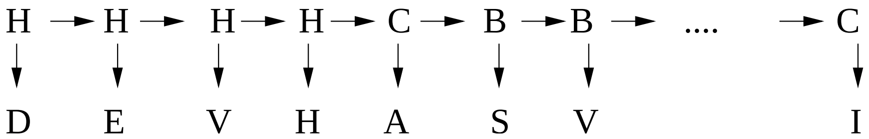


Fig. 2. Secondary structure prediction *via* a hidden Markov model. The upper line represents the secondary structure along a protein sequence: H for a residue in α-helix, B for β-strand, C for coil. The arrows between symbols symbolize the first order dependency of the *hidden process*. The lower line represents the amino-acid sequence of the protein. This is the *observed sequence*. Arrows between the two lines symbolize the dependency between the observed sequence and the hidden chain. The forward/backward algorithm will be used to recover the hidden process from the observed sequence.

The hidden process to be recovered is the secondary structure of the protein. The observed process is the amino-acid sequence. The hidden chain process is a first order Markov chain. Each hidden state is characterized by a distribution of amino-acids. Due to the large alphabet size, the order of the observed chain is 0, which means that amino-acids are independent conditionally on the the hidden process. We use the software called SHOW[1][Nicolas *et al.*, 2002] to design and train the model and to recover the hidden process. The prediction is achieved with the forward/backward algorithm. Note that this algorithm provides the probability associated to each hidden states at each position.

[1] http://www-mig.jouy.inra.fr/ssb/SHOW/

The simplest model for three-classes prediction is a HMM with three hidden states, each state accounting for a secondary structure class. Parameter estimation of such a model is straightforward because the segmentation is fully determined. But the performance of this model is limited: the Q3 score (proportion of residues with correct prediction) is 58.3%. A random prediction gives a Q3 score equals to 34.5%.

We thus want to design a model that takes into account the specific features of secondary structures.

4 Model of α-helices

A well-characterized sequence motif in α-helices is the amphiphilic motif, i.e., a succession of two polar residues and two apolar residues. This motif occurs when an helix has a side facing the solvent (thus preferentially supporting polar residues) while the other side faces the core of the protein (preferentially supporting apolar residues). This motif is very frequent. With the amino-acids classification; A,V,L,I,F,M,W,C=hydrophobic (h), S,T,Y,N,Q,-H,P,D,E,K,R=polar (p), the motif hhpphh or pphhpp is found in 24% of the helices in our cross-validation set. Glycine (G) residues do not exhibit strong preference for either polar or apolar environment. It is thus considered as a special type of residue and left apart. When reduced to hhpp or pphh, the motif is found in 69% helices. Figure 3 shows the model we propose to take into account the amphiphilic nature of α-helices. States H5 and H6 help to fit the periodicity of an α-helix which is 3.6 residues.

States with hydrophobic preference favour amino-acids A, V, L, I, F, P and M. States with polar preference favour S, T, N, Q, H, D, E, K and R.

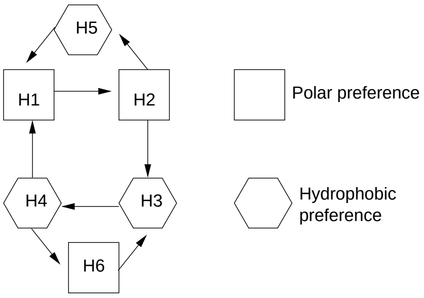


Fig. 3. Model for amphipatic helices

5 Model of β-strands

There is no strong motif characterizing β-strands similar to the amphipatic motif for α-helices. Characteristic motifs are found using a statistical approach based on exceptional words. A word is over (resp. under)-represented if its frequency in the data is significantly greater (resp. lower) than its expected frequency under some Markovian model. The R'MES software[2] [Bouvier *et al.*, 1999] is dedicated to this task. Amino-acids are grouped as before, the G is put into the hydrophobic group. Sequences of β-strands and α-helices in the cross-validation set are analyzed with R'MES using the Gaussian approximation.

Because the HMM uses a zero order for the observed chain, exceptional words when compared to a zero order Markov model are interesting. Interesting words should also be frequent in absolute (over-represented words are not necessarily frequent) and must not be over-represented in α-helices. We also consider some frequent words, although not over-represented, if they are under-represented in α-helices. Table 1 contains interesting words found in β-strand with R'MES. The over-representation is assessed by R'MES. The relative abundance is evaluated by looking at rank of the word when sorted according to the frequency.

Motif	Occurrence in β-strands	Occurrence in α-helices
hphp	over-represented and frequent	under-represented and not frequent
phph	over-represented and frequent	under-represented and not frequent
pphhh	over-represented and very frequent	under-represented and not frequent
pphph	over-represented and very frequent	under-represented and not frequent
hhhhp	not over-represented, but very frequent	under-represented and not frequent
phhhhp	not over-represented but very frequent	under-represented

Table 1. Interesting motifs in β-strands

Figure 4 shows the model we propose to take into account these words in β-strands. Words hphp and phphp are favoured by the alternation between states b1 and b2. This alternation corresponds to the case of β-strands at the solvent interface with one side facing the solvent and one side facing the core of the protein. The transition from state b4 to itself favours long runs of hydrophobic amino-acids in words pphhh, hhhhp, phhhhp. Long runs of hydrophobic residues are seen when β-strands are buried in the core of proteins. The transition between b2 and b3 favours the apparition of two polar amino-acids surrounded by hydrophobic ones appearing in words pphhh and pphph.

Note that the study of exceptional words on α-helices reveals that the motifs occurring in amphipatic α-helices are over-represented.

[2] http://www-mig.jouy.inra.fr/ssb/rmes/

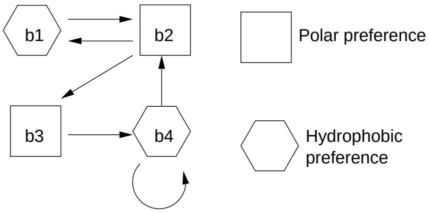


Fig. 4. Model for β-strands

6 Complete HMM for secondary structures

Models of β-strands and α-helices are merged to form a full model of secondary structures.

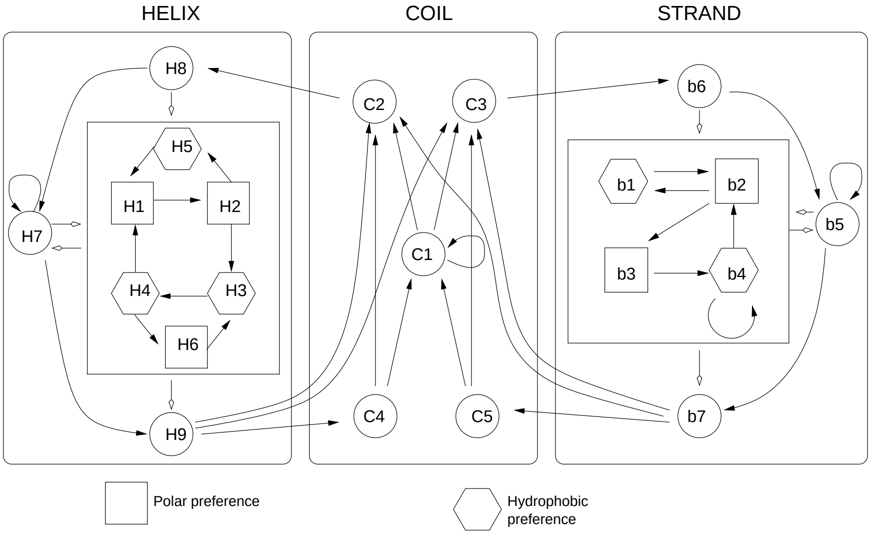


Fig. 5. Full model for secondary structure

Figure 5 shows the full model. Models of α-helices and β-strands integrate informations about frequent/over-represented words, but they don't necessarily reflect the totality of motifs in periodic structures. To allow the presence of β-strands and α-helices that do not fit well in the constrained models, two "generic" states were added (H7, b5). These states show no prior preference for polar or hydrophobic amino-acids. Transitions are allowed between all states of the constrained models and the "generic states". Specific states are added at secondary structure ends (H8, H9, b6, b7), as it is known that there are specific signals such as helix-caps. The coil is not well characterized yet, except the states preceding and following regular secondary structures.

Initial parameters for estimation by the EM algorithm are set as follows:

- Initial transition probabilities are set to $\frac{1}{n}$, with n the number of outgoing states.
- Initial emission probabilities are derived from those obtained on a simple 3-states model. Emission probabilities are manually modified to favour the apparition of polar amino-acids and penalize the emission of hydrophobic amino-acids in polar-preferring states (and vice-versa). No such bias is introduced in other states.

Prediction of the three structural classes (α-helix, β-strand, coil) is achieved by the forward-backward algorithm. The predicted structure is the one with the greatest posterior probability.

7 Integrating information from homologous sequences in the prediction

Protein structures are more conserved than sequences during evolution. Thus different sequences can have the same structure. This information has been successfully used in secondary structure prediction methods [Rost, 2003]. To integrate this information, the prediction is done independently on each sequence of a family. These sequences are detected using a search with PSI-BLAST against a database were the redundancy is reduced to 80% sequence identity. This search generates an average number of 60 sequences per family. Independent predictions are combined with a weighting scheme to generate a prediction for the sequence family using the formula

$$P(state = S/family) = \sum_i \lambda_i \times P(state = S/sequence_i)$$

with $P(state = S/sequence_i)$ provided by the forward-backward procedure and λ_i the weight of sequence i in the family. Sequence weights are computed as proposed in Henikoff and Henikoff [Henikoff and Henikoff, 1994].

Prediction on single sequence provides an accuracy of 65.2% residues correctly classified when compared to our secondary structure assignment, on the cross-validation test set. This score is 65.3% on the learning set and 65.6% on the independent test set. When compared with stride assignment, the accuracy is around 66.3% for all data sets. Hence, we experienced no over-fitting on the training data.

With the family sequence information, the percentage of correct prediction is in the range 71.3 to 72%. Best available methods, that also use sequence families, have achieved accuracy in the range of 78% (reported for reasonnably big datasets on the continuous evaluation server EVA, [Koh *et al.*, 2003]). Thus our results are not fully satisfying yet. However we think that our approach is promising because our model is relatively small, statistically speaking: the number of independent parameters is only 471. Most of existing methods use neural networks. The number of parameters, when reported, seems to be of the order of thousands [Pollastri *et al.*, 2002]. Moreover, the graphical nature of hidden Markov models allows intuitive data modeling. Along this line, an important perspective of this work is to introduce a geometrical description of coil. The coil class represents about 50% of residues in proteins. Even a perfect three state prediction would leave half of the data with no structural clue. We also think that the sequence family information could be taken into account more efficiently that it is done here. This is another of our perspectives.

References

[Bouvier *et al.*, 1999]A. Bouvier, F. Gélis, and S. Schbath. *RMES : Programs to Find Words with Unexpected Frequencies in DNA Sequences, User Guide (in french)*, 1999.

[Brenner *et al.*, 2000]S.E. Brenner, P. Koehl, and M. Levitt. The astral compendium for protein structure and sequence analysis. *Nucleic Acids Res*, 28(1):254–6, Jan 2000.

[Frishman and Argos, 1995]D. Frishman and P. Argos. Knowledge-based protein secondary structure assignment. *Proteins*, 23(4):566–79, Dec 1995.

[Henikoff and Henikoff, 1994]S. Henikoff and JG. Henikoff. Position-based sequence weights. *J Mol Biol*, 243(4):574–8, Nov 1994.

[Koh *et al.*, 2003]I.Y. Koh, V.A. Eyrich, M.A. Marti-Renom, D. Przybylski, M.S. Madhusudhan, N. Eswar, O. Grana, F. Pazos, A. Valencia, A. Sali, and B. Rost. Eva: evaluation of protein structure prediction servers. *Nucleic Acids Res*, 31(13):3311–5, Jul 2003.

[Nicolas *et al.*, 2002]P. Nicolas, A.S. Tocquet, and F. Muri-Majoube. *SHOW : Structured HOmogeneities Watcher. User Guide*, 2002.

[Pollastri *et al.*, 2002]G. Pollastri, D. Przybylski, B. Rost, and P. Baldi. Improving the prediction of protein secondary structure in three and eight classes using recurrent neural networks and profiles. *Proteins*, 2002.

[Rost, 2003]B. Rost. Prediction in 1d: secondary structure, membrane helices, and accessibility. *Methods Biochem Anal*, 44:559–87, 2003.

Part IV

Knowledge Management and Data Mining

Assessing rule interestingness with a probabilistic measure of deviation from equilibrium

Julien Blanchard, Fabrice Guillet, Henri Briand, and Régis Gras

LINA – FRE 2729 CNRS
Polytech'Nantes
La Chantrerie – BP 50609
44306 – Nantes cedex 3 – France
julien.blanchard@polytech.univ-nantes.fr

Abstract. Assessing rule interestingness is the cornerstone of successful applications of association rule discovery. In this article, we present a new measure of interestingness named $IPEE$. It has the unique feature of combining the two following characteristics: first, it is based on a probabilistic model, and secondly, it measures the deviation from what we call *equilibrium* (maximum uncertainty of the consequent given that the antecedent is true). We study the properties of this new index and show in which cases it is more useful than a measure of deviation from independence.

Keywords: Data mining, Association rules, Interestingness measures, Statistical significance, Deviation from equilibrium.

1 Introduction

Among the knowledge models used in Knowledge Discovery in Databases (KDD), association rules [Agrawal *et al.*, 1993] have become a major concept and have received significant research attention. Association rules are implicative tendencies $a \rightarrow b$ where a and b are conjunctions of items (boolean variables of the form $databaseAttribute = value$). Such a rule means that if a record verifies the antecedent a in the database then it certainly verifies the consequent b.

A crucial step in association rule discovery is post-processing, i.e. the interpretation, evaluation, and validation of the rules in order to find interesting knowledge for decision-making. Indeed, due to their unsupervised nature, the data mining algorithms can produce a great many rules, many of which have no interest. To help the user (a decision-maker specialized in the data studied) to find relevant knowledge in this mass of information, one of the main solutions consists in evaluating and sorting the rules with interestingness measures. There are two kinds of measures: the subjective (user-oriented) ones and the objective (data-oriented) ones. Subjective measures take into account the user's goals and user's *a priori* knowledge of the data [Liu *et al.*, 2000] [Padmanabhan and Tuzhilin, 1999] [Silberschatz and

Tuzhilin, 1996]. On the other hand, only the data cardinalities appear in the calculation of objective measures [Tan *et al.*, 2004] [Bayardo and Agrawal, 1999] [Guillet, 2004] [Lenca *et al.*, 2004] [Lallich and Teytaud, 2004]. In this article, we are interested in the objective measures.

We have shown in [Blanchard *et al.*, 2004] that there exist two different but complementary aspects of the rule interestingness: the deviation from independence and the deviation from what we call *equilibrium* (maximum uncertainty of the consequent given that the antecedent is true). Thus, the objective measures of interestingness are divided into two groups:

- the measures of deviation from independence, which have a fixed value when the variables a and b are independent $(n.n_{ab} = n_a n_b)$[1];
- the measures of deviation from equilibrium, which have a fixed value when examples and counter-examples are equal in numbers $(n_{ab} = n_{a\bar{b}} = \frac{1}{2}n_a)$.

The objective measures can also be classified according to their descriptive or statistical nature [Lallich and Teytaud, 2004] [Gras *et al.*, 2004]:

- The descriptive (or frequential) measures are those which do not vary with the cardinality expansion (when all the data cardinalities are increased or decreased in equal proportion).
- The statistical measures are those which vary with the cardinality expansion. Among them, one can find the probabilistic measures, which compare the observed distribution to an expected distribution, such as the implication intensity [Gras, 1996] [Blanchard *et al.*, 2003b] or the likelihood linkage index [Lerman, 1991].

	Measures of deviation from equilibrium	**Measures of deviation from independence**
Descriptive measures	– confidence, – Sebag et Schoenauer index, – example and counter-example ratio, – Ganascia index, – *moindre-contradiction*, – inclusion index...	– correlation coefficient, – lift, – Loevinger index, – conviction, – J-measure, – *TIC*, – odds ratio, – *multiplicateur de cote...*
Statistical measures		– implication intensity, – implication index, – likelihood linkage index, – oriented contribution to χ^2, – rule-interest...

Table 1. Classification of the objective measures of rule interestingness

With these two criteria, we classify the objective measures of rule interestingness into four categories. As shown in table 1 (cf. [Guillet, 2004] for

[1] The notations are defined in section 2

the references), there are no statistical measures which evaluate the deviation from equilibrium. Nevertheless, the statistical measures have the advantage of taking into account the size of the phenomena studied. Indeed a rule is statistically all the more reliable since it is assessed on a large amount of data. Moreover, when based on a probabilistic model, a statistical measure refers to an intelligible scale of values (a scale of probabilities); this is not the case for many interestingness measures. Also, such a measure facilitates the choice of a threshold for filtering the rules, since the complement to 1 of the threshold has the meaning of the significance level of a hypothesis test (generally in a test, one chooses $\alpha \in \{0.1\%, 1\%, 5\%\}$).

In this article, we propose a new measure of rule interestingness which evaluates the deviation from equilibrium while having a statistical nature. More precisely, this index is based on a probabilistic model and measures the statistical significance of the deviation from equilibrium (whereas implication intensity or likelihood linkage index, for example, measure the statistical significance of the deviation from independence). In the next section, we present a probabilistic index of deviation from equilibrium named *IPEE*, and then study in section 3 its properties. Section 4 is devoted to the comparison between the measures of deviation from equilibrium and the measures of deviation from independence.

2 Measuring the statistical significance of the deviation from equilibrium

We consider a set O of n objects described by boolean variables. In the association rule terminology, the objects are transactions stored in a database, the variables are called items, and the conjunctions of variables are called itemsets. Given an itemset a, we note A the set of the objects which verify a, and n_a the cardinality of A. The complement of A in O is the set $\overline{A}$ of cardinality $n_{\overline{a}}$. An association rule is a couple (a, b) noted $a \rightarrow b$ where a and b are two itemsets which have no items in common. The rule examples are the objects which verify the antecedent a and the consequent b (objects in $A \cap B$), while the rule counter-examples are the objects which verify a but not b (objects in $A \cap \overline{B}$). In the following, we call "variables" the itemsets.

2.1 Random model

Given a rule $a \rightarrow b$, we want to measure the statistical significance of the rule deviation from equilibrium. As the equilibrium configuration is defined by the equidistribution in A of examples $A \cap B$ and counter-examples $A \cap \overline{B}$, the null hypothesis is the hypothesis H_0 of equiprobability between the examples and counter-examples. So, let us associate to the set A a random set X of cardinality n_a drawn in O under this hypothesis: $\mathrm{P}(X \cap B) = \mathrm{P}(X \cap \overline{B})$ (cf. figure 1). The number of counter-examples expected under H_0 is the

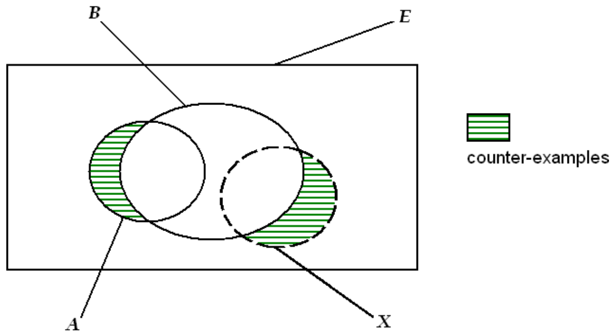


Fig. 1. Random draw of a set X under the equiprobability hypothesis between the examples and counter-examples

cardinality of $X \cap \overline{B}$, noted $\left|X \cap \overline{B}\right|$. It is a random variable whose $n_{a\overline{b}}$ is an observed value. The rule $a \rightarrow b$ is even better since there is a high probability that chance creates more counter-examples than data.

Définition 1 The **probabilistic index of deviation from equilibrium (*IPEE*^{2})** of a rule $a \rightarrow b$ is defined by:

$$IPEE(a \rightarrow b) = \mathrm{P}(\left|X \cap \overline{B}\right| > n_{a\overline{b}} \ \mid \mathrm{H}_0)$$

A rule $a \rightarrow b$ is said to be acceptable with the confidence level $1 - \alpha$ if $\delta(a \rightarrow b) \geq 1 - \alpha$.

Therefore, *IPEE* quantifies the unlikelihood of the smallness of the number of counter-examples $n_{a\overline{b}}$ with respect to the hypothesis H_0. In particular, if $\delta(a \rightarrow b)$ is close to 1 then it is unlikely that the features $(a \textit{ and } b)$ and $(a \textit{ and } \overline{b})$ are equiprobable. This new index can be seen as the complement to 1 of the p-value of a hypothesis test (and α as the significance level of this test). However, following the implication intensity and the likelihood linkage index (where H_0 is the hypothesis of independence between a and b), the aim is not testing a hypothesis but actually using it as a reference to evaluate and sort the rules.

2.2 Analytical expression

In the case of drawing random sets with replacement, $\left|X \cap \overline{B}\right|$ is binomial with parameters n_a and $\frac{1}{2}$:

$$\delta(a \rightarrow b) = 1 - \frac{1}{2^{n_a}} \sum_{k=0}^{n_{a\overline{b}}} \binom{n_a}{k}$$

[2] *IPEE* is for *Indice Probabiliste d'Ecart à l'Equilibre* in French

IPEE depends neither on n_b (it does not increase with the rarity of the consequent), nor on n since the equilibrium hypothesis H_0 is not defined by means of n_b and n (contrary to the independence hypothesis). It must be noticed that the statistical significance of the deviation from equilibrium could be measured by comparing not the counter-examples but the examples: $\widehat{IPEE}(a \to b) = \mathrm{P}(|X \cap B| < n_{ab} \mid \mathrm{H}_0)$. However, since the binomial distributions with parameter $\frac{1}{2}$ are symmetrical, the two indexes are identical:

$$IPEE(a \to b) = 1 - \frac{1}{2^{n_a}} \sum_{K=n_{ab}}^{n_a} \binom{n_a}{n_a - K} = 1 - \frac{1}{2^{n_a}} \sum_{K=n_{ab}}^{n_a} \binom{n_a}{K} = \widehat{IPEE}(a \to b)$$

where $K = n_a - k$.

When $n_a \geq 10$, the binomial distribution can be approximated by the normal distribution with mean $\frac{n_a}{2}$ and standard deviation $\sqrt{\frac{n_a}{4}}$. The standardized number of counter-examples $\widetilde{n}_{a\bar{b}}$ can be interpreted as the oriented contribution to the χ^2 of goodness-of-fit between the observed distribution of examples/counter-examples, and the uniform distribution: $\chi^2 = \widetilde{n}_{a\bar{b}}{}^2$ This constitutes a strong analogy with the implication intensity and the likelihood linkage index, since in the poissonian models of these two measures, the standardized values of $n_{a\bar{b}}$ and n_{ab} can be seen as oriented contributions to the χ^2 of independence between a and b [Lerman, 1991].

3 *IPEE* properties

Range	$[0; 1]$
Value for logical rules	$1 - \frac{1}{2^{n_a}}$
Value for equilibrium	0.5
Variation w.r.t. $n_{a\bar{b}}$ with fixed n_a	↘
Variation w.r.t. n_a with fixed $n_{a\bar{b}}$	↗

Table 2. *IPEE* properties

The properties and the graph of *IPEE* are given respectively in table 2 and figure 2. We can observe that :

- *IPEE* varies slightly with the first counter-examples (slow decrease). This behavior is intuitively satisfactory since a small number of counter-examples do not question a rule [Gras *et al.*, 2004].
- The discarding of the rules gets quicker in an uncertainty range around the equilibrium $n_{a\bar{b}} = \frac{n_a}{2}$ (fast decrease).

As shown in figure 3, with a ratio examples/counter-examples which is constant, the values of *IPEE* are all the more extreme (close to 0 or 1) since n_a is large. Indeed, owing to its statistical nature, the measure takes into account the size of the phenomena studied: the larger n_a is, the more one can trust the imbalance between examples and counter-examples observed in the data, and the more one can confirm the good or bad quality of the rule deviation from equilibrium. In particular, for *IPEE*, the quality of a logical rule (rule with no counter-examples, i.e. $n_{a\overline{b}} = 0$) depends on n_a (cf. table 2). Thus, contrary to the other measures of deviation from equilibrium (cf. table 1), *IPEE* has the advantage of systematically attributing the same value to the logical rules. This allows to differentiate and sort the logical rules.

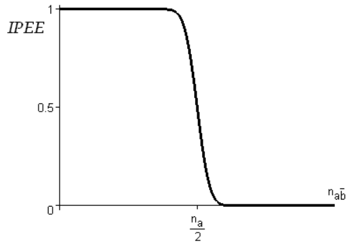


Fig. 2. Plot of *IPEE* w.r.t. the number of counter-examples $n_{a\overline{b}}$

It must be noticed that *IPEE* has no symmetry: it does not assign the same value to a rule $a \longrightarrow b$ and to its converse $b \longrightarrow a$, or to its contrapositive $\overline{b} \longrightarrow \overline{a}$, or to its opposite $a \longrightarrow \overline{b}$. Nevertheless, we have the following relation: $\delta(a \longrightarrow \overline{b}) = 1 - \delta(a \longrightarrow b) - \frac{C_{n_a}^{n_{ab}}}{2^{n_a}}$ (the last term is negligible when n_a is large).

We have seen that the strength of statistical significance measures lies in the fact that they take into account the size of the phenomena studied. On the other hand, it is also their main limit: the measures have a low discriminating power when the size of the phenomena is large (beyond around 10^4) [Elder and Pregibon, 1996]. Indeed, with regard to large cardinalities, even minor deviations can be statistically significant. *IPEE* does not depart from this: when n_a is large, the measure tends to evaluate the rules as either very good (values close to 1), or very bad (values close to 0). In this case, to fine-tune the filtering of the best rules, it is necessary to use a descriptive measure (cf. table 1) such as the inclusion index [Blanchard *et al.*, 2003b] in addition to *IPEE*. On the other hand, contrary to the implication intensity or the likelihood linkage index, *IPEE* does not depend on n. Therefore, the measure is sensitive to both the specific rules ("nuggets") and the general rules ; it can be used on either small or large databases.

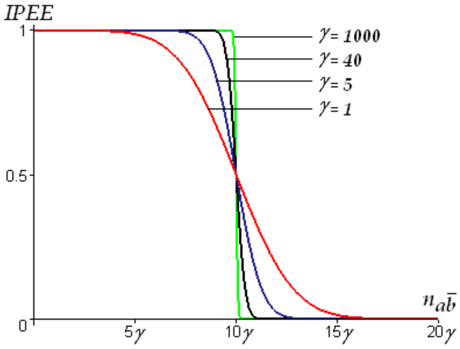


Fig. 3. Plot of *IPEE* w.r.t. cardinality expansion
($n_a = 20 \times \gamma,\ n_{a\bar{b}} \in [0 \times \gamma\ ;\ 20 \times \gamma],\ \gamma \in \{1;\ 5;\ 40;\ 1000\}$)

4 Measures of deviation from equilibrium and independence: a comparison

Let us consider a rule with the cardinalities $n_{a\bar{b}},\ n_a,\ n_b,\ n$. By varying $n_{a\bar{b}}$ with fixed n_a, n_b, and n, one can distinguish two different cases [Blanchard *et al.*, 2004] :

- If $n_b \geq \frac{n}{2}$ (case 1), then $\frac{n_a n_{\bar{b}}}{n} \leq \frac{n_a}{2}$, so the rule goes through the independence before going through the equilibrium when $n_{a\bar{b}}$ increases.
- If $n_b \leq \frac{n}{2}$ (case 2), then $\frac{n_a n_{\bar{b}}}{n} \geq \frac{n_a}{2}$, so the rule goes through the equilibrium before going through the independence when $n_{a\bar{b}}$ increases.

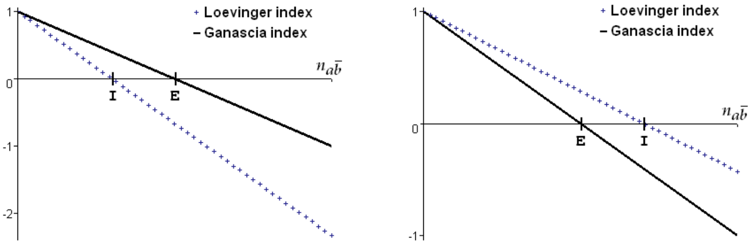
Let us now compare a measure of deviation from equilibrium M_{eql} and a measure of deviation from independence M_{idp} for these two cases. In order to have a fair comparison, we suppose that the two measures have similar behaviors:

- same value for a logical rule,
- same value for equilibrium/independence,
- same decrease speed with regard to the counter-examples.

For example, M_{eql} and M_{idp} can be the Ganascia and Loevinger indexes [Ganascia, 1991] [Loevinger, 1947] (cf. the definitions in table 3), or *IPEE* and the implication intensity. As shown in figures 4 and 5, M_{idp} is more filtering than M_{eql} in case 1, whereas M_{eql} is more filtering than M_{idp} in case 2. In other words, in case 1, it is M_{idp} which contributes to rejecting the bad rules, while in case 2 it is M_{eql}. This confirms that the measures of deviation from equilibrium and the measures of deviation from independence have to be regarded as complementary, the second ones not being systematically "better" than the first ones. In particular, the measures of deviation from equilibrium must not be neglected when the realizations of the studied variables are rare. Indeed, in this situation, should the user not take an interest in the rules having non-realizations (which is confirmed in practice), case 2 is more frequent than case 1.

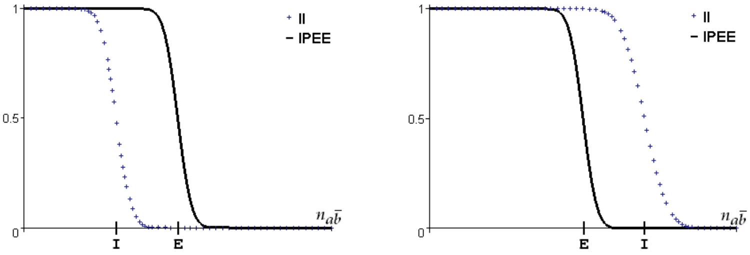
Ganascia index	**Loevinger index**
$\frac{2n_{ab}-n_a}{n_a}$	$1-\frac{nn_{a\overline{b}}}{n_a n_{\overline{b}}}$

Table 3. Ganascia and Loevinger indexes for a rule $a \longrightarrow b$



(a) case 1 ($n_b \geq \frac{n}{2}$) (b) case 2 ($n_b \leq \frac{n}{2}$)

Fig. 4. Comparison of the Ganascia and Loevinger indexes (E: equilibrium, I: independence)



(a) case 1 ($n_b \geq \frac{n}{2}$) (b) case 2 ($n_b \leq \frac{n}{2}$)

Fig. 5. Comparison of the measures *IPEE* and implication intensity (*II*)

5 Conclusion

In this article, we have presented a new measure of rule interestingness which evaluates the deviation from equilibrium with respect to a probabilistic model. Due to its statistical nature, this measure has the advantage of taking into account the size of the phenomena studied, contrary to the other measures of deviation from equilibrium. Moreover, it refers to an intelligible scale of values (a scale of probabilities). Our study shows that *IPEE* is efficient to assess logical rules, and well adapted to the search for specific rules ("nuggets").

IPEE can be seen as the counterpart of the implication intensity [Gras, 1996] [Blanchard *et al.*, 2003b] for the deviation from equilibrium. Used together, these two measures allow an exhaustive statistical evaluation of the rules. To continue this research work, we will integrate *IPEE* into our rule validation system *ARVis* [Blanchard *et al.*, 2003a] in order to experiment with the couple (*IPEE*, implication intensity) on real data.

References

[Agrawal *et al.*, 1993]Rakesh Agrawal, Tomasz Imielienski, and Arun Swami. Mining association rules between sets of items in large databases. In *Proceedings of the 1993 ACM SIGMOD international conference on management of data*, pages 207–216. ACM Press, 1993.

[Bayardo and Agrawal, 1999]Roberto J. Bayardo and Rakesh Agrawal. Mining the most interesting rules. In *Proceedings of the fifth ACM SIGKDD international conference on knowledge discovery and data mining*, pages 145–154. ACM Press, 1999.

[Blanchard *et al.*, 2003a]Julien Blanchard, Fabrice Guillet, and Henri Briand. A user-driven and quality-oriented visualization for mining association rules. In *Proceedings of the third IEEE international conference on data mining ICDM'03*, pages 493–496. IEEE Computer Society, 2003.

[Blanchard *et al.*, 2003b]Julien Blanchard, Pascale Kuntz, Fabrice Guillet, and Régis Gras. Implication intensity: from the basic statistical definition to the entropic version. In Hamparsum Bozdogan, editor, *Statistical Data Mining and Knowledge Discovery*, pages 473–485. Chapman and Hall/CRC Press, 2003. chapter 28.

[Blanchard *et al.*, 2004]Julien Blanchard, Fabrice Guillet, Régis Gras, and Henri Briand. Mesurer la qualité des règles et de leurs contraposées avec le taux informationnel tic. *Revue des Nouvelles Technologies de l'Information*, E-2:287–298, 2004. Actes des journées Extraction et Gestion des Connaissances (EGC) 2004.

[Elder and Pregibon, 1996]John F. Elder and Daryl Pregibon. A statistical perspective on knowledge discovery in databases. In Usama M. Fayyad, Gregory Piatetsky-Shapiro, Padhraic Smyth, and Ramasamy Uthurusamy, editors, *Advances in knowledge discovery and data mining*, pages 83–113. AAAI/MIT Press, 1996.

[Ganascia, 1991]J.-G. Ganascia. Charade : apprentissage de bases de connaissances. In Y. Kodratoff and E. Diday, editors, *Induction symbolique et numérique à partir de données*, pages 309–326. Cépaduès Editions, 1991.

[Gras *et al.*, 2004]Régis Gras, Raphaël Couturier, Julien Blanchard, Henri Briand, Pascale Kuntz, and Philippe Peter. Quelques critères pour une mesure de qualité de règles d'association. *Revue des Nouvelles Technologies de l'Information*, E-1:3–31, 2004. numéro spécial Mesures de qualité pour la fouille de données.

[Gras, 1996]Régis Gras. *L'implication statistique : nouvelle méthode exploratoire de données.* La Pensée Sauvage Editions, 1996.

[Guillet, 2004]Fabrice Guillet. Mesures de la qualité des connaissances en ecd, 2004. Tutoriel des journées Extraction et Gestion des Connaissances (EGC) 2004, www.isima.fr/˜egc2004/Cours/Tutoriel-EGC2004.pdf.

[Lallich and Teytaud, 2004]Stéphane Lallich and Olivier Teytaud. Evaluation et validation de l'intérêt des règles d'association. *Revue des Nouvelles Technologies de l'Information*, E-1:193–218, 2004. numéro spécial Mesures de qualité pour la fouille de données.

[Lenca *et al.*, 2004]Philippe Lenca, Patrick Meyer, Benoît Vaillant, Philippe Picouet, and Stéphane Lallich. Evaluation et analyse multicritère des mesures de qualité des règles d'association. *Revue des Nouvelles Technologies de l'Information*, E-1:219–246, 2004. numéro spécial Mesures de qualité pour la fouille de données.

[Lerman, 1991]I.C. Lerman. Foundations in the likelihood linkage analysis classification method. *Applied Stochastic Models and Data Analysis*, 7:69–76, 1991.

[Liu *et al.*, 2000]Bing Liu, Wynne Hsu, Shu Chen, and Yiming Ma. Analyzing the subjective interestingness of association rules. *IEEE Intelligent Systems*, 15(5):47–55, 2000.

[Loevinger, 1947]J. Loevinger. A systematic approach to the construction and evaluation of tests of ability. *Psychological Monographs*, 61(4), 1947.

[Padmanabhan and Tuzhilin, 1999]Balaji Padmanabhan and Alexander Tuzhilin. Unexpectedness as a measure of interestingness in knowledge discovery. *Decision Support Systems*, 27(3):303–318, 1999.

[Silberschatz and Tuzhilin, 1996]Avi Silberschatz and Alexander Tuzhilin. What makes patterns interesting in knowledge discovery systems. *IEEE Transactions on Knowledge and Data Engineering*, 8(6):970–974, 1996.

[Tan *et al.*, 2004]Pang-Ning Tan, Vipin Kumar, and Jaideep Srivastava. Selecting the right objective measure for association analysis. *Information Systems*, 29(4):293–313, 2004.

Implicative statistical analysis applied to clustering of terms taken from a psychological text corpus

Jérôme David[1], Fabrice Guillet[1], Vincent Philippé[2], and Régis Gras[1]

[1] LINA - École Polytechnique de l'université de Nantes
La Chantrerie, BP 50609
44 306 NANTES Cedex 3, France
(e-mail:
jerome.david,fabrice.guillet,regis.gras@polytech.univ-nantes.fr)

[2] PerformanSe S.A.S.
Atlanpôle La Fleuriaye
44 470 CARQUEFOU, France
(e-mail: vincent.philippe@performanse.fr)

Abstract. In order to validate a textual base contained in a behavioural skill-testing software, we suggest a methodology which can extract subsets of characteristic terms used to describe personality traits. Our approach permits, after an automatic language processing task, to evaluate the association rules between terms and descriptors (personality traits) structuring the corpus with the help of the theory of statistic implication.

Keywords: data-mining, association rule, terminology, statistical implication analysis.

1 Introduction

Text-mining consists in finding knowledge structures in large text databases. To accomplish this work, text-mining uses some methods of data-mining such as association rule discovery ([Maedche and Staab, 2000], [Janetzko *et al.*, 2004], [Roche, 2003]).
Association rule discovery aims at finding implicative relations between boolean items. To evaluate these associations, the measures of support and confidence are commonly used, despite some deficiencies. In addition, many other measures are proposed ([Tan *et al.*, 2004], [Guillet, 2004], [Lenca *et al.*, 2004]). In this article, we are focusing on the implicative statistical analysis ([Gras, 1979], [Gras and others, 1996]) which offers measures such as implication intensity and entropic implication intensity ([Gras *et al.*, 2001]).
Nevertheless, we must perform some automatic language processing tasks in order to obtain a strutured list of terms representing the textual base. Many approaches are available : statistic approaches ([Salem, 1986]), linguistic approaches (([David and Plante, 1990], [Bourigault and Fabre, 2000], [Jacquemin, 1997]) and combined approaches of these two ([Smadja, 1993],

[Daille, 1994]).
In this paper, we suggest a methodology to associate each descriptor of an indexed corpus with a set of terms describing the descriptor studied. In other words, our method clusters terms with the help of other variables. First of all, we present the data, the problem and our general methodology. Then, we briefly introduce the principles of implication intensity. Next, we describe our method more precisely. Finally, we evaluate and analyse our results.

2 Methodology.

2.1 Analysed data and the problem

Our approach has been designed for a textual database extracted from a personality test, DIALECHO software, distribued by PERFORMANSE SAS company. This program is used by human resources managers. After a binary-choice questionnaire of 70 questions, this program provides a scored evaluation of 10 personality variables and a behavioural assessment report. The generation process of the report is perform in 2 steps : (1) discretisation on 3 modalities of the scored personality variables;(2) by parsing and selecting the annotated paragraphs by a conjunction of modalities named personality traits. Examples of traits: Extroversion/Introversion (discrete values : EXT+, EXT0, EXT-), Anxiety/Relaxation (ANX+, ANX0, ANX+).
Our corpus is a set of 12 805 documents. Each document is made of a paragraph (text) and a rule (conjunction of traits) as shown in figure 1. According to DIALECHO software, a document implies: "If a psychologic profile matches the rule (the conjunction) then the paragraph below will be included in the personality report". We have extracted and selected 6 977 terms from the paragraphs.

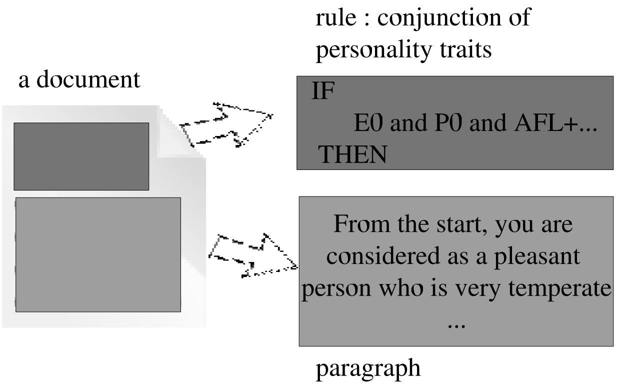

rule : conjunction of personality traits
a document
IF
E0 and P0 and AFL+...
THEN
From the start, you are considered as a pleasant person who is very temperate ...
paragraph

Fig. 1. The strucure of a document.

The finality of this approach is to enabled the author of the expert analysis to verify if the vocabulary used is in adequation with the personality traits described in the paragraphs. Our problem consists in finding for each item, the set of terms which best describes them.

First, we represent the paragraphs by a list of binary terms. These binary terms are noun phrases composed of two meaningful words. This terminological process (step 1, figure 2) is performed by ACABIT, an automatic term acquisition software program ([Daille, 1994]). Next, we add the personality traits set to the paragraph representation (step 2, figure 2). At this stage, a document is represented by a set of terms and personality traits. Then, we consider the set of association rules "term ⇒ trait" whose validity depends on their intensity of implication value (step 3, figure 2). For each distinct rule head (i.e. for each traits), we aim at all their bodies (terms) (step 4, figure 2). This last stage generates one cluster of terms by personality trait. Finally, the expert (author of the texts) evaluates the quality of the clusters (step 5, figure 2). According to this last stage, the expert can verify if the vocabulary used matches the personality traits.

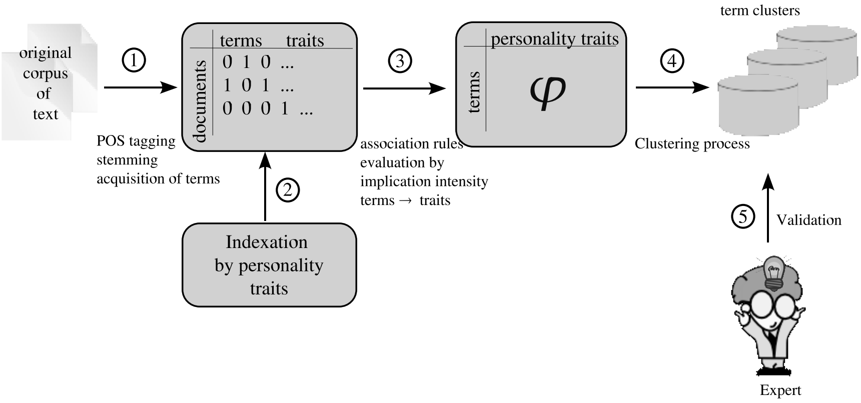


Fig. 2. Process sequence.

2.2 Rules evaluation using the implication intensity.

Association rules ([Agrawal *et al.*, 1993]) are almost like logic implications but admit some counter-examples. The quality of such rules is usually evaluated by two measures : support and confidence. Nevertheless, we intend to evaluate infrequent but interesting rules. Indeed Y. Kodratoff has said ([Kodratoff, 2001]) that "the best rules are often the least frequent". The confidence measure is not quite perfect: it cannot reject the statistical independence ([Blanchard *et al.*, 2004]).

The rule is retained for a given threshold $1 - \sigma$ if $\varphi(a \Rightarrow b) \geq 1 - \sigma$. Let us now consider a finite set T of n transactions described by a set I of p items. Each transaction t can be considered as an itemset so that $t \subseteq I$. A transaction t is said to contain an itemset a if $a \subseteq t$ and we denote by $A = \{t \in T; a \subseteq t\}$ the transaction set in T which contains a and by $\overline{A}$ its complementary set in T.

An association rule is an implication of the form $a \Rightarrow b$, where a and b are disjoined itemsets ($a \subset I$, $b \subset I$, and $a \cap b = \emptyset$). In practice, it is quite common to observe a few transactions which contain a and not b without contesting the general trend to have b when a is present. Therefore, with regards to the cardinal n of T but also to the cardinals n_A of A and n_B of B, the number $n_{A\cap\overline{B}} = card(A \cap \overline{B})$ of counter-examples must be taken into account to statistically accept to retain or not the rule $a \Rightarrow b$. Following the likelihood linkage analysis of Lerman [Lerman, 1981], the implication intensity expresses the unlikelihood of counter-examples $n_{A\cap\overline{B}}$ in T.

More precisely, we compare the observed number of counter-examples to a probabilistic model. Let us assume that we randomly draw two subsets X and Y in T which respectively contain n_A and n_B transactions. The complementary sets $\overline{Y}$ of Y and $\overline{B}$ of B in T have the same cardinality $n_{\overline{B}}$. In this case, $N_{X\cap\overline{Y}} = card(X \cap \overline{Y})$ is a random variable and $n_{A\cap\overline{B}}$ an observed value. The association rule $a \Rightarrow b$ is acceptable for a given threshold $1 - \sigma$ if σ is greater than or equal to the probability that the number of counter-examples in the observations is greater than or equal to the number of expected counter-examples in a random drawing, i.e. if $\Pr(N_{X\cap\overline{Y}} \leq n_{A\cap\overline{B}}) \leq \sigma$.

The distribution of the random variable $N_{X\cap\overline{Y}}$ depends on the drawing mode [Gras and others, 1996]. In order to explicitly take into account the asymmetry of the relationships between itemsets, we here restrict ourselves to the Poisson distribution with $\lambda = n_A n_{\overline{B}}/n$. For cases where the approximation is justified (e.g. $\lambda > 3$), the standardized random variable $\widetilde{N}_{X\cap\overline{Y}} = (card(X \cap \overline{Y}) - \lambda)/\sqrt{\lambda}$ is approximately $N(0,1)$-distributed. The observed value of $\widetilde{N}_{X\cap\overline{Y}}$ is $\widetilde{n}_{A\cap\overline{B}} = (n_{A\cap\overline{B}} - \lambda)/\sqrt{\lambda}$.

The implication intensity of the association rule $a \Rightarrow b$ is defined by $\varphi(a \Rightarrow b) = 1 - \Pr(\widetilde{N}_{X\cap\overline{Y}} \leq \widetilde{n}_{A\cap\overline{B}})$ if $n_B \neq n$; otherwise $\varphi(a \Rightarrow b) = 0$.

The rule is retained for a given threshold $1 - \sigma$ if $\varphi(a \Rightarrow b) \geq 1 - \sigma$.

3 Detailed clustering process.

We choose to define the studied database by $B = (D, T, C)$ where $D = \{d_1, .., d_m\}$ is representative of the paragraph set, $T = \{t_1, ..., t_n\}$ concerns the term set and $C = \{c_1, ..., c_y\}$ express as the item set. By asserting $A = C \cup T$, the value of an item a, for a paragraph d_x, is equal to 1 if the attribute describes the document or if not to 0. The following example (table 1) shows the values of the documents over the term set (in French): "conscience professionelle" (conscienciousness), "sens de la méthode" (rigour), "preuve de créativité" (creativity), "attrait de la nouveauté" (appeal of novelty), and the personality trait set : "Extroversion", "Medium extroversion", "Rigour", "Intellectual dynamism".

id_doc	conscience professionnelle	sens de la méthode	preuve de créativité	attrait de la nouveauté	Extroversion	Medium extroversion	Rigour	Intellectual dynamism
d1	1	1	0	0	0	1	1	0
d2	0	0	1	1	1	0	0	1

Table 1. Extract from the table representing the documents.

In order to build sets of terms which best describe personality traits, we evaluate for each term $t \in T$ and for each personality trait $c \in C$, the rule $t \Rightarrow c$. This rule means "if a paragraph holds the term t then this paragraph describes (at least) a person who has the personality trait c". To do this, we define the matrix $\mathcal{M}_\varphi$ of dimension $n \times m$ where rows denote terms, columns personality traits and whose values are $\psi_{t \Rightarrow c} = \begin{cases} \varphi(t \Rightarrow c)\, if\, \varphi(t \Rightarrow c) \geq 0 \\ 0\, if not \end{cases}$.

The following example denotes the implication intensity of the French terms ("conscience professionnelle", "sens de la méthode", ...) toward the personality traits ("Extroversion", "Medium Extroversion", "Rigour", "Intellectual dynamism").

$t \Rightarrow c$	Extroversion	Medium extroversion	Rigour	Intellectual dynamism
conscience professionnelle	0.0	0.63	0.99	0.0
sens de la méthode	0.77	0.0	0.92	0.0
preuve de créativité	0.0	0.0	0.0	0.94
attrait de la nouveauté	0.0	0.0	0.0	0.94
domaine de la communication	0.0	0.0	0.86	0.86

Table 2. Extract from the implication intensities matrix $\mathcal{M}_\varphi$.

Finally, we define the most representative term set of a personality trait with a threshold $\varphi_{threshold}$ by the following formula :
$T_x = \{t_y \mid \varphi(t_y \Rightarrow c_x) \geq \varphi_{threshold}\}$. The choice of a threshold is not easy because it depends on the database studied. We suggest to choose, firstly, $\varphi_{threshold} = 0,5$ because a rule begin to be interesting from this threshold. After, the expert could increase this value until he/she is statisfied.

4 Results.

We have tried our method over the 30 personality traits and the expert evaluated the accuracy of each set of terms. Each set is divided into two groups by the expert (decision maker): the relevant terms for the cluster studied and the others. The accuracy value is defined as the proportion of relevant terms. The following table shows some accuracy values for groups of terms generated by our process with a threshold value $\varphi_{threshold} > 0.5$.

Class	Accuracy
Rigour (CON+)	1
Combativeness (P+)	0.9
Anxiety (N+)	0.9
Intellectual dynamism (CLV+)	0.9
Assertion (EST+)	0.9
Questioning (EST-)	0.9
Motivation of power (LED+)	0.9
Motivation of protection (LED-)	0.9
Relaxation (N-)	0.8
Improvisation (CON-)	0.8
Motivation of belonging (AFL+)	0.8
Conciliation (P-)	0.7
Motivation of independence (AFL-)	0.7
Medium Anxiety (N0)	0.6
Intellectual conformism (CLV-)	0.6
Introversion (E-)	0.5
Extroversion (E+)	0.4
Medium extroversion (E0)	0

Table 3. Accuracy of the cluster.

Results show that some sets have bad accuracy. Indeed, these clusters describe personality traits which are not directly described in the text but their occurrence will modulate other traits. For example, the personality traits "E+", "E0", "E-" are not directly described in text but they are used to reinforce or moderate the expression of other personality traits. However, we obtain good accuracy values for most clusters. We have 8 good clusters, that is to say they have an accuracy value superior or equal to 90%. And we have only 3 bad clusters (accuracy value < 50%)

5 Conclusion.

In this paper, we have presented a clustering method which matches descriptors with sets of terms based on association rules between terms and descriptors. We have designed it for a psychological corpus in order to study the adequation between terms and personality traits.
This process is divided into three steps : first, we extract a selection of relevant terms from the corpus, second, we evaluate all association rules between terms and descriptors (personality traits) with the help of implication intensity, and last, we generate sets of terms from the results obtained in the second step.
Our proposal is original in the sense that, it permits to put together terms and indexation descriptors extracted from a corpus. A prototype software program has been implemented and tested on the psychological corpus with good results.
However, we do not currently consider the relationships between descriptors or between terms. We plan to study this question in the near future in order to consider taxonomies or assimilated structures. Therefore, we intend to try our method on other corpuses.

References

[Agrawal *et al.*, 1993]R. Agrawal, T. Imielinski, and A.N. Swami. Mining association rules between sets of items in large databases. In Buneman P. and Jajodia S., editors, *Proceedings of the 1993 ACM SIGMOD ICMD*, pages 207–216, 1993.

[Blanchard *et al.*, 2004]J. Blanchard, F. Guillet, R. Gras, and H. Briand. Mesurer la qualité des règles et de leur contraposées avec le taux informationnel TIC. *RNTI E-2 Extraction et gestion des connaissances*, 1:287–298, 2004.

[Bourigault and Fabre, 2000]D. Bourigault and C. Fabre. Approche linguistique pour l'analyse syntaxique de corpus. *Cahiers de Grammaires*, 25:131–151, 2000.

[Daille, 1994]B. Daille. *Approche mixte pour l'extraction automatique de terminologie : statistique lexicale et filtres linguistiques.* PhD thesis, University Paris 7, 1994.

[David and Plante, 1990]S. David and P. Plante. De la nécessité d'une approche morpho-syntaxique dans l'analyse de textes. *ICO*, 2(3):140–155, 1990.

[Gras and others, 1996]R. Gras et al. *L'implication statistique, une nouvelle méthode exploratoire de données.* La pensée sauvage, 1996.

[Gras *et al.*, 2001]R. Gras, P. Kuntz, R. Couturier, and F. Guillet. Une version entropique de l'intensité d'implication pour les corpus volumineux. *ECA Extraction et Gestion de Connaissances*, 1(1–2):69–80, 2001.

[Gras, 1979]R. Gras. Contribution à l'étude expérimentale et à l'analyse de certaines acquisitions cognitives et de certains objectifs didacticques mathématiques, 1979. Thèse d'Etat, Université de Rennes.

[Guillet, 2004]F. Guillet. Mesure de la qualité des connaissances en ecd. In *Tuturiels de la 4ème Conf. Francophone d'extraction et gestion des connaissances*, pages 1–60, Clermond-Ferrand, 2004.

[Jacquemin, 1997]C. Jacquemin. Variation terminologique : Reconnaissance et acquisition automatique de termes et de leurs variantes, 1997. Mémoire d'HDR, IRIN - Université de Nantes.

[Janetzko *et al.*, 2004]D. Janetzko, H. Cherfi, R. Kennke, A. Napoli, and Y. Toussaint. Knowledge-based selection of association rules for text mining. In *ECAI'04*, pages 485–489. IOS Press, 2004.

[Kodratoff, 2001]Y. Kodratoff. On the induction of interesting rules. *Noesis*, XXVI:103–124, 2001.

[Lenca *et al.*, 2004]P. Lenca, P. Meyer, B. Vaillant, P. Picouet, and S. Lallich. Evaluation et analyse multicritère des mesures de qualité des règles d'association. *RNTI-E-1 Mesures de qualité pour la fouille de données*, pages 219–246, 2004.

[Lerman, 1981]I.C. Lerman. *Classification et analyse ordinale des données.* Dunod, Paris, 1981.

[Maedche and Staab, 2000]A. Maedche and S. Staab. Semi-automatic engineering of ontologies from text. In KSI, editor, *the 12th Internationnal Conference SEKE*, 2000.

[Roche, 2003]M. Roche. L'extraction paramétrée de la terminologie du domaine. *RSTI Extraction et Gestion des Connaissances*, 17:295–306, 2003.

[Salem, 1986]A. Salem. Segments répétés et analyse statistique des données textuelles. Etude quantitative à propos du Père Duchesne de Hébert. *Histoire et Mesure*, 1(2):5–28, 1986.

[Smadja, 1993]F. Smadja. Retrieving collocations from text : Xtract. *Computational linguistics*, 19:143–177, 1993.

[Tan *et al.*, 2004]P.N Tan, V. Kumar, and J. Srivastava. Selecting the right objective measure for association analysis. *Inf. Syst.*, 29(4):293–313, 2004.

Preference Learning in Terminology Extraction: A ROC-based approach

Jérôme Azé, Mathieu Roche, Yves Kodratoff, and Michèle Sebag

LRI – Laboratoire de Recherche en Informatique
UMR8623, CNRS, Université Paris Sud, 91405 Orsay Cedex, France
(e-mail: {aze,roche,yk,sebag}@lri.fr)

Abstract. A key data preparation step in Text Mining, Term Extraction selects the terms, or collocation of words, attached to specific concepts. In this paper, the task of extracting relevant collocations is achieved through a supervised learning algorithm, exploiting a few collocations manually labelled as relevant/irrelevant. The candidate terms are described along 13 standard statistical criteria measures. From these examples, an evolutionary learning algorithm termed ROGER, based on the optimization of the Area under the ROC curve criterion, extracts an order on the candidate terms. The robustness of the approach is demonstrated on two real-world domain applications, considering different domains (biology and human resources) and different languages (English and French).
Keywords: Text Mining, Terminology, Evolutionary algorithms, ROC Curve.

1 Introduction

Besides the known difficulties of Data Mining, Text Mining presents specific difficulties due to the structure of natural language. In particular, the polysemy and synonymy effects are dealt with by constructing ontologies or terminologies [Bourigault and Jacquemin, 1999], structuring the words and their meanings in the domain application. A preliminary step for ontology construction is to extract the terms, or word collocations, attached to the concepts defined by the expert [Bourigault and Jacquemin, 1999, Smadja, 1993]. Term Extraction actually involves two steps: the detection of the relevant collocations, and their classification according to the concepts.

This paper focuses on the detection of relevant collocations, and presents a learning algorithm for ranking collocations with respect to their relevance, in the spirit of [Cohen *et al.*, 1999]. An evolutionary algorithm termed ROGER, based on the optimization of the Receiver Operating Characteristics (ROC) curve [Ferri *et al.*, 2002, Rosset, 2004], and already described in previous works [Sebag *et al.*, 2003a, Sebag *et al.*, 2003b], is applied to a few collocations manually labelled as relevant/irrelevant by the expert. The optimization of the ROC curve is directly related to the recall-precision tradeoff in Term Extraction (TE).

The paper is organized as follows. Section 2 briefly reviews the main criteria used in TE. Section 3 presents the ROGER (ROc-based GEnetic learneR) algorithm for the sake of self-containedness, and describes the bagging of the diverse hypotheses constructed along independent runs. Sections 4 et 5 report on the experimental validation of the approach on two real-world domain applications, and the paper ends with some perspectives for further research.

2 Measures for Term Extraction

The choice of a quality measure among the great many criteria used in Text Mining (see e.g., [Daille *et al.*, 1998, Xu *et al.*, 2002, Roche *et al.*, 2004b]) is currently viewed as a decision making process; the expert has to find the criterion most suited to his/her corpus and goals. The criteria considered in the rest of the paper are:

- Mutual Information (MI) [Church and Hanks, 1990]
- Mutual Information with cube (MI^3) [Daille *et al.*, 1998]
- Dice Coefficient ($Dice$) [Smadja *et al.*, 1996]
- Log-likelihood (L) [Dunning, 1993]
- Number of occurrences + Log-likelihood (Occ_L)[1] [Roche *et al.*, 2004a]
- Association Measure (Ass) [Jacquemin, 1997]
- Sebag-Schoenauer ($SeSc$) [Sebag and Schoenauer, 1988]
- J-measure (J) [Goodman and Smyth, 1988]
- Conviction ($Conv$) [Brin *et al.*, 1997]
- Least contradiction (LC) [Azé and Kodratoff, 2004]
- Cote multiplier (CM) [Lallich and Teytaud, 2004]
- Khi2 test used in text mining ($Khi2$) [Manning and Schütze, 1999]
- T-test used in text mining ($Ttest$) [Manning and Schütze, 1999]

Vivaldi *et al.* [Vivaldi *et al.*, 2001] have shown that the search for a quality measure can be formalized as a supervised learning problem. Considering a training set, where each candidate term is described from its value for a set of statistical criteria and labelled by the expert, they used Adaboost [Schapire, 1999] to automatically construct a classifier.

The approach presented in next section mostly differs from [Vivaldi *et al.*, 2001] as it learns an ordering function (term t_1 is more relevant than term t_2) instead of a boolean function (term t is relevant/irrelevant).

3 Learning ranking functions

This section first briefly recalls the ROGER (*ROc-based GEnetic learneR*) algorithm, used for learning a ranking hypothesis and first described in [Sebag *et al.*, 2003b, Sebag *et al.*, 2003a]. The N'ROGER variant used in this paper involves two extensions: i) the use of non-linear ranking hypotheses; ii) the

[1] Occ_L is defined by ranking collocations according to their number of occurrences, and breaking the ties based on the term Log-likelihood.

exploitation of the ensemble of hypotheses learned along independent runs of ROGER. Using the standard notations, the dataset $\mathcal{E} = \{(\mathbf{x}_i, y_i), i = 1..n, \mathbf{x}_i \in \mathbb{R}^d, y_i \in \{-1, +1\}\}$ includes n collocation examples, where each collocation $\mathbf{x}_i$ is described by the value of d statistical criteria, and its label y_i denotes whether collocation $\mathbf{x}_i$ is relevant.

3.1 ROGER

The learning criterion used in ROGER is the Wilcoxon rank test, measuring the probability that a hypothesis h ranks $\mathbf{x}_i$ higher than $\mathbf{x}_j$ when $\mathbf{x}_i$ is a positive and $\mathbf{x}_j$ is a negative example:

$$\mathcal{W}(h) = Pr(h(x_i) > h(x_j) \mid y_i > y_j) \tag{1}$$

This criterion, with quadratic complexity in the number n of examples[2] offers an increased stability compared to the misclassification rate ($Pr(h(x_i).y_i > 0$), with linear complexity in n); see [Rosset, 2004] and references therein. The Wilcoxon rank test is equivalent to the area under the ROC (Receiver Operating Characteristics) curve [Jin *et al.*, 2003]. This curve, intensively used in medical data analysis, shows the trade-off between the true positive rate (the fraction of positive examples that are correctly classified, aka recall) and the false positive rate (the fraction of negative examples that are misclassified) achieved by a given hypothesis/classifier/learning algorithm. Therefore, the area under the ROC curve (AUC) does not depend on the imbalance of the training set [Kolcz *et al.*, 2003], as opposed to other measures such as Fscore [Caruana and Niculescu-Mizil, 2004]. The ROC curve also shows the misclassification rates achieved depending on the error cost coefficients [Domingos, 1999]. For these reasons, [Bradley, 1997] argues the comparison of the ROC curves attached to two learning algorithms to be more fair and informative, than comparing their misclassification rates only. Accordingly, the area under the ROC curve defines a new learning criterion, used e.g. for the evolutionary optimization of neural nets [Fogel *et al.*, 1995], or the greedy search of decision trees [Ferri *et al.*, 2002].

In an earlier step [Sebag *et al.*, 2003b], the search space $\mathcal{H}$ considered is that of linear hypotheses ($\mathcal{H} = \mathbb{R}^d$). To each vector w in $\mathbb{R}^d$ is attached hypothesis h_w with $h_w(x) =< w, x >$, where $< w, x >$ denotes the scalar product of w and x. Hypothesis h defines an order on $\mathbb{R}^d$, which is evaluated from the Wilcoxon rank test on the training set $\mathcal{E}$ (Eq. 1), measured after cross-validation.

The combinatorial optimization problem defined by Eq. 1, thus mapped onto a numerical optimization problem, is tackled by Evolution Strategies (ES). ES are the Evolutionary Computation algorithms that are best suited to parameter optimization; the interested reader is referred to [Bäck, 1995]

[2] Actually, the computational complexity is in $\mathcal{O}(n \log n)$ since $\mathcal{W}(h)$ is proportional to the sum of ranks of the positive examples.

for an extensive presentation. In the rest of the paper, ROGER employs a $(\mu + \lambda)$-ES, involving the generation of λ offspring from μ parents through uniform crossover and self-adaptive mutation, and deterministically selecting the next μ parents from the best μ parents + λ offspring.

3.2 Extensions

An extension first presented in [Jong *et al.*, 2004] concerns the use of non-linear hypotheses. Exploiting the flexibility of Evolutionary Computation, the search space $\mathcal{H}$ is set to $\mathbb{R}^d \times \mathbb{R}^d$; each hypothesis $h = (w, c)$, composed of a weight vector w and a center c, associates to x the weighted L_1-distance of x and c:

$$h(x = (x_1, ..., x_d)) = \sum_{i=1}^{d} w_i |x_i - c_i|$$

It must be noted that this representation allows ROGER for searching (a limited kind of) non linear hypotheses, by (only) doubling the size of the linear search space. Previous work has shown that non-linear ROGER significantly outperforms linear ROGER for some text mining applications [Roche *et al.*, 2004a].

A new extension, inspired from ensemble learning [Breiman, 1998], exploits the hypotheses $h_1, \ldots, h_T$ learned along T independent runs of ROGER. The aggregation of the (normalised) h_i, referred to as H, associates to each example x the median value of $\{h_1(x), \ldots, h_T(x)\}$.

4 Goals of Experiments and Experimental Setting

The goal of experiments is twofold. On one hand, the ranking efficiency of N'ROGER will be assessed and compared to that of state-of-the-art supervised learning algorithms, specifically Support Vector Machines with linear, quadratic and Gaussian kernels, using SVMTorch implementation[3] with default options. Due to space limitations, only ensemble-based non-linear ROGER, termed N'ROGER, will be considered.

On the other hand, the results provided by N'ROGER will be interpreted and discussed with respect to their intelligibility. The experimental setting is as follows. An experiment is a 5-fold stratified cross-validation process; on each fold, i) SVM learns a hypothesis h_{SVM}; ii) ROGER is launched 21 times, and the bagging of the 21 learned hypotheses constitutes the hypothesis $h_{n'Roger}$ learned by N'ROGER; iii) both hypotheses are evaluated on the fold test set and the associated ROC curve (True Positive Rate *vs* False Positive Rate) is constructed. The AUC curves are averaged over the 5 folds.

[3] http://www.idiap.ch/machine_learning.php?content=Torch/en_OldSVMTorch.txt

The overall results reported in the next section are averaged over 10 experiments (10 different splits of the dataset into 5 folds).

The ROGER parameters are as follows: $\mu = 20$; $\lambda = 100$; the self adaptative mutation rate is 1.; the uniform crossover rate is .6.

5 Empirical validation

After describing the datasets, this section reports on the comparative performances of the algorithms, and inspects the results actually provided by N'ROGER.

5.1 Datasets

In both domains, the data preparation step [Roche *et al.*, 2004b] allows for categorizing the word collocations depending on the grammatical tag of the words (e.g. Adjective, Noun).

A first corpus related to Molecular Biology involves 6119 paper abstracts in English (9,4 Mo) gathered from queries on Medline[4]. The 1028 Noun-Noun collocations occurring more than 4 times are labelled by the expert; the dataset includes a huge majority of relevant collocations (Table 1).

A second corpus related to Curriculum Vitae[5] involves 582 CVs in French (952 Ko). The "Frequent CV" dataset includes the 376 Noun-Adjective collocations with at least 3 occurrences (two hours labelling required), with a huge majority of relevant collocations. The "Infrequent CV" dataset includes the 2822 Noun-Adjective collocations occurring once or twice (two days labelling required), with a significantly different distribution of relevant/irrelevant collocations (Table 1). Examples of relevant *vs* irrelevant collocations are respectively *compétences informatiques* and *euros annuels*;

although both collocations make sense, only the first one conveys useful information for the management of human resources.

Collocations	# collocations	Relevant	Irrelevant
Molecular Biology	1028	90.9%	9.1%
CV, Frequent collocations	376	85.7%	14.3%
CV, Infrequent collocations	2822	56.6%	43.4%

Table 1. Relevant and irrelevant collocations.

5.2 Ranking accuracy

After the experimental setting described in section 4, Table 2 compares the average AUC achieved for N'ROGER and SVMTorch with linear, Gaussian

[4] http://www.ncbi.nlm.nih.gov/entrez/query.fcgi

[5] Courtesy of the VediorBis Foundation.

and quadratic kernels. On these domain applications, both supervised learning approaches significantly improve on the statistical criteria standalone (Table 3). Further, N'ROGER improves significantly on SVM using any kernel, excepted on the *Infrequent CV* dataset. A tentative interpretation for this result is based on the fact that this dataset is the most balanced one; SVM has some difficulties to cope with imbalanced datasets.

Corpus	N'ROGER	SVM ($\sim$ 1.5s/fold)		
	($\sim$ 17s/fold)	Linear	Gaussian	Quadratic
Molecular Biology (MB)	0.73 ± 0.05	0.50 ± 0.08	0.46 ± 0.08	0.59 ± 0.08
Frequent CV (F-CV)	0.64 ± 0.08	0.48 ± 0.08	0.48 ± 0.08	0.50 ± 0.10
Infrequent CV (I-CV)	0.73 ± 0.01	0.72 ± 0.01	0.72 ± 0.02	0.71 ± 0.02

Table 2. Ranking accuracy (Area under the ROC curve) of learning algorithms.

Corpus	MI	MI^3	$Dice$	L	Occ_L	Ass	J	$Conv$	$SeSc$	CM	LC	$Ttest$	$Khi2$
MB	0.30	0.35	0.31	0.42	0.57	0.31	**0.59**	0.35	0.43	0.31	0.46	0.31	0.31
F-CV	0.31	0.40	0.39	0.43	**0.58**	0.32	**0.58**	0.39	0.40	0.31	0.44	0.36	0.36
I-CV	0.29	0.30	0.33	0.30	0.37	0.29	**0.50**	0.40	0.39	0.30	0.45	0.30	0.30

Table 3. Ranking accuracy (Area under the ROC curve) of statistical criteria.

A more detailed picture is provided by Fig. 1, showing the ROC curve associated to SVM, N'ROGER and the Occ_L and J measures on the *Frequent CV* dataset on a representative fold (termed *RF* in this paper). Interestingly, the major differences between N'ROGER and the other measures are seen at the beginning of the curve, i.e. they concern the top ranked collocations. Typically, a recall (True Positive Rate) of 50% is obtained for 18% false positive with N'ROGER, against 23% with Occ_L, 31% with J measures and 68% for quadratic SVM[6].

In summary, N'ROGER improves the accuracy of the top-ranked collocations, and therefore the satisfaction and productivity of the expert if he/she only examines the top results. A proof of principle of the generality of the approach has been presented in [Roche *et al.*, 2004b], as the ranking function learned from one corpus, in one language, was found to outperform the standard statistical criteria when applied on the other corpus, in another language.

5.3 Analysis of a ranking function

As shown in [Jong *et al.*, 2004], the weights associated to distinct features by ROGER can provide some insights into the relevance of the features. Accordingly, the hypotheses constructed by N'ROGER are examined.

Fig. 2 displays the weights and center coordinates of all 13 features (section 2) for a representative ROGER hypothesis h (closest to the ensemble

[6] SVM ROC Curves is not significant as its AUC is lower than .5 on this test fold.

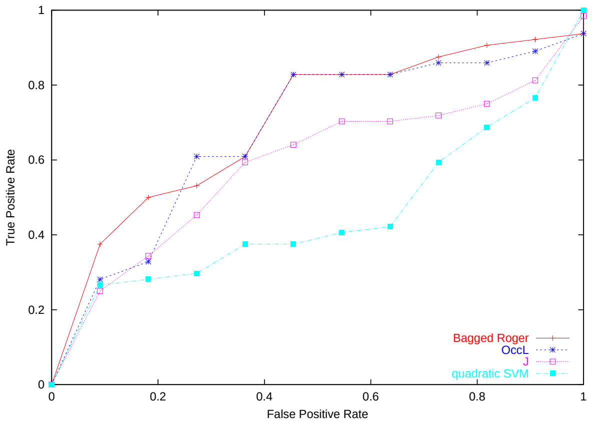


Fig. 1. ROC Curves on Frequent Collocations of CV corpus (for the test set of RF).

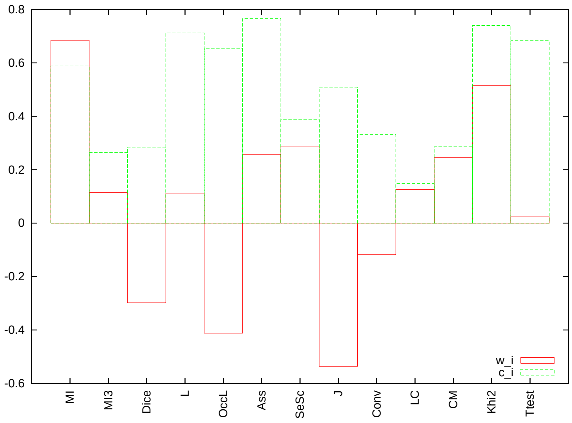


Fig. 2. Weights (w_j, c_j) on the frequent CVs (for the learn set of RF).

N'Roger hypothesis H) learned on a fold of the *Frequent CV* dataset. Although AUC(h) is lower than that of H (.61 *vs* .64), it still outpasses that of standalone features (statistical criteria).

As could have been expected, Roger detects that the mutual information (MI) criterion does badly (AUC(MI)= .31, Table 3), with a high center c_{MI} and weight w_{MI} values (collocations with high MI are less relevant, everything else being equal). Inversely, as the Occ_L criterion does well (AUC(Occ_L) $=$ .58), the center c_{Occ_L} is high associated with a highly negative weight w_{Occ_L} (collocations with low Occ_L are less relevant, everything else being equal) (see Tab. 4).

Although these tendencies could have been exploited by linear hypotheses, this is no longer the case for the J criterion (AUC(J) $=$ .58): interestingly, the center c_J takes on a medium value, with a high negative weight w_J. This is interpreted as collocations with either too low *or too high* values of J, are less relevant everything else being equal. The current limitation of the approach is to provide a “conjunctive” description of the region of relevant collocations[7].

	MI	Occ_L	N'ROGER
	w_{MI} **= 0.68**	w_{Occ_L} **= -0.41**	
	c_{MI} **= 0.59**	c_{Occ_L} **= 0.65**	
Collocation	**Rank**	**Rank**	**Rank**
expérience commerciale	297	258	1
formation informatique	300	123	2
société informatique	298	299	3
gestion informatique	299	76	4
colonne morris	1	211	90
bouygue telecom	2	213	298
fromagerie riches-mont	3	212	297
sauveteur secouriste	4	151	296
expérience professionelle	146	1	300
ressource humaine	44	2	299
baccalauréat professionel	193	3	22
baccalauréat scientifique	148	4	58

Table 4. Rank of relevant collocations given with 2 measures (MI and Occ_L) and N'ROGER. For each measure the weights (w_i, c_i) used by N'ROGER are given (on the learn set of RF).

6 Discussion and Perspectives

The main claim of the paper is that supervised learning can significantly contribute to the Term Extraction task in Text Mining. Some empirical evidence supporting this claim have been presented, related to two corpora with different domain applications and languages. Based on a domain- and language-independent description of the collocations along a set of standard statistical criteria, and on a few collocations manually labelled as relevant/irrelevant by the expert, a ranking hypothesis is learned.

The ranking learner N'ROGER used in the experiments is based on the optimization of the combinatorial Wilcoxon rank test criterion, using an evolutionary computation algorithm. Two new features, the use of non-linear hypotheses and the exploitation of the ensemble of hypotheses learned along independent runs of ROGER, have been exploited in N'ROGER.

Further research is concerned with enriching the description of collocations, e.g. adding typography-related indications (e.g. distance to the closest typographic signs) or distance to the closest Noun, possibly providing additional cues on the role of relevant collocations. Another perspective is to

[7] In the sense that a single center c is considered, though the condition *far from* c_i actually corresponds to a disjunction.

extend ROGER using multi-modal and multi-objective evolutionary optimization [Deb, 2001], e.g. enabling to characterize several types of relevant collocations in a single run. A long-term goal is to study along a variety of domain applications and expert goals, the eventual regularities associated to i) the (domain and language independent) description of the relevant collocations; ii) the ranking hypotheses.

Acknowledgment: We thank Oriane Matte-Tailliez for her expertise and labelling of the Molecular Biology dataset and Mary Felkin who did her best to improve the readability of this paper. The authors are partially supported by the PASCAL Network of Excellence, IST-2002-506778.

References

[Azé and Kodratoff, 2004]J. Azé and Y. Kodratoff. Extraction de "pépites" de connaissance dans les données : une nouvelle approche et une étude de la sensibilité au bruit. *Revue RNTI*, E-1:247–270, 2004.

[Bäck, 1995]T. Bäck. *Evolutionary Algorithms in theory and practice.* New-York:Oxford University Press, 1995.

[Bourigault and Jacquemin, 1999]D. Bourigault and C. Jacquemin. Term extraction + term clustering: An integrated platform for computer-aided terminology. In *Proc. of EACL'99, Bergen.*, pages 15–22, 1999.

[Bradley, 1997]A.P. Bradley. The use of the area under the roc curve in the evaluation of machine learning algorithms. *Pattern Recognition*, 30(7):1145–1159, 1997.

[Breiman, 1998]L. Breiman. Arcing classifiers. *Annals of Statistics*, 26(3):801–845, 1998.

[Brin *et al.*, 1997]S. Brin, R. Motwani, and C. Silverstein. Beyond market baskets: generalizing association rules to correlations. In *Proc. of ACM SIGMOD'97*, pages 265–276, 1997.

[Caruana and Niculescu-Mizil, 2004]R. Caruana and A. Niculescu-Mizil. Data mining in metric space: An empirical analysis of supervised learning performance criteria. In *Proc. of "ROC Analysis in AI" Workshop (ECAI)*, pages 9–18, 2004.

[Church and Hanks, 1990]K.W. Church and P. Hanks. Word association norms, mutual information, and lexicography. *Computational Linguistics*, 16:22–29, 1990.

[Cohen *et al.*, 1999]W. Cohen, R. Schapire, and Y. Singer. Learning to order things. *Journal of Artificial Intelligence Research*, 10:243–270, 1999.

[Daille *et al.*, 1998]B. Daille, E. Gaussier, and J.M. Langé. An evaluation of statistical scores for word association. In *Proc. of The Tbilisi Symposium on Logic, Language and Computation, CSLI Publications*, pages 177–188, 1998.

[Deb, 2001]K. Deb. *Multi-Objective Optimization using Evolutionary Algorithms.* John Wiley & Sons, Chichester, 2001.

[Domingos, 1999]P. Domingos. Meta-cost: A general method for making classifiers cost sensitive. In *Knowledge Discovery from Databases*, pages 155–164, 1999.

[Dunning, 1993]T. E. Dunning. Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, 19(1):61–74, 1993.

[Ferri *et al.*, 2002]C. Ferri, P. Flach, and J. Hernandez-Orallo. Learning decision trees using the area under the ROC curve. In *Proc. of ICML'02*, pages 139–146, 2002.

[Fogel *et al.*, 1995]D.B. Fogel, E.C. Wasson, and E.M. Boughton. Evolving neural networks for detecting breast cancer. *Cancer Letters*, 96:49–53, 1995.

[Goodman and Smyth, 1988]M.F.R. Goodman and P. Smyth. Information-theoretic rule induction. In *Proc. of ECAI'88*, pages 357–362, 1988.

[Jacquemin, 1997]C. Jacquemin. Variation terminologique : Reconnaissance et acquisition automatiques de termes et de leurs variantes en corpus. In *Mémoire d'Habilitation à Diriger des Recherches, Université de Nantes*, 1997.

[Jin *et al.*, 2003]R. Jin, Y. Liu, L. Si, J. Carbonell, and A. Hauptmann. A New Boosting Algorithm Using Input-Dependent Regularizer. In *ICML 2003*. AAAI Press, 2003.

[Jong *et al.*, 2004]K. Jong, J. Mary, A. Cornuéjols, E. Marchiori, and M. Sebag. Ensemble feature ranking. *Proc. of PKDD-2004*, pages 20–24, 2004.

[Kolcz *et al.*, 2003]A. Kolcz, A. Chowdhury, and J. Alspector. Data duplication: An imbalance problem ? In *Workshop on Learning from Imbalanced Data Sets II (ICML)*, 2003.

[Lallich and Teytaud, 2004]S. Lallich and O. Teytaud. évaluation et validation de l'intérêt des règles d'association. *Revue RNTI*, E-1:193–217, 2004.

[Manning and Schütze, 1999]C. Manning and H. Schütze. *Collocations*, pages 165–184. Cambridge, MA: MIT Press, 1999.

[Roche *et al.*, 2004a]M. Roche, J. Azé, Y. Kodratoff, and M. Sebag. Learning interestingness measures in terminology extraction. a roc-based approach. In *Proc. of "ROC Analysis in AI" Workshop (ECAI)*, pages 81–88, 2004.

[Roche *et al.*, 2004b]M. Roche, J. Azé, O. Matte-Tailliez, and Y. Kodratoff. Mining texts by association rules discovery in a technical corpus. In *Proc. of IIPWM'04, Springer Verlag*, pages 89–98, 2004.

[Rosset, 2004]S. Rosset. Model Selection via the AUC. In *Proc. of the Twenty-First International Conference on Machine Learning (ICML'04)*, 2004.

[Schapire, 1999]R.E. Schapire. Theoretical views of boosting. In *Proc. of EuroCOLT-99*, pages 1–10, 1999.

[Sebag and Schoenauer, 1988]M. Sebag and M. Schoenauer. Generation of rules with certainty and confidence factors from incomplete and incoherent learning bases. In *Proc. of EKAW'88*, 1988.

[Sebag *et al.*, 2003a]M. Sebag, J. Azé, and N. Lucas. Impact studies and sensitivity analysis in medical data mining with ROC-based genetic learning. In *Proc. of ICDM 2003*, pages 637–640, 2003.

[Sebag *et al.*, 2003b]M. Sebag, N. Lucas, and J. Azé. ROC-based Evolutionary Learning: Application to Medical Data Mining. In *Proc. of EA 2003*, pages 384–396, 2003.

[Smadja *et al.*, 1996]F. Smadja, K. R. McKeown, and V. Hatzivassiloglou. Translating collocations for bilingual lexicons: A statistical approach. *Computational Linguistics*, 22(1):1–38, 1996.

[Smadja, 1993]F. Smadja. Retrieving collocations from text: Xtract. *Computational Linguistics*, 19(1):143–177, 1993.

[Vivaldi *et al.*, 2001]J. Vivaldi, L. Màrquez, and H. Rodríguez. Improving term extraction by system combination using boosting. *Proc. of ECML*, 2167:515–526, 2001.

[Xu *et al.*, 2002]F. Xu, D. Kurz, J. Piskorski, and S. Schmeier. A Domain Adaptive Approach to Automatic Acquisition of Domain Relevant Terms and their Relations with Bootstrapping. In *Proc. of LREC*, 2002.

Parametrised measures for the evaluation of association rules interestingness

Stéphane Lallich[1], Benoît Vaillant[2], and Philippe Lenca[2]

[1] Laboratoire ERIC, Université Lyon 2, France
(e-mail: stephane.lallich@univ-lyon2.fr)
[2] GET - ENST Bretagne -CNRS TAMCIC, France
(e-mail: forname.name@enst-bretagne.fr)

Abstract. In this paper, we present a original and synthetical overview of most commonly used association rule interestingness measures. These measure usually relate the confidence of a rule to an independency reference situation. Others relate it to indetermination, or impose a minimum confidence threshold. We propose a systematic generalisation of these measures, taking into account the reference point choosen by an expert in order to apprehend the confidence of a rule. This generalisation introduces new connections between measures, leads to the enhancement of some of them, and we propose new parametrised possibilities.
Keywords: interestingness measure, independency, indetermination.

1 Motivations

In this paper, we focus on the generalisating objective interestingness measures. We will consider association rule intersetingness measures, which aim at quantifying the quality of rules extracted from binary transactional datasets. In such datasets, each row is representing an object of the data mined, and consists of binary attributes, relating each object with properties that it may have or not. In this context, an association rule is an implication $\mathtt{A} \rightarrow \mathtt{B}$, where A and B (also called *itemsets*) are conjunctions of attributes. We denote by n the total number of transactions in the database, n_a (resp. n_b, n_{ab}, $n_{a\bar{b}}$) the number of transactions matching A (resp. B, A and B, A but not B), and by p_a (resp. p_b, p_{ab}, $p_{a\bar{b}}$) the corresponding relative frequencies. Most objective measures are expressed as real valued functions of n, of the marginal frequencies p_a, p_b, and either p_{ab} or $p_{a\bar{b}}$, *i.e.* as functions of n, and of the *confidence* (Conf) p_{ab}/p_a and marginal frequency counts of the considered rule since $p_{a\bar{b}}$ =p_a-p_{ab}. The higher the value of the measure, the better the rule is expected to be. Considering that the more counter-examples to a rule there are, the worst it is, we restrict our set of measures to those decreasing with $p_{a\bar{b}}$ (see table 1, references may be found in [Lenca *et al.*, 2004]). For a larger list of measures the reader should refer to [Guillet, 2004].

Support (Sup) and confidence (Conf) are the most famous of such measures, being the fundamentals principles of Apriori-like algorithms [Agrawal and Srikant, 1994]. These algorithms extract rules such that their Sup and Conf is above given thresholds, σ_s and σ_c. They are deterministic [Freitas, 2000], and produce a large number of rules which may not be interessting:

- one would expect from a rule that its Conf should be above a reference value, but the later seldom if ever equals σ_c. Two main references are clearly identified as worthy from a user point of view. The first one is p_b, which corresponds to the independence of the itemsets A and B. In this case the user wishes to focus on rules such that the prior knowledge of A increases the knowledge of B, *i.e.* rules having a confidence $p_{b/a}$ above the *a priori* frequency p_b. An alternative reference sometimes used is 0.5, as in [Blanchard *et al.*, 2005]. In our opinion, the first reference is to be taken within a *targeting* strategy, and the second one when considering a *predictive* strategy. More generally, a user may be interested in taking into account a reference value θ, $0 < \sigma_c \leq \theta \leq 1$, and will consider only rules having a Conf greater than θ. Fukuda Gain (Fuku) is an example of such a measure, where $\theta = \sigma_c$.
- what is more, the data mined is often subject to some sampling scheme. In order to take that into account, a special kind of measures have been proposed. They are called "statistical" in the sense that, unlike the others (also called "descriptive" measures), their value rises with n, the relative frequencies being fixed. This consideration accounts for developping an inferential approach, and retaining only rules that are significantly well evaluated by measures, comparison to the reference choosen. Amongst the issues that arise from this approach, validating a large number of rules through the control of false rules discovery is assessed in [Lallich *et al.*, 2004].

Various properties of interestingness measures have been investigated, in particular in [Piatetsky-Shapiro, 1991], [Hilderman and Hamilton, 1999], [Freitas, 1999], [Lallich, 2002], [Lallich and Teytaud, 2004], [Gras *et al.*, 2004] and [Lenca *et al.*, 2004]. One of these properties deals with the reference value to which the measure compares confidence, that is to say p_b (independency), 0.5 (indetermination), or some other value.

In this paper, we present a general survey of association rule interestingness measures and parametrise the reference value to which the measures will compare the confidence of a rule in order to estimate its quality. Such a consideration leads to an organised review of classical measures, the introduction of new ones, and enables us to enhance the coherence of some of them. We will first focus on descriptive measures, and then look at the statistical ones.

2 Descriptives measures

2.1 Reference to independency

Amongst frequently used measures added to Sup and Conf in order to capture the interestingness of a rule, are those taking the independence of the

	Authors	Relative definitions
Sup	(Agrawal and Srikant, 1994)	p_{ab}
Conf	(Agrawal and Srikant, 1994)	$p_{b/a}$
r	(Pearson, 1896)	$\frac{p_{ab}-p_a p_b}{\sqrt{p_a p_{\bar{a}} p_b p_{\bar{b}}}}$
CenConf		$p_{b/a} - p_b$
PS	(Piatetsky-Shapiro, 1991)	$np_a \left(p_{b/a} - p_b\right) = np_a p_b \,(\text{Lift} - 1)$
Loe	(Loevinger, 1947)	$\frac{p_{b/a}-p_b}{p_{\bar{b}}} = \frac{1}{p_{\bar{b}}}\text{CenConf} = 1 - \frac{1}{\text{Conv}}$
- ImpInd	(Lerman *et al.*, 1981)	$\sqrt{n}\frac{p_{a\bar{b}}-p_a p_{\bar{b}}}{\sqrt{p_a p_{\bar{b}}}}$
Lift	(Brin *et al.*, 1997)	$\frac{p_{b/a}}{p_b}$
LC	(Azé and Kodratoff, 2002)	$\frac{p_{ab}-p_{a\bar{b}}}{p_b} = 2\frac{p_a}{p_b}\,(\text{Conf} - 0.5)$
Seb	(Sebag and Schoenauer, 1988)	$\frac{p_{ab}}{p_{a\bar{b}}} = \frac{\text{Conf}}{1-\text{Conf}}$
OM	(Jeffreys, 1935)	$\frac{p_{b/a}/p_{\bar{b}/a}}{p_b/p_{\bar{b}}} = \frac{p_{ab}}{p_b}\frac{p_{\bar{b}}}{p_{a\bar{b}}} = \text{Lift} \cdot \text{Conv}$
Conv	(Brin *et al.*, 1997)	$\frac{p_a p_{\bar{b}}}{p_{a\bar{b}}}$
ECR		$1 - p_{a\bar{b}}/p_{ab} = 1 - 1/Seb$
IG	(Church and Hanks, 1990)	$\log \frac{p_{ab}}{p_a p_b} = \log(\text{Lift})$
IntImp	(Gras *et al.*, 1996)	$P\left[Poi\left(np_a p_{\bar{b}}\right) \geq np_{a\bar{b}}\right]$
EII	(Gras *et al.*, 2001)	$\left\{[(1 - h_1(p_{ab})^2)(1 - h_2(p_{ab})^2)]^{1/4}\varphi\right\}^{1/2}$
PDI	(Lerman and Azé, 2003)	$P\Big[\mathcal{N}(0,1) > \text{ImpInd}^{RC/\mathcal{B}}\Big]$
Fuku	(Fukuda *et al.*, 1996)	$np_a\left(p_{b/a} - \sigma_c\right)$
Gan	(Ganascia, 1988)	$2p_{b/a} - 1$

- $h_1(t) = -(1 - \frac{t}{p_a})\log_2(1 - \frac{t}{p_a}) - \frac{t}{p_a}\log_2(\frac{t}{p_a})$ if $t \in [0, p_a/2[$; else $h_1(t) = 1$
- $h_2(t) = -(1 - \frac{t}{p_{\bar{b}}})\log_2(1 - \frac{t}{p_{\bar{b}}}) - \frac{t}{p_{\bar{b}}}\log_2(\frac{t}{p_{\bar{b}}})$ if $t \in [0, p_{\bar{b}}/2[$; else $h_2(t) = 1$
- Poi stands for Poisson and $\mathcal{N}(0,1)$ for the standard normal distribution
- ImpInd $^{CR/\mathcal{B}}$ corresponds to ImpInd, centred reduced (CR) for a rule set $\mathcal{B}$

Table 1. List of measures

itemsets A and B as reference. This is the case of many linear transformation of Conf: the centered confidence (CenConf), Piatetsky-Shapiro (PS), Loevinger (Loe), the implication index (ImpInd), and the lift (Lift). All these measures additively centre confidence on p_b from $p_{b/a}-p_b$, save Lift for which the centring is multiplicative and based on $\frac{p_{b/a}}{p_b}$. Other monotonically increasing transformations of confidence making reference to independency are the odd multiplier ($OM = \frac{1-p_b}{p_b} \times \frac{Conf}{1-Conf}$), the conviction ($Conv = \frac{1-p_b}{1-Conf}$), whereas the information gain ($IG = \log Lift$) is a transformation of Lift.

2.2 Reference to indetermination

Some measures may (explicitly or not) refer to the indetermination situation, when the number of examples and counter-examples is balanced for a given n_a [Blanchard *et al.*, 2005]. This is the case of Conf and the two linear transformation: least confidence ($LC = 2 \times (p_{b/a} - 0.5) \times \frac{p_a}{p_b}$) and the Ganascia measure ($Gan = 2 \times (Conf - 0.5)$) that both additively centre Conf at 0.5. Other transformations can be listed, in particular the Sebag and Shoenauer measure ($Seb = \frac{Conf}{1-Conf}$) and the examples and counter-examples rate ($ECR = \frac{2\times(Conf-0.5)}{Conf}$).

2.3 Reference at θ

In order to generalise the expression of interestingness measures with respect to θ, *i.e.* rules such that $1 \geq Conf(\mathtt{A} \rightarrow \mathtt{B}) \geq \theta(\mathtt{A} \rightarrow \mathtt{B})$, we will alternatively consider the quantities $Conf - \theta$, $\frac{Conf}{\theta}$ and $\frac{Conf-\theta}{1-\theta}$. Descriptive interestingness measures are generalised as follows:

$$CenConf_{|\theta} = Conf - \theta$$

$$Gan_{|\theta} = \frac{Conf-\theta}{1-\theta} = Loe_{|\theta} = \frac{1}{1-\theta} CenConf_{|\theta}$$

$$Fuku_{|\theta} = PS_{|\theta} = np_a\left(Conf - \theta\right)$$

$$Lift_{|\theta} = \frac{Conf}{\theta}$$

$$IG_{|\theta} = \log(Lift_{|\theta})$$

$$Conv_{|\theta} = \frac{1-\theta}{1-Conf}$$

$$OM_{|\theta} = Seb_{|\theta} = \frac{Conf}{\theta} \times \frac{1-\theta}{1-Conf} = Lift_{|\theta} \times Conv_{|\theta}$$

$$LC_{|\theta} = \frac{Conf-\theta}{1-\theta} \times \frac{p_a}{p_b} = Loe_{|\theta} \times \frac{p_a}{p_b}$$

Some measures in table 1 are particular instances of several generalised expression:

$$OM_{|\theta=p_b} = Seb_{|\theta=0.5}, \quad Gan_{|\theta=0.5} = Loe_{|\theta=p_b}, \quad Fuku_{|\theta=\sigma_c} = PS_{|\theta=p_b}$$

3 Statistical measures

3.1 Intrinsics of statistic and probabilistic measures

As mentioned previously, a statistic measure takes into account the size of the sampling scheme. It is qualified of "probabilistic" when expressed as the complement of the p-value of the test under $p_{b/a} \leq p_b$ hypothesis. Classical approaches use the independence of itemsets $\mathtt{A}$ and $\mathtt{B}$ hypothesis as reference. The modelling of this hypothesis realised in [Lerman *et al.*, 1981] can be done in three different ways, with respectively 1, 2 and 3 hazard levels. We introduce model $1'$ which is an alternative to model 1 where p_a is fixed, rather than n_a (table 2).

We denote by N_{ab} the random variable generating n_{ab}, and H and B refer respectively to the hypergeometric and binomial laws. The statistic and probabilistic index based on $n_{a\bar{b}}$ are built as follows: by establishing the law of N_{ab} et $N_{a\bar{b}}$ under null hypothesis (H_0) following the choosen modelling, we can express a centered and reduced index under H_0, noted $N_{a\bar{b}}^{CR}$. Under standard conditions, the law of this index can be approximated to the normal distribution, leading to the definition of a probabilistic measure, defined as the surprise of observing such a high value of the index under H_0. The choosen modelling does not affect the expectation, but does modify the variance. [Gras, 1979] and [Lerman *et al.*, 1981] prefer the third modelling, that dissociates most rules $\mathtt{A} \rightarrow \mathtt{B}$ and $\overline{\mathtt{B}} \rightarrow \overline{\mathtt{A}}$ whereas the first modelling makes no dinstinction between these rules. The measure hence obtained is the implication intensity (IntImp), which is most satisfying on properties one expects a measure should have [Lenca *et al.*, 2004], [Gras *et al.*, 2004].

	Modelling 1 and 1′	Modelling 2	Modelling 3
Principle	1.1 n_a fixed, N_{ab} randomised 1.1' p_a fixed N_{ab} randomised	2.1 $N_a \equiv B(n, p_a)$ 2.2 $/N_a = n_a$, $N_{ab} \equiv B(n_a, p_b)$	3.1 $N \equiv P(n)$ 3.2$/N = n$, $N_a \equiv B(n, p_a)$ 3.3 $/N = n$, $N_a = n_a$, $N_{ab} \equiv B(n_a, p_b)$
Law N_{ab} under H_0	1.1 $H(n, n_a, p_b)$ 1.1' $B(n_a, p_b)$	$B(n, p_a p_b)$	$Poi(np_a p_b)$
Law $N_{a\overline{b}}$ under H_0	1.1 $H(n, n_a, p_{\overline{b}})$ 1.1' $B(n_a, p_{\overline{b}})$	$B(n, p_a p_{\overline{b}})$	$Poi(np_a p_{\overline{b}})$
Statistical index $N_{a\overline{b}}^{CR}$	1.1 $\frac{N_{a\overline{b}} - np_a p_{\overline{b}}}{\sqrt{np_a p_{\overline{a}} p_b p_{\overline{b}}}}$ $= -r\sqrt{n}$ 1.1' $\frac{N_{a\overline{b}} - np_a p_{\overline{b}}}{\sqrt{np_a p_b p_{\overline{b}}}}$	$\frac{N_{a\overline{b}} - np_a p_{\overline{b}}}{\sqrt{np_a p_{\overline{b}}(1 - p_a p_{\overline{b}})}}$	$IndImp$ $\frac{N_{a\overline{b}} - np_a p_{\overline{b}}}{\sqrt{np_a p_{\overline{b}}}}$
Probabilistic index $P(N(0,1) > N_{a\overline{b}}^{CR})$	1.1 $P(N(0,1) < r)$		IntImp $P(N(0,1) > IndImp)$

Table 2. Modelling of the various statistical and probabilistic index

3.2 Retaining the discriminating power

Although having many good properties, one of the major drawbacks of IntImp (drawback shared by the other statistic and probabilistic measures) is the loss of discriminating power. By its definition, it will evaluate rules significantly different from independency between 0.95 and 1. If n becomes important, which is particularly true in a data mining context, the slightest divergence from an independency situation becomes highly significant, thus leading to high and homogeneous values of the measure, close to 1.

In order to counter-balance this loss in discriminating power, [Lerman and Azé, 2003] introduce a contextual approach where ImpInd is centered and reduced (^{CR} notation) on a case database $\mathcal{B}$, thus leading to the definition of the probabilistic discriminant index, $PDI = P\left[N(0,1) > ImpInd^{CR/\mathcal{B}}\right]$.

[Gras *et al.*, 2001] propose an alternative solution by wheighting IntImp through the use of an inclusion index. This index is based on the entropy of experiments `B/A` and $\overline{\texttt{A}}/\overline{\texttt{B}}$. We denote by $H(X) = p_x \log_2 p_x + p_{\overline{x}} \log_2 p_{\overline{x}}$ the entropy associated with an event X. In [Blanchard *et al.*, 2004] the most general form of the inclusion index is given as:

$$i(A \subset B) = \left[\left(1 - H^*(B/A)^\alpha\right)\left(1 - H^*(\overline{A}/\overline{B})^\alpha\right)\right]^{\frac{1}{2\alpha}}$$

where $H^*(X) = H(X)$ if $p_x > 0.5$, $H^*(X) = 1$ otherwise. The α parameter is choosen by the user. The value $\alpha = 2$ is advised if one wants that this index should be tolerant to initial counter-examples, and we will use this value from now on. Hence, [Gras *et al.*, 2001] define the entropic intensity of implication as $EII = [IntImp \cdot i(A \subset B)]^{\frac{1}{2}}$

The shift from $H(X)$ to $H^*(X)$ aims at discarding uninteresting situations, such as $p_{b/a} < 0.5$ or $p_{\overline{a}/\overline{b}} < 0.5$, and complies with a predictive strategy. In a targeting strategy, the value of $p_{b/a}$ should have been compared to p_b, and the value of $p_{\overline{a}/\overline{b}}$ to $p_{\overline{a}}$.

The wheighting of the implication of intensity by the inclusion index, although effective, is problematic. The inclusion index is a measure of the distance to indetermination based on entropy, thus being null when $p_{b/a} = 0.5$, and so is EII. Still, IntImp values 0.5 at independency. Hence EII is not always null at independency: $EII = \sqrt[8]{\frac{(1-H(A)^2)(1-H(B)^2)}{16}}$ if $p_a < 0.5$ and $p_b > 0.5$, and is null otherwise.

3.3 Revised entropic intensity of implication

We will now propose two adaptations if EII in order to cope with the above mentioned issues: *REII* (Revised EII) et TEII (Troncated EII). Our first proposal consists in replacing IntImp by $IntImp^* = \max\{2IntImp - 1; 0\}$ in EII. This will solve the issues pointed out, but has the inconvenient of modifying the entire spectrum of values taken by EII:

$REII = [IntImp^* \cdot i(A \subset B)]^{\frac{1}{2}}$

Our second proposal solely nullifies the values of EII when $\frac{n_a n_{\bar{b}}}{n} \leq n_{a\bar{b}} \leq \min\{\frac{n_a}{2}, \frac{n_{\bar{b}}}{2}\}$, whithout modifying its values otherwise. To achieve this, we introduce an adequate version of $H(X)$. In order to take into account both predictive and targeting strategies, a rule will have a null evaluation by the inclusion index, and hence by TEII when the following conditions are jointly met:

- $p_{b/a} > 0.5$ (prediction) and $p_{b/a} > p_b$ (targeting); *i.e.* $p_{b/a} > \max(0.5, p_b)$
- $p_{\overline{a}/\overline{b}} > 0.5$ (prediction) and $p_{\overline{a}/\overline{b}} > p_{\overline{a}}$ (targeting); *i.e.* $p_{\overline{a}/\overline{b}} > \max(0.5, p_{\overline{a}})$

With these new conditions, TEII is null whenever the number of counter-examples is above $\min\left(\frac{n_a n_{\bar{b}}}{n}; \frac{n_a}{2}; \frac{n_{\bar{b}}}{2}\right)$. $TEII = [IntImp(A \rightarrow B) \times i_t(A \subset B)]^{\frac{1}{2}}$, with:

- $i_t(A \subset B) = \left[(1 - H_t^*(B/A)^\alpha)\left(1 - H_t^*(\overline{A}/\overline{B})^\alpha\right)\right]^{\frac{1}{2\alpha}}$,
- $H_t^*(B/A) = H(B/A)$ if $p_{b/a} > \max(0.5, p_b)$, $H_t^*(B/A) = 1$ otherwise,
- $H_t^*(\overline{A}/\overline{B}) = H(\overline{A}/\overline{B})$ if $p_{\overline{a}/\overline{b}} > \max(0.5, p_{\overline{a}})$, $H_t^*(\overline{A}/\overline{B}) = 1$ otherwise.

3.4 Measures making reference to indetermination

[Blanchard *et al.*, 2005] propose IPEE, a probabilistic measure of deviation from equilibrium. The authors implicitly use modelling $1'$ since they consider $N_{a\bar{b}} \equiv B(n_a, 0.5)$ under indetermination hypothesis, *i.e.* $N_{a\bar{b}}^{CR} = \frac{N_{a\bar{b}} - 0.5n_a}{0.5\sqrt{n_a}}$. They introduce $IPEE = P\left[B(n_a, 0.5) > n_{a\bar{b}}\right] \approx P\left[N(0,1) > \frac{n_{a\bar{b}} - 0.5n_a}{0.5\sqrt{n_a}}\right]$. Under normal approximation, IPEE equals 0.5 at indetermination. This measure corresponds to the probalistic index associated to modelling $1'$ (see table 2), where p_b is replaced by 0.5. IPEE will hence inherit of the weak discriminating power of this kind of measures, thus leading the authors to propose that it should be modulated by the inclusion index, which is all the most coherent, since both index make reference to indetermination.

3.5 Generalised intensity of implication

Using the same approach as with descriptive measures, we can generalise statistical measures and evaluate the interestingness of a rule by comparing its CONF to θ. This is done by considering in table 2 that for each modelling under H_0, the probability of an example, conditionally to n_a, of an example is θ: $N_{ab} \equiv B(n_a, \theta)$.

The results of the hence adapted modelling 1 is immediate, and those of modelling 2 and 3 are easily obtained through the use of the probability generating functions. If $X \equiv B(m,p)$, its generating function then is $G(s) = E(s^X) = (1-p+ps)^m$, and if $X \equiv Poi(\lambda)$, it is $G(s) = E(s^X) = e^{-\lambda(1-s)}$.

- In modelling 2, n is fixed, $N_a \equiv B(n, p_a)$ and $N_{ab}/(N_a = n_a) \equiv B(n_a, \theta)$. Since $G_{N_{ab}}(s) = E(s^{N_{ab}}) = E(E(s^{N_{ab}}/N_a)) = E\left((1-\theta+\theta s)^{N_a}\right)$, we have:
$$N_{ab} \equiv B(n, \theta p_a) \text{ and } N_{a\bar{b}} \equiv B(n, (1-\theta)p_a)$$
- In modelling 3, we have $N \equiv Poi(n)$, $N_a/(N = n) \equiv B(n, p_a)$, and $N_a/(N = n \text{ and } N_a = n_a) \equiv B(n_a, \theta)$.
As $G_{N_a}(s) = E(s^{N_a}) = E(E(s^{N_a}/N)) = E((1-p_a+p_a s)^N) = e^{-np_a(1-s)}$, then $N_a \equiv Poi(np_a)$.
Similarly, since $G_{N_{ab}}(s) = E(s^{N_{ab}}) = E(E(s^{N_{ab}}/N_a)) = E((1-\theta+\theta s)^{N_a}) = e^{-n\theta p_a(1-s)}$, we have:
$$N_{ab} \equiv Poi(n\theta p_a) \text{ and } N_{a\bar{b}} \equiv Poi(n(1-\theta)p_a)$$

From these results, we propose a range of generalised measures (see table 1), and will focus on two of them. The first one, $GIPE_{|\theta}$, associated to modelling $1'$ and generalises IPEE. It corresponds to the χ^2 adjustment of B/A distribution and $(\theta;\ 1-\theta)$. The second one, $GIntImp_{|\theta}$, associated to modelling 3 generalises INTIMP.

	Modelling 1 and $1'$	Modelling 2	Modelling 3
Principle	1.1 n_a fixed, N_{ab} randomised 1.1' p_a fixed, N_{ab} randomised	2.1 $N_a \equiv B(n, p_a)$ 2.2 $/N_a = n_a$, $N_{ab} \equiv B(n_a, \theta)$	3.1 $N \equiv Poi(n)$ 3.2$/N = n$, $N_a \equiv B(n, p_a)$ 3.3 $/N = n$, $N_a = n_a$, $N_{ab} \equiv B(n_a, \theta)$
Law N_{ab}	1.1 $H(n, n_a, \theta)$ 1.1' $B(n_a, \theta)$	$B(n, \theta p_a)$	$Poi(np_a\theta)$
Law $N_{a\bar{b}}$	1.1 $H(n, n_a, 1-\theta)$ 1.1' $B(n_a, 1-\theta)$	$B(n, (1-\theta)p_a)$	$Poi(np_a(1-\theta))$
Statistical index $N_{a\bar{b}}^{CR}$	1.1 $\frac{N_{a\bar{b}} - np_a(1-\theta)}{\sqrt{np_a p_{\bar{a}} \theta(1-\theta)}}$ 1.1' $\frac{N_{a\bar{b}} - np_a(1-\theta)}{\sqrt{np_a \theta(1-\theta)}}$	$\frac{N_{a\bar{b}} - np_a(1-\theta)}{\sqrt{np_a(1-\theta)(1-p_a(1-\theta))}}$	$GIndImp_{\lvert\theta}$ $\frac{N_{a\bar{b}} - np_a(1-\theta)}{\sqrt{np_a(1-\theta)}}$
Probabilistic index $P(N(0,1) > N_{a\bar{b}}^{CR})$	1.1' $GIPE_{\lvert\theta}$		$GIntImp_{\lvert\theta} =$ $P(N(0,1) > GIndImp_{\lvert\theta})$

Table 3. Modelling of the various generalised index

3.6 Discriminant power of the generalised measures

The generalised statistical or probabilistic measures have, as the original ones do, a weak discriminating power. In order to enhance these measures, we will consider two approaches, one being contextual, like [Lerman and Azé, 2003], the other one relying on a weighting through the use of an inclusion index, like [Gras *et al.*, 2001].

In the contextual approach, GIndImp $_{|\theta}$ (or its equivalent following modelling 1 and 2) is centred and reduced on a case database $\mathcal{B}$, and thus define a generalised probabilistic discriminant index, GIPD $_{|\theta}$, as follows.

$$GIPD_{|\theta} = P\left(N(0,1) > GIndImp_{|\theta}^{CR/\mathcal{B}}\right)$$

This way, we also define the generalised entropic intensity of implication, GEII$_{|\theta}$, as the product of GIndImp$_{|\theta}$ and an inclusion index. In order to remain coherent, we think advisable to use a generalised inclusion index $i_{|\theta}$, using θ as reference value and not 0.5. This can be achieved by replacing in the original formula $H(B/A)$ by $\widetilde{H}_{|\theta}(B/A)$ and $H(\overline{A}/\overline{B})$ by $\widetilde{H}_{|\theta}(\overline{A}/\overline{B})$ where:

- $\widetilde{H}_{|\theta}(B/A)$ is expressed as $H(B/A)$, in which we replace $p_{b/a}$ by $\widetilde{p}_{b/a}$ defined as follows:

$$\widetilde{p}_{b/a} = \frac{p_{b/a}}{2\theta} \text{ if } p_{b/a} \leq \theta,\ \widetilde{p}_{b/a} = \frac{p_{b/a}+1-2\theta}{2(1-\theta)} \text{ otherwise}$$

- $\widetilde{H}_{|\theta}(\overline{A}/\overline{B})$ can be expressed either:
  - by considering θ as reference, in which case we form $\widetilde{H}_{|\theta}(\overline{A}/\overline{B})$ as we did for $\widetilde{H}_{|\theta}(B/A)$, by replacing $p_{\overline{a}/\overline{b}}$ by $\widetilde{p}_{\overline{a}/\overline{b}}$ in $H(\overline{A}/\overline{B})$, with:

$$\widetilde{p}_{\overline{a}/\overline{b}} = \frac{p_{\overline{a}/\overline{b}}}{2\theta} \text{ if } p_{\overline{a}/\overline{b}} \leq \theta,\ \widetilde{p}_{\overline{a}/\overline{b}} = \frac{p_{\overline{a}/\overline{b}}+1-2\theta}{2(1-\theta)} \text{ otherwise}$$

    This first possibility generalises the inclusion index proposed in [Gras *et al.*, 2001], and can be found back using $\theta = 0.5$.
  - or using $1 - \frac{p_a}{p_{\overline{b}}} \times (1-\theta)$ as reference, since $p_{\overline{a}/\overline{b}} = 1 - \frac{p_a}{p_{\overline{b}}} \times (1 - p_{b/a})$. In this case, when considering independancy (*i.e.* $\theta = p_b$), the reference value for $\widetilde{H}_{|\theta}(\overline{A}/\overline{B})$ is $p_{\overline{a}}$.

$\widetilde{H}^*_{|\theta}(B/A)$ and $\widetilde{H}^*_{|\theta}(\overline{A}/\overline{B})$, are defined as:

$$\widetilde{H}^*_{|\theta}(X) = \widetilde{H}_{|\theta}(X) \text{ if } p_x > \theta,\ \widetilde{H}^*_{|\theta}(X) = 1 \text{ otherwise}$$

and $i_{|\theta}$ as:

$$i_{|\theta} = \left[\left(1 - \widetilde{H}^*_{|\theta}(B/A)^{\alpha}\right)\left(1 - \widetilde{H}^*_{|\theta}(\overline{A}/\overline{B})^{\alpha}\right)\right]^{\frac{1}{2\alpha}}, \text{ with } \alpha = 2.$$

From this, we deduce $\mathrm{GEII}_{|\theta}$ as $GEII_{|\theta} = \left[IntImp_{|\theta} \times i_{|\theta}\right]^{\frac{1}{2}}$, which is a more discriminating version of GIntImp. A similar approach leads to the definition of a generalised probabilistic measure of deviation, $\mathrm{GEIPE}_{|\theta}$, as $GEIPE_{|\theta} = \left[GIPE_{|\theta} \times i_{|\theta}\right]^{\frac{1}{2}}$.

Their behaviour, compared to their original counterparts, is represented figure 1. They were obtained using 3 different values for θ, $\theta = 0.1$ (thus targeting at independency), $\theta = 0.2$ (targeting for situations such that B happens twice more often when A is true) and $\theta = 0.5$ (prediction).

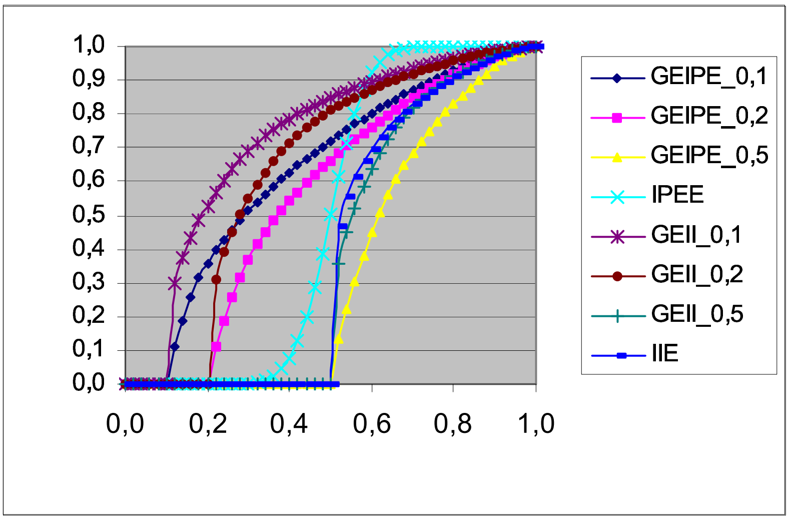


Fig. 1. Behaviour of the measures, in function of $p_{b/a}$ for $n = 1000$, $p_a = 0.05$ and $p_b = 0.10$

4 Conclusion

Following modelling and coherence principles, we proposed in this paper an innovating framework, from which a unified view of a large number of interestingness measures can be drawn, and which clarifies some of the links between these measures. Moreover, this framework is at the basis of the definition of new measures, namely the generalised intensity of implication, generalised probabilistic discriminant index, generalised entropic intensity of implication and the generalised probabilistic measure of deviation from equilibrium, that all compare the confidence of a rule to a reference parameter.

References

[Agrawal and Srikant, 1994]R. Agrawal and R. Srikant. Fast algorithms for mining association rules. In Jorge B. Bocca, Matthias Jarke, and Carlo Zaniolo, editors, *Proceedings of the 20th VLDB Conference*, pages 487–499, 1994.

[Blanchard *et al.*, 2004]J. Blanchard, P. Kuntz, , F. Guillet, and R. Gras. Mesure de la qualité des règles d'association par l'intensité d'implication entropique. *Revue des Nouvelles Technologies de l'Information*, 1(RNTI-E):33–43, 2004.

[Blanchard *et al.*, 2005]J. Blanchard, F. Guillet, H. Briand, and R. Gras. IPEE : Indice Probabiliste d'Ecart à l'Equilibre pour l'évaluation de la qualité des règles. In *Atelier DKQ*, pages 26–34, 2005.

[Freitas, 1999]A. Freitas. On rule interestingness measures. *Knowledge-Based Systems journal*, pages 309–315, 1999.

[Freitas, 2000]A. Freitas. Understanding the crucial differences between classification and discovery of association rules - a position paper. In *ACM SIGKDD Explorations*, volume 2, pages 65–69, 2000.

[Gras *et al.*, 2001]R. Gras, P. Kuntz, R. Couturier, and F. Guillet. Une version entropique de l'intensité d'implication pour les corpus volumineux. *Extraction des connaissances et apprentissage (EGC 2001)*, 1(1-2):69–80, 2001.

[Gras *et al.*, 2004]R. Gras, R. Couturier, J. Blanchard, H. Briand, P. Kuntz, and P. Peter. Quelques critères pour une mesure de qualité de règles d'association. *RNTI-E-1*, 2004.

[Gras, 1979]R. Gras. *Contribution à l'étude expérimentale et à l'analyse de certaines acquisitions cognitives et de certains objectifs didactiques en mathématiques.* PhD thesis, Université de Rennes I, 1979.

[Guillet, 2004]F. Guillet. Mesure de la qualité des connaissances en ECD. In *Tutoriel de la 4e Conf. Extraction et Gestion des Connaissances (60 p.)*, 2004.

[Hilderman and Hamilton, 1999]Robert J. Hilderman and Howard J. Hamilton. Knowledge discovery and interestingness measures: A survey. Technical Report 99-4, Dpt. of Computer Science, University of Regina, october 1999.

[Lallich and Teytaud, 2004]S. Lallich and O. Teytaud. Évaluation et validation de l'intérêt des règles d'association. *RNTI-E*, 1:193–217, 2004.

[Lallich *et al.*, 2004]S. Lallich, E. Prudhomme, and O. Teytaud. Contrôle du risque multiple en sélection de règles d'association significatives. In *EGC 04*, volume 2, pages 305–316, 2004.

[Lallich, 2002]S. Lallich. Mesure et validation en extraction des connaissances à partir des données. HDR – Université Lyon 2, 2002.

[Lenca *et al.*, 2004]P. Lenca, P. Meyer, B. Vaillant, P. Picouet, and S. Lallich. Evaluation et analyse multi-critères des mesures de qualité des règles d'association. *RNTI-E*, 1:219–246, 2004.

[Lerman and Azé, 2003]I.C. Lerman and J. Azé. Une mesure probabiliste contextuelle discriminante de qualité des règles d'association. *RSTI-RIA (EGC 2003)*, 1(17):247–262, 2003.

[Lerman *et al.*, 1981]I.C. Lerman, R. Gras, and H. Rostam. Elaboration d'un indice d'implication pour les données binaires, i et ii. *Mathématiques et Sciences Humaines*, (74, 75):5–35, 5–47, 1981.

[Piatetsky-Shapiro, 1991]G. Piatetsky-Shapiro. Discovery, analysis and presentation of strong rules. In G. Piatetsky-Shapiro and W.J. Frawley, editors, *Knowledge Discovery in Databases*, pages 229–248. AAAI/MIT Press, 1991.

Visualisation and exploration of high-dimensional data using a “force directed placement”method: application to the analysis of genomic signatures

Sylvain Lespinats, Alain Giron, and Bernard Fertil

INSERM Unité 678, CHU Pitié-Salpêtrière
91 bd de l'hôpital, 75634 PARIS (France)

Abstract. Visualization of high-dimensional data is generally achieved by projection in a low dimensional space (usually 2 to 3 dimensions). Visualization is designed to facilitate the understanding of data sets by preserving some “essential”information. We have designed a non-linear multi-dimensional-scaling (MDS) tool relying on the force directed placement (FDP) algorithm to help dynamically discover features of interest in data sets. A user-driven relaxation of constraints built on the preservation of pairwise distances between data allows getting subjective representations of data that meet some specific angle. In a context of classification, we examine the impact of metric, sample size, and neighborhood preservation on the mapping of genomic signatures.

Keywords: Multi-Dimensional Scaling, Force Directed Placement, Classification, Proximity visualisation, Metric.

1 Introduction

High dimensional data raise unusual problems of analysis, given that some properties of the spaces they live in cannot be extrapolated from our current experience [Verleysen, 2001]. The notion of neighborhood in particular must be revised to take into account the number of dimensions. In particular (notably in the case of Euclidean spaces), we often face the problems of empty space and concentration of measure: when the number of dimensions is high, the neighborhood of each object is scarcely filled whereas most of the other objects are found in a thin outer shell. Distances between high dimensional objects are usually very concentrated around their average.

Exploration and analysis of high dimensional data are often made by means of dimension reduction techniques. Since human experience mostly deals with 3D space (and most data display devices are two-dimensional), finding a meaningful mapping of data in such low dimensional spaces is the issue. Principal component analysis (PCA), multidimensional scaling (MDS) [Cox and Cox, 1994], Kohonen maps (SOM) [Kohonen, 1997] are classic approaches in this context. In general, a loss function is defined to characterize the error in representing the dissimilarity between objects. It allows building the rules of projection from the original space of the data on to a smaller dimensional space. It is important to realize that any reduction of dimension leads to a subjective data representation. Depending on the purpose, different mappings may be achieved for the same set of data. For classification tasks, for example, the preservation of the neighborhood appears one of the aspects important to master. In this work, we examine some interesting mappings obtained by means of a nonlinear MDS-based projection. In particular, the consequences of dimension reduction on the classification of genomic signatures (256-dimension data originally) are analyzed.

2 Reduction of data dimension: principles ruling the present study

The approach that is presented here belongs to the MDS group of methods. It is thus advisable to define metrics for the original data space and for the target space (called output space thereafter), a lost function and a mapping algorithm. Usually, the characteristics of the data to be analyzed are to be considered to choose these various elements.

2.1 Data, metric and lost function

Data under investigation in this work concern the genomic signature. The whole set of short oligonucleotide frequencies observed in a DNA sequence is species-specific and is thus considered as a genomic signature (Deschavanne et al., Karlin et al.). The genomic signature characterizes the DNA molecule by 256 frequency variables, defined in the range [0-1]. Counts (and frequencies) of oligonucleotides can be displayed as parametric images allowing fast visual examination and comparison (http://genstyle.imed.jussieu.fr). It has been observed that the genomic signature results from a species-specific "writing style"[Deschavanne *et al.*, 1999]. Indeed, on one hand, the genomic signatures of species differ from one another; on the other hand, the majority of DNA segments isolated from the genome of a given species have comparable signatures. As a consequence, each species is given a genomic signature that can be derived from most of its available DNA fragments. The DNA style is obtained from the examination of relatively small chains of the genetic material. In practice, a sequence as short as 2000 nucleotides usually provides a good estimate.

The Euclidean metric allows showing statistically significant differences between species' genomic signatures [Deschavanne *et al.*, 1999]. This metric will thus be chosen to illustrate the method, for typical examples at first (projection from a 3D space towards a 2D space), then for the problem of classification of genomic signatures. In some instances, we may consider preserving only the rank order of distances between objets, not the exact values.

Such a procedure is found useful when the projection provides “unsatisfactory”results. The projection should then try matching the rank order of distances between objects in the two-dimensional output space to the rank order in the original space.

The lost function is defined as a weighted sum of errors over dissimilarities (distances or ranks) between all pairs of objects in the original space and the output space. Eventually, subsets of data may be considered to test the robustness of projection. A part of data is used to define the mapping whereas the remaining part serves checking representiveness of output space. In order to preferentially favor close proximity, a weighting scheme reducing the impact of errors related to large dissimilarities may be gradually applied during the phase of optimization. This approach takes benefit from the work by P. Demartine and J. Herault [Demartine and Herault, 1997] and T. Kohonen [Kohonen, 1997].

2.2 Loss function minimization algorithm

In general, the optimal position of data in the output space cannot be obtained analytically. It is necessary to implement a function minimization algorithm with widely recognized robustness and convergence aptitudes. Classically, in the context of MDS, one alternatively uses the generalized Newton-Raphson algorithm, TABU Search [Glover and Laguna, 1995], genetic algorithms [Goldberg, 1989] or simulated annealing [Dowsland, 1995].

Regarding our model (called FDP-MDS thereafter), we propose to set up a dynamic algorithm grounded on the “Force Directed Placement”paradigm (FDP) [Fruchterman and Reingold, 1999]. Firstly described at the beginning of the Eighties, the FPD method is yet popular in only a limited number of fields. In particular, it is extensively used for the design of printed circuits. It is on the other hand little known in the field of data analysis. The force directed placement metaphor may be clarified in the following way: the data to place in the output space are bounded by forces (materialized by springs for example) the magnitude of which are related to the satisfaction of dissimilarities. In the case of springs, length at rest corresponds to the dissimilarity between the connected objects in data space. Any departure from the resting value consequently results in a recall force contributing in the movement of object and accounting for the energy of the system. Starting from an initial state with the objects placed the most judiciously possible in output space, the system is allowed to relax towards a minimum state of energy for which the constraints of dissimilarities between objects are satisfied as much as possible. FPD algorithm is very interesting in the case of MDS, considering its speed of convergence and its possibilities to escape from local minima.

For problems dealing with few thousands of objects, it is possible to directly run the FDP algorithm with the whole set of data. For larger data collections, it is often interesting to select a subset of objects to coarsely define the topology of the output space, in a first step. Remaining data are subsequently positioned with respect to preceding ones, by preferentially satisfying local constraints. In our hands, the incremental approach shows up very effective, especially when initial objects are selected after clustering.

2.3 Non-linear projection achieved by FDP-MDS: examples

Two boxes: Data to be projected have three dimensions. Objects are organized to represent 2 cubic boxes with an open side not pointing in the same direction. Projection onto a 2D space with FDP-MDS correctly develops the 2 boxes and carries out a twist on a large scale (fig. 1). Relations of vicinity are satisfactorily preserved.

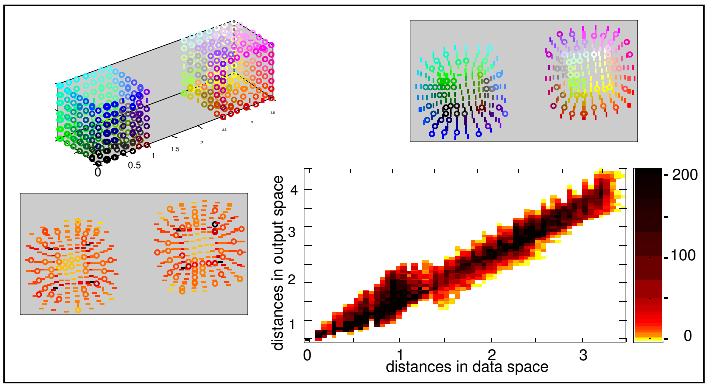


Fig. 1. Mapping of 2 3D open boxes in a 2D space. Upper left: original data (3D space), upper right: mapping (2D space), lower left, satisfaction of constraints on objects (satisfaction increases from black to white, LUT of fire), lower right, pairwise distances preservation (color codes for density of distances). NB: Colored figures are available from our WEB site <http://e6.imed.jussieu.fr/afficherpub.php/ASMDA05.pdf>

Earth globe: Data to be projected are the big cities around the word (3D). Projection accounts for local density of cities. The north hemisphere is properly developed (Fig. 2). Cities-free areas are distorted although continuity is preserved in most places (The grid is not used during the mapping construction).

2.4 Mapping high dimensional data: the genomic signature issues

The data concerned with this study belong to two families; the signatures of 5000 species constitute a subset of the diversity of ADN molecules on earth. The signature of a species, *B. subtilis*, is studied in detail. One thousand

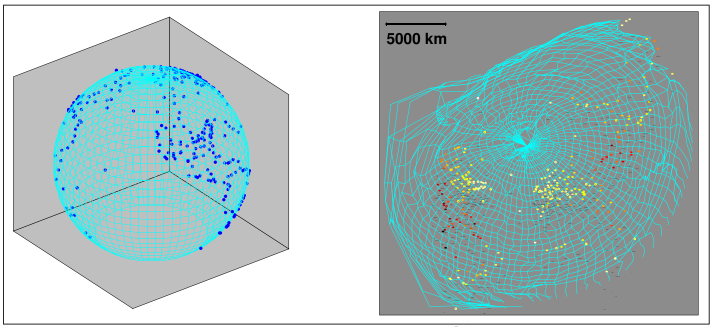


Fig. 2. Mapping of the earth globe (defined by the big cities) in a 2D space. Color indicates satisfaction of pairwise distances for the corresponding city (Color scheme is similar to Fig. 1).

eight hundred and twenty four signatures corresponding to the analysis of *B. subtilis* genome through a sliding window of preset size (5000 nucleotides) are calculated. The signature of each of these windows (called local signatures thereafter) generally displays the characteristics of *B. subtilis*. All signatures are defined by 256 frequency variables.

The first issue to be addressed in this work concerns the effect of sampling on the mapping of high dimensional data. Five hundred local signatures of *B. subtilis* are randomly selected to build a proximity preserving 2D output space (Fig 3, left panel). The 1324 remaining signatures are subsequently placed, using the FDP algorithm. It appears clearly that the mapping is not suitable to handle the diversity of local signatures of *B. subtilis*. Most of the signatures that were not considered for the mapping are concentrated around the center of the space, whereas a randomly placement would be expected. Obviously, pairwise distances between 500 local signatures are not enough to properly describe the proximity characteristics of these highly dimensional objects. New objects cannot fit in the output space. It must be pointed out that this peculiar behavior is not observed for the 5000 genomic signatures although their dimension is the same (result not shown). It is suggested that the intrinsic dimension of signatures is the key to explain this surprising result. Local signatures may stretch over most of the avaible dimensions (sampling effect) whereas variations among genomic signatures only concern specific directions characterizing the restricted set of possible pathways for species differentiation.

Surprisingly, switching from the Euclidean metric to the rank pseudo-metric solves the problem (Fig. 3, right panel)! It may be considered that the mapping obtained using the rank pseudo-metric is robust to sampling size, but additional experiments and theoretical developments are required to firmly conclude on this point.

The second issue deals with classification. Local signatures are expected similar to the genomic signature of the species they come from. It should be

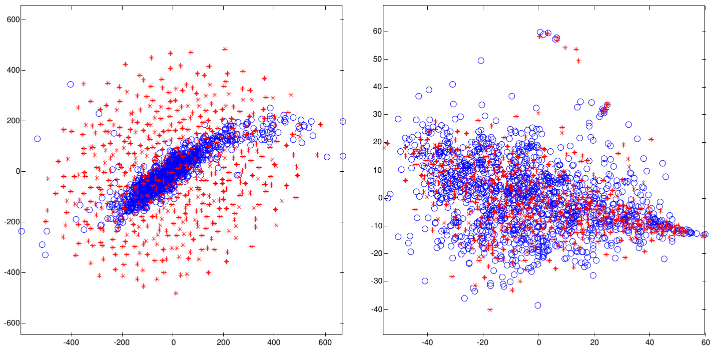


Fig. 3. Mapping of *B. subtilis* local signatures. Red crosses (500) are for signatures used to construct the mapping, blue circles (1324) are for additional local signatures placed afterwards.

subsequently possible to search for the species of origin of any local signature of *B. subtilis*, using a nearest neighbor classifier exploring the 5000 genomic signature set. Within the framework of this paper, 2 situations are considered: i) the mapping is learned using the species' signatures, ii) the mapping is learned with all available signatures, including *B. subtilis*' local signatures.

In data space (256 dimensions), only 64% of local signatures are correctly assigned to *B. subtilis*. In fact there are about one hundred of species in the hyper-sphere holding 95% of local *B. subtilis*' signatures, some of them being even very close to *B. subtilis*. It should be noted that an important subset of local signatures is misclassified for known biological reasons. When the space of projection is learned from the species' signatures, the rate of good classification falls to 0,7%(fig. 4, left panel). It is 24% when the space of projection is learned from the whole set of signatures (species and local, fig. 4, right panel). The zone devoted to local signatures in the output space is extended to satisfy constraints of distances between local signatures when they are included in the training sample. Even so, quality of classification remains poor.

3 Discussion et conclusion

The nonlinear approach of mapping described in this article was designed to preferentially preserve proximity. For small dimension problems, it appears that its effectiveness is quite good. It is unfortunately not the case for high dimension data where the learning sample size seems to be a critical parameter and the efficiency of local signature nearest neighbor classification is strongly reduced in the output space. The method of classification used in this work is particularly sensitive to "errors"of placement since only one "mis-placed"species may cause multiple classification errors. However, this

situation is likely to occur many times in such dramatic reductions of dimension (256 towards 2-3). Considering that the growth of neighborhood with increasing radius (around every object) in a high dimensional space cannot be effectively matched in a low dimensional space, only data with a small intrinsic dimension may be properly mapped in a small dimensional Euclidean space.
An interesting alternative is proposed by H. Ritter and J. Walter [Ritter, 1999] [Walter and Ritter, 2002]: they use a 2-dimensional hyperbolic plane as output to simulate the singular growth of neighborhood of high dimensional space. The approach seems very promising. The learning sampling size is also an important parameter to master. Obviously, the conjunction of the empty space phenomenon with the singular growth of neighborhood in high dimensional space make the sampling phase (when required) particularly tricky. All together, it seems useful to recall that the analysis of the data resulting from consequent compression ratios must be carried out with infinite precautions.

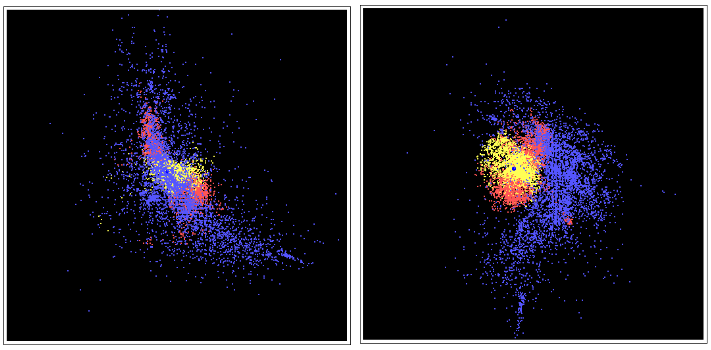


Fig. 4. mapping of genomic signatures in a small dimensional space: Species' signatures are in blue (dark), well-classified local *B. subtilis*' signatures (in the data space) are in yellow (light), mis-classified signatures are in red (see text). Left panel: mapping obtained with species' genomic signatures, right panel: mapping obtained with the full set of available signatures (species and local).

4 Acknowledgments

INSERM (Institutional funding), Action INTER EPST Bio-informatique 2001 Grant, contract N° 120910

References

[Cox and Cox, 1994]T. Cox and M. Cox. *Multidimensional Scaling.* Chapman & Hall, London, 1994.

[Demartine and Herault, 1997]P. Demartine and J. Herault. Curvilinear component analysis: a self-organizing neural network for nonlinear mapping of data sets. *IEEE Trans. Neural Networks*, 8:148–154, 1997.

[Deschavanne *et al.*, 1999]P. J. Deschavanne, A. Giron, J. Vilain, G. Fagot, and B. Fertil. Genomic signature: characterization and classification of species assessed by chaos game representation of sequences. *Mol. Biol. Evol.*, 16(10):1391–9, 1999.

[Dowsland, 1995]K.A. Dowsland. Simulated annealing. In C.R. Reeves, editor, *Modern techniques for combinatorial problems*, chapter 2. McGraw-Hill Book Company, Berkshire, 1995.

[Fruchterman and Reingold, 1999]T. Fruchterman and E. Reingold. Graph drawing by force-directed placement. *Software-Practice and Experience*, 21:1129–64, 1999.

[Glover and Laguna, 1995]F. Glover and M. Laguna. Tabu search. In C.R. Reeves, editor, *Modern euristic techniques for combinatorial problems*, chapter 3. McGraw-Hill Book Company, Berkshire, 1995.

[Goldberg, 1989]D.E. Goldberg. *Genetic algorithms in search, optimization and machine learning.* Addison-Wesley, Reading, Massachusetts, 1989.

[Kohonen, 1997]T. Kohonen. *self organizing maps.* Springer-Verlag, 1997.

[Ritter, 1999]H. Ritter. Self-organizing maps on non-euclidean spaces. In S. Oja and E. Kaski, editors, *Kohonen maps*, pages 97–110. Elsevier, Amsterdam, 1999.

[Verleysen, 2001]M. Verleysen. Learning high-dimensional data. In *NATO advance research workshop on limitation and future trends in neural computing*, Siena, Italy, 2001.

[Walter and Ritter, 2002]J.A. Walter and H. Ritter. On interactive visualization of high-dimensional data using the hyperbolic plane. In *SIKDD '02*, Edmonton, Alberta, Canada, 2002.

About the locality of kernels in high-dimensional spaces

Damien Francois[1], Vincent Wertz[1], and Michel Verleysen[2]

[1] UCL - CESAME
Avenue G. Lemaitre, 4
B-1348 Louvain-la-Neuve, Belgium
(e-mail: {francois,wertz}@auto.ucl.ac.be)
[2] UCL - Machine Learning Group
Place du Levant, 3,
B-1348 Louvain-la-Neuve, Belgium
(e-mail: verleysen@dice.ucl.ac.be)

Abstract. Gaussian kernels are widely used in many data analysis tools such as Radial-Basis Function networks, Support Vector Machines and many others. Gaussian kernels are most often deemed to provide a local measure of similarity between vectors. In this paper, we show that Gaussian kernels are adequate measures of similarity when the representation dimension of the space remains small, but that they fail to reach their goal in high-dimensional spaces. We suggest the use of p-Gaussian kernels that include a supplementary degree of freedom in order to adapt to the distribution of data in high-dimensional problems. The use of such more flexible kernel may greatly improve the numerical stability of algorithms, and also the discriminative power of distance- and neighbor-based data analysis methods.
Keywords: High dimensional spaces, Local Models, Gaussian Kernels.

1 Introduction

Data analysis is one of the areas where artificial neural networks and machine learning techniques in general, have the most impact. During the last twenty years, there has been a considerable effort to develop data analysis techniques that are adapted to the abundance of data in nowadays information society. Although those tools are different in many aspects, be it from theoretical, technical or historical point of view, many of them share a common characteristic: for one reason or another, they use kernels. This is for example the case for Radial-Basis Function Networks (RBFN) [Bishop, 1995], for Support Vector Machines (SVM) [Cristianini and Shawe-Taylor, 2000], but also

[1] The work of D. Francois is funded by a grant from the Belgian FRIA. Part of the work of D. Francois and of the work of V. Wertz is supported by the Interuniversity Attraction Pole (IAP), initiated by the Belgian Federal State, Ministry of Sciences, Technologies and Culture. The scientific responsibility rests with the authors.

[2] M. Verleysen is a Senior Research Associate of the Belgian F.N.R.S. (National Fund For Scientific Research).

for the more traditional Parzen estimators of probability densities [Parzen, 1962], for mixtures of Gaussians [McLachlan and Peel, 2000], etc.

Kernels can be defined in various ways. In most cases however, Kernel means a function whose value only depends on a distance between the input and a constant, named center; the input and the center may be vectors. Naturally, the most often used kernel is the Gaussian one. There are several good justifications to using Gaussian kernels. The first one is that the Gaussian function is a natural one: by the Central Limit Theorem, the sum of independent variables having the same distribution, whatever the distribution is, tends to a Gaussian distribution as the number of terms in the sum tends to infinity. The Gaussian function or distribution is also the only one that can be described without loss of information by its two first moments; it is therefore of particular interest for second-order statistics, including all linear data analysis methods.

Besides these general considerations, Gaussian kernels are most often used for their locality property: it is obvious that the Gaussian output may be considered as high when the input is close from the center and low (or even negligible) when the argument is far from the center. Locality is a primary importance concept for many reasons that range from the interpretability of the models to their numerical stability, through experimentally observed advantages with specific types of data.

This paper aims to show that the use of Gaussian kernels may be valid when the data are represented in low-dimensional spaces, but fails to reach its objectives in high-dimensional spaces. It is shown that high-dimensional Gaussian kernels are usually not local, and cannot be made local through scaling factors. This paper suggests using the so-called Generalized p-Gaussian kernel, which can be made local in any-dimensional space through the adaptation of a supplementary parameter.

This paper is organized as follows. Section 2 briefly recalls why the concept of locality is important in data analysis methods. Section 3 shows that Gaussian kernels are not local functions in high-dimensional spaces. Finally, in Section 4 Generalized p-Gaussian kernels are introduced as a possible alternative to Gaussian kernels for high-dimensional data analysis methods.

2 Why is locality so important?

While the locality property seems important in many algorithms, few papers address the reasons why it is indeed important. In the following, some intuitive arguments in the favor of local kernels are developed, without any attempt to be exhaustive.

2.1 Interpretability

The main argument for locality is interpretability. In most if not all applications, practitioners are not happy about responses given by blind models, i.e.

models that do not provide interpretability of their outputs. Nevertheless several algorithms are mostly blind, or at least have the reputation to be blind; examples are feed-forward artificial neural networks such as the Multi-Layer Perceptron (MLP) and RBFN. Interpretation in the latter models can however come from an examination of their hidden units outputs.

Indeed, the Kernel function can be seen as measure of similarity. The range of the kernel is between zero and one (note that when kernels are used for density estimation they are normalized so that their integral equals one ; this is not the case here. In any case, scaling does not change the arguments below). An input may be considered as close to the kernel center when the kernel output is near 1, and far when the output is near 0 ; indeed the ouptu of a Gaussian kernel decreases from 1 to 0 according to a negative exponential of the squared Euclidean distance between the vectors. Kernels may then be used to express in a numerical form the intuitive notion of closeness, i.e. similarity, with the continuity and derivability properties that are necessary in most algorithms. Regions spanned by kernels up to the limits defined (in afuzzy way) by the notion of closeness may help to the interpretation of the model.

The closeness concept is essential in local mmodels. For instance, RBFN and SVM models build the output corresponding to a new input x as a weighted sum of the output values associated to certain entities living in the input space (respectively called centroids and support vectors) ; while the weight is the similarity measurement between x and those entities. In other words, the more similar the new input is to a given entity, the more importance that entity has in computing the predicted value. Many Lazy Learning methods can be interpreted this way too.

2.2 Numerical stability

For RBFN as for SVM, the values of each kernel at each data point is gathered into a matrix which is used to formulate the corresponding optimization problem. The conditioning, and thus the sensitivity and numerical stability of the problem, depends on the condition number of that matrix. This section illustrates the fact that building a kernel-based model leads to an ill-formed optimization problem when locality of the kernels is not ensured.

Suppose N points randomly drawn according to a uniform distribution in the $[0, 1]^d$ d-dimensional cube. A vector quantization is then performed on these N points to obtain M centroids, representative on the initial distribution. A traditional RBFN learning consists in placing Gaussian kernels on each of the M centroids, and evaluating the scalar RBFN output as a linear combination of the kernel outputs [Hwang and Bang, 1997]. The M linear coefficients are found by least squares. The matrix of the system is the $N \times M$ matrix built by evaluating each kernel on each data point. It is known that the numerical stability of the system depends on the condition

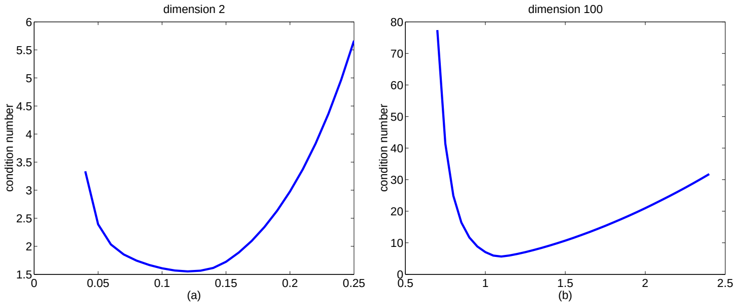


Fig. 1. Condition number of the system matrix in a RBF network, with respect to the standard deviation (width) of the kernels.

number of the matrix, which is defined as the ratio between the largest and smallest singular value of the matrix [Golub and van Loan, 1996].

Figure 1 shows an example of this condition number, for a system with 200 data points and 10 centroids. The condition number is plotted versus the common standard deviation (width) of the kernels. As the centroids learned by vector quantization have a distribution equal to the distribution of the initial data, i.e. they are uniformly distributed, it is natural to assume that all kernel standard deviations are equal.

One can see on Figure 1 that an optimum exists in the condition number of the system matrix, corresponding to an optimal standard deviation. While the exact value does not matter here, one can easily see that deviations much smaller or larger than the optimal lead to ill-conditioned matrices.

If the standard deviation is too small, the Gaussian kernels will not reach (with a significant value) the data points, even those that are close to the centroids. Very large coefficients will thus result from the system solution, both in positive and negative values, in order to both include all data into the radius of attraction of at least one Gaussian kernel, and at the same time keeping a weighted sum into a small range (corresponding to a smooth function to approximate).

On the contrary, if the standard deviation is too large, the Gaussian kernels will be very flat, leading to having most or all points into their respective radius of attraction. Approximating a smooth but non-flat (constant) function therefore also results in very large, both positive and negative, model coefficients.

Both situations therefore lead to ill-defined systems. Locality (not too large standard deviation) is thus also important for the numerical stability of the algorithms. Of course, too narrow kernels should be avoided too, as this corresponds to a kind of overfitting.

3 Gaussian kernels are not adequate in high-dimensional spaces

At first sight, the objective, i.e. measuring the similarity between two vectors, and the way to reach the goal, i.e. using a Gaussian kernel, perfectly match. However, without reference to Gaussian kernels, one could define an ideal kernel as a kernel whose output gives an acceptable measure of the similarity between two vectors; acceptable means for example that among a finite distribution, the closest vectors to a query should be evaluated as similar to the query, while vectors that are far from the query should be evaluated as non similar. In other words, among a finite distribution, the selected similarity measure should be able to find in acceptable proportions both similar and non similar vectors to a query point. In the next section, it will be shown that Gaussian kernels fit with this definition in low-dimensional spaces, while they do not fit it in high-dimensional spaces. To illustrate this problem, let us imagine that data have a Gaussian distribution centered at C (the following is qualitatively valid for any distribution though). We will compare the distribution of distances between any point and C, to the shape of a kernel centered on C too. As the kernel will be used to assess if points are close or not from C, this experiment allows to verify that the kernel is discriminative (is not too flat) in the effective range of the distance distribution. On Figure 2, the thick line represents the kernel value, while the thin line (and grayed area) represent the distance distribution. One easily sees on graphs (a) and (b) that, in low dimension, for a well-chosen kernel width value, the small (resp. large) distances in the distribution will be mapped onto kernel values close to one (resp. zero). This matches the definition of an ideal kernel as detailed in the previous paragraph.

However when the space dimension increases, the correspondence between the range of distances in the histogram, and the range of the decreasing slope in the Gaussian kernel cannot be guaranteed anymore. Graphs (c) and (d) refer to space dimensions 10 and 100 respectively, for several kernel width values. It is seen that it is more difficult to adjust the value of the kernel width is in order to cope with the ideal kernel definition: in all cases, there is a large part of the Gaussian kernel decreasing slope that falls out of the range of distances in the histogram. This means that close distances (left queue of the distribution) and large distances (right queue of the distribution) are hardly distinguishable from their kernel values; the notion of similarity itself (are data close or far one from another) looses its significance. Needless to say, the consequences in methods based on nearest neighbors are dramatic.

Another view of the same phenomenon comes from the following experiment. Let us imagine a d-dimensional uniform distribution, quantized into a predefined number M of centroids. A Gaussian kernel is centered on each initial point of the distribution; the kernel is evaluated on the furthest and closest centroids. Then the difference between the two Gaussian outputs is taken, and averaged over all points of the distribution. The result is repre-

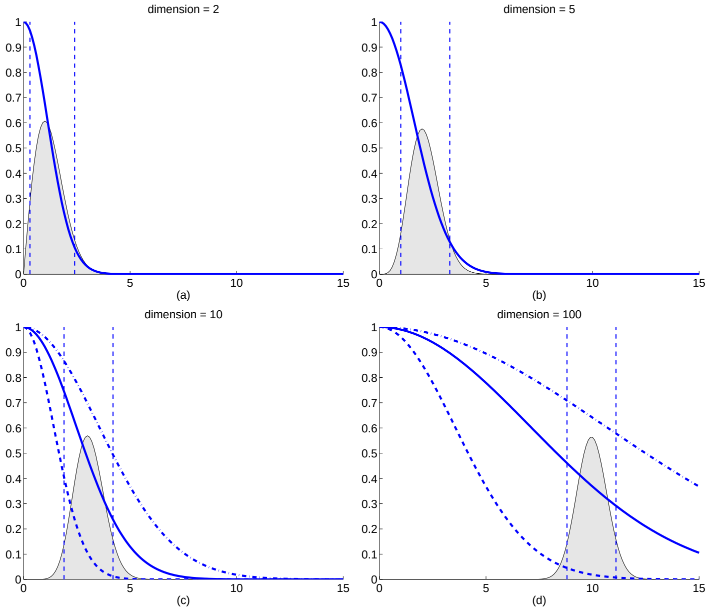


Fig. 2. Kernel values as a function of the distance to their centers for several space dimensions, along with the distribution of distances for normally distributed data. Vertical lines correspond to 5 and 95 percentile resp.

sentative of the contrast between the similarity of a point to its closest and furthest away centroids; if the contrast is large, a model built with such a kernel can be considered 'local' ; if it is small, the notion of neighborhood looses its significance.

Figure 3 shows this contrast with respect to the width of the kernels. In dimension 2 (left), the contrast is close to 1 for a well-chosen value of the kernel; distances are easily distinguishable. Note that the ideal kernel standard deviation is relatively small, which corresponds to a kernel having a local character. In dimension 100, the contrast hardly reaches 0.2; distances are far less distinguishable, whatever the kernel standard deviation is.

4 Recovering locality in HD spaces

The necessity to more or less span the effective range of distances between data in a real distribution setting, by the effective part of the kernel (i.e. the part with the decreasing slope), requires to add a parameter with respect to the Gaussian kernel. Besides the width that controls the slope of the kernel, there is a need for a supplementary parameter that controls the smallest distance corresponding to the decreasing part of the kernel. An example of

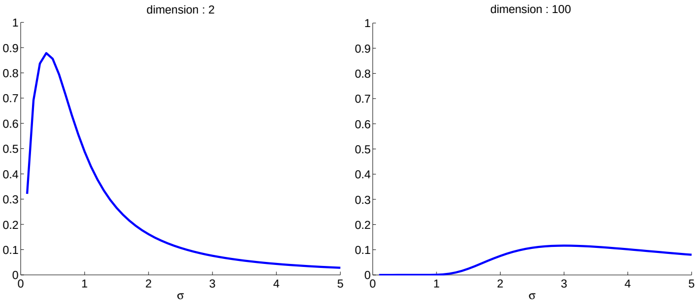


Fig. 3. Contrast (see definition in text) in a 2-dimensional (left) and a 100-dimensional (right) uniform distribution, with respect to the kernel standard deviation.

kernel that fulfills this requirement is the p-Gaussian kernel :

$$K(x, y) = \exp(-d(x, y)^p / \sigma^p),$$

where p and σ are the two parameters. Normalizing coefficient for density estimation can be found in [Kassam, 1988], but once again this is not needed for measuring similarities. Figure 4 (left) shows an example of p-Gaussian kernel, width $p = 11$ and $\sigma = 4.3$. It is seen that the kernel slope effectively covers the range of distances, according to the definition of ideal kernel

The method to set adequate values to p and σ can easily be deduced from the same requirements. As the decreasing slope of the kernel has to cover the effective range distances in the histogram built on the sample distribution, two equations can be deduced once this range is known: one for the lowest value of the range, one for the highest one. Of course, as we are speaking about distributions, taking extreme values is not a good idea; rather, for example, the 5% and 95% percentiles of the distribution should be estimated. Let d_N and d_F be these two values respectively. Then two equations can be written by making the p-Gaussian kernel evaluated at d_N (resp. d_F) equal to 95% (resp. 5%) of the full kernel range :

$$p = \frac{\ln\left(\frac{\ln(0.05)}{\ln(0.95)}\right)}{\ln \frac{d_F}{d_N}} \quad ; \quad \sigma = \frac{d_N}{(-\ln(0.05))^{1/p}} = \frac{d_F}{(-\ln(0.95))^{1/p}}$$

Figure 4 (right) shows the results of the experiment described earlier to estimate the contrast, with respectively, the Gaussian kernel and a kernel with optimized p and σ values.

5 Conclusion

Local kernels or functions are used in many data analysis paradigms and algorithms, such as Radial-Basis Function networks, Support Vector Machines,

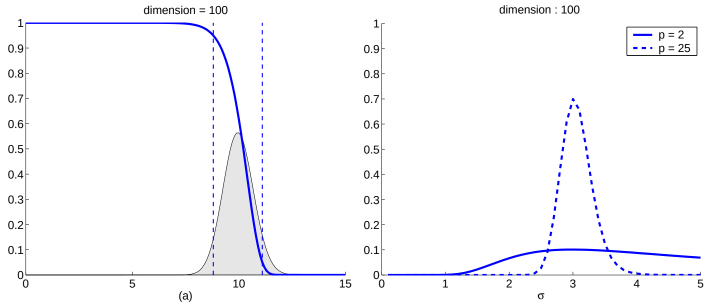


Fig. 4. (left) Kernel values along with distance distribution for the ideal kernel, $p = 11$; (right) contrast for Gaussian kernel and ideal kernel in dimension 100.

some Vector Quantization methods, etc. Locality is used as a way of interpretation, and also to provide measures of similarities between data. In this paper, we show that the widely used Gaussian Kernel is appropriate to represent similarities in low-dimensional spaces, but fails to fulfill this goal in high-dimensional ones. When similarities cannot be expected anymore to be measured adequately, many problems may be expected, for example in nearest neighbor search. The numerical stability of the methods may be lost.

p-Gaussian kernels are presented as an alternative to Gaussian kernels. An additional parameter makes it possible to keep the effective part of the Gaussian slope in the effective part of the distribution of distances between data. In this way, p-Gaussian kernels will adequately discriminate small and large distances between pairs of data even in a high-dimensional setting, a task that Gaussian kernel fails to fulfill. A methodology is presented to set the parameters according to a specific data sample. Future work will consist in using such flexible kernels in learning algorithms for high-dimensional data.

References

[Bishop, 1995]C. M. Bishop. *Neural Networks for Pattern Recognition.* Oxford university press, 1995.

[Cristianini and Shawe-Taylor, 2000]N. Cristianini and J. Shawe-Taylor. *An Introduction to Support Vector Machines.* Cambridge University Press, 2000.

[Golub and van Loan, 1996]G. H. Golub and C. F. van Loan. *Matrix Computations.* Johns Hopkins University Press, 3rd edition, 1996.

[Hwang and Bang, 1997]Y.-S. Hwang and S.-Y. Bang. An efficient method to construct a radial basis function neural network classifier. *Neural Networks*, 10(8):1495–1503, 1997.

[Kassam, 1988]S. A. Kassam. *Signal Detection in Non-Gaussian Noise.* Springer-Verlag, 1988.

[McLachlan and Peel, 2000]G. J. McLachlan and D. Peel. *Finite Mixture Models.* Wiley, 2000.

[Parzen, 1962]E. Parzen. On estimation of a probability density function and mode. *Ann. Math. Stat.*, 33:1065–1076, 1962.

Quality measure based on Kohonen maps for supervised learning of large high dimensional data

Elie Prudhomme and Stéphane Lallich

Laboratoire E.R.I.C, Université Lumière Lyon 2
5, avenue Pierre Mendès-France,
69676 BRON Cedex, France
(e-mail: stephane.lallich@univ-lyon2.fr,
elie.prudhomme@etu.univ-lyon2.fr)

Abstract. In supervised learning, the prediction of the class is the ultimate goal. On a broader basis, a good learning methodology is expected to (1) enable a representation of the data in order to facilitate user's navigation within the data set and (2) contribute to the choice of examples and attributes, while ensuring a structured, understandable prediction. Various studies have shown how the so-called neighbourhood graph, from the predictors, gives ground to such a methodology (e.g.: the relative neighbourhood graph of Toussaint). However, the construction of such a graph ($\mathrm{O}(n^3)$) remains complex. Moreover, when the number of dimensions increases, distance becomes hard to compute and lose their selectivity.

In the case of large high dimensional dataset, we propose to substitute a self-organized map built on the predictors to the neighbourhood graph. After a short reminder on the principles of the SOM for unsupervised learning, we analyse how it can found an optimized strategy of learning. Then we propose to use original statistics (narrowly correlated with the error in generalization) in order to assess the level of quality of this strategy. Diverse experiments highlight the feasibility of this approach, therefore reliable criterion are available for us to select relevant examples and attributes.

Keywords: supervised learning, Kohonen maps, statistical validation.

1 Motivation

Supervised learning methods of a categorical variable aim at predicting the class of a new instance from a sample of labelled examples. Indeed, prediction is only a step in the learning process, which is enriched through the exploratory analysis of the data. This allows to clean and transform the data, to select features and subsets of records, and to detect outliers, while integrating possible contextual information.

In such a perspective, resorting to neighbourhood graphs brings an effective solution. One builds the neighbourhood graph based on the predictors, for example the Relative neighbourhood Graph of Toussaint [1980] (*RNG*). The vertices of the graph are then colored according to the class they belong to. To find the class of a new instance, it is first inserted in the neighbourhood graph and then it is attributed the majority class among its neighbours on the graph. Various studies proposed a statistic: the cut edge weight statistic. This statistic evaluates the predictive capacity of a neighbourhood graph. It also allows for the selection of relevant variables or for the detection of outliers by spotting the impact of an example or of a variable on the predictive capacity of the graph ([Sebban, 1996], [Zighed *et al.*, 2001], [Lallich, 2002], [Muhlenbach *et al.*, 2003], [Zighed *et al.*, 2004]).

In comparison with the k-Nearest Neighbour method (kNN), neighbourhood graphs adapt the number of nearest neighbours to the local topology. On those graphs, the cut edge weight statistic that evaluates their predictive capacity is strongly correlated to their error rate in generalization. Their results in generalization are at least as good and they have the advantage of establishing an effective procedure of navigation within the basis of examples. Furthermore, the neighbourhood graph allows to navigate efficiently in the database, making the exploratory analysis of the data easier.

Neighbourhood graphs present a double difficulty when confronted to large high-dimensional datasets. Firstly, their great complexity - $O(n^3)$ for Relative Neighbourhood Graphs of Toussaint - makes them poorly adapted to very large datasets. The second issue is linked to the curse of dimensionality which triggers a loss of selectiveness of euclidean distance.

Faced with this double difficulty, we propose to replace the *RNG* issued from the predictors with a Self Organized Map (*SOM*). We thus get a representation of the information given by the predictors. That method has the advantage of preserving the local topology in case of high dimensional data while using a complexity which varies linearly with the number of examples. The advantages of neighbourhood graphs are also maintained in the *SOMs*: especially the spatialisation of the information obtained from the predictors and the efficient navigation in the database.

In this article, we show that it is possible to construct a cross-product statistic which is closely linked to the predictive ability of the map in generalization. This statistic has the advantage of helping us in data preparation, especially to select relevant variables or detect outliers. After presenting the notations we used (see section below), we introduce the SOM algorithm and its use in supervised learning (section 3). Then we present our cross-product statistics estimates in *SOM* (section 4). Their validation on different datasets is presented in section 5.

2 Notations

- m: number of examples, d: number of predictors, p: number of classes, n: number of neurons.
- X: (m, d) matrix of data; line i corresponds to example i and column j to predictor j.
- y: vector with m components indicating the class of each example.

- W: (n,d) matrix of general term w_{ij}, designating the weight of neuron i for predictor j.
- c: vector with n components indicating the class of each neuron; $c_i = 0$ if neuron i is ambiguous, $c_i = -1$ if neuron i is empty.
- $bmu_i = \arg\min_r \|w_r - x_i\|$, index of *best matching unit*, the nearest neuron to example i.
- $dist_c(r,q)$: distance between neurons r and q, according to the map.
- $dist_p(r,q)$: Euclidean distance between the weights of neurons r and q.
- PPV: (n,n) symmetrical matrix of general term ppv_{ij}, worth 1 if $dist_c(i,j) \leq \max\left(dist_c(i,k), dist_c(j,k)\right), \forall k, k \neq i, k \neq j; c_i, c_j, c_k \neq -1$ $(i,j$ connected), 0 otherwise; ppv_{r+} represents the number of neurons connected to neuron r.

3 *SOM* and supervised learning

The Self Organized Map allows i) a fast unsupervised learning of input examples and ii) their representation. The map is built on a uniform distribution of neurons in 2 or 3 dimensions. Each neuron is associated to a vector in the space of the example. Originally, that association was called a model. During the learning, the input examples are successively presented to the map. Assuming a general distance measure between inputs and models (usually euclidian distance), the neuron the nearest to the input (called the Best Matching Unit) is modified with its neighbourhood so that all of them get closer to the input example.

The iterative algorithm for the input example i at time t is summarized by the following formula updating the weights W of the neuron r:

$$w_r^{t+1} = w_r^t + h_r^t \times (x_i - w_r^t)$$

where $h_r^t = \alpha^t \times v_r^t$, with α^t the learning-rate factor and v_r^t the neighbourhood function which represents the size of the modified neighbourhood. Both α^t and v_r^t are monotonically decreasing as a function of time.

This algorithm ensures a local preservation of the topology through a non linear projection. Thus, after learning, two close input examples will have close models on the *SOM*. Nevertheless this non linear projection is particular in the sense that it does not preserve the distances from the input space.

Because of those properties (fast algorithm and topology preservation) some authors have adapted them to a supervised learning. The most popular of those algorithms is the *LVQ* proposed by Kohonen [1988]. Here, the classes of the input examples are used to control the modification of the models. Another idea is used by Midenet [1994] in the LASSO model. In that case, the classes are used during the learning phase in the same way as other input variables. Two phenomena result in the use of classes during learning. First, the prediction is more robust: more information is used. But at the same time the local topology preservation is changed. It is not simply a function of the input variables (as in the original *SOM* algorithm) but also a function of the classes.

To avoid that problem some authors have proposed a different approach. On that account, the class of the input example is only used after a classical learning of the *SOM* on the input variables. During that second step, the neurons take the class of the inputs they represent. The reverse happens during prediction: the class of a new input example is determined by the class of the best matching unit of that example. Three methods use that principle: Kohonen-KNN [Zupan *et al.*, 1994], Kohonen-WI [Song and Hopke, 1996] and Kohonen-Opt [Prudhomme and Lallich, 2005]. There are at least two cases which show the difference between those three approaches. First empty neurons: after the learning phase some neurons do not match any input example. Secondly ambiguous neurons: after the learning phase some neurons match the same proportion of examples from different classes. So each method proposes a way of predicting a new example which matches one of those two type of neurons. Prudhomme and Lallich [2005] have shown that Kohonen-Opt generally gives better results in generalisation than the others. Moreover, the results obtained with Kohonen-Opt on different datasets are almost equivalent to those obtained by the ID3 method of classification.

Consequently, *SOMs* could be used in supervised learning. In that case there is a double advantage. First the non linear projection is particulary adapted to high dimensional spaces. It allows a dimension reduction based on the most significant feature. Secondly the examples are synthetically represented by the models. Thus the *SOM* representation is well adapted to large datasets. In the rest of the document we propose a statistic which takes advantage of those two points in order to assess the predictive capacity of the *SOM*. Because this statistic is based on the neighbourhood, distance preservation is not mandatory.

4 Quality measures for SOM under supervised learning

We therefore suggest a learning strategy that relies on the construction of the *SOM*. The reliability of the *SOM*, reagrdless of any consideration of class, can be assessed through various statistical tools proposed notably by [Bodt *et al.*, 2002]. We suggest here an assessment of the predictive ability of the *SOM* through different statistics. We will experimentally show the strong correlation of those statistics with the precision in generalization. Similarly to the cut edge weight statistic worked out for neighbourhood graphs [Lallich, 2002], those different statistics are based on the notion of cross-product statistic [Mantel, 1967]. Thus they are constructed as the scalar product of two proximity measures, the first one depending on the predictors and the other one depending from the class.

4.1 Definition of J type statistics

To assess the strength of the link between proximity in the sense of the map and proximity in the sense of the classes, one can reason about examples or neurons. Reasoning about neurons helps to deal with a large amount of examples.

When reasoning on examples, the proximity between examples based on the map is assessed by the matrix T' of general term t'_{ij}, which is worth 1 if the examples i and j are represented by the same neuron, and 0 otherwise. In order to take into account the topological properties of the map, one also can resort to the matrix T'' of general term t''_{ij}, which is worth 1 if the examples i and j are represented by the same neuron, $norm(dist_p(w_{bmu_i}, w_{bmu_j}))$ if i and j are represented by adjacent neurones ($i.e.\ dist_c(bmu_i, bmu_j) = 1$), and 0 otherwise. The proximity between examples based on the class is assessed by the matrix U of general term U_ij, which is worth 1 if the examples i and j do not have the same class (i.e $c_i \neq c_j$), and 0 otherwise.

When reasoning on neurons, the proximity between neurons based on the map is assessed by the matrix T''', of general term t'''_{ij}, which is worth $norm(dist_p(w_i, w_j))$ if $ppv_{ij} = 1$, and 0 otherwise. The proximity between neurons based on the class is assessed by the matrix R, of general term r_{ij}, which is worth 1 if the neurons i and j do not have the same class, 0 otherwise.

As a result, one will obtain three different statistics, J', J'' and J''' which are defined below.

J’	**J”**	**J”’**
$\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} T'_{ij}U_{ij}$	$\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} T''_{ij}U_{ij}$	$\frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m} T'''_{ij}R_{ij}$

The following simplifying notations are used, where T can take the value of T',T'' or T''' and the sums finishing respectively in m for the two former cases and in n for the latter one:

S_0	S_1	S_2
$\sum_{i=1}\sum_{j=1} t_{ij}$	$\frac{1}{2}\sum_{i=1}\sum_{j=1}(t_{ij} + t_{ji})^2$	$\sum_{i=1}(t_{i+} + t_{+i})^2$

J Type statistics vary between 0 and $\frac{1}{2}S_0$. They are the weakest when the link between proximity according to the class and proximity according to the map is strongly positive. They may be standardized by forming $2J/S_0$ which varies between 0 and 1.

Dataset	**Variables**	**Classes**	**Example**	**Dimensions**	**Times**
(1) Abalone	8	29	4177	25 × 25	90000
(2) Balance Scale	4	3	625	15 × 15	60000
(3) Breast Cancer	9	2	699	20 × 20	90000
(4) Glass Indent	9	6	214	10 × 10	10000
(5) Haberman	3	2	306	10 × 10	20000
(6) Ionosphère	34	2	351	10 × 10	20000
(7) Iris	4	3	150	10 × 10	2000
(8) Italian Olive Oil	9	9	572	15 × 15	45000
(9) Liver	6	2	345	10 × 10	35000
(10) Yeast	8	10	1484	25 × 25	90000

Table 1. dataset and associate parameters

4.2 Meaning of J type statistics

In order to know to which extent the evaluation given by J is not due to chance, a random multinomial outline was defined. The null hypothesis (H_0) was that the examples (the neurons) are labelled independently from each other, with the same probability distribution $(\pi_r)_r$ where π_r denotes the frequency of the class $y_r, r = 1, 2, \ldots, p$.

The significance of the observed value of J is appraised with its left unilateral p-value. This is the probability of getting a value of J as extreme as or more extreme than the observed one if H_0 is true. That calculation can be done either by simulation or more quickly by normal approximation [Cliff and Ord, 1981]. In the last case, we have to calculate $\mu = E(J/H_0)$ and $\sigma^2 = Var(J/H_0)$. It is easy to calculate $\mu = S_0 \sum_{r=1}^{p-1} \sum_{s=r+1}^{p} \pi_r \pi_s$. One can find in Lallich [2002], following Cliff and Ord [1981], the calculation of the variance σ^2, which depends on S_0, S_1 and S_2.

5 Experiment

Those different statistics were tested on 10 datasets, coming from the repository of the University of Irvine [Blake and Merz, 1998] (except for one *Italian Olive Oil* which is from [Hopke and Massart, 1993]). Table 1 details those datasets in terms of the number of input variables for each example, the number of classes and the number of input examples in each dataset. This table also summarizes some parameters of the *SOM* used for learning: the total number of input examples presented (called time) and the size of the *SOM.* The algorithm used for learning is the classical one presented in section 3. Table 2 shows the value of each statistic, their associated *p-value* and the error rate in generalization with the *Kohonen-Opt* method.

The *p-values* are significant ($p < 0,05$) for J'', and for J' (except for *Haberman* where $p = 0.08$). Thus they are sufficiently robust to assess the

Base	$2J'/S_0$	**p-value**	$2J''/S_0$	**p-value**	$2J'''/S_0$	**p-value**	**Error**
(1)	79,96	0	79,88	0	81,97	1	73,86
(2)	21,66	0	23,68	0	22,28	0	17,3
(3)	0,40	0	1,10	0	8,30	0	3,21
(4)	43,29	0	44,64	0	61,65	0,02	34,21
(5)	34,57	0,078	32,51	0,005	28,20	0,022	24,06
(6)	10,92	0	14,76	0	36,11	0,022	11,6
(7)	3,60	0	4,70	0	8,50	0	4,67
(8)	41,40	0	8,05	0	20,06	0	7,69
(9)	60,58	0,0031	0,3750	0	58,28	0,9989	37,53
(10)	50,00	0	52,40	0	64,52	0	47,53
Means	31,64	0,0081	29,92	0,0005	27,99	0,2045	

Table 2. Statistic, their associated *p-value* and error rate in test with Kohonen-Opt

quality of the representation built by the *SOM*. In the case of J''', two *p-values* are almost equal to 1. For this statistic, the link between two ambiguous neurons is a cut edge one. In the two cases, the graph extracted from the *SOM* has a many ambiguous neurons. So, in the statistic sense, the class of the neuron is independent from the topology of the map. For that reason, the *p-value* is high. In fact, this happened only when the error rate was high too.

A more interesting property is the correlation between that statistic and the error rate in generalization. r^2 of this correlation is respectively 0.78, 0.98 and 0.88 for J', J'' and J'''. J' just takes into account the input example of different classes matching the same neurons. So this statistic does not use the information contained in the local topology of the *SOM*. That information is used by J''. For that reason that statistic has a better correlation with the error rate. The correlation between J''' and the error rate is intermediate. That statistic takes into account the local topology of the *SOM* thanks to the the neighbourhood graph which was built on the map. On the other hand, the input examples are not used. Therefore some information is lost during the projection of the input space on the map. However the estimation of that statistic has a low complexity as only the neurons are used. In the case of datasets composed by a high number of examples, it is an interesting property. On the contrary, J'' needs the examples.

Moreover, we have tested the capacity of that approach to be applied on large datasets. Therefore, we used Wave [Blake and Merz, 1998], which allows to randomly generate a user fixed number of input examples. For each generated dataset, the error rate in generalization is know and constant. We applied Kohonen-Opt and our statistics on different datasets containing 5 000 to 1 280 000 examples. The learning time was reported on table 3. The *SOM* used for each dataset is the same and the test was made on the same dataset of 100 000 examples, never used in learning.

The table 3 shows the results. First, the error rate in generalization is stable regardless of the number of input examples (approximatively 15%). Secondly the time needed for learning increases linearly from a factor 2 (like the number of examples). Finally, the statistics ($2J/S0$) are stable too with a little decrease when the number of examples increases. Their *p-values* are always equal to 0.

This experiment shows that the quality of the learning by the *SOM* does not decrease when the number of examples increases. Thus they could be used in the case of large datasets. This is also the case for the proposed statistics which are relatively stable.

Size	$2J'/S_0$	$2J''/S_0$	$2J'''/S_0$	**Error**	**Time (s)**
1250	26,67	26,27	17,80	16,03	2
2500	26,08	26,09	13,39	15,70	5
5000	24,57	24,92	9,87	15,46	10
10000	24,45	24,36	9,44	15,57	20
20000	23,25	23,68	7,42	14,92	41
40000	22,65	23,04	7,78	14,84	78
80000	22,64	23,22	7,40	15,25	127
160000	22,53	23,03	7,45	14,93	245
320000	22,94	23,37	7,92	15,04	500
640000	22,54	23,09	7,18	15,17	1073
1280000	22,32	22,94	7,98	15,22	2014

Table 3. Statistic, their associate *p-value* and error rate in test with Kohonen-Opt on different Waves dataset

Finally, we have tested the capacity of that approach on high dimensional datasets. Here we use the Forest CoverType dataset [Blake and Merz, 1998]. That dataset presents 54 input variables for 8 classes. Moreover the classification performance on that dataset is known. It was obtained by Blackard [1998] for neural networks and linear discriminant analysis.

Table 4 shows those results and those obtained with Kohonen-Opt. A direct application of Kohonen-Opt on this dataset gives poor results. To avoid that problem, a normalization of the attributes was carried out i) with the Milligan and Cooper (*MC*) procedure [1988] and ii) with a standardization by removing the mean and dividing by the standard deviation (s). Since attributes are both boolean and continuous, the MC procedure gives better results. In that case, the error rate is in the same order as the one obtained by the neural network. That result tends to show that the learning based on the *SOM* is robust when the number of input variables increases.

	Kohonen-Opt			**Other**	
Method	None	s	MC	ANN	linear discriminant
Error Rate	45,7	43,4	32,2	30	42

Table 4. Result in classification task on Forest CoverType dataset

6 Conclusion

SOMs are popular algorithm in unsupervised learning. Their complexity is linear with the number of example and they allow for a data exploration [Lechevallier, 2002]. In that paper we suggested that they can be used in supervised learning. In that case *SOMs* synthesize the information of the predictors through a non linear projection and enable a navigation through the dataset. Even if that non linear projection does not maintain the distance, it is nevertheless a way to assess our statistic $(2J/S_0)$ which is correlated to the error rate.

In further work we want to use that statistic for outliers detection and feature selection from large high dimensional datasets. In addition, we want to test the effect of the choice of the distance on the learning process. We hope to show that fractional distance metrics are more useful than euclidian distance to learn high dimensional datasets with *SOMs*, as it is the case for k-means [Aggarwal *et al.*, 2001].

References

[Aggarwal *et al.*, 2001]Charu C. Aggarwal, A. Hinneburg, and D. A. Keim. On the surprising behavior of distance metrics in high dimensional space. *Lecture Notes in Computer Science*, 1973:420–434, 2001.

[Blackard, 1998]Jock A. Blackard. *Comparison of Neural Networks and Discriminant Analysis in Predicting Forest Cover Types.* Ph.d. dissertation, Department of Forest Sciences. Colorado State University., Fort Collins, Colorado, 1998.

[Blake and Merz, 1998]C.L. Blake and C.J. Merz. UCI repository of machine learning databases, 1998.

[Bodt *et al.*, 2002]E. Bodt, M. Cottrell, and M. Verleysen. Statistical tools to access the reliability of the self organizing maps. *Neural Network*, 15:967–978, 2002.

[Cliff and Ord, 1981]A. D. Cliff and J. K. Ord. *Spatial processes, models & applications.* London, 1981.

[Hopke and Massart, 1993]P. K. Hopke and D. L. Massart. Reference data sets for chemometrical methods testing. *Chemometrics and Intelligent Laboratory Systems*, 19:35–41, 1993.

[Kohonen, 1988]T. Kohonen. Learning vector quantization. *Neural Network*, 1:303, 1988.

[Lallich, 2002]S. Lallich. *Mesure et validation en extraction des connaissances à partir des données.* Habilitation à diriger les recherches, Université Lumière Lyon 2, Lyon: France, 2002.

[Lechevallier, 2002]Y. Lechevallier. Construction de super-classes à partir de la carte de kohonen et indicateurs de qualité de cette carte, séminaire laboratoire eric, http ://www-sop.inria.fr/axis/talks/~eric/, 2002.

[Mantel, 1967]N. Mantel. The detection of disease clustering and a general regression approach. *Cancer Res.*, 27:209–220, 1967.

[Midenet and Grumbach, 1994]S. Midenet and A. Grumbach. Learning associations by self-organisation : the lasso model. *Neurocomputing*, 6:343–361, 1994.

[Milligan and Cooper, 1988]G. W. Milligan and M. C. Cooper. A study of standardization of variables in cluster analysis. *Journal of Classification*, 5:181–204, 1988.

[Muhlenbach *et al.*, 2003]F. Muhlenbach, S. Lallich, and D.A. Zighed. Identifying and handling mislabelled instances. *Journal of Information Intelligent Systems*, 22:89–109, 2003.

[Prudhomme and Lallich, 2005]E. Prudhomme and S. Lallich. Validation statistique des cartes de kohonen en apprentissage supervisé. In *5èmes Journées Francophones d'Extraction et de Gestion des Connaissances (EGC 05), Paris*, Revue des Nouvelles Technologies de l'Information, Janvier 2005.

[Sebban, 1996]M. Sebban. *Modèles théoriques en reconnaissance des formes et architecture hybride pour machine perceptive.* Thèse de doctorat en informatique, Université Lumière Lyon 2, Lyon: France, 1996.

[Song and Hopke, 1996]X-H. Song and P. K. Hopke. Kohonen neural network as a pattern recognition method based on the weight interpretation. *Analytica Chimica Acta*, 334:57–66, 1996.

[Toussaint and Menard, 1980]G. T. Toussaint and R. Menard. Fast algorithms for computing the planar relative neighborhood graph. In *Methods of Operations Research, Proceedings of the Fifth Symposium on Operations Research*, pages 425–428, 1980.

[Zighed *et al.*, 2001]D. A. Zighed, S. Lallich, and F. Muhlenbach. Séparabilité des classes dans r^p. In *Actes du VIIIème Congrès de la Société Francophone de Classification (SFC'01)*, pages 356–363, 2001.

[Zighed *et al.*, 2004]D. A. Zighed, S. Lallich, and F. Muhlenbach. A statistical approach of classes separability. In H. Mannila et H. Toivonen T. Elomaa, editor, *Revue Applied Stochastic Models in Business and Industry*, pages 475–487. Springer-Verlag, 2004.

[Zupan *et al.*, 1994]J. Zupan, M. Novic, X. Li, and J. Gasteiger. Classification of multicomponent analytical data of olive oils using different neural networks. *Analytica Chimica Acta*, 292:219–234, 1994.

Exploratory Data Analysis Leading towards the Most Interesting Binary Association Rules

Alfonso Iodice D'Enza[1], Francesco Palumbo[2⋆], and Michael Greenacre[3]

[1] Department of Mathematics and Statistics - Università di Napoli Federico II
Complesso Universitario di Monte S. Angelo, Via Cinthia
I-80126 Napoli, Italy
(e-mail: iodicede@unina.it)
[2] Department of Economics and Finance - Università di Macerata
Via Crescimbeni, 20
I-62100 Macerata, Italy
(e-mail: francesco.palumbo@unimc.it)
[3] Department of Economics and Business - Universitat Pompeu Fabra
Ramon Trias Fargas, 25-27
E-08005 Barcelona, Spain
(e-mail: michael.greenacre@upf.edu)

Abstract. Association Rules (AR) represent one of the most powerful and largely used approach to detect the presence of regularities and paths in large databases. Rules express the relations (in terms of co-occurence) between pairs of items and are defined in two parts: *support* and *confidence*. Most techniques for finding AR scan the whole data set, evaluate all possible rules and retain only rules that have support and confidence greater than thresholds, which should be fixed in order to avoid both that only trivial rules are retained and also that interesting rules are not discarded. This paper proposes a two steps interactive, graphical approach that uses factorial planes in the identification of potentially interesting items.
Keywords: Association Rules, Classification, Binary variables.

1 Introduction

Association rules (AR) [Agrawal *et al.*, 1993] represent a suitable data mining tool to identify frequently occurring patterns of information in large data bases. The classic field of application of AR mining is *market basket analysis* (MBA); in this context, data are stored as *transactions*: each transaction is a binary sequence that records the presence/absence of a set of p features. The basic data structure consists of an $n \times p$ Boolean matrix; the terms of the table are 1 and 0, which correspond to the states *presence* and *absence*, respectively. MBA is one of the first and better known application fields of AR mining: however, the binary structure of the data makes AR applicable in

[⋆] *Acknownledgements:* The paper was financially supported by PRIN2003 grant: "*Multivariate Statistical and Visualization Methods to Analyze, to Summarize, and to Evaluate Performance Indicators*" research unit of University of Macerata, responsible F. Palumbo.

many different contexts. AR are largely used in text mining, image analysis and microarray data analysis.

An AR in its simplest form involves a pair of items playing different roles: the antecedent part of the rule (*body*), and the consequent part of the rule (*head*). The relation characterizing the considered items is usually expressed in two different measures: *support* and *confidence*. The support is the intensity of the association between the considered items; the confidence measures the strength of the logical dependence expressed by the rule.

An example illustrates an AR in its simplest form: let A and B be a pair of items, *beer* and *chips* for example, the following notation represents the achieved AR:

$$A \Longrightarrow B = \{\text{support} = 20\%, \text{confidence} = 80\%\} .$$

The rule shows that 20 percent of customers buy both beer and chips, and if a customer buys beer, in 80 percent of cases buys chips too.

Simple rules does not represent the whole output of a mining process, since complex AR are extracted too. In its most general definition a complex AR is a rule in which there are one or more antecedent items and one or more consequents. Complex AR represent a more powerful tool, but they are even harder to handle.

The large size of the starting data matrix implies very high computational efforts in mining rules, and it can lead to a massive quantity of rules. Many proposed algorithms lead to mine rules in more and more efficient ways, but the identification and selection of interesting rules remains a still opened problem. It requires the definition of consistent criteria to avoid the risk that the truly interesting information is hidden by the presence of trivial and redundant rules.

The great part of the contributions in the AR literature is aimed to the implementation of algorithms for the generation of AR in reduced computational costs and output amount, as well as for the generalization of AR to categorical and numerical data.

The reference point among these algorithms is the *Apriori*, introduced by [Agrawal and Srikant, 1994]. This algorithm consists of two phases: the *frequent itemset mining* phase and the properly defined *association rule mining* phase. An itemset is frequent if the items involved in it co-occur with a frequency greater than a user-defined threshold, in other words the itemset support is greater than the assigned minimal support threshold. In the second phase, there is the generation of all the possible rules deriving from the itemsets previously selected: the output rules provided by the procedure are those characterized by confidence exceeding a minimal confidence threshold.

Based on the same idea of the Apriori, the AprioriTid [Agrawal and Srikant, 1994] introduces an encoding of the founded frequent itemsets in order to reduce the computational effort.

The AprioriHybrid is a combination of both Apriori and AprioriTid, in the earlier and the latter iterations respectively [Agrawal and Srikant, 1994].

Other interesting algorithms are the DHP (*Direct Hashing and Pruning*) [Park *et al.*, 1995], the Partition algorithm [Savasere *et al.*, 1995], the DIC (*Dynamic Itemset Counting*) algorithm [Brin *et al.*, 1997] and the FP-growth (*Frequent Pattern growth*) algorithm [Han *et al.*, 2000] for which we refer the reader to the bibliography. The procedures above are applicable to Boolean data. To generalize the algorithms to numerical and categorical AR, a previous recoding of the data is required: in this sense the contribution of [Srikant and Agrawal, 1996] and [Miller and Yang, 1997]. A last class of proposals is aimed to solve a common problem of AR mining, the huge number of rules generated, selecting and identifying the interesting generated AR. The contributions belonging to this class, like [Zaki, 2000] and [Liu *et al.*, 1999], are characterized by a same target as well as approaches of different nature.

Almost all AR mining algorithms use thresholds whose settings are related to a trade-off: tight thresholds can cause loss of interesting information; otherwise, loose thresholds cause an excessive number of uninteresting rules to be selected. In addition, notice that the reduction of output rules is anyway linked to the generation of all potential rules.

The present paper proposes an exploratory strategy to identify *a priori* items that are potentially interesting as antecedent or consequent parts of rules. The method does not produce AR but provides the user with information about the most probably interesting items. One of the above mentioned algorithms must be used, focusing the attention only to those items that the procedure *indicated* as interesting, to obtain the AR set. Reducing the number of considered items, the user can define looser thresholds, avoiding as well the risk of a huge amount of rules. In addition, information about the items help the user to pay greater attention to the rules containing the previously identified items.

The proposed strategy exploits the graphical and analytical capabilities of multidimensional data analysis (MDA), in particular the paper will focus on the computational aspects when n and p are large.

The procedure deals with the following data structures: let $\mathbf{Z}$ be a $(n \times p)$ presence/absence matrix characterized by n binary sequences considered with respect to p Boolean variables; in our application the n rows refer to baskets of items purchased and the p columns refer to the items (with coding $1 =$ *buy* and $0 =$ *not buy*). Let $\mathbf{S} = n^{-1}\mathbf{Z}^{\mathsf{T}}\mathbf{Z}$ be a symmetric contingency table, whose general extra-diagonal term $s_{jj'} = s_{j'j}$ represents the support with respect to the items j and j', in our case, the relative frequency of purchasing pairs of items. The asymmetric square matrix $\mathbf{C}$ is defined as $\mathbf{C} = \mathbf{Z}^{\mathsf{T}}\mathbf{Z}\mathbf{D}^{-1}$, where $\mathbf{D}$ is a diagonal matrix having general term $d_{jj} = s_{jj}$ $(j = 1, \ldots, p)$. Matrix $\mathbf{C}$ has the general diagonal term $c_{jj} = 1$ while for $j \neq j'$ the term $c_{jj'}$ corresponds to the confidence of the rule $\{A_j \Longrightarrow A_{j'}\}$.

In Section 2 we present the steps of the strategy and the related tools: clustering phase (subsection 2.1), items selection according to the supports (subsection 2.2), identification of rules bodies and heads (subsection 2.3). In section 3 we present an example on the BMS-Webview-2 dataset.

2 Multidimensional data analysis (MDA) approach

The proposed strategy consists of two main phases and aims at generating a reduced number of AR; in particular the phases are:

i) partitioning of the considered binary sequences (transactions) in homogeneous classes;
ii) selecting interesting items and visually representing the interesting rules using MDA techniques.

In the case of huge data sets the applicability of the whole procedure strongly depends on the the first phase of classification: this step is very expensive in terms of time elapsed and memory required. Thus it is necessary to choose an algorithm increasing speed and efficiency of the whole procedure.

Once the group are determined, the attention is focused on the supports and confidences matrices of order $p \times p$ related to each group. In addiction, our proposal is based on a suitable factorial analysis on these matrices that requires short computing time and low memory usage: the most time consuming phase consists in the singular value decomposition of symmetric matrices. Moreover, partitioning the data in groups permits to perform the analysis on parallel computing architectures. The aim is to select the most occurring pairs of items in each group and to assign the role of antecedent or consequent to the set of selected items.

2.1 Clustering transactions

The general aim of clustering techniques is to partition the statistical units of a given data set in disjoint classes such that similar units are grouped together. Dealing with large, high dimensional and sparse data sets, classic clustering techniques like K-means algorithm [Hartigan, 1975] and *agglomerative* algorithm require very high computational costs and do not guarantee reliable solutions. Thus, in the literature, there are many contributions proposing algorithms optimized for massive amounts of binary data: the ROCK algorithm proposed by [Guha *et al.*, 2000], that is based on *links* and represents an agglomerative hierarchical clustering; QROCK that is a speeded up version of ROCK, while a density based algorithm considering links is the SNN (*shared nearest neighbors*). Two of the non-hierarchical clustering algorithms for binary data are LWC (*light weight clustering*) and *incremental* K-means proposed by [Gaber *et al.*, 2004] and [Ordonez, 2003], respectively: the main aim of both these procedures is clustering of data streams, which are flows of binary sequences. In this paper, however, the procedure is applied on a finite set of binary sequences.

The first step of the proposed strategy is then implemented exploiting the features of one of the above procedures, incremental K-means that is a non-hierarchical algorithm and it is characterized by doing a single iteration to get the partition of the rows of the binary matrix $\mathbf{Z}$ in K disjoint classes. The logical distance between the binary rows of $\mathbf{Z}$ is measured trough the Jaccard coefficient. The incremental K-means takes as input the number of clusters that is hence user-defined. The output provided by the procedure consists of: the partition matrices $\mathbf{Z}_k$ $(n_k \times p)$, where $k = 1, \ldots, K$; the centroid matrix $\mathbf{C}$ that is $(K \times p)$, with cluster centroids on rows; a K-elements vector $\mathbf{w}$ of cluster weights such that $\mathbf{w}_k = \frac{n_k}{n}$; a $(K \times p)$ matrix $\mathbf{R}$ of squared distances.

The initialization phase of *Incremental* K-means presents a difference with Standard K-means: instead of using k random sequences as centers, this algorithm exploits global statistics (mean and variance) of the input indicator matrix to obtain the starting centers. Furthermore, the procedure does not update the centroid matrix $\mathbf{C}$ and the cluster weights vector $\mathbf{w}$ at every binary sequence but every (n/L) times, where n and L are the number of considered sequences and an initialization parameter, respectively. The reader is referred to [Ordonez, 2003] for further details about the procedure.

2.2 Selection of interesting items

The previous step defined a partition of $\mathbf{Z}$ in $\mathbf{Z}_k$, with $k = 1, \ldots, K$; for each of the $\mathbf{Z}_k$ matrices, supports $(\mathbf{S}_k)$ and confidences $(\mathbf{C}_k)$ are defined. AR mining is hence referred to each group of homogeneous sequences. In particular, through the analysis of each $\mathbf{S}_k$, the most occurring pairs of items within the *k-th* group are selected; while through the analysis of each $\mathbf{C}_k$, the procedure assigns the role of antecedent or consequent to each one of the selected items. The selected pairs of items resulting by the analysis of $\mathbf{S}_k$ represent evident relations characterizing groups of considered binary sequences: these hidden patterns can be missed using general support thresholds. The pairs of items characterized by a degree of co-occurrence that is high in one or more of the K groups and low with respect to the whole data are then considered non-trivial. The criteria used to define what is "high" or "low" are based on the ratio between the most occurring supports inside the groups and the total supports. Proper statistical tests can be adopted to exploit the task; however, this paper does not focus on this aspect.

2.3 Items roles in the rules

The confidence table $\mathbf{C}$ is square and asymmetric, a characteristic that has to be taken into account in analyzing the matrix. The features of $\mathbf{C}$ can be extended to each $\mathbf{C}_k = \mathbf{Z}_k^{\mathsf{T}}\mathbf{Z}_k\mathbf{D}_k^{-1}$. In the context of multidimensional data analysis, different proposals in the literature extend well-known methods like correspondence analysis (CA) [Greenacre, 2000] and multidimensional scaling (MDS) to square asymmetric tables [Bove, 1989]. A common aspect of these methods is in the decomposition of the asymmetric table into two components: *symmetric* and *skew-symmetric*. Applying this decomposition to a table $\mathbf{C}$, it results: $\mathbf{C} = \mathbf{C}_s + \mathbf{C}_{sk}$, where $\mathbf{C}_s = \frac{1}{2}\left(\mathbf{C} + \mathbf{C}^{\mathsf{T}}\right)$ represents the *symmetric* component of $\mathbf{C}$, and $\mathbf{C}_{sk} = \frac{1}{2}\left(\mathbf{C} - \mathbf{C}^{\mathsf{T}}\right)$ represents the *skew-symmetric* component of $\mathbf{C}$. The separate analyses of the two components lead to obtain a representation of the symmetric and skew-symmetric characteristics of table $\mathbf{C}$: the methodology applied on $\mathbf{C}_s$ and $\mathbf{C}_{sk}$, and the corresponding representation display, characterize the different proposals treating square asymmetric tables.

Taking into account the general pair of items j and j', identified by the analysis of $\mathbf{S}$, the role of j and j' depends on the values $c_{(j,j')sk}$ and $c_{(j,j')s}$. Remark that the diagonal elements of $\mathbf{C}$ are constant values equal to 1. These trivial values are completely irrelevant to the analysis and their presence introduces noise in the representations. In order to cut off noise from $\mathbf{C}$ according to Greenacre [Greenacre, 1984], these values can be ignored and replaced by an iterative alternating procedure.

The matrix $\mathbf{C}$ is decomposed in symmetric and skew-symmetric components and then treated to replace the diagonal elements, by the following procedure:

i) decomposition $\mathbf{C}$ in $\mathbf{C}_s$ and $\mathbf{C}_{sk}$;
ii) correspondence analysis of $\mathbf{C}_s$ and $\mathbf{C}_{sk}$;
iii) reconstruction of the main diagonal of $\mathbf{C}_s$ through the general reconstruction formula:

$$np_{ij} = nr_i r_j (1 + \sum_{f=1}^{F} \lambda_f^{-0.5} \varphi_{if} \varphi_{jf}), \tag{1}$$

with $i, j = 1, \ldots, p$ and $f = 1, \ldots, F$. F is the number of considered factors, p is the number of considered items, φ_{if} is the principal coordinate of the i-th item on the factor f; r_i and r_j represent the row margins of $\mathbf{C}$;

iv) comparison of the reconstructed diagonal with the previous main diagonal for $\mathbf{C}_s$: if there is no difference then the whole matrix $\mathbf{C}_s^*$ is reconstructed using formula (1); else $\mathbf{C}_s$ is updated with the obtained diagonal and repeat the previous steps;
v) rebuild the confidence $\mathbf{C}^* = \mathbf{C}_s^* + \mathbf{C}_{sk}$;

Each of the previously selected items is considered as an antecedent or consequent part of interesting rule depending on its deviation from symmetry: items having a positive deviation are considered interesting *heads*, the remaining items are then the *bodies*.

3 Example

In this section the procedure is applied to the BMS-Webview-2 data set: this data set was used in KDD-cup 2000 competition and refers to the transactions associated to an e-commerce company. The whole data set is available at KDD-cup 2000 home page (`url: http://www.ecn.purdue.edu/KDDCUP`). This data set was already used in many applications of data mining procedures proposals and it represents a qualifying benchmark. Being the presented approach complementary to the computer science based proposals, a direct result comparison would not make any sense.

In the BMS-Webview-2 data set each statistical unit represents the click-streams of a single session of a visitor in the e-commerce web-site; each item corresponds to a single product. The raw data set is characterized by 77512 web click-streams and 3340 product-pages (items). After a pre-processing phase the incremental K-means algorithm is applied with different numbers of classes. As shown in figure 1, the best partition is obtained for $K = 10$, since the quality level of the classification becomes stable. The quality level measure for the classification is proportional to the reciprocal of the mean distance between each unit and its related cluster centroid. Following the procedure, once the partition of $\mathbf{Z}$ in K groups is determined, the items that mostly characterize each group are then selected. Interesting items selection and the consequent determination of the selected items roles can be completely automated: however expert users can interact and iteratively set up different parameters.

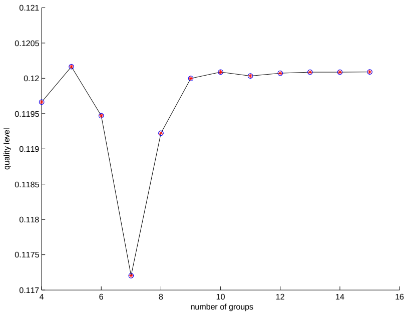


Fig. 1. *Classification quality versus number of classes*

The lack of space does not allow us to represent the graphical displays associated for all ten classes. We just shall give an interpretation key of the whole procedure output.

The procedure output is mainly graphical and consists of various representations of the reconstructed confidence matrix. The paper proposes two of them. The first one (see figure 2) represents the items in the principal factorial space of $\mathbf{C}_k$ and $\mathbf{C}_{sk}$ according to Greenacre (2000). The other one, which is more simple, represents the values of the reconstructed $\mathbf{C}^*$. In the symmetric display, the closeness of two points/items indicates a high degree

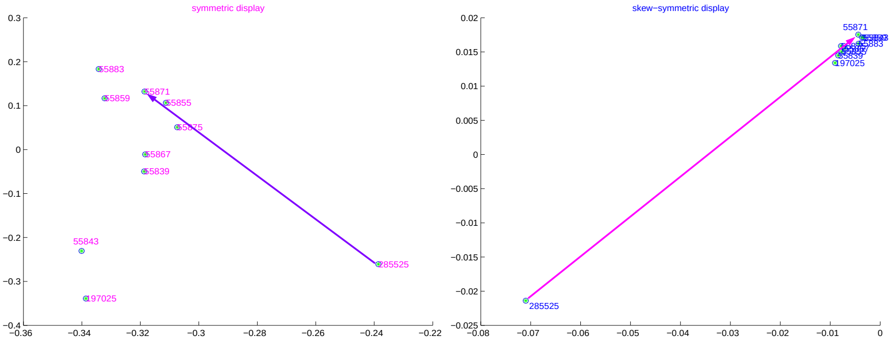


Fig. 2. a) representation of the symmetric component of confidence matrix; b)representation of the skew-symmetric component of the confidences.

of co-occurrence that is a support-like information; in the skew symmetric display, points far from the center of the map are characterized by a high deviation from symmetry. For each pair of items, the amount of the deviation from symmetry is proportional to the area of the triangle formed by the considered pair of points and the axis origin. The sign of the deviation from symmetry depends on oriented triangles: positive deviations correspond to clockwise oriented triangles. The latter display could be quite difficult for a non-expert user to interpret.

The left part of figure 2 shows the item '285525' to be positioned far from the others, that means a different degree of occurrence. The skew-symmetric display confirm the different behavior of '285525'. On the basis of such considerations, a possible rule could be in this case '285525'⟶ '55871', and it is highlighted in both the sides of figure 2.

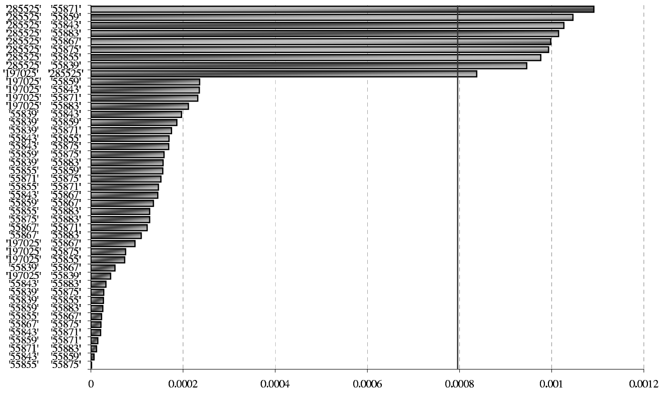


Fig. 3. *Confidences representation*

The second approach is much more understandable, it displays, by a bar-chart, the $n \times (n-1)$ quantities $q_{j,j'} = 1 - \frac{c^*_{(j,j')}}{c^*_{(j',j)}}$ (with $j, j' = 1, \dots, n$ and $j > j'$). The consideration of confidence ratios $q_{j,j'}$ lead to identify the pairs of items j and j' with differing $c_{(j,j')}$ and $c_{(j',j)}$.

Bars are sorted in decreasing order, according to the confidence ratios. A bar represents an interesting rule $j \longrightarrow j'$ if it has an high value, being $c^*_{(j,j')} < c^*_{(j',j)}$, and the ordered pair (j, j') has a positive deviation from symmetry.

Figure 3 confirms the "importance" of the rule '285525'⟶ '55871' that is associated to the first bar in the bar-graph. Furthermore, the different behavior of the item '285525' is evident: the first-ranked bars are all associated to rules having '285525' as an antecedent part.

4 Conclusion and perspectives

Since the AR were introduced for the analysis of large (and huge) data bases, most of the computational aspects, in term of speed and memory, have been successfully solved. However, it is still an open problem how to interpret the massive output. J. Edler and D. Pregibon in 1996 [Elder and Pregibon, 1996] foresaw that the KDD would have been a field for important challenges for the statistical community. They wrote: "*The statistician's tendency to avoid complete automation out of respect for the challenges of the data, and the historical emphasis on models with interpretable structure, has led that community to focus on problems with a more manageable number of variables (a dozen, say) and cases (several hundred typically) than may be encountered in KDD problems, which can be orders of magnitude larger at the outset. With increasingly huge and amorphous databases, it is clear that methods for automatically hunting down possible patterns worthy of fuller, interactive attention, are required*". This paper, eight years later, goes in the direction indicated by Edler and Pregibon. Nowadays, the more and more increased power of modern computer makes easier to achieve this task.

Future enhancements of the proposed approach are into different directions. From a computational point of view, the aim is to improve the classification step in order to obtain higher quality solution in less time. Another important aspect is the generalization of the selection criteria from pairs of items to pairs of itemsets, or rather to generalize the procedure to complex rules. Furthermore, according to the exploratory nature of the procedure, it is necessary to improve the visualization tools and introduce interactive capabilities.

References

[Agrawal and Srikant, 1994]R. Agrawal and R. Srikant. Fast algorithms for mining association rules. In *Proc. of the 20th Very Large Data Bases Conference*, pages 487–499, Santiago, Chile, 1994.

[Agrawal *et al.*, 1993]R. Agrawal, T. Imielinski, and A. Swami. Mining association rules between sets of items in large databases. In *ACM SIGMOD int. conf. on Management of data*, volume 22, pages 207–216, N.Y., 1993. ACM Press.

[Bove, 1989]G. Bove. *Nuovi metodi di rappresentazione di dati di prossimità.* Tesi di dottorato, Univ. di Roma La Sapienza, Roma, 1989.

[Brin *et al.*, 1997]S. Brin, R. Motwani, J. Ullman, and S. Tsur. Dynamic itemsets and implication rules in market basket data. In *ACM SIGMOD int. conf. on Management of data*, pages 255–264, Tucson, Arizona, USA, May 1997.

[Elder and Pregibon, 1996]J. F. IV Elder and D. Pregibon. A statistical perspective on knowledge discovery in databases. In U. M. Fayyad et *al.*, editors, *Advances in knowledge discovery and data mining*, pages 83–113, Am. Ass. for AI, 1996.

[Gaber *et al.*, 2004]M. M. Gaber, S. Krishnaswamy, and A. Zaslavsky. Cost-efficient mining techniques for data streams. In *Proc. of the second workshop on Australasian information security, Data Mining and Web Intelligence, and Software Internationalisation*, pages 109–114, Dunedin, New Zealand, 2004.

[Greenacre, 1984]M. Greenacre. *Theory and Applications of Correspondence Analysis.* Academic Press, London, 1984.

[Greenacre, 2000]M. Greenacre. Correspondence analysis of square asymmetric matrices. *Applied Statistics*, 49(3):297–310, 2000.

[Guha *et al.*, 2000]S. Guha, R. Rastogi, and K. Shim. Rock: A robust clustering algorithm for categorical attributes. In *Proc. of the 15^{th} Int. Conf. on Data Engineering.* IEEE Computer Society, 2000.

[Han *et al.*, 2000]J. Han, J. Pei, and Y. Yin. Mining frequent patterns without candidate generation. In *ACM SIGMOD int. conf. on Management of data*, pages 1–12, Dallas, Texas, May 2000. ACM SIGMOD.

[Hartigan, 1975]J. A. Hartigan. *Clustering Algorithms.* John Wiley & Sons, 1975.

[Liu *et al.*, 1999]H. Liu, W. Hsu, and Y. Ma. Pruning and summarizing the discovered associations. In *Proc. of KDD*, pages 125–134, San Diego, CA, 1999.

[Miller and Yang, 1997]R. J. Miller and Y. Yang. Association rules over interval data. In *ACM SIGMOD int. conf. on Management of data*, pages 452–461, Tucson, AZ, USA, 1997.

[Ordonez, 2003]C. Ordonez. Clustering binary data streams with k-means. In *Proc. of the 8th ACM SIGMOD workshop on Research issues in data mining and knowledge discovery*, pages 12–19, San Diego, CA, 2003. ACM Press.

[Park *et al.*, 1995]J. Park, M. Chen, and P. Yu. An effective hash based algorithm for mining association rules. In M. J. Carey and D. A. Schneider, eds, *ACM SIGMOD Int. Conf. on Management of Data*, pp. 175–186, San Jose, 1995.

[Savasere *et al.*, 1995]A. Savasere, E. Omiecinski, and S. Navathe. An efficient algorithm for mining association rules in large data bases. In *21^{st} Conference on Very Large Data Bases*, pages 407–419, Zurich, Switzerland, September 1995.

[Srikant and Agrawal, 1996]R. Srikant and R. Agrawal. Mining quantitative association rules in large relational tables. In *ACM SIGMOD int. conf. on Management of data*, pages 1–12, Montreal, Quebec, Canada, June 1996.

[Zaki, 2000]M. J. Zaki. Generating non-redundant association rules. In *Knowledge Discovery and Data Mining*, pages 34–43, 2000.

Dimension Reduction for Visual Data Mining

Edwige Fangseu Badjio and François Poulet

ESIEA Recherche
Parc Universitaire de Laval-Changé,
38, Rue des Docteurs Calmette et Guérin
53000 Laval, France
(e-mail: fangseubadjio@esiea-ouest.fr, poulet@esiea-ouest.fr)

Abstract. We present a method for dimension reduction applied to visual data mining in order to reduce the user cognitive load due to the density of data to be visualized and mined. We use consensus theory to address this problem: the decision of a committee of experts (in our case existing attribute selection methods) is generally better than the decision of a single expert. We illustrate the choices operated for our algorithm and we explain the results. We compare successfully these results with those of two widely used methods in attribute selection, a filter based method (LVF) and a wrapper based method (Stepclass).
Keywords: visual data mining, dimension reduction, feature selection, filter, wrapper, consensus.

1 Introduction

The quantity of stored data doubles every 9 months, these data are not useful if at least a part of information they contain is not extracted. It is the goal of knowledge discovery in the databases (KDD) which can be defined as the non trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data [Fayyad *et al.*, 1996]. In most of data mining (a step of KDD) approaches, the process of discovering correlations in data sets is performed in an automatic way. For users, understanding and explaining data with only automatic algorithms results can be difficult. Visual data mining is a new data mining approach using visualization as a communication channel for data mining. It lies in tightly coupling the visualizations and analytical process into one data mining tool that takes advantage of the strengths of all worlds [Wong, 1999]. Visualization is the process of transforming information into a graphical representation allowing the user to perceive and interact with the information.

Visual representation allows understanding data, determining what should be done about it. The human eye can capture complex patterns and relationships. Compared to data mining, the advantages of visual data mining are:

- the confidence in the results is improved, the KDD process is not just a "black box" giving more or less comprehensible results,

- the quality of the results is improved by the use of human pattern recognition capabilities,
- if the user is the data specialist, we can use the domain knowledge during the whole process (and not only for the interpretation of the results).

Computer devices can display vast amount of information with various techniques. This information must be appropriately communicated to us in order to make the best use of it. According to [Ware, 2000], in order to be visualized, data are passed through four basic stages : independently of any visualization technique, the first step of visualization is data collection and storage. Secondly, there is a data pre-processing which goal is to transform the data into a comprehensive form. At the third step, display hardware and software are used to produce a visual representation of the data. Lastly, the users perceive, interact with the visual representation and mine it. It is necessary to address the limits of human perception. When the collected data are multidimensional, there are some limits in the third and fourth steps.

For [Ferreira and Levkowitz, 2003], the conceptual boundary between low and high-dimensional data is round three to four data attributes. Their suggested guideline for characterizing dimensionality is the following: low: up to four attributes, medium: five to nine attributes and high: 10 or more. When the number of dimensions is over some dozen, the large number of axes needed to create these displays tends to overcrowd the figure, limiting the value of the plot for detecting patterns or other useful information.

We are interested in visual data mining methods performing supervised classification. Our objective is to select some dimension of a data set in order to create a visualization from which relevant information can be extracted. We want to identify attributes that are significant in order to reduce dimensionality. Dimension reduction can be used to improve the efficiency of visualization of large, multidimensional data sets and may be the accuracy of algorithms used for classification in visual data mining.

Knowing that:

1. an optimal subset of attributes is not necessarily unique,
2. the visualization of more than a dozen attributes is unusable for visual data mining,
3. without investigation, it is not possible to determine a dimension reduction method that can perfectly reduce the set of attributes (by taking account of different trade-offs between performance and complexity (tolerate lower performance in a model that also require less features)),
4. the decision of a committee of experts is generally better than the decision of a single expert,

we use a meta-analysis algorithm based on consensus theory for dimension reduction in visual data mining. The proposed algorithm combines decisions of several experts (in our case feature selection algorithms). More precisely, it maps a given set of dimension subsets to a single dimension subset.

The rest of this paper is organized as followed: section 2 explains the context of dimension reduction. In section 3, we present the visual data mining domain, the specificities related to this domain and the task analysis. Next, there is an explanation of the specificities of dimension reduction applied to visual data mining which allow us to design our dimension reduction method. Section 4 introduces this method before experiments, conclusion and future work.

2 Dimension reduction

Many techniques for the visualization of multidimensional data have been developed: pixel oriented techniques, parallel coordinates, survey plot, etc. With visualization techniques, large amount of data can be displayed on the screen, colors allow the users to instantly recognize similarities or difference of thousands of data items, the data items may be arranged to express some relationship. We try to solve the following problem: how can we select from a set of candidate dimensions, a subset that performs the best under visual tools and visual data mining and discard the others? We use visual data mining in order to find an accurate decision tree by using a visualization technique with interaction capabilities. The decision tree is interactively constructed by the user who uses his perception and data domain knowledge. This kind of interactive decision tree construction algorithm can only be used if the number of dimensions of the data is small enough (less than dozen).

Dimension reduction and attribute selection aim at choosing a small subset of attributes that is sufficient to describe the data set. It is the process of identifying and removing as much as possible the irrelevant and redundant information. Sophisticated attribute selection methods have been developed to tackle three problems: reduce classifier cost and complexity, improve model accuracy (attribute selection), improve the visualization and comprehensibility of induced concepts. There are two major components in a attribute selection/dimension reduction algorithm: the generation procedure and the evaluation function [Dash and Liu, 1997].

2.1 Generation procedure

Let N denote the number of varia in the original data set, attribute selection requires to test 2^N different subsets to find the optimal one. A solution in order to avoid this search is to proceed to random search or to use one of the following search strategies: backward, forward or both. After the generation of feature subsets, an evaluation function measures the goodness of the subset and this value is compared with the previous best subset of attributes [Dash and Liu, 1997]. The following section presents the available evaluation functions.

2.2 Evaluation functions

Two types of evaluation functions are used in attribute selection: in the first one, filter-based approach the dimensions are filtered independently of the induction algorithm. The relevance of each dimension is computed with some statistical information calculated from the training data set. Examples of statically measures used: information gain [Dumais *et al.*, 1998], [Quinlan, 1993], correlation [Hall, 2000], etc.

The other type is the wrapper approach [Kohavi and John, 1997]: a learning algorithm is used in order to select the subset of features, while discarding the rest. For any iteration of the wrapper algorithm, the quality of the feature subset is evaluated by an inductive learning algorithm.

Attribute wrappers often achieve better results than filters due to the fact they are tuned to the specific interaction between an induction algorithm and its training data. However, they tend to be much slower than attribute filters because they must repeatedly call the induction algorithm and must be run when a different induction algorithm is used [Kotsiantis and Pintelas, 2004].

2.3 Problems encountered in attribute selection

At the initialization step, the attribute selection algorithms require many parameters. In order to lead to best results, it is necessary to choose the most relevant parameters. Knowing that an attribute selection process may stop under one of the following reasonable criteria: a defined set of dimensions are selected, a defined number of iterations are reached, addition (or deletion) of any dimension does not produce a better result, an optimal subset according to the evaluation criteria is obtained.

3 Applying selection to visual data mining

As we said, if the data dimension is high (figure 1), the human cognitive task for detecting correlations or discover hidden patterns in data is very hard.

The figure presents a sequence of $\frac{n-1}{2}$ two-dimensional matrices (like scatter plot matrices [Chambers *et al.*, 1983]) generated by CIAD [Poulet, 2002], n represents the number of attributes. In order to deal with high dimensional data, the above approach of data exploration has been proposed, CIAD supports the user in selecting one representation which matches the best with his mining objective. The focus presents details of the most suitable view. Figure 1 does not allow distinguishing visually colors in order to mine the data set. This is because the number of attributes and the number of instances in the data set are too large. The following paragraph briefly presents the visual data mining task analysis.

3.1 Visual data mining task analysis

In order to mine a data set, the user interacts with a graphical representation (chart) of the data. The data model (knowledge) is built in an interactive and iterative way.

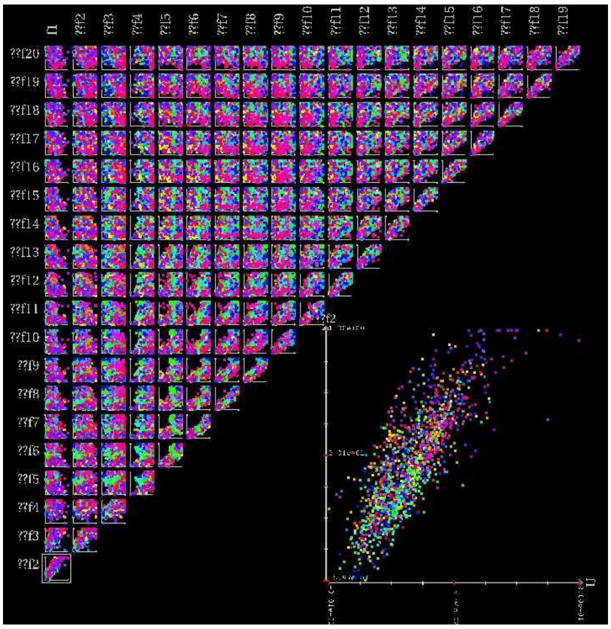


Fig. 1. Isolet data set visualization with CIAD

3.2 User categorization

A visual data mining environment can be used by several type of users:

- data domain specialist: according to his knowledge about data, this type of user can select the best subset of attributes or request the support of an automatic tool for attribute selection.
- data analysis specialist: in this category, the user can be a statistician or a machine learning expert.
  - the statistician expert can adequately use filter approach and determine the appropriate parameters for the initialization of attribute selection algorithm.
  - the machine learning expert can perfectly initialize supervised classification algorithms used by wrapper approach. This type of user is able to choose a supervised classification algorithm to be used in order to evaluate the selected attributes in the attribute selection algorithm and to choose a best set of criteria for the evaluation of selected attributes.

  These users can also be interested in wrapper or filter based approach advantages and need to be supported by an automatic tool.

The automatic dimension reduction framework in all these cases will require a great accuracy of the results.

4 New dimension reduction algorithm

To obtain the best accuracy in attribute selection, the best is to operate an exhaustive search among the 2^N possible combinations of attribute subsets and to use a wrapper-based approach as evaluation function. For a large value of N, this approach is computationaly prohibitive. We propose to use random search and (backward, forward ((like sequential floating selection), knowing that the function used is non monotonic [Pudil *et al.*, 1994])). We believe that this procedure will allow us to treat a large number of attribute subsets.

The wrapper approach allows rising to interesting details for the data analysis specialist (data mining domain). Knowing that the classifier error rate capture two basic performance aspects: class separability ability and any structural error imposed by the form of the classifier. Other types of details, namely, properties that good dimension sets are presumed to have (class separability or a high correlation between the attributes) are more appropriate to statistician. These details could not be highlighted at all by the wrapper methods. In order to take this fact into consideration, we have added some filter-based criteria (consistency, entropy, distance) to our attribute subset selection method.

In input, there is a data set and the output is a subset of attributes of this data set. The generation procedure uses a combination of random search and sequential floating selection. Concerning the evaluation functions, we use a combination of filter (consistency, entropy, distance) and wrapper ((LDA, QDA, KNN) [Ripley, 1996]). LDA, QDA, KNN executions use ten fold cross validation. At each step of the execution of these algorithms, the following evaluation criteria are used: the correctness of the classification rule, the accuracy, the ability to separate classes, and the confidence. Next, we combine their selected attribute subset in order to derive a consensus of the most suitable subset of attributes. For this purpose, a learning step, based on the results of generation procedures evaluated by filter-based criteria and wrapper based approaches enables us to lead to final results.

More precisely, the domain we consider consist of a set of $N = 6$ experts (consistency, entropy, distance, LDA, QDA, KNN evaluation functions) $E = \{e_1, ..., e_N\}$, a set of dimension subsets $DS = \{D_1, ..., D_K\}$, where K is not a constant. Attribute subsets are available for expert/subset pairs $\{e, D\}$, where $e \in E$ and $D \in DS$. We define preference of a dimension d as the probability that the dimension appears in the experts feature subsets, $p(d) = \sum p_i(d)$. $p_i(d)$ represents the probability that expert i selects dimension d .

$p_i(d) = \frac{y}{Z}$ if expert i has selected feature$d$, 0 otherwise. y is the number of selected dimensions. Z represents the number of attributes in the original data set. The preference value of features is used in order to pool together the selected features and to rank them. Next, if the pool number of dimensions is greater than twenty (number of attributes which can be correctly display and visually mine), it is divided into relevant attributes (consensus) and less relevant attributes. At the cutting point, if some features have the same preference value (we consider these attributes as conflicting attributes), we use expert relevance score (ERS) in order to determine which features match the best. For each feature in the conflicting part, the decision to add it in consensus part of the pool or not is made according to the relevance score of the experts who choose the feature. The selected features are those with great expert relevance score computed as following: $ERS = \frac{g}{T}$, where g represents the number of attributes in the consensus part which have been selected by the expert and T the total number of features selected by that expert.

As we will see in the case study part of this paper, the main advantage of this approach is the combination of feature subsets from various feature selection algorithms.

5 Experiments

The purpose of this study was to see if the method would be able to effectively reflect the performance differences among experts.

In order to test proposed approach, we compare its results with the results of two widely used attribute selection methods. Namely, R language implementations of: Las Vegas Filter [Liu and Setiono, 1996] (package dprep) and a wrapper based feature selection algorithm (Stepclass, package klaR). Our consensus based algorithm is also implemented in R. We use a PC pentium IV, 1,7 GHz, Windows to perform these tests. The data sets (from the UCI [Blake and Merz, 1998] and the Kent Rigde Bio-Medical Data Set repositories [Jinyan and Huiqing, 2002] were chosen because of their large number of attributes (table 1).

Name	NbAt	NbInst	NbClass
Lung-Cancer	57	32	3
Promoter	59	106	2
Sonar	60	208	2
Arrhythmia	280	452	16
Isolet	618	1560	26
ColonTumor	2000	62	2
CentralNervSyst	7129	60	2

Table 1. Data set description

The final results of LVF, stepclass and consensus based algorithm were evaluated by IBk, a K nearest neighbor algorithm (KNN) found in WEKA, a free Java-based, open source, that provide a variety of machine learning algorithms.

Table 2 shows the difference (attribute size and KNN accuracy) between the original and the final data sets. The attribute subset selected by the consensus based approach (less or equal to 20) allows visualizing and mining the whole data sets. The changes in the accuracies of KNN classifier is minimal or there is no change. This is not the case of LVF or stepclass (table 3). The data set Arrhythmia for example has a subset with 109 attributes (LVF results) and for the data set Promoter, stepclass does not reduce the dimension.

Name	Initial NbAt	Final NbAt	Acc before	Acc after
Lung-Cancer	57	4	37.5%	75%
Promoter	59	9	85.84%	68.87%
Sonar	60	8	86.54%	71.15%
Arrhythmia	280	4	53.44%	59.96%
Isolet	618	14	85.57%	70.24%
ColonTumor	2000	19	77.42%	79.03%
CentralNervSyst	7129	20	56.67%	60%

Table 2. Comparison of number of attributes and accuracy with KNN algorithm before and after reduction

Feature selection frameworks as we said aim at reducing classifier cost and complexity, improving model accuracy. Our goal is firstly to reduce the number of dimensions in order that the data set could be visualized. Table 3 shows that we attend our principal goal and we obtain results that are comparable to those of the attribute selection algorithms which objective is to improve classifiers accuracy. Indeed, the consensus based approach allows obtaining the best result for data set Lung-Cancer and about the same accuracy rate for the data sets Sonar, Arrhythmia and colonTumor. It should be noted that two cases arise: either the attributes of the data set to be treated are redundant or irrelevant and then the results are comparable with those of filters or wrappers based approaches or it does not exist redundancy in the attributes and dimension reduction implies a loss of accuracy. The data sets in this category are: Isolet (best accuracy with LVF for 268 attributes) and Promoter (best accuracy with Stepclass for 59 attributes). For these data sets, the number of selected dimensions in spite of the best accuracy remains unusable for visual data mining.

Name	Final NbAt	Lvf NbAt	Wrap NbAt	Final Acc	Lvf Acc	Wrap Acc
Lung-Cancer	4	17	4	75%	62.5%	71.87%
Promoter	9	16	59	68.87%	80.19%	85.85%
Sonar	8	18	4	71.15%	82.21%	71.63%
Arrhythmia	4	109	4	59.95%	54.65%	60.84%
Isolet	14	268	8	70.24%	83%	57.98%
ColonTumor	19	918	5	79.03%	77.42%	79.03%
CentralNervSyst	20	3431	8	60%	58.33%	71.67%

Table 3. Comparison of our method with LVF and stepclass

6 Conclusion

The data visualization, the performance of classification algorithms are affected by attributes. When a data set has a large number of attributes, it is impossible to perform visual data mining. Irrelevant, redundant features have a negative effect on the accuracy of a classifier and on visual representations. We have defined a dimension reduction method for visual data mining. Then we have compared successfully the results of this framework to two widely used attribute selection algorithms. The data visualization (figure 1) which represents a visualization in which the relationships within the data are unclear is replaced by another visualization (figure 2) which is more usable and much more appropriate to visual data mining.

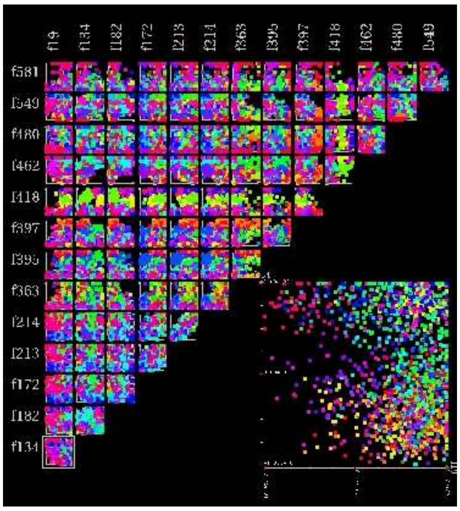


Fig. 2. Isolet Reduced data set visualization with CIAD

Our dimension reduction framework reduces the number of attributes. However, we remark that with a low number of attributes and a high number of instances, it is not easy to represent and mine data perfectly. We plan to develop some methods for reducing the number of instances in a data set to be treated.

References

[Blake and Merz, 1998]C. Blake and C. Merz. *UCI Repository of machine learning databases.* Irvine, University of California, Department of Information and Computer Science, from www.ics.uci.edu/~ mlearn/MLRepository.html, 1998.

[Chambers *et al.*, 1983]J. Chambers, W. Cleveland, and P. Turkey. *Graphical Methods for Data Analysis.* Wadsworth, 1983.

[Dash and Liu, 1997]M. Dash and H. Liu. Feature selection methods for classification. In *Intelligent Data Analysis: An International Journal*, pages 1–2, 1997.

[Dumais *et al.*, 1998]S. Dumais, J. Platt, D. Heckerman, and M. Shahami. Inductive learning algorithms and representation for text categorisation. In *Proc, The International Conference on Information and Knowledge management*, pages 148–155, 1998.

[Fayyad *et al.*, 1996]U. M. Fayyad, G. Piatetsky-Shapiro, and G. Smyth. *Advances in Knowledge Discovery and Data Mining.* AAAI Press / MIT Press, Menlo Park, CA, 1996.

[Ferreira and Levkowitz, 2003]d.O. Ferreira and Levkowitz. From visual data exploration to visual data mining: a survey, visualization and computer graphics. *IEEE Transactions*, pages 378–394, 2003.

[Hall, 2000]M. Hall. Correlation-based feature selection for discrete and numeric class machine learning. In *Proc, International Conference on Machine Learning*, pages 359–366, 2000.

[Jinyan and Huiqing, 2002]L. Jinyan and L. Huiqing. *Kent Ridge Bio-medical Data Set Repository.* http://sdmc.lit.org.sg/GEDatasets, 2002.

[Kohavi and John, 1997]R. Kohavi and G. H. John. Wrappers for feature subset selection. In *Artificial Intelligence*, pages 273–324, 1997.

[Kotsiantis and Pintelas, 2004]S. B. Kotsiantis and P. E. Pintelas. Hybrid feature selection instead of ensembles of classifiers in medical decision support. In *Proc, IPMU*, 2004.

[Liu and Setiono, 1996]H. Liu and R. Setiono. A probabilistic approach to feature selection: a filter solution. In *Proc, The 13th International Conference on Machine Learning*, pages 319–327, 1996.

[Poulet, 2002]F. Poulet. Cooperation between automatic algorithms, interactive algorithms and visualization tools for visual data mining. In *Proc, Visual Data Mining workshop, PKDD2002*, 2002.

[Pudil *et al.*, 1994]P. Pudil, J. Novovicova, and J. Kittler. Floating search meathods in feature selection. *Pattern Recognition Letters*, pages 1119–1125, 1994.

[Quinlan, 1993]J. R. Quinlan. *C4.5: Programs for Machine Learning.* Morgan-Kaufman, San Mateo, CA, 1993.

[Ripley, 1996]B. D. Ripley. *Pattern Recognition and Neural Networks.* Cambridge University Press, 1996.

[Ware, 2000]C. Ware. *Information Visualization, Perception for design.* Morgan-Kaufman Publishers, San Diego, USA, 2000.

[Wong, 1999]P.C. Wong. Visual data mining. *IEEE Computer Graphics and Applications*, pages 20–21, 1999.

Data Visualizations on small and very small screens

Monique Noirhomme-Fraiture[1], Frédéric Randolet[1], Luca Chittaro[2], and Grégory Custinne

[1] Pôle IHM
Institut d'informatique, Université de Namur (FUNDP)
21, rue Grandgagnage, 5000 Namur, Belgique
(e-mail: mno@info.fundp.ac.be, fra@info.fundp.ac.be)

[2] HCI Lab
Dept. of Math and Computer Science, University of Udine
via delle Scienze, 206, 33100 Udine, Italy
(e-mail: chittaro@dimi.uniud.it)

Abstract. We can now access to information almost everywhere, at any time, using mobile phone or PDA with connection to the Web. A good way to resume data is visualization but the size of the screen and bad environment generate huge constraints.

We first present some general recommandations for small screens. Then we study the case of time series visualization for more particularly of stock values and describe two solutions fot PDA and two solutions for mobile.

Keywords: Data Visualizations, Time series, PDA, Mobile Phones, Small screens, Clustering.

1 Introduction

Access to information has been improved considerably the last years. Through internet, we can connect to large data bases which are updated continually, most often on the Web.

Moreover, with mobile technology, we can consult this information even if we are not in our office or at home, in front of a standard PC. Indeed we can get information when using a mobile phone or a PDA. But these tools are much more limited in term of memory, power and screen resolution. If we can expect that in the next years, memory, power and screen quality definition will be improved, the size of the screen will not be enlarged because the intrinsic characteristic of a mobile device is to be small, in order to carry it, in the pocket or the handbag.

Whereas human computer interaction was concerned up to now on how present information on screen which enlarged and improved, from year to year, we discover now new (or old) problems due to the size and the bad quality of the screens. The bad quality of the vision is also due to the fact that with mobile devices, we can get information in areas not really well suited for reading: bad lighting, bad posture, using sometimes only one hand.

We have thus to be more innovative and find solutions which are usable, with strong technological constraints.

Visualization of data is a very powerful tool for summarizing information on input data or on results of analysis. It seems thus a good way to communicate information on a small screen but limitation is so high that the solution is not evident. Representations cannot just be adapted from standard PC to mobile devices. We have to find new solutions.

We have also to point out guidelines in order to improve the usability of information representation on mobile. The following paragraph will describe such recommendations.

As a case study for data visualisation, we have developed systems of time series visualization, because the problem is very common, met by many persons in management and finance. Whereas we can assume that results of data analysis like clustering, factorial analysis, regression analysis, can be consulted at work office, time dependant data can be needed in more varied environment. It is why we have oriented our study on this kind of data. As a particular case, stock values are a good example of such series: they interest a lot of people, need to be updated every day and are crucial in our economy. Users will then be very motivated to use good visualization systems.

2 State of the art

Literature about visualization on mobile concerns mostly applications like tourism application [Bornträger *et al.*, 2003][Pospischil *et al.*, 2002], nomadic guide [Schmidt-Belz and Hermann, 2004] [Fithian *et al.*, 2003] [Chinchille *et al.*, 2002], mobile gaming [Sanneblad and Holmquist, 2003], remote control [Tarrini *et al.*, 2002]

Some papers concern also guidelines for tasks on small screens. Many general guidelines for HCI with mobile and PDA performed by mobile industry concern the arrangement of keyboard or the shape and size of the mobile. We will not consider this kind of problems here and will look only at the screen usability.

Some guidelines for small screens have been verified by experimental studies. Most identify good procedures in searching tasks [Giller *et al.*, 2003], [Jones *et al.*, 2002].

Few papers have been published in the field of data visualization on small screen but some systems are available for dedicated data like stock information.

In data visualization field, [Chittaro and Camaggio, 2002] presents solutions for bar charts format representation on WAP phones. For stock information, let us reference specialized sites (www.wap.boursorama.com, www.cprbourse.tm.fr/wap, www.firstinvest.waptoo.com). They give general information, with some visualization like tables and graphics. Those are usually standard time series representing the value of a stock or indices with time in the X-axis. Time vary in hours, days or years, with of course different level of precision. On mobile phone, these graphics are generally badly readable, due to the lack of definition of the screen.

In what follow, we summarize recommendations for small screens. They are derived and updated from general usability principles [Scapin, 1986] and from some dedicated papers [Jones *et al.*, 2002], [Giller *et al.*, 2003].

2.1 Conciseness, precision and consistency

In order to save space, give precise, concise information: use as few words as possible but enough precise. Precision does not mean necessarily to use technical word. The vocabulary must be understood by the user. This recommendations have been established for usability on standard PC but conciseness has more importance in the context of small screens.

2.2 Navigation

As few information can be displayed on a screen, when information is voluminous, the solution is to split it on several screens. This gives problems of navigation and orientation. Considering searching tasks, [Jones *et al.*, 2002] have shown that screen size has a major impact on user performance. They propose following guidelines which can be interesting in our data visualization context:

- reduce the amount of page-to-page navigation needed to view search. They have observed that additional user effort when vertical scrolling affects performance to a lesser extension that the page-to-page navigation. [Giller *et al.*, 2003] found the same result.
- adapt for vertical scrolling. Users tend to scroll vertically rather than horizontally. [Giller *et al.*, 2003] have shown that in a selection task in a list of items when the users know the target item, 15 items still seems to be OK, whereas for an unknown target, items should lie around 8.

3 Solutions for optimisation of screen use

Even if preceding recommendations are valid for small screen (pocket pc) and very small screen (mobile phone), problems appear differently. They are much more accurate on mobile phone and must be solved appropriately. On PDA, we can try to summarize information on one screen, with opening of windows for more details using at the best colour features and screen definition quality. We have to be innovative in the visualization. We will present some new time series visualization for PDA in paragraph 4.

On the contrary, on mobile phone, we are obliged to simplify the visualization. When the quantity of information is large, we suggest to use analytical method to reduce the amount of information before visualizing the data. In paragraph 5, we present two examples of such reduction using clustering.

4 Examples of visualization on PDA

Most of the solutions for time series representations developped for PDA are a miniaturisation of the common visualizations used for PC. If an overview of a time serie can be viewed by this method, it is not sufficient because the elements displayed are too small and the visualization is not clear. Thus we provide to find solutions to that problems. The two visualizations presented here are very different. They don't use the same scale, one is common and the other uses colour. More, they don't present the information in the same way because the representation of the days is different.

4.1 Brick Wall Chart

The Brick Wall Chart is a very rich visualization, it can display the evolution in one year of 4 different attributes (see fig.1). The Brick Wall Chart uses a scale of colour to represent the data values. The scale goes from the white, yellow, red and black. As recommended by [Spence, 2000] white color represents the minimum value, black represents the maximum value, and between these two colors, we use the interpolation to find the color.

The screen is divided in 12 columns representing months from the left to the right, and each column is vertically divided in rectangles representing days (the number of days depends on the month) from the bottom to the top of the screen. The rectangle is also horizontally divided in 4 small rectangles to represent each attribute. The rectangles are filled with color to represent the data values.

If the user want more precision, he can switch to the same visualization but with only one attribute. The Brick Wall Chart is the same but the days are not divided in small rectangles. Because the days are bigger, the color can be seen with less difficulty, the visualization is then more precise and clear.

The user can obtain the precise values of the attributes for one day, he just has to tap on the sreen over the corresponding day. A box appears on the screen providing the values of the 4 attributes.

4.2 Stacked Bar Chart

The Stacked Bar Chart uses the metaphore of the calendar because it divides the screen in 12 parts representing the months (see fig.2). The scale is classical and use the height of the elements to reprensent their values. Stacked Bar Chart provide showing two different attributes. It is less richer than the previous visualization but is more common and usual. Actually, the visualization is represented by 12 small graphics. Each graphic is a classical one with the time on the X-Axis and the value on the Y-Axis. The first attribute is display in blue, the second in orange. The smaller attribute appears in front of the screen, and the second appears behind the smaller attribute.

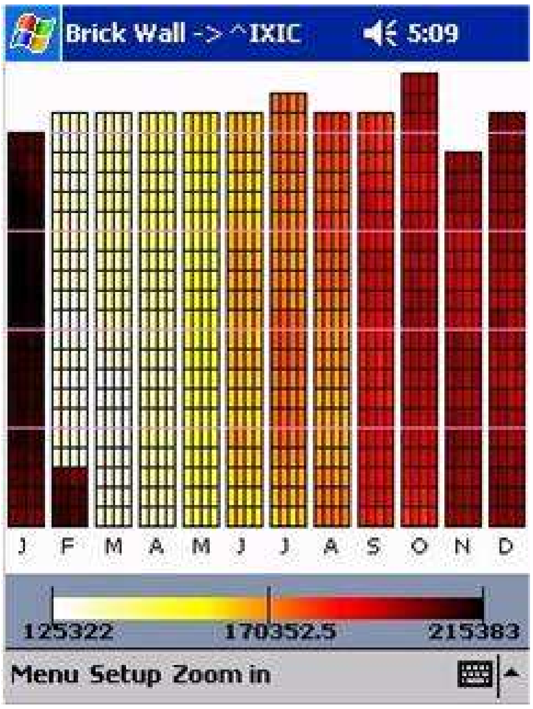


Fig. 1. Brick Wall Chart.

4.3 Evaluation

The evaluation had two goals: compare the efficiency of the visualizations and the preferences of the users.

First, we tested the performance, the time to complete a task, the quality of the representation of the days, the quality of the reprensentation of the values. The results cannot affirm that one visualization is better than another. So we can expect that the users wont prefer any visualization and in this case, the preference would be homogeneous.

But 17 of the 20 users chose the Brick Wall Chart as their favourite visualization and the appreciation for this visualization is better than the one for the Stacked Bar Chart. So, the preferences of the users go mainly to the Brick Wall Chart that use a color scale even if their are both efficient. More information on that evaluation can be found in [Noirhomme-Fraiture *et al.*, 2005].

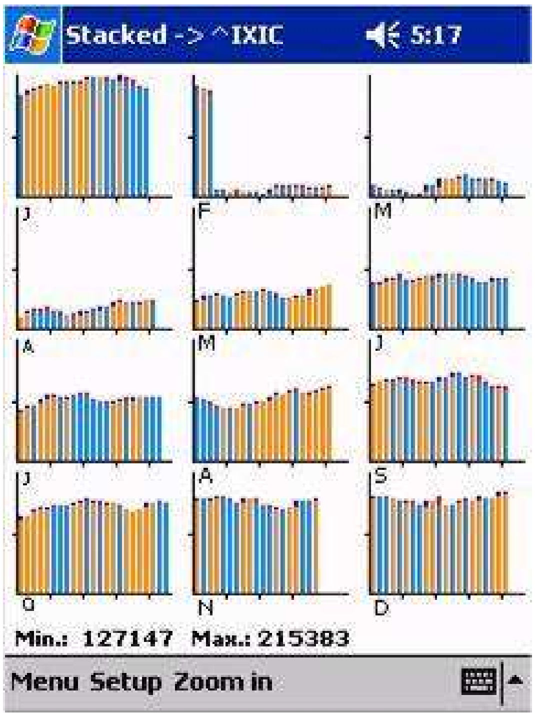


Fig. 2. Stacked Bar Chart.

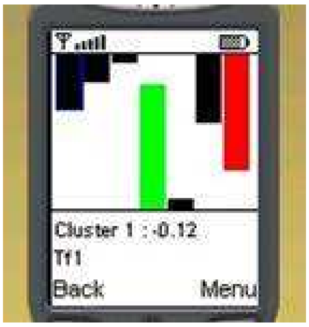
5 Examples of visualization on mobile phone

On very small screen, superposition of several windows is not advisable. Due to lack of definition, too detailed graphics are not readable. We have to avoid outlines and to privilege pictures with colored or greyed surfaces, because lines are not distinguished. Using Gestalt principles, [Easterby] has shown that a contrast boundary is better than a line boundary for making a shape stand out. We present here two visualization that we have designed to represent stock values. The first one is dedicated to visualization of different values at a given time, the second one is dedicated to visualization of the evolution of such values with time. More visualization and details can be found in [Custinne *et al.*, 2004]. If the values are too numerous, we suggest to summarize them by a clustering. For example, when representing values of stock portfolio, stocks can be very numerous so that it is impossible to represent all the stocks on the same graph. What stock sites do usually is to represent stocks separately. This method is tiresome for the user, does not allow comparing the stocks between them and does not allow having a global view of the portfolio. Many stocks have common behaviour so it is possible to assemble them in classes with homogneous behaviour. As example, we have operated a K-Means clustering to merge 40 stocks in 7 clusters. The clustering was done on base of the mean value on one month or on variation during one month (max-min).

We could also perform a clustering to reduce the number of days, when we represent the variation of stocks on a long period. In this case, it seems more appropriate to use clustering methods which try to agglomerate days which are contiguous in order to obtain clusters of periods. In what follows, we present the two visualization adapted to clusters (but they can also be used for individual stocks).

5.1 Alternate Bar Chart

This visualization is closed to the kind of representation that workers in stock exchange are used to. The height of the bars are proportional to the value to be represented: maximum variation of stock value during a period. As this variation can be positive or negative, we represent positive value from bottom to top on the bottom line and negative value from top to bottom on the upper line (see fig. 3). Each bar has a different color according to the cluster of stocks. Rectangles are full colored. Details can be obtained when pointing a particuliar bar. This kind of optimise the use of the screen space and allows very quick visualization of the importance of positive/negative variations.



Cluster 1 : -0.12

Tf1

Fig. 3. Alternate Bar Chart.

5.2 Pixel bar Chart

In this representation we visualize the evolution of values during a period for different clusters (or stocks). Each column is dedicated to a cluster, its width is proportional to the number of stocks in the cluster. Each daily value is mapped on a color scale. The different days of the period are represented from bottom to top. We use the same convention of colour as the one used in the Brick bar Chart on PDA (see fig.4).

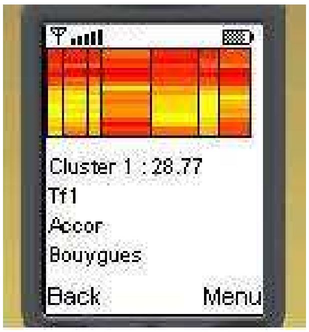

Cluster 1 : 28.77
Tf1
Accor
Bouygues

Fig. 4. Pixel Bar Chart.

As in the preceding visualization, the representation uses maximum of the available space on the screen. Whereas the alternate bar chart is commonly appreciated and easy to use, people with weak sight have some difficulty with Pixel Bar Char. We hope that improvement in screen quality will reduce this problem.

5.3 Evaluation

We have evaluated the two types of representation with 20 subjects, computer scientists or economists.

Performance were rather good with both visualization but on our test the tasks to be performed were rather simple.

We have thus more analysed preferences and qualitive remarks done by the subjects. Alternate Bar Chart is better perceived, because very close to standard representation. Pixel Bar Chart needs some time for training.

Classification method need also training. Some persons in the sample did not know about K-Means so that preliminary explanations were necessary. Some experts in economy found the method interesting, for use on very small screens.

6 Conclusion

We have presented solutions for time series visualization on PDA and on mobile phone, and considered the particular application of stock market. In our system we have applied the following principles:

- use of the maximum screen space

- do not charge the display
- use contrast boundary instead of line boundary
- give information on demand, interactively.

A substantial improvement will be obtained in linking together the different representations. We are working on such a global system. This system should be tested with real users, in their working environment.

More generally many other applications need to find adapted visualizations on small screens. The research has only started to be explored.

References

[Bornträger *et al.*, 2003]C. Bornträger, K. Cheverst, N. Davies, A. Dix, A. Friday, and J. Seitz. Experiments with multi-modal interfaces in a contex-aware guide city. *Mobile HCI 2003 Udine*, pages 116–130, 2003.

[Chinchille *et al.*, 2002]D. Chinchille, M. Goldstein, M. Nyberg, and M. Eriksson. Lost or found? a usability evaluation of a mobile navigation and location-based service. *Mobile HCI 2002 Pisa*, pages 224–240, 2002.

[Chittaro and Camaggio, 2002]L. Chittaro and A. Camaggio. Visualizing bar charts on wap phones. *Mobile HCI 2002 Pisa*, pages 224–240, 2002.

[Custinne *et al.*, 2004]G. Custinne, M. Noirhomme-Fraiture, and L. Chittaro. Visualisation d'informations boursières sur téléphones mobiles. *IHM 2004 Namur*, pages 224–240, 2004.

[Fithian *et al.*, 2003]R. Fithian, G. Iachello, J. Moghazy, Z. Pousman, and J. Stasko. The design and evaluation of a mobile-location awarehandeld event planner. *Mobile HCI 2003 Udine*, pages 145–160, 2003.

[Giller *et al.*, 2003]V. Giller, R. Melcher, J. Schrammel, R. Sefelin, and M. Tscheligi. Usability evaluations for multi-device application development three example studies. *Mobile HCI 2003 Udine*, pages 224–240, 2003.

[Jones *et al.*, 2002]M. Jones, G. Buchanan, and H. Thimblely. Sorting out searching on small screen devices. *Mobile HCI 2002 Pisa*, pages 224–240, 2002.

[Noirhomme-Fraiture *et al.*, 2005]M. Noirhomme-Fraiture, F. Randolet, and L. Chittaro. Visualisation of annual time series on pda's. *HCI International 2005 Las Vegas*, 2005.

[Pospischil *et al.*, 2002]G. Pospischil, M. Umlauft, and E. Michlmayr. Designing lol@, a mobile tourist guide for umts. *Mobile HCI 2002 Pisa*, pages 224–240, 2002.

[Sanneblad and Holmquist, 2003]J. Sanneblad and L.E. Holmquist. Opentrek: A platform for developing interactive networked games on mobile devices. *Mobile HCI 2003 Udine*, pages 224–240, 2003.

[Scapin, 1986]L. Scapin. *Guide ergonomique de conception des interfaces Homme-Machine.* INRIA, Paris, 1986.

[Schmidt-Belz and Hermann, 2004]B. Schmidt-Belz and F. Hermann. User validation of a nomadic exhibition guide. *Mobile HCI 2004 Glasgow*, pages 86–97, 2004.

[Spence, 2000]R. Spence. *Information Visualization.* Adison Wesley, New-York, 2000.

[Tarrini *et al.*, 2002]L. Tarrini, T. Bianchi Bandinelli, V. Miori, and G. Bertini. Remote control of home automation systems with mobile devices. *Mobile HCI 2002 Pisa*, pages 224–240, 2002.

Expert consulting and information combining: a sequential model

Paola Monari, Patrizia Agati, and Luisa Stracqualursi

Dipartimento di Scienze Statistiche
Università di Bologna
via delle Belle Arti, 41
I-40126 Bologna, Italy
(e-mail: paola.monari@unibo.it)
(e-mail: patrizia.agati@unibo.it)
(e-mail: stracqualursi@stat.unibo.it)

Abstract. In many research fields, where valuable information about a random phenomenon may come from different, possibly heterogeneous sources of knowledge ("experts"), the combining of the available information is a powerful uncertainty-reducing process. As efficiency reasons often suggest to perform a *sequential* procedure, in this paper some informativeness-founded selecting and stopping rules are proposed; their performance is discussed in a case-study.
Keywords: sequential consulting, Kullback-Leibler divergence, curvature.

1 Introduction

In many research fields, particularly in decision making and risk analysis, valuable information about a random phenomenon may come from different, possibly heterogeneous, sources of knowledge: information systems (such as, for example, sensor fusion systems), theoretical or empirical models, privileged witnesses. In a single, conventional word: 'experts'. So, the combining of the available information — especially once they were modelled in form of probability distributions — become a powerful uncertainty-reducing process: for example, to assess the entity of an environmental risk or the probability of a space probe malfunctioning, or forecast hurricane track, or classify biological samples, such as fossils. The output of the process — a final probability distribution on the investigated random variable — can be viewed as representing a synthesis of the current state of knowledge regarding the uncertainty of interest: a 'sufficient' synthesis, which must not involve loss of any relevant information.

Numerous algorithms for *simultaneous* combining have been proposed in literature (for a critical review, [Genest and Zidek, 1986] and [Cooke, 1991]). It's not so about *sequential* algorithms. And it is a fact that the investigator often prefers to consult the experts in successive stages rather than simultaneously. So, s/he avoids wasting time and money by consulting a too large sample of experts: at each stage, depending on the amount of information reached, s/he can choose whether to stop or to continue the process and, depending on the answers obtained from the experts already contacted, s/he can select the 'best' expert to be consulted on the subsequent stage.

The aim of this work is to propose some selecting and stopping rules which can be suitable to be used in a sequential consulting process. The substance of such rules is almost independent of the procedure chosen for combining information from the experts; not so their mathematical form. The reference, in the present work, is the Bayesian aggregation model suggested by Morris (1977), reviewed in a recursive form.

The paper is organized as follows. Section 2, in writing Morris' aggregation algorithm in a recursive form, gives the notation for the successive sections. In Section 3, some stopping and selecting criteria are suggested. Their performance is discussed in a real data based case-study which, together with some concluding remarks, is presented in Section 4.

2 A recursive algorithm for the sequential knowledge updating

In a context of uncertainty about the value of a random quantity $\theta \in \Theta \subset \Re$, let's denote with $h_0(\theta)$ the prior probability distribution which reflects the initial state of information. With the aim to acquire knowledge (so reducing the uncertainty) about θ, an investigator A performs a sequential consulting of (at most n) experts Q_j: at each stage k ($k = 1, 2, \ldots, K$; $K \leq n$), the selected expert $Q^*_{j;k}$ (or, more briefly, Q_k) answers by giving his/her/its own density $g_k(\theta)$. Treating each expert's density as result of an experiment, the investigator can revise the initial distribution $h_0(\theta)$ via Bayes' theorem.

Assuming that [Morris, 1977]:

a) each $g_k(\cdot)$ is parameterized with a location parameter m_k and a shape parameter v_k;

b) for each k, the probability which A assigns to the event $v^{(k)} = \bigcap_{i=1}^{k} v_i$ — that is, the event "the shape parameter values the experts will give are $[v_1, ..., v_i, ..., v_k]' = \mathbf{v}$" — does not depend on θ: in symbols, $\ell\left(v^{(k)}|\theta\right) = \ell\left(v^{(k)}\right)$;

Morris shows that the posterior density can be written as[1],

$$h\left(\theta|m^{(k)}, v^{(k)}\right) = \frac{\ell\left(m^{(k)}|v^{(k)}, \theta\right) \cdot h_0(\theta)}{\int_\Theta \ell\left(m^{(k)}|v^{(k)}, \theta\right) \cdot h_0(\theta)\, d\theta} \tag{1}$$

where:

[1] It can be shown that these assumptions can be relaxed without changing substantially the results [Morris, 1977].

- $\ell\left(m^{(k)}|v^{(k)},\theta\right)$, denoted in the following by $\ell_k(\theta)$ for notational convenience, indicates the conditioned likelihood function of θ for the data $m^{(k)}=\bigcap_{i=1}^k m_i$, given $v^{(k)}$: it represents — for θ varying — A's probabilities that the location parameter values the experts will provide are $\mathbf{m}=[m_i]'_{i=1,\ldots,k}$;
- the posterior $h\left(\theta|m^{(k)},v^{(k)}\right)$ or, more briefly, $h_k(\theta)$, represents the synthesis distribution at stage k.

If the following assumption holds too:

c) for each k, the conditional probability which A assigns to the event "$g_k(\cdot)$ shape parameter will be v_k", given $m^{(k-1)},v^{(k-1)}$ and θ, does not depend on θ — that is, $\ell\left(v_k|m^{(k-1)},v^{(k-1)},\theta\right)=\ell\left(v_k|m^{(k-1)},v^{(k-1)}\right)$—;

then Morris' (simultaneous) aggregation algorithm (1) can be written in a recursive form as,

$$h_k(\theta)=\frac{\ell\left(m_k|v_k,m^{(k-1)},v^{(k-1)},\theta\right)\cdot h_{k-1}(\theta)}{\int_\Theta \ell\left(m_k|v_k,m^{(k-1)},v^{(k-1)},\theta\right)\cdot h_{k-1}(\theta)\,d\theta} \tag{2}$$

where $\ell\left(m_k|v_k,m^{(k-1)},v^{(k-1)},\theta\right)$ is the conditioned likelihood function of θ for the only observation m_k, given v_k and also the location and shape values provided by the $k-1$ previously consulted experts.

As regards the arduous assessment of the function $\ell(\cdot)$ in (2), the relation $\ell\left(m_k|v_k,m^{(k-1)},v^{(k-1)},\theta\right)=\ell_k(\theta)/\ell_{k-1}(\theta)$ allows to use Morris' (simultaneous) result,

$$\ell_k(\theta)\propto C_k(\theta)\cdot\prod_{i=1}^k g_i(\theta) \tag{3}$$

where the *calibration function* $C_k(\theta)$ encapsulates the state of knowledge about each expert's performance and the degree of dependence among the k experts. Briefly [Morris, 1977], let τ_i denote the i-th *performance indicator*, defined as Q_i's cumulative function $G_i(\cdot|m_i,v_i)$ evaluated at the true value of θ: $C_k(\theta)$ expresses the admissibility degrees which the investigator assigns to each possible θ value looked at as the realization of the k-dimensional quantile vector $\tau=[\tau_i]'_{i=1,\ldots,k}$. Technically, $C_k(\cdot)$ is nothing but a subjectively assessed density $\phi_k(\cdot)$ of τ, conditioned on $\mathbf{v}$ and θ, looked at as a function of θ (for fixed $\mathbf{m}$): in symbols, the relation between the so-called *performance function* $\phi_k(\cdot)$ and the calibration function $C_k(\theta)$ is,

$$\phi_k(\tau|\mathbf{v},\theta)=\phi_k\left[\mathbf{G}(\theta|\mathbf{m},\mathbf{v})\,|\mathbf{v},\theta\right]=C_k(\theta) \tag{4}$$

where $\mathbf{G}(\theta|\mathbf{m},\mathbf{v})$ — briefly, $\mathbf{G}(\theta)$ — denotes the vector $[G_i(\theta|m_i,v_i)]'_{i=1,\ldots,k}$.

Whenever only some pieces of information about the experts are available — an 'information block' which is not adequate to construct an empirically founded probability distribution of their performance indicators — the fiducial argument [Fisher, 1956] can be used for inductively modelling the calibration function, enabling it to be specified with a relatively small number of assessments [Monari and Agati, 2001]. With the following notation:

- $\tilde{\mathbf{G}}(\theta) = \left[\tilde{G}_i(\theta)\right]'_{i=1,\ldots,k}$, with $\tilde{G}_i(\theta) = \ln\left[G_i(\theta)/\left(1 - G_i(\theta)\right)\right]$;
- $\tilde{\mathbf{t}} = \left[\tilde{t}_i\right]'_{i=1,\ldots,k}$, with $\tilde{t}_i = \ln\left[t_i/\left(1 - t_i\right)\right]$;
- c as normalization constant;

the resulting fiducial calibration function can be written as,

$$C_k(\theta) = C_k(\theta; \mathbf{t}, \mathbf{S}) =$$

$$= c \cdot \prod_{i=1}^{k} \left\{G_i(\theta) \cdot \left[1 - G_i(\theta)\right]\right\}^{-1} \cdot \exp\left\{-\frac{1}{2}\left[\tilde{\mathbf{G}}(\theta) - \tilde{\mathbf{t}}\right]' \mathbf{S}^{-1}\left[\tilde{\mathbf{G}}(\theta) - \tilde{\mathbf{t}}\right]\right\} \quad (5)$$

It's worth noting that function (5) is univocally defined by the following two quantities:

- A's assessment $\mathbf{t} = [t_i]'_{i=1,\ldots,k}$ of the performance indicator τ;
- the subjective variance-covariance matrix $\mathbf{S}$, reflecting A's information about the variability and the reciprocal dependence of the experts' performance indicators.

3 Selecting and stopping rules

The purpose of expert consulting is reducing the uncertainty about the unknown quantity θ. So, in designing and performing the sequential process, it is reasonable to found the selecting and stopping rules on some criterion of informativeness. In particular, though no single number can convey the amount of information encapsulated in a density function, a synthetic measure of the (*expected*) additional informative value of a not-yet-consulted expert $Q_{j;k}$ is indispensable for selecting the one to be consulted at stage k, especially when the investigator's calibration assessments, together with the shape parameters provided by the experts, lead to not-coinciding preference orderings. And, analogously, as likelihood functions and posterior densities can display a wide variety of form, a synthetic measure of the reached knowledge degree about θ is needed for picking out the 'optimal' stage k^* at which data acquiring can be stopped.

Let's suppose the investigator A is performing the process of revising beliefs in light of new data according to the algorithm described in Section 2. The prior $h_0(\theta)$ has already been specified; each of n contacted experts Q_j has revealed the variance v_j — assumed as uninformative about θ: see *b)* in Section 2) — of his/her/its own density $g_j(\theta)$, and A has already consulted $k-1$ of them, so obtaining the locations of $k-1$ expert densities: A is

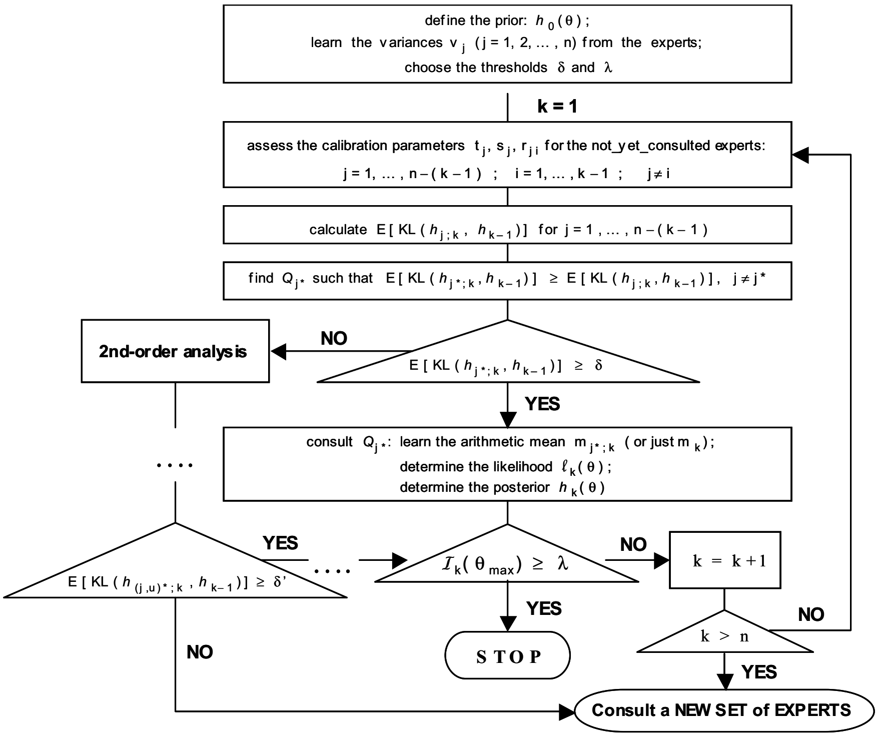

define the prior: $h_0(\theta)$;
learn the variances v_j ($j = 1, 2, \ldots, n$) from the experts;
choose the thresholds δ and λ
k = 1
assess the calibration parameters t_j, s_j, r_{ji} for the not_yet_consulted experts:
$j = 1, \ldots, n-(k-1)$; $i = 1, \ldots, k-1$; $j \neq i$
calculate $E[KL(h_{j;k}, h_{k-1})]$ for $j = 1, \ldots, n-(k-1)$
find Q_{j^*} such that $E[KL(h_{j^*;k}, h_{k-1})] \geq E[KL(h_{j;k}, h_{k-1})]$, $j \neq j^*$
NO
2nd-order analysis
$E[KL(h_{j^*;k}, h_{k-1})] \geq \delta$
YES
consult Q_{j^*}: learn the arithmetic mean $m_{j^*;k}$ (or just m_k);
determine the likelihood $\ell_k(\theta)$;
determine the posterior $h_k(\theta)$
....
YES
$E[KL(h_{(j,u)^*;k}, h_{k-1})] \geq \delta'$
....
$\mathcal{I}_k(\theta_{max}) \geq \lambda$
NO
k = k+1
NO
k > n
YES
NO
STOP
YES
Consult a NEW SET of EXPERTS

Fig. 1. Flow-chart of the sequential procedure.

now at stage k of the process (figure 1), and must select one among the not-yet-consulted experts $Q_{j;k}$ $(j = 1, 2, \ldots, n-k+1)$.

For each $Q_{j;k}$, the investigator A assesses — conditionally on v_j, on the basis of the information at his disposal (including all the expert locations m_i revealed up to stage $k-1$) — the parameters of the k-stage calibration function $C_{j;k}(\theta)$: that is, t_j, s_{jj} and the covariances s_{ji} (or the linear correlations r_{ji}) between $Q_{j;k}$ and each already-consulted expert Q_i, $i = 1, 2, \ldots, k-1$. At this point of the procedure, no $Q_{j;k}$ has revealed the location value m_j of his own $g_j(\theta)$: the several 'answers' m_j which each can virtually give are not all equally informative, so the (informative) value of each expert at the k-stage — to be measured with regard to A's current knowledge[2] of θ reflected in the posterior density $h_{k-1}(\theta)$ of the previous stage — is an *expected* value,

[2] In fact, all the other elements being equal, the more A is uncertain about θ, the more an answer m_j is worthy.

calculated by averaging a selected measure of relevant information about θ in $Q_{j;k}$'s answer over the space M_j of the virtually possible m_j values.

By reasoning in a *knowledge* context — which is an *inductive* context, where an expert opinion is more relevant the more it is able to modify the posterior distribution on the unknown quantity — a suitable measure of $Q_{j;k}$'s informative value can be the *expected Kullback-Leibler divergence* of the density $h_{j;k}(\theta)$ with respect to the previous stage posterior $h_{k-1}(\theta)$,

$$\mathrm{E}\left[KL\left(h_{j;k}, h_{k-1}\right)\right] := \int_{M_j} f\left(m_{j;k}|v_{j;k}, m^{(k-1)}, v^{(k-1)}\right) \cdot KL\left(h_{j;k}, h_{k-1}\right) dm_j \tag{6}$$

where the KL-divergence [Kullback, 1959],

$$KL\left(h_{j;k}, h_{k-1}\right) := \int_{\Theta} h_{j;k}(\theta) \cdot \ln\left[h_{j;k}(\theta) / h_{k-1}(\theta)\right] d\theta \tag{7}$$

measures indirectly the information provided by an answer $m_{j;k}$ in terms of the changes it yields on the density $h_{k-1}(\theta)$. The conditional density $f(\cdot)$ in (6) is equal to the denominator of (2) read as a function of $m_{j;k}$ and normalized; when assumptions *a)*, *b)* and *c)* hold, it can be determined as

$$f\left(m_{j;k}|v_{j;k}, m^{(k-1)}, v^{(k-1)}\right) = f\left(m^{(j;k)}|, v^{(j;k)}\right) / f\left(m^{(k-1)}|v^{(k-1)}\right) \tag{8}$$

where the density $f\left(m^{(j;k)}|, v^{(j;k)}\right)$ — and analogously $f\left(m^{(k-1)}|v^{(k-1)}\right)$ — is equal, up to the normalization term, to the denominator $\int_{\Theta} \ell\left(m^{(k)}|v^{(k)}, \theta\right) \cdot h_0(\theta)\, d\theta$ of (1), read as a function of $m^{(k)}$.

The expert $Q^*_{j;k}$ presenting the greatest expected KL-divergence is, at stage k, the most informative: but is he/she/it an expert worth consulting? The answer is yes, if the information he provides is, on average, *enough* different from what A already knows about θ, *i.e.* if the expected divergence of $h_{j^*;k}(\theta)$ with respect to $h_{k-1}(\theta)$ is not less than a predetermined value δ ($0 \leq \delta < \infty$). About the choose of the threshold δ, a very useful tool is the scheme proposed by McCulloch for deciding whether a KL-divergence value is a large or a small one [McCulloch, 1989].

So the *selecting rule* can be expressed as follows. *Consult the expert* $Q^*_{j;k}$ *such that*

$$\mathrm{E}\left[KL\left(h_{j^*;k}, h_{k-1}\right)\right] \geq \mathrm{E}\left[KL\left(h_{j;k}, h_{k-1}\right)\right] \qquad j \neq j^* \tag{9}$$

on condition that

$$\mathrm{E}\left[KL\left(h_{j^*;k}, h_{k-1}\right)\right] \geq \delta \tag{10}$$

If $Q^*_{j;k}$ *does not satisfy* (10), *then proceed to a 2nd order analysis: that is, consult the pair* $\left(Q_{j;k}, Q_{u;k}\right)^*$ *presenting the greatest expected KL-divergence, provided that it is* $\mathrm{E}\left[KL\left(h_{(j,u)^*;k}, h_{k-1}\right)\right] \geq \delta$*; otherwise contact a new set of experts and perform a new process by using the posterior* $h_{k-1}(\theta)$ *as a new prior* $h_0^{'}(\theta)$.

The expert $Q^*_{j;k}$ satisfying (10) becomes just Q_k, the "k-stage expert". By consulting him, A learns the location m_k of the density $g_k(\cdot)$: now, the k-stage calibration function $C_k(\theta)$ is univocally defined, and consequently, the likelihood function $\ell_k(\theta)$ and the posterior density $h_k(\theta)$ too.

In theory, the investigator should stop the process only when the knowledge about θ, reflected in the posterior density, is 'inertially stable': *i.e.*, only when additional experts, even if jointly considered, are not able to modify appreciably the synthesis distribution, on the contrary they contribute to its inertness. But too many experts could be needed for realizing such a stopping condition. It can be weakened by requiring just the knowledge about θ deriving from expert answers to be enough for A's purposes. A measure encapsulating the strength of the experimental data in determining a preference ordering among 'infinitesimally close' values of θ is Fisher's notion of information. The value of the *observed information* $I(\cdot)$ at the maximum of the log-likelihood function,

$$I_k(\theta_{\max}) := -\partial^2/\partial\theta^2 \ln \ell_k(\theta_{\max}) \tag{11}$$

is a second-order estimate of the spherical curvature of the function at its maximum: within a second-order approximation, it corresponds to the KL-divergence between two distributions that belong to the same parametric family and differ infinitesimally over the parameter space.

So, the *stopping rule* may be defined as follows. *Stop the consulting at stage k^* at which a pre-selected observed curvature λ of the log-likelihood valued at $\theta := \theta_{\max}$ has been reached,*

$$I_{k^*}(\theta_{\max}) \geq \lambda \tag{12}$$

For deciding whether a curvature value $I(\theta_{\max}) = w$ is a large or a small one, a device could be the following. Let's think of a binomial experiment where a number $x = n/2$ of successes is observed in n trials and find x such that $I(\hat{p}_{ML} = 0.5) = w$, where $\hat{p}_{ML} = 0.5$ is the maximum likelihood estimate of the binomial parameter p. Table 1 shows a range of x values with the corresponding w curvature values. The simple relation $x = w/8$ holds: so, for example, if $w = 120$, the width of the curve $\ln \ell_k(\theta)$ near $\theta := \theta_{\max}$ is the same as the curve $\ln \ell(p)$ at $\hat{p}_{ML} = 0.5$ when $x = 15$ and $n = 30$.

x	1	2	5	10	15	20	25	30	40	50
w	8	16	40	80	120	160	200	240	320	400

Table 1. Large or small curvature values? Relation between x and w values.

4 Case-study and concluding remarks

The behavior of the algorithms proposed in the previous section — and implemented [Agati and Stracqualursi, 2001] in MATHEMATICA — has been investigated in simulation and experimental studies. In this section, the results from medical data are synthetically presented to exemplify how the selecting and stopping rules work. Particularly, data in table 2 regard a sequential consulting process of $n = 4$ orthopaedists, performed by an Italian research laboratory about the long-term failure log-odds θ of a new hip prosthesis. A fifth surgeon has assessed the calibration parameters, without modifying them in proceeding from a stage to the successive one. He has also (subjectively) chosen the following thresholds:

- $\delta = 0.02$: by reading this value in McCulloch's scale, at stage k the most informative expert $Q^*_{j;k}$ is consulted only if the expected KL-divergence of $h_{j^*;k}(\theta)$ with respect to $h_{k-1}(\theta)$ is not less than the KL-divergence of a Bernoulli distribution $B(p)$ with $p = 0{,}5$ from a Bernoulli distribution with $p = 0.65$; or, in other words, only if stopping the process at stage $k-1$ instead of proceeding to stage k involves, on average, an information loss larger than that one yielded by using a $B(0{,}65)$ instead of a $B(0{,}5)$;
- $\lambda = 120$: by using the scale proposed in Section 3, the consulting process is stopped at stage k* at which the observed curvature of the log-likelihood function $\ln \ell(\theta)$ valued at $\theta := \theta_{\max}$ is the same as the function $\ln \ell(p)$ at $\hat{p}_{ML} = 0.5$ when, in a binomial experiment, $n = 30$ and $x = 15$.

Q_j	v_j	t_j	s_{jj}	r_{j1}	r_{j2}	r_{j3}	r_{j4}
Q_1	0.150	0.45	1.20	1			
Q_2	0.145	0.65	1.50	+0.20	1		
Q_3	0.120	0.75	1.70	−0.05	+0.50	1	
Q_4	0.110	0.45	1.10	+0.10	+0.10	+0.10	1

Table 2. Input data for the sequential consulting of four orthopaedists about long-term failure log-odds of a new hip prosthesis.

In this study, the conditions *a)*, *b)* and *c)* mentioned in Section 2 can be held to be satisfied. In fact: *a)* it rests on empirical evidence — and the experts confirm it — that the failure log-odds θ can be supposed as Gaussian; *b)* it is reasonable to think the probability the fifth orthopaedist assigns to the event "the experts will give the variances $[v_1, ..., v_4]' = \mathbf{v}$" is the same for all θ values: so the surgeons' stated variances alone give no information able to change the investigator's beliefs about θ; *c)* it is reasonable as well to assume the conditional probability the investigator assigns to the event "the expert $Q_{j;k}$ will give the variance v_k", given the shape and location values provided by the $k-1$ previously consulted experts, is the same for all θ values. So the combining algorithm outlined in Section 2 has been applied, as well as the selecting and stopping rules suggested in Section 3.

Q_j	*Stage* $k=1$ $\mathrm{E}\left[KL\left(h_{j;1},h_0\right)\right]$	*Stage* $k=2$ $\mathrm{E}\left[KL\left(h_{j;2},h_1\right)\right]$	*Stage* $k=3$ $\mathrm{E}\left[KL\left(h_{j;3},h_2\right)\right]$
Q_1	1.41487	**1.92935**	—
Q_2	1.35582	1.52293	1.42842
Q_3	1.42427	1.72981	**1.93624**
Q_4	**1.60348**	—	—
	↓	↓	↓
$Q^*_{j;k}$	Q_4	Q_1	Q_3
m_k	-1.208	-1.992	-2.752
$I_k\left(\theta_{\max}\right)$	18.713	53.492	138.984 $(> 120 = \lambda)$

Table 3. Output of the proposed sequential procedure in the consulting of four orthopaedists about long-term failure log-odds of a new hip prosthesis.

Table 3 summarizes the results of the sequential process, while figure 2 shows the posterior distributions $h_k\left(\theta\right)$ at each stage.

For $k=1$, the selecting rule proposed in Section 2 chooses the expert Q_4: really he offers the smallest variance ($v_4 = 0.110$), and also the investigator's uncertainty about his performance indicator is assessed to be the smallest ($s_{44} = 1.10$). Q_4's answer ($m_{4;1} = -1,208$) leads to a curvature value $I_1\left(\theta_{\max}\right) = 18.713 < 120 = \lambda$: so the process goes on.

At stage 2, the selecting rule shows its usefulness: in fact, the v_j, t_j and s_{jj} values [3] don't lead to a unique preference ordering. The most informative expert Q_1 is selected by the algorithm[4] and $m_{1;2}$ is observed. The curvature value is $I_2\left(\theta_{\max}\right) = 53.492 < 120 = \lambda$: the consulting proceeds.

At stage 3, the preference for Q_3 instead of Q_2 is also (but not only) motivated by the correlations with Q_1: a negative correlation ($r_{31} = -0.05$) is more informative than a weak positive one ($r_{21} = 0.20$). The observed $m_{3,3}$ leads to $I_3\left(\theta_{\max}\right) = 138.984 > 120 = \lambda$. The process is stopped: the expert Q_2 is left out of the consulting and stage-3 posterior $h_3\left(\theta\right)$ — whose location and shape values are, respectively, -1.873 (the median, here coinciding with the arithmetic mean and the mode) and 0.084 (the standard deviation) — can be regarded as the synthesis expression of the expert knowledge about the long-term failure log-odds θ of the new hip prostheses.

[3] The correlations between Q_4 and the other experts are all equals: so they don't come into play.

[4] It's worth noting that the value $m_{4;1}$ observed at stage 1 has modified, at stage 2, the previous-stage preference ordering: for this reason, the selecting at each stage one only expert is to be preferred to selecting a set of experts (simultaneously.

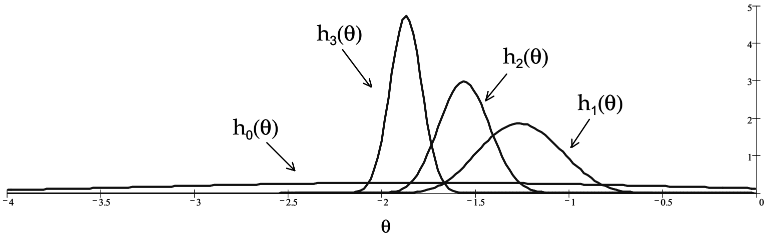


Fig. 2. Posterior distributions at stages 0 (*i.e.*, the prior), 1, 2 and 3 of the sequential procedure.

By looking at this selecting and stopping output, the behavior of the informativeness criteria appears to be coherent with the intuition, so giving an empirical support about the soundness of the proposed selecting and stopping algorithms in performing an efficient sequential consulting process. At present, our research efforts are focused on the combining of information from hurricane track prediction models: so, with the aim of assessing the calibration parameters for each model (an 'expert', in our framework), simulations were performed on a training-set of North Atlantic historical hurricane data regarding the location of specific storms at prefixed time intervals. Successively, separately for each time interval, each track prediction model with its own parameters entered in the informativeness-founded sequential algorithm and, on the basis of the selecting and stopping output, a Bayesian combined track prediction model for each prefixed time interval was proposed: the research — still in progress — promises interesting results.

References

[Agati and Stracqualursi, 2001]P. Agati and L. Stracqualursi. *Algoritmi computazionali per l'aggregazione di opinioni esperte.* Clueb, Bologna, 2001.

[Cooke, 1991]R.M. Cooke. *Experts in Uncertainty. Opinion and subjective probability in science.* Oxford University Press, Oxford, 1991.

[Fisher, 1956]R.A. Fisher. *Statistical Methods and Scientific Inference.* Oliver & Boyd, Edinburgh, 1956.

[Genest and Zidek, 1986]C. Genest and J. V. Zidek. Combining probability distributions: a critique and an annotated bibliography. *Statistical Science*, pages 114–148, 1986.

[Kullback, 1959]S. Kullback. *Information Theory and Statistics.* Wiley, New York, 1959.

[McCulloch, 1989]R.W. McCulloch. Local model influence. In *Journal of the American Statistical Association*, pages 473–478, 1989.

[Monari and Agati, 2001]P. Monari and P. Agati. Fiducial inference in combining expert judgements. *Journal of the Italian Statistical Society*, pages 81–97, 2001.

[Morris, 1977]P. A. Morris. Combining expert judgments: a bayesian approach. *Management Science*, pages 679–693, 1977.

Analysis of Multinomial Response Data: a Measure for Evaluating Knowledge Structures

Ali Ünlü

Department of Psychology
University of Graz
Universitätsplatz 2/III
A-8010 Graz, Austria
(e-mail: ali.uenlue@uni-graz.at)

Abstract. Multinomial response data obtained from nominally and dichotomously scored test items in knowledge space theory are explained by knowledge structures. A central problem is the derivation of a "realistic" explanation, i.e., knowledge structure, representing the organization of "knowledge" in a domain and population of reference. In this regard, often, one is left with the problem of selecting among candidate competing explanations for the data. In this paper, we propose a measure for the selection among competing knowledge structures. The approach is illustrated with simulated data.

Keywords: Discrete multivariate response data, Qualitative test data analysis, Knowledge space theory, Selection measure, Simulation.

1 Knowledge space theory (KST)

This section reviews basic deterministic and probabilistic concepts of KST. For details, refer to [Doignon and Falmagne, 1999].

Definition 1 *A knowledge structure is a pair* $(Q, \mathcal{K})$*, with* Q *a non-empty, finite set, and* $\mathcal{K}$ *a family of subsets of* Q *containing at least the empty set* $\emptyset$ *and* Q*. The set* Q *is called the domain of the knowledge structure. The elements* $q \in Q$ *and* $K \in \mathcal{K}$ *are referred to as (test) items and (knowledge) states, respectively. We also say that* $\mathcal{K}$ *is a knowledge structure on* Q*.*

The general definition of a knowledge structure allows for infinite item sets as well. However, throughout this work, we assume that Q is finite.

The set Q is supposed to be a set of *dichotomous* items. In this paper, we interpret Q as a set of dichotomous questions/problems that can either be *solved* (coded as 1) or *not solved* (coded as 0). Here, "solved" and "not solved" stand for the observed responses of a subject (*manifest level*). This has to be distinguished from a subject's true, unobservable knowledge of the solution to an item (*latent level*). In the latter case, we say that the subject is *capable of mastering* (coded as 1) or *not capable of mastering* (coded as 0) the item. For a set X, let 2^X denote its *power-set*, i.e., the set of all subsets of X. Let $|X|$ stand for the *cardinality* (*size*) of X. The observed responses of a subject to the items in Q are represented by the subset $R \subset Q$ containing exactly the items that are solved by the subject. This subset R is called the *response pattern* of the subject. Similarly, the true latent state of knowledge of a subject with respect to the items in Q is represented by the subset $K \subset Q$ containing exactly the items the subject is capable of mastering. This subset K is called the *knowledge state* of the subject. Given a knowledge structure $\mathcal{K}$, we assume that the only states of knowledge possible are the ones in $\mathcal{K}$. In this sense, $\mathcal{K}$ captures the organization of knowledge in the domain and population of reference. Idealized, if no response errors, i.e., careless errors and lucky guesses, would be committed, the only response patterns possible would be the knowledge states in $\mathcal{K}$.

Let $\mathbb{N}$ stand for the set of natural numbers (without 0). We fix a population of reference, and examinees are drawn from this population randomly. Let the sample size be $N \in \mathbb{N}$. The data is constituted by the observed absolute counts $N(R) \in \mathbb{N}_0 := \mathbb{N} \cup \{0\}$ of response patterns $R \in 2^Q$. The data, $\mathbf{x} = (N(R))_{R\in 2^Q}$, are assumed to the realization of a random vector $\mathbf{X} = (X_R)_{R\in 2^Q}$, which is distributed *multinomially* over 2^Q. That is,

$$\begin{aligned}\mathbb{P}(\mathbf{X} = \mathbf{x}) &:= \mathbb{P}(X_\emptyset = N(\emptyset), \ldots, X_Q = N(Q)) \\ &= \frac{N!}{\prod_{R\in 2^Q} N(R)!} \prod_{R\in 2^Q} \rho(R)^{N(R)}.\end{aligned}$$

Here, $\rho(R) \in [0,1]$ for any $R \in 2^Q$, $\sum_{R\in 2^Q} \rho(R) = 1$, and $N(R) \in \mathbb{N}_0$ with $0 \leq N(R) \leq N$ for any $R \in 2^Q$, $\sum_{R\in 2^Q} N(R) = N$.

Let the maximum probability of occurence be denoted by $\rho(R_m)$, i.e.,

$$\rho(R_m) = \max_{R\in 2^Q} \rho(R),$$

for some appropriate response pattern $R_m \in 2^Q$.

Maximum likelihood estimates (briefly, MLEs) for the population probabilities $\rho(R)$ ($R \in 2^Q$) are $\widehat{\rho(R)} = N(R)/N$. The MLE for $\rho(R_m)$ is $\widehat{\rho(R_m)} = N(R'_m)/N$, where $N(R'_m)$ denotes the maximum absolute count $N(R'_m) = \max_{R\in 2^Q} N(R)$, for some appropriate response pattern $R'_m \in 2^Q$.

We will simulate multinomial response data in accordance with a *basic local independence model.*

Definition 2 *A quadruple $(Q, \mathcal{K}, p, r)$ is called a basic local independence model (BLIM) iff*

1 *$(Q, \mathcal{K})$ is a knowledge structure;*

2 *p is a probability distribution on $\mathcal{K}$, i.e., $p : \mathcal{K} \to [0,1], K \mapsto p(K)$, with $p(K) \geq 0$ for any $K \in \mathcal{K}$, and $\sum_{K\in\mathcal{K}} p(K) = 1$;*

3 *r is a response function for $(Q, \mathcal{K}, p)$, i.e., r is a function $r : 2^Q \times \mathcal{K} \to [0,1]$, $(R, K) \mapsto r(R, K)$, with $r(R, K) \geq 0$ for any $R \in 2^Q$ and $K \in \mathcal{K}$, and $\sum_{R\in 2^Q} r(R, K) = 1$ for any $K \in \mathcal{K}$;*

4 *r satisfies local independence, i.e.,*

$$r(R,K) = \left\{ \left[\prod_{q \in K \setminus R} \beta_q \right] \cdot \left[\prod_{q \in K \cap R} (1-\beta_q) \right] \right.$$
$$\left. \cdot \left[\prod_{q \in R \setminus K} \eta_q \right] \cdot \left[\prod_{q \in Q \setminus (R \cup K)} (1-\eta_q) \right] \right\},$$

with two constants $\beta_q, \eta_q \in [0,1[$ for each $q \in Q$, respectively called careless error probability and lucky guess probability at q.

A probability distribution p on $\mathcal{K}$ (point **2**) is interpreted as follows. To each knowledge state $K \in \mathcal{K}$ is attached a probability $p(K) \in [0,1]$ measuring the likelihood that a randomly sampled subject is in state K. Further, any randomly sampled subject is necessarily in exactly one of the states of $\mathcal{K}$. A response function r (point **3**) is interpreted as follows. For $R \in 2^Q$ and $K \in \mathcal{K}$, $r(R,K) \in [0,1]$ specifies the conditional probability of response pattern R for an examinee in state K. Given the probability distributions p on $\mathcal{K}$ and $r(\,.\,,K)$ on 2^Q $(K \in \mathcal{K})$, a BLIM takes into account the two ways in which probabilities must supplement deterministic knowledge structures. For one, knowledge states will occur with different proportions in the population of reference. For another, response errors (careless errors and lucky guesses) will render impossible the a-priori specification of the observable responses of a subject, given her/his knowledge state. The condition of *local independence* (point **4**) states that the item responses of an examinee are assumed to be independent, given the knowledge state of the examinee, and the response error probabilities $\beta_q, \eta_q \in [0,1[$ $(q \in Q)$ are attached to the items and do not vary with the knowledge states.

The BLIM is a *multinomial* probability model.

Corollary 1 *Given a BLIM, the occurence probabilities of response patterns are parameterized as*

$$\rho(R) = \sum_{K \in \mathcal{K}} \left\{ \left[\prod_{q \in K \setminus R} \beta_q \right] \cdot \left[\prod_{q \in K \cap R} (1-\beta_q) \right] \right.$$
$$\left. \cdot \left[\prod_{q \in R \setminus K} \eta_q \right] \cdot \left[\prod_{q \in Q \setminus (R \cup K)} (1-\eta_q) \right] \right\} p(K).$$

□

2 Measure $\boldsymbol{\kappa}$

In this section, we propose a measure, κ, for the selection among competing explanations, i.e., knowledge structures, for the multinomial response data. For details, refer to [Ünlü, 2004].

2.1 Prediction paradigm

The derivation of κ heavily rests on the following *prediction paradigm.*

The *prediction problem* considered is this. An individual is chosen randomly from the population of reference, and we are asked to guess his/her response pattern, given, either

(no info). no further information (than the multinomial distribution), or

(info). the knowledge structure $\mathcal{K}$ assumed to underlie the responses of the individual.

The *prediction strategies* in both cases are as follows. In the "no info" case, we *optimally* guess some response pattern $R_m \in 2^Q$, which has the largest probability of occurence $\rho(R_m) = \max_{R\in 2^Q} \rho(R)$. In the "info" case, we *proportionally* guess the knowledge states with their probabilities of occurence. That is, if $\mathcal{K} = \{K_1, K_2, \dots, K_{|\mathcal{K}|}\}$, we guess K_1 with probability $\rho(K_1)$, K_2 with probability $\rho(K_2)$, …, $K_{|\mathcal{K}|}$ with probability $\rho(K_{|\mathcal{K}|})$. Since these probabilities may not add up to one, in general, there is a non-vanishing *residual* probability $\{1 - \sum_{K\in\mathcal{K}} \rho(K)\} > 0$. Thus, in order to complete the prediction strategy, we abstain from any guessing with probability $1-\sum_{K\in\mathcal{K}} \rho(K)$, and, in the sequel, view this as a prediction error.

The probabilities of a *prediction error* in both cases are as follows. In the "no info" case, the probability is $1 - \rho(R_m)$, and, in the "info" case, it is $1 - \sum_{K\in\mathcal{K}} \rho(K)^2$. Of course, the probabilities of a *prediction success* are $\rho(R_m)$ and $\sum_{K\in\mathcal{K}} \rho(K)^2$, respectively.

2.2 First constituent of κ: measure of fit

The measure κ consists of two constituents. The first constituent of κ captures the (descriptive) *fit* of a knowledge structure $\mathcal{K}$ to the response data. This constituent is derived based on the method of *proportional reduction in predictive error* (PRPE)—the method of PRPE was introduced originally by [Guttman, 1941], and it was applied systematically in the series of papers by [Goodman and Kruskal, 1954, 1959, 1963, 1972]. The general probability formula of the method of PRPE quantifies the *predictive utility*, *PU*(*info*), of given information. Informally,

$$PU(\mathit{info}) := \frac{\text{Prob. of error (no info)} - \text{Prob. of error (info)}}{\text{Prob. of error (no info)}}.$$

Inserting the previous prediction error probabilities into the PRPE formula, we obtain the population analogue of the first constituent, m_1, of κ.

Definition 3 *Let $\rho(R_m) \neq 1$. The measure m_1 is defined as*

$$\begin{aligned} m_1 &:= \frac{(1-\rho(R_m)) - (1-\sum_{K\in\mathcal{K}} \rho(K)^2)}{1-\rho(R_m)} \\ &= \frac{\sum_{K\in\mathcal{K}} \rho(K)^2 - \rho(R_m)}{1-\rho(R_m)}. \end{aligned}$$

In the sequel, we assume that $\rho(R_m) \neq 1$! Inserting MLEs, we obtain the MLE, $\widehat{m_1}$, for m_1 (We assume that $1 - N(R'_m)/N \neq 0$!):

$$\widehat{m_1} = \frac{\sum_{K \in \mathcal{K}} N(K)^2 - N \cdot N(R'_m)}{N^2 - N \cdot N(R'_m)}.$$

2.3 Second constituent of κ: measure of size

The second constituent of κ captures the *size* of a knowledge structure $\mathcal{K}$. For the definition of it, we need the concept of a *truncation* of $\mathcal{K}$.

Definition 4 *Let $M \in \mathbb{N}$ be a truncation constant. An M-truncation of $\mathcal{K}$ is any subset, $\mathcal{K}_{M\text{-}trunc}$, of $\mathcal{K}$ which is derived in the following way.*

1 Order the knowledge states $K \in \mathcal{K}$ according to their probabilities of occurence $\rho(K)$, say, from left to right, ascending with smaller ρ values to larger ones. Knowledge states with equal probabilities of occurence are ordered arbitrarily.

2 Starting with the foremost right knowledge state, i.e., a knowledge state with largest probability of occurence, take the first $\min(|\mathcal{K}|, M)$ *knowledge states, descending from right to left. The set of these knowledge states is $\mathcal{K}_{M\text{-}trunc}$.*

The definition of the second constituent, m_2, of κ is this.

Definition 5 *Let $\sum_{K \in \mathcal{K}} \rho(K) \neq 0$. Let $M \in \mathbb{N}$ be a truncation constant, and let $\mathcal{K}_{M\text{-}trunc}$ denote an M-truncation. The measure m_2 is defined as*[1]

$$m_2 := \frac{\sum_{K \in \mathcal{K}} \rho(K)^2}{\sum_{K \in \mathcal{K}_{M\text{-}trunc}} \rho(K)^2}.$$

In the sequel, we assume that $\sum_{K \in \mathcal{K}} \rho(K) \neq 0$ for any knowledge structure $\mathcal{K}$. Inserting MLEs, we obtain the MLE, $\widehat{m_2}$, for m_2 (We assume that $\sum_{K \in \mathcal{K}} N(K) \neq 0$!):

$$\widehat{m_2} = \frac{\sum_{K \in \mathcal{K}} N(K)^2}{\sum_{K \in \mathcal{K}_{\widehat{\text{M-trunc}}}} N(K)^2},$$

where $\mathcal{K}_{\widehat{\text{M-trunc}}}$ is defined analogously as in Definition 4, where we have to replace $\rho(K)$ with its MLE $N(K)/N$ for any $K \in \mathcal{K}$.

[1] m_2 is invariant with respect to the choice of a particular M-truncation.

2.4 κ: size trading-off fit measure

The measure κ is (more or less) the product of m_1 and m_2.

Definition 6 *Let $M \in \mathbb{N}$ be a truncation constant, and let $C \in [0, 0.01]$ be a small, fixed non-negative correction constant.*[2] *The measure κ is defined as*

$$\kappa := m_2 \cdot (m_1 - C).$$

The MLE for κ is $\widehat{\kappa} := \widehat{m_2} \cdot (\widehat{m_1} - C)$.

The measure κ may be interpreted as a *performance measure* for the evaluation of knowledge structures. The two (performance) criteria being merged and traded-off are "(descriptive) fit" and "(structure) size", respectively measured by its constituents m_1 and m_2. The *decision rule* important for applications of κ is this. *The greater the value of κ is, the "better" a knowledge structure "performs" with respect to a trade-off of the criteria.* The (unknown) ordering of the population κ values is "estimated" by the ordering of the corresponding MLEs.

2.5 Model selection and truncation constant

Finally, we describe a special choice for the truncation constant in the context of model selection among competing knowledge structures $\mathcal{K}_1, \mathcal{K}_2, \ldots, \mathcal{K}_n$ ($n \in \mathbb{N}$, $n \geq 2$) on (same) domain Q.

Definition 7 *Let $v_i := |\{K \in \mathcal{K}_i : \rho(K) \neq 0\}|$ be the match of candidate model $\mathcal{K}_i$ ($1 \leq i \leq n$). Let $\mathbf{v} := (v_1, v_2, \ldots, v_n)^T \in \mathbb{N}^n$ be the match vector. The (empirical) median of the matches $v_i \in \mathbb{N}$ ($1 \leq i \leq n$) is denoted by median($\mathbf{v}$) and called the median match of the competing models. Formally,*

$$\mathit{median}(\mathbf{v}) := \begin{cases} v_{(\frac{n+1}{2})} & : \ \mathit{odd}\ n \\ v_{(\frac{n}{2})} & : \ \mathit{even}\ n, \end{cases}$$

where $v_{(1)}, v_{(2)}, \ldots, v_{(n)}$ with $v_{(1)} \leq v_{(2)} \leq \cdots \leq v_{(n)}$ is the ordered list of matches v_i ($1 \leq i \leq n$).

The special truncation constant, M_s, is this.

Definition 8 *The special truncation constant M_s is defined as*[3]

$$M_s := \min\left([2^{|Q|/2}], \mathit{median}(\mathbf{v})\right).$$

[2] C is introduced to compensate for a zero value of m_1.

[3] The meaning of term $2^{|Q|/2}$ is clarified in the context of knowledge assessment procedures (for details, see [Ünlü, 2004]). For any real $x \geq 0$, $[x]$ denotes the entier of x, i.e., the integer $I \in \mathbb{N} \cup \{0\}$ with $I \leq x < I + 1$.

3 Simulation example

In this section, we apply κ to data simulated in accordance with a specific BLIM. For details (including software), refer to [Ünlü, 2004].

We consider the knowledge structure

$$\mathcal{H} := \Big\{\emptyset, \{a\}, \{b\}, \{a,b\}, \{a,b,c\}, \{a,b,d\}, \{a,b,c,d\}, \{a,b,c,e\}, Q\Big\}$$

on domain $Q := \{a,b,c,d,e\}$. We suppose that the knowledge states of $\mathcal{H}$ occur in a population of reference with the probabilities

$$\begin{aligned} p(\emptyset) &:= 0.04, \\ p(\{a\}) &:= 0.10, \\ p(\{b\}) &:= 0.06, \\ p(\{a,b\}) &:= 0.12, \\ p(\{a,b,c\}) &:= 0.11, \\ p(\{a,b,d\}) &:= 0.07, \\ p(\{a,b,c,d\}) &:= 0.13, \\ p(\{a,b,c,e\}) &:= 0.18, \\ p(Q) &:= 0.19. \end{aligned}$$

Let the careless error and lucky guess probabilities β_q and η_q at items $q \in Q$, respectively, be specified as

$$\begin{aligned} \beta_a &:= 0.16, \ \eta_a := 0.04, \\ \beta_b &:= 0.18, \ \eta_b := 0.10, \\ \beta_c &:= 0.20, \ \eta_c := 0.01, \\ \beta_d &:= 0.14, \ \eta_d := 0.02, \\ \beta_e &:= 0.24, \ \eta_e := 0.05. \end{aligned}$$

Based on this BLIM, we simulated a binary (of type 0/1) $1\,200 \times 5$ data matrix representing the response patterns for 1 200 fictitious subjects. The collection of competing models (knowledge structures) for model selection was obtained from the multinomial response data data-analytically, based on a modified version of the *Item Tree Analysis* (ITA; see [Leeuwe, 1974]) described in [Ünlü, 2004]. A modified ITA of the BLIM data resulted in a collection of fifteen knowledge structures, which contained the true knowledge structure $\mathcal{H}$ underlying the data.

From this collection, we selected an optimal model based on maximum κ. Table 1 lists the values of κ (for $M := M_s$, and $C := 0.01$) for the fifteen competing knowledge structures. In Table 1, models are labeled by their respective *tolerance levels* $0 \leq L \leq 1200$ of the modified ITA, and L_κ denotes the optimal (maximum κ) solution. The true model is labeled by "(true)".

L	$\boldsymbol{\kappa}$
0–58	−0.098487
59–62	−0.098591
63–71	−0.098672
72–77	−0.098807
78–88	−0.098880
89–95	−0.098931
96–100	−0.099029
101–150 (true)	−0.099040
151–191	−0.098871
$\boxed{\mathsf{L}_\kappa = \mathsf{192–213}}$	$\boxed{\mathsf{-0.097610}}$
214–236	−0.098913
237–239	−0.102439
240–285	−0.108678
286–394	−0.118036
395–1 200	−0.133919

Table 1. κ (for $M := M_s$, and $C := 0.01$)

Measure κ assumed its maximum value at tolerance range $L_\kappa = 192–213$, i.e., for the candidate knowledge structure $\mathcal{K}_{192-213}$,

$$\mathcal{K}_{192-213} := \left\{\emptyset, \{a\}, \{a,b\}, \{a,b,c\}, \{a,b,c,d\}, \{a,b,c,e\}, Q\right\}.$$

Compared to the true model $\mathcal{K}_{101-150} = \mathcal{H}$, this "best" solution was quite acceptable. In $\mathcal{H}$, the subsets $\{b\}$ and $\{a,b,d\}$ were knowledge states, whereas, in $\mathcal{K}_{192-213}$, they were not. In all other respects, both the models were identical. We had $|\mathcal{H}| = 9$ versus $|\mathcal{K}_{192-213}| = 7$ ($\mathcal{K}_{192-213} \subset \mathcal{H}$).

Acknowledgements

This work was supported by the Austrian Science Fund (FWF) through the project grant P17071-N04 to Dietrich Albert.

References

[Doignon and Falmagne, 1999]J.-P. Doignon and J.-Cl. Falmagne. *Knowledge Spaces.* Springer, Berlin, 1999.

[Goodman and Kruskal, 1954]L.A. Goodman and W.H. Kruskal. Measures of association for cross classifications. *Journal of the American Statistical Association*, pages 732–764, 1954.

[Goodman and Kruskal, 1959]L.A. Goodman and W.H. Kruskal. Measures of association for cross classifications, II: Further discussion and references. *Journal of the American Statistical Association*, pages 123–163, 1959.

[Goodman and Kruskal, 1963]L.A. Goodman and W.H. Kruskal. Measures of association for cross classifications, III: Approximate sampling theory. *Journal of the American Statistical Association*, pages 310–364, 1963.

[Goodman and Kruskal, 1972]L.A. Goodman and W.H. Kruskal. Measures of association for cross classifications, IV: Simplification of asymptotic variances. *Journal of the American Statistical Association*, pages 415–421, 1972.

[Guttman, 1941]L. Guttman. An outline of the statistical theory of prediction. In P. Horst et al., editors, *The Prediction of Personal Adjustment*, volume 48 of *Bulletin*, pages 253–318. Social Science Research Council, New York, 1941.

[Leeuwe, 1974]J.F.J. van Leeuwe. Item tree analysis. *Nederlands Tijdschrift voor de Psychologie*, pages 475–484, 1974.

[Ünlü, 2004]A. Ünlü. *The Correlational Agreement Coefficient CA and an Alternative κ.* PhD thesis, Graz University of Technology, Graz, Austria, 2004.

Feature selection and preferences aggregation

Gaelle Legrand and Nicolas Nicoloyannis

Laboratoire ERIC
Université Lumière Lyon 2
Bat. L ; 5, av. Pierre Mendès-France
69676 Bron Cedex - France
(e-mail: glegrand@eric.univ-lyon2.fr;nicolas.nicoloyannis@univ-lyon2.fr)

Abstract. The feature selection allows to choose P features among M ($P < M$) and thus to reduce the representation space. This process gets more and more useful because of the databases size increases. Therefore we propose a method based on preferences aggregation. It is an hybrid method that lies filter and wrapper approaches.
Keywords: Feature selection, wrapper approach, filter approach, preferences aggregation.

1 Introduction

Across a wide variety of fields, data are being collected and accumulated at a dramatic pace. The Knowledge Discovery in Databases (KDD) process can extract useful knowledge and patterns from the rapidly growing volumes of data to improve the performance of various classifiers and to reduce the running time. Feature selection is an essential step of the KDD process: it eliminates irrelevant, noisy and redundant features, it selects the most relevant features and it reduces the effective number of features under consideration, the data mining step is then accelerated and the calculative cost may be reduced (see [Sémani *et al.*, 2005]).
This paper addresses feature selection for supervised learning. We propose a new feature selection algorithm which is situated at the intersection of filter and wrapper approaches. It uses preferences aggregation to determine an ordered list of features subsets. The next section reviews existing feature selection methods. The third section presents our starting point. Section 4 presents our feature selection method. Experimental evaluations are presented in section 5.

2 Existing feature selection methods

Feature selection methods are gathered in two approaches: wrapper approach, [John *et al.*, 1994], and filter approach, [Kira and Rendell, 1992a]. Wrapper approach takes the influence of selected features subset on the performances of the learning algorithm into account. The learning algorithm is used as an evaluation function to test different features subsets. However, its computational cost is too important in most cases.

Filter approaches are grouped into 5 categories:

Complete methods test all possible features subsets. Their computational cost is very high: MDLM [Sheinvald *et al.*, 1990]...

Heuristic methods have many representatives like Relief, an iterative feature weight-based algorithm inspired by instance-based learning algorithms, (see [Kira and Rendell, 1992b]). These methods require several accesses to databases.

Random methods main representative is LVF, [Liu and Setiono, 1996]. Because of their probabilistic property, the number of selected features tends towards the half of the initial features number. Like previous methods, these methods require several accesses to databases.

Fast sequential selection method principle is an iterative feature selection with a single access to databases. In order to have a single data scan, fast correlation measures must be used such as Kendall rank correlation coefficient. This kind of methods is represented by MIFS [Battiti, 1994], or the method proposed by Lallich and Rakotomalala (see [Lallich and Rakotomalala, 2000]). These methods are the fastest and quite efficient.

Step-by-step methods use short-sighted criteria to select features. This type of methods is effective and very rapid particularly for problems with many features and objects.

Each approach is characterized by a search procedure to generate the next candidate subset (see [Langley, 1994]) and an evaluation criterion to evaluate the subset under consideration. There are 4 categories of criteria which measure various feature specifications: **Information measures**: these measures determine the information gain: Shannon entropy [Shannon, 1948], gain ratio [Quinlan, 1986],...; **Distance measures**: they evaluate the separability of classes: Gini coefficient [Breiman *et al.*, 1984], Mantaras distance measure [De Mantaras, 1991]...; **Dependence measures** are the whole correlation or association measures: Tschuprow coefficient [Hart, 1984]...; **Consistency measures**: These measures detect redundant features: τ of Zhou [Zhou and Dillon, 1991].

3 Starting point

We start from the following observation: step by step methods using short-sighted criteria are fast and have good results. However, the use of a step-by-step method generates two problems: The choice of criterion is delicate, which criterion is the most effective? and the form of result (a list of sorted features) doesn't provide us with the optimal features subset.

The method we propose solves these problems in the following way:

- There is no criterion better or more effective than others. Each criterion emphasizes some specific feature qualities. It seems to be interesting to

obtain a result which takes the opinion of different criteria into consideration. So to obtain this type of results, we use a set of criteria and a preferences aggregation method.
- Obtaining a sorted list of features limits the interest of feature selection : we parameterize the aggregation method so that it doesn't provide with an ordering on the features but a preordering. Also, we don't add features one by one but features subset by features subset.

4 Presentation of our method

Our feature selection method is at the intersection of filter and wrapper approaches which makes the features classification possible with the use of short-sighted criteria. This method has 3 steps:

- Calculus and discretization of the different criteria for each feature (filter approach),
- Application of a preferences aggregation method on the results obtained at the previous stage (filter approach),
- Research of the optimal features subset (wrapper approach).

4.1 Calculus and discretization of criteria

We let users choose the short sighted criteria set. The only condition is that there must be a representative of each criteria categories. For experiments and tests, we choose a set of 10 short-sighted criteria: Shannon entropy, gain ratio, normalized gain, Mantaras distance measure, Gini coefficient, chi-squared, Tschuprow coefficient, Cramer coefficient, and τ of Zhou.

Each criterion for all features are calculated parallelly. The result is a set of 10 ordered lists (order descending) of feature relevance.

A feature can be as relevant as another one even if the two features don't bring the same information type. So, we introduce the concept of features equivalence.

In order to define this concept, we consider a set of objects $O = \{o_1, ..., o_n\}$ described by the initial features set $X = \{x_1, ..., x_i, ..., x_p\}$, and a set of K short-sighted criteria $CR = \{cr_1, ..., cr_k, ..., cr_K\}$ with $cr_k = \{cr_{k1}, ..., cr_{kp}\}$, the set of criterion values for each feature.

Values for each criterion are normalized with the following transformation: $cr_{ki,N} = (cr_{ki} - Min(cr_k)) \setminus (Max(cr_k) - Min(cr_k))$ for a feature x_i and a criterion cr_k.

After their normalization, these values are discretized in deciles. The discretization assigns to each feature a rank R_{ki} for each criterion. This rank is such that the most relevant feature has the smallest rank (For a criterion which must be minimized: If $cr_{ki,N} \in [0; 0.1]$ then $R_{ki} = 1$... If $cr_{ki,N} \in [0.9; 1]$ then $R_{ki} = 10$; For a criterion which must be maximized: If $cr_{ki,N} \in [0; 0.1]$ then $R_{ki} = 10$... If $cr_{ki,N} \in [0.9; 1]$ then $R_{ki} = 1$).

Thus the equivalence concept is defined as follows: two features are equivalent according to a criterion if and only if they have the same rank for this criterion. We tested various combination of normalization and discretization methods. The combination described here gave us the most interesting and general results on the tested datasets. It could be interesting to modify this combination according to data structure.

4.2 Aggregation of the results of criteria

For all aggregation methods (see [Vincke, 1982], [Tanguiane, 1991]), the set of judges and the set of objects must be defined. In our case, the objects correspond to features and the judges correspond to criteria.
We use the aggregation method developed in [Nicoloyannis *et al.*, 1998] and [Nicoloyannis *et al.*, 1999] based on pairwise comparison concept developed in [Marcotochino, 1984a] and [Marcotochino, 1984b]. We don't describe in details this method but we present its subjacent principle.

For each features pair (x_i, x_j), each judge (criterion) states its opinion $A_k(i,j)$. A_k, the opinion of a judge k is an application of $X \times X$ in $\{Pref, NPref, EQ\}$. Thus,
$A_k(i,j) = Pref$: the judge k prefers x_i to x_j, $R_{ki} < R_{kj}$
$A_k(i,j) = NPref$: the judge k prefers x_j to x_i, $R_{ki} > R_{kj}$
$A_k(i,j) = EQ$: the judge k considers x_j and x_i as equivalent, $R_{ki} = R_{kj}$.
The result we wish to obtain is an opinion OP called opinion of broad preferences and which generates a preordering relation on X. OP is an application of $X \times X$ in $\{Pref, NPref, EQ\}$.
Definition 1: The degree of agreement $\rho_{ij}(OP, A_k)$ between $OP(i,j)$ and $A_k(i,j)$ is defined in Table 1.

OP/A_k	$Pref$	$NPref$	EQ
$Pref$	1	0	1/2
$NPref$	0	1	1/2
EQ	1/2	1/2	1

Table 1. Degree of agreement

Definition 2: The degree of agreement $DA(OP, A_k)$ is $DA(OP, A_k) = \sum \rho_{i,j}(OP, A_k)$.

Definition 3: The degree of agreement between the opinion OP and the opinion of all judges is $DA(OP) = \sum DA(OP, A_k)$.

Their problem consists in building an opinion OP which generates a preordering on X and which maximizes $DA(OP)$. The corresponding optimization problem is NP-hard, hence requires the use of a meta-heuristic. The simulated annealing method [Kirkpatrick *et al.*, 1983] is used for maximization. The simulated annealing method is used because it's a rapid and easy to use method by [Nicoloyannis *et al.*, 1998]. But , they can use another methods. The parameters are : the decay rate is set to 0.98, the halting condition is a number of iterations which is set to $10 \times |X|$. The neighbourhood of the current solution is defined as follows: a preordering $\acute{L}$ belongs to the neighbourhood of a preordering $L = \{l,, ..., l_m, ..., l_M\}$, $(\acute{L} \in V(L))$, if and only if $\acute{L}$ derives from L by the movement of only one object $x_i \in l_m, l_m \subset L$: x_i is flipped into l_{m+1}, $(m < M)$ or into l_{m-1}, $(m = M)$; Or x_i constitutes a group by itself.

After the application of this aggregation method, we obtain an ordered list of disjoint features subsets $L = \{l_1, ..., l_h, ..., l_H\}$.

4.3 Optimal features subset

Until this step, our method belong to filter approach. From this step, our method belong to wrapper approach. The advantage of using a wrapper approach is to take into consideration the influence of the features subset on the learning algorithm performances. The detection of the optimal subset is carried out as follows: within the h^{th} iteration, the features subset l_h is added to the optimal features subset. The optimal features subset is the one having the smallest error rate on the learning set.

5 Experimentations

For our experiments we used 11 databases from the UCI repository (see [Merz and Murphy, 1996]). Quantitative features are discretized with Fusinter method developed in [Zighed *et al.*, 1996]. The feature selection is carried out on 30% of the initial set of objects keeping initial classes distribution. The 70% remaining are used for the learning phase. For that, we use a 10-fold-cross-validation and the learning algorithms are ID3 and Naive Bayesian. The tests without selection are also carried out on these same 70% of studied base. After the application of our selection method, we can see some improvements in error rate with ID3 and the Naive Bayesian (Tables 2 and 3). Our method is comparable with MIFS and ReliefF and sometimes better. Tables 2 and 3 show the number of iteration carried out by our method. The maximum number of iterations is about 9 (for Vehicle). The number of learning algorithm runs in our method is then smaller than in pure wrapper methods. For our method, the number of selected features depends on the learning algorithm (Table 4). This number is often smaller than the number of features selected by MIFS et ReliefF.

6 Conclusion

In this article, we present a feature selection method based on preferences aggregation. It is a hybrid method between filter and wrapper approaches having the advantages of each approach and reducing their disadvantages:

Bases	Our method	MIFS	ReliefF	Without selection	Number of iterations
	Error (Sd)	Error (Sd)	Error (Sd)	Error (Sd)	with our method
Austra	15,29 (3,48)	17,17 (4,12)	15,31 (5,23)	16,6 (4,57)	2
Breast	4,27 (2,8)	5,9 (2,64)	5,29 (3,16)	5,95 (1,95)	3
Cleve	21,9 (8,67)	24,68 (10,27)	40,54 (7,77)	18,53 (8,68)	5
CRX	15,7 (3,1)	16,12 (6,7)	17,54 (5,88)	14,73 (5,68)	2
German	26,14 (4,87)	27,43 (5,06)	30,14 (6,01)	31,86 (7,53)	5
Heart	26,32 (11,04)	28,42 (9,76)	27,38 (9,06)	27,05 (10,29)	2
Iono	11,73 (5,59)	15,75 (8,71)	11,78 (3,94)	21,37 (8,39)	3
Iris	4,73 (4,74)	4,82 (6,58)	3,73 (4,57)	3,73 (4,57)	3
Monks-1	25,18 (7,56)	25,2 (7,71)	55,52 (3,34)	25,22 (8,3)	2
Monks-2	34,89 (6,71)	34,91 (6,7)	34,9 (8,63)	34,91 (6,79)	2
Monks-3	3,88 (2,69)	3,86 (2,86)	3,88 (3,34)	1,28 (1,28)	2
Pima	24,5 (5,15)	24,87 (4,83)	25,05 (7,69)	26,11 (5,43)	3
Tic Tac Toe	25,16 (6,31)	30,81 (7,11)	30,51 (5,9)	33,43 (5)	4
Vehicle	28,75 (5,44)	40,62 (7,39)	42,25 (6,52)	34,24 (4,96)	9

Table 2. Test with ID3

Bases	Our method	MIFS	ReliefF	Without Selection	Number of iterations
	Error (Sd)	Error (Sd)	Error (Sd)	Error (Sd)	with our method
Austra	15,27 (3,61)	14,28 (3,08)	15,28 (5,15)	16,6 (4,57)	3
Breast	2,65 (2,05)	2,86 (1,87)	3,45 (2,56)	5,95 (1,95)	5
Cleve	17,77 (6,14)	20,52 (11,34)	40,67 (4,33)	18,53 (8,68)	4
CRX	15,69 (3,99)	14,66 (5,7)	16,53 (2,8)	14,73 (5,68)	3
German	23,43 (4,62)	26,29 (3,63)	30,71 (4,96)	31,86 (7,53)	7
Heart	17,89 (7,14)	17,89 (10,04)	21,05 (10,53)	27,05 (10,29)	4
Iono	7,25 (5,88)	5,22 (4,4)	9,32 (6,22)	21,37 (8,39)	6
Iris	2,82 (4,31)	4,64 (6,17)	6,45 (7,14)	3,73 (4,57)	3
Monks-1	25,19 (4,68)	25,2 (7,18)	51,9 (8,2)	25,22 (8,3)	2
Monks-2	34,92 (5,11)	34,92 (6,24)	34,92 (6,65)	34,91 (6,79)	2
Monks-3	3,85 (3,67)	3,86 (2,87)	3,85 (3,85)	1,28 (1,28)	2
Pima	22,83 (5,73)	21,33 (4,3)	25,04 (3,41)	26,11 (5,43)	4
Tic Tac Toe	27,83 (3,92)	28,87 (5,42)	27,97 (4,19)	33,43 (5)	4
Vehicle	33,95 (4,18)	39,85 (8,01)	45,82 (8,78)	34,24 (4,96)	7

Table 3. Test with Naive Bayesian

- The influence of the selected features on the learning algorithm is taken into account. Thus, the selected features are different according to the used algorithm.
- The computational cost is largely lower than the computational cost of pure wrapper methods due to the use of a preordering.

We plan to improve our method according to two aspects. The discretization method used for the criteria values must be better. Also we would like the result of the preferences aggregation method to be the optimal features subset.

Bases	Without selection	Our method with ID3	Our method with BN	ReliefF	MIFS
Austra	14	1	2	2	13
Breast	9	3	7	6	9
Cleve	13	7	5	6	8
CRX	15	3	5	2	7
German	20	5	9	14	3
Heart	13	2	8	2	13
Iono	34	2	26	25	8
Iris	4	3	2	4	3
Monks-1	6	1	1	2	1
Monks-2	6	1	1	2	2
Monks-3	6	2	2	2	3
Pima	8	2	5	7	4
Tic Tac Toe	9	7	7	5	3
Vehicle	18	14	12	18	6

Table 4. Number of selected features

References

[Battiti, 1994]R. Battiti. Using mutual information for selecting features in supervised neural net learning. 5:537–550, July 1994.

[Breiman *et al.*, 1984]L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone. *Classification and Regression trees, The Wadsworth Statistics/Probability Series, Wadsworth, Belmont, CA*. 1984.

[De Mantaras, 1991]R.L. De Mantaras. A distance-based attribute selection measure for decision tree induction. In *Machine Learning*, volume 6, pages 81–92, 6-9 1991.

[Hart, 1984]A. Hart. Experience in the use of an inductive system in knoowledge eng. In M. Bramer, editor, *Research and Development in Expert Systems.* Cambridge Univ. Press, Cambridge, MA, 1984.

[John *et al.*, 1994]George H. John, Ron Kohavi, and Karl Pfleger. Irrelevant features and the subset selection problem. In *Int'l Conf. on Machine Learning*, pages 121–129, 1994.

[Kira and Rendell, 1992a]K. Kira and L. A. Rendell. The feature selection problem: Traditional methods and a new algorithm. In MIT Press, editor, *Tenth Nat. Conf. on Artificial Intelligence*, pages 129–134, 1992.

[Kira and Rendell, 1992b]K. Kira and L.A. Rendell. A practical approach to feature selection. In *Proc. of the Tenth Int'l Conf. on Machine Learning*, pages 500–512, 1992.

[Kirkpatrick *et al.*, 1983]S. Kirkpatrick, C. D. Gelatt, and M. P. Vecchi. Optimization by simulated annealing. *Science, Number 4598, 13 May 1983*, 220, 4598:671–680, 1983.

[Lallich and Rakotomalala, 2000]S. Lallich and R. Rakotomalala. Fast feature selection using partial correlation for multi-valued attributes. In *Proc. of the 4th European Conf. on Knowledge Discovery in Databases, PKDD 2000*, pages 221–231, 2000.

[Langley, 1994]P. Langley. Selection of relevant features in machine learning. In *Proc. of the AAAI Fall Symposium on Relevance*, pages 1 – 5, 1994.

[Liu and Setiono, 1996]Huan Liu and Rudy Setiono. A probabilistic approach to feature selection - a filter solution. In *Int. Conf. on Machine Learning*, pages 319–327, 1996.

[Marcotochino, 1984a]F. Marcotochino. Utilisation des comparaisons par paires en statistique des contingences, Étude n°f-071. Technical report, Centre Scientifique IBM-France, Février 1984.

[Marcotochino, 1984b]F. Marcotochino. Utilisation des comparaisons par paires en statistique des contingences, partie ii Étude n°f-071. Technical report, Centre Scientifique IBM-France, Mai 1984.

[Merz and Murphy, 1996]C. Merz and P. Murphy. Uci repository of machine learning databases. *http://www.ics.uci.edu/ mlearn/MLRepository.html*, 1996.

[Nicoloyannis *et al.*, 1998]N. Nicoloyannis, M. Terrenoire, and D. Tounissoux. An optimisation model for aggregating preferences: A simulated annealing approach. *Health and System Science*, 2(1-2):33–44, 1998.

[Nicoloyannis *et al.*, 1999]N. Nicoloyannis, M. Terrenoire, and D. Tounissoux. Pertinence d'une classification. *READ*, 3(1):39–49, 1999.

[Quinlan, 1986]J.R. Quinlan. Introduction of decision trees. In *Machine Learning*, volume 1, pages 81–106, 1986.

[Sémani *et al.*, 2005]Dahbia Sémani, Carl Frélicot, and Pierre Courtellemont. Un critère d'évaluation pour la sélection de variables. In *EGC*, pages 91–102, 2005.

[Shannon, 1948]C.E. Shannon. A mathematical theory of communication. In *Bell System Technical Journal*, 1948.

[Sheinvald *et al.*, 1990]Sheinvald, Dom, Niblack, and Rendell. A modeling approach to feature selection. In *Tenth Int. Conf. on Pattern Recognition*, 1990.

[Tanguiane, 1991]A. S. Tanguiane. *Agregation and representation of preferences: Introduction to Mathematical Theory of Democracy, Springer Verlag.* 1991.

[Vincke, 1982]Ph. Vincke. Aggregation of preferences: a review. *European Journal of Operational Research*, 9:17–22, 1982.

[Zhou and Dillon, 1991]X. Zhou and T.S. Dillon. A statistical–heuristic feature selection criterion for decision tree induction. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, volume 13, pages 834–841, 1991.

[Zighed *et al.*, 1996]D. A. Zighed, R. Rakotomalala, and S. Rabasééda. A discretization method of continous attributes in induction graphs. *Proc. Of the 13th European Meetings on Cybernetics and System Research*, 3(1):997–1002, 1996.

Indexing by Isotopy

Christian Mauceri

ENST Bretagne (e-mail: mauceri@fr.ibm.com)

Abstract. Within the Semantic Web initiative, Topic Maps have enabled a common architecture for indexing applications. In this paper, we present shallow reading methods aimed at semi automatic indexing through the integration of a finite state technology in the framework of Topic Maps.
Keywords: Finite state technology, Indexing, Isotopy, Semantic, Topic Maps.

Introduction

Anybody who has ever looked for precise information in a book knows how valuable an index can be. Indexing is a very ancient activity: in the antique Rome an index was a little slip attached to scrolls on which information about the work was written in order to easily identify its content without having to read it [Wellisch, 1991]. There are several types of indexes with different levels of complexity in the way they are structured. Indexes can be the mere positions of relevant words in a book (the good back of book index) or terms structured in complex thesauri used to describe the subject matter of documents in a library. The term 'index' spread over and may refer to tree structures used to speed up retrieval of records in databases; it is very often this accepted meaning which is used in the electronic document word[1]. Conversely, traditional indexing is too often perceived as a dusty technique because of its cost and slowness. Besides, it varies from an indexer to another and therefore suffers from a reputation of subjectivity. However human indexing is a mind production which, goes far beyond what a program in a machine will ever able to do. Instead of running after hypothetical objectivity, it should be accepted that human indexing is a hermeneutical activity [Mai, 2000]; it supposes an interpretation by a reader of what an author wrote depending on both the specific cultural and social context.
The objective of this paper is to show a way to reconcile automatic and human indexing through the integration of Semantic Web technologies, robust parsing techniques and a shallow reading method. A survey analysis application will be used as reference example in the remaining chapters.
This paper is divided into four sections. The first section presents Topic Maps, a Semantic Web technology used to organize information. The second section presents shallow reading methods aimed at gathering relevant vocabulary on particular semantic objects called isotopies. The third section describes a finite state technology implementation allowing for automatically

[1] Such an index is the list of positions in a document of each word occurring in it.

proposing candidate occurrences of isotopies according to the context they occur in. The fourth section deals with a real case application on survey analysis.

Topic Maps

The Semantic Web initiative shows the growing interest in sharing information between people which can be intelligible by computers. In this framework, Topic Maps is the technology addressing document indexing. Their core concepts are Topics, Associations, and Occurrences (TAO [Pepper, 2000]). Each of these objects can have a type; there are, hence, topic types, association types, occurrence types, all of these types being themselves topics. Topics represent subjects of anything one can imagine. The only constraint on topics is that one topic must refer to one and only one subject. Topic refers to subject by the mean of Universal Resource Identifiers (URI). In the survey analysis application we are interested in, the main topic types are:
Survey, Interview, Section, Question, Comments, Score, Person, Company, Term, Triggers, Contexts, . . .
The main association types are:
• **Contains**; a survey **contains** interviews, an interview **contains** sections, a section **contains** questions.
• **Respondent**; persons are **respondent** of an interview.
• **Interviewer**; a person is the **interviewer** in an interview.
• **Rated**; a question is **rated** with a score,
• **Belongs to**; a person **belongs to** a company,
• **Expressed**; a comment is **expressed** about a question,
• **Is indexed by**; a comment **is indexed by** a term,
• **Is triggered**; a term **is triggered** by a trigger,
Once defined, the topic types and the association types can be used to describe surveys. Hence "Satisfaction Survey 2004" can be a topic of type survey, "Interview IBM 1756" can be a topic of type interview and the association "Satisfaction Survey 2004" **contains** "Interview IBM 1756" is an instance of the association type contains. The mechanism that allows for referring to real objects is called 'occurrence.' Similar to associations and topics, occurrences can have type (these types being topics as well). In the example considered here, topic occurrence types can be:
• **Described by**; the survey "Satisfaction Survey 2004" is **described by**:
"http://www.hermeneutician.com/#satisfaction%20survey%20guidelines"
• **Transcript**; the comment "Comment on Overall Section of Interview IBM 201756" has a **transcript** in:
"http://www.hermeneutician.com/Interview%20IBM%201756/#Overall"
• **Value**; the score "Score of Overall Section of Interview IBM 201756" has **value**:

"Satisfied"

- **Email**; the person "Christian Mauceri" has **Email**: "mauceri@hermeneutician.com" .

The next question is why using topics to index documents? Saying they were designed for this purpose is obviously not the right answer, but their authors had in mind a number of common indexing issues:

- Flexibility; it is very easy to add new topics, associations and occurrences to a Topic Map,
- Ease of use in querying and browsing; Topic Maps are well suited to dynamically generate HTML pages allowing intuitive navigation in the map in order to rapidly find information (have a look at http://www.ontopia.net/operamap/).
- Serialization; Topic Maps can be serialized via an XML format named XTM,
- Standard; Topic Maps are a standard (see [ISO, 2000]),
- Collaboration; Topic Maps can quite easily be merged and exchanged.

What Topic Maps do not address is the way to define indexes and to how find them in unrestricted text. We will focus on this in the remaining chapters.

Shallow Reading

The inherent defects of variation and slowness in human indexing can be partially overcome by indexing a corpus rather than isolated documents. The term corpus (see F. Rastier [Rastier, 2002]) is related to law, religion, and linguistics; tt refers to a set of texts. In recent years, a regain of interest for corpus linguistics has been at the origin of a number of papers. In general, a corpus is merely seen as a set of texts or even sentences no matter how they are related to each other.

Originally, however, this term was set up by disciplines like hermeneutic and philology, where the relations between texts were taken into account. In this tradition, the corpus is a structured set of texts sharing characteristics of genre and discourse (legal discourse, medical discourse, etc.). Besides, a corpus is built for a particular purpose, and corresponds to a need. It is this definition of a corpus that will be used in the remaining paragraphs.

There is not isolated text, as a text respects a set of social norms shared by other texts. From an indexing perspective, it is an important point to take into account, because:

- It allows for a global view of text, and therefore reduces the indexing variation by considering what texts share and what makes them different.
- It reduces semantic variation. Indeed, words and expressions tend to have narrower meaning. Syntactic constructions are more regular and are very often almost frozen.

However, it is very often impossible to read, in the traditional way, an entire corpus or even a representative sample of it. (How does one read all of the articles of the Wall Street Journal over the past 10 years or thousands of commercial reports in a short period of time?). Nevertheless, the indexer has an a priori knowledge of the content of the corpus that he/she is supposed to process; its literary kind, the writing style, what the texts are talking about. This knowledge is very important and is the basis of the reading method proposed here.

Interpretative Semantics gives a theoretical framework of what reading is and how indexing is connected to it. Interpretative Semantics takes its roots in the Saussurian statement that human languages are made of oppositions. People perceive similarities and differences between linguistic objects, A.J. Greimas, in [Greimas, 1986] gives an example of such an opposition, 'bet' Vs 'pet', (he actually used 'pas' Vs 'bas' in French) in fact the fundamental structure in this opposition is given by 'b' Vs 'p' or 'voiced' Vs 'non-voiced'. At the semantic level, these oppositions allow for defining semantic relations; hence the opposition, 'girl' Vs 'boy', defines the semantic axis 'sex'.

The minimal units of sense used to describe these structural relations within a corpus are called semes (see for instance, [Pincemin, 1999]). Semes are not used to describe isolated words but rather are defined as sets of words related to them. For instance, instead of describing a priori 'chair' as {/furniture/, /for sitting/,etc...}, /furniture/ is described by the set {'chair', 'closet', 'table', 'sofa'}. Semes depend on context, are not universal truth; and are defined by the corpus reading. Restricting semantic relations definition at the corpus level avoids looking for an improbable universal ontology of semes.

In Interpretative Semantics, reading is the result of an interpretation, an operation specifying the meaning of a text. An interpretation can add or remove semes to words, depending on the context. A.J. Greimas gives the example of the sentence "The superintendent barks" to show how a seme can be added by the interpretation process. In this sentence a seme /animal/ is added to the word 'superintendent' because it is the subject of the verb 'bark' obviously bearing the seme /animal/, giving the sentence its pejorative interpretation. This example introduces a central concept in Interpretative Semantic, the notion of isotopy (see for instance, [Sonesson, 2004]); an isotopy is the effect produced by a same seme recurrence in a corpus[2]. An isotopy analysis produces a list of words having some contextual semes in common.

Isotopies are useful for word sense disambiguation. For instance, let's consider the word 'bugs' in the two sentences:

- Bugs were crawling everywhere in the room.
- Bugs were found in the program.

In the first sentence there is an isotopy /animal/ between 'bugs' and 'crawling', in the second sentence there is an isotopy /computer/ between

[2] On isotopy in a corpus A.J. Greimas give in [Greimas, 1986] an interesting example of the isotopy /death/∼/life/ in Bernanos' work

‘bugs’ and ‘programs.’
Isotopies are given a priori and come before the semes definition; they are expected by the reader/indexer, they ensure the corpus coherence.
As isotopies are given a priori and entail semes characterization it is much more productive to gather the corpus vocabulary related to an isotopy rather than to build an a priori hierarchy of semes to describe the entire corpus vocabulary. Isotopy recognition is triggered by keywords depending on the context they appear in. A person reading a list of words easily detects potential triggers, so a very productive approach is to detect these triggers in the corpus vocabulary without having to read the entire corpus, it is what we call a shallow reading of the corpus; gathering isotopy trigger candidates by reading the corpus vocabulary, looking at them in their contexts and expressing rules inhibiting or refining them when they occur in particular contexts.
But why read all the vocabulary? Indeed, as an isotopy is a seme recurrence in the corpus it can be thought that only words occurring more than twice are of interest; it is precisely because a seme occurring very often in the corpus can be borne by words occurring only once it is important to look at rare words which represent the major part of the vocabulary, in average words occurring only once represents more than 50% of the vocabulary. For instance, in an application aimed at detection of commercial reports offending the European Privacy Regulation in a French bank the word ‘fingernails’ appeared only once in the context “her only project is to paint her fingernails which is a clearly sexist statement. This example is interesting because beyond the argument in favor of considering the scarce words, it shows traditional lexicons cannot help in detecting such cases heavily depending on contexts.
Once the candidates are gathered they must be allocated to the isotopies they are supposed to trigger and checked in context in order to write rules inhibiting or refining the triggers. A trivial example of such rules is given by the occurrence of the trigger ‘cost’ of the isotopy ‘Pricing’ in the context: **“Mr. Redford was happy with fisher.com commitment to reach their objectives at all cost”**, obviously the indexer doesn’t want to trigger the ’Pricing’ isotopy in such a case and would like to write something like “I don’t want the word ’cost’ to trigger the isotopy ‘Pricing’ when it is preceded by ‘at all.”’ In summary, shallow reading consists in:
• Locating in the corpus vocabulary words triggering isotopies,
• Building a concordance of these words and their contexts in the corpus,
• Writing rules inhibiting or refining the triggers according to these contexts.
Many of these rules can be expressed by the means of regular expressions and integrated in the Topic Maps frameworks; it is the subject of the next chapters.

Finite State Machines

A finite state machine consists of a set of states, a start state, a final state, an input alphabet, and a transition function. A transition function maps an element of the input alphabet and a current state to another state. At the beginning of a computation the machine's current state is the start state, and changes of state depend on an input string and the transition function.

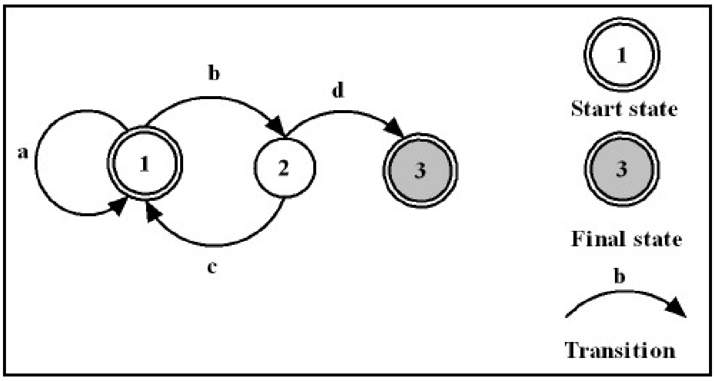


Fig. 1. An automaton example.

The figure above represents an automaton with state set {1, 2, 3}, an input alphabet **{a, b, c, d}**, a start state 1, a finale state 3 and a function transition mapping (1,a) to 1, (1, b) to 2, (2, c) to 1 and (2,d) to 3. This automaton recognizes the strings **{"aaabcbd", "abcbcbcbd", "bcbd", "bd", etc.}**. Finite state machines can be described by regular expressions whose principal operators are union, intersection, complementation, concatenation, and Kleen star on the input alphabet. (See Aho, Sethi and Ullman [Alfred V. Aho, 1986]). For instance, the automaton given in example is described by the regular expression **"a"* "b" "cb"* "d"**; the Kleen star operator means no or many occurrences of the expression it applies to, hence **"a"*** means no or many occurrences of the character **"a"** and **"cb"*** no or many occurrences of the character **"c"** followed by the character **"b"** so **"a"* "b" "cb"* "d"** defines strings beginning by zero or many characters **"a"** followed by the character **"b"** followed by zero or many substrings **"cb"** and ended by the character **"d"**. In the same way the expression **"ab" | "bc"** defines the strings **"ab"** or **"bc"** (| is the union operator). The expression **@* ("ab" | "bc") @*** defines all the strings containing the substrings **"ab"** or **"bc"** (@ means any character). The expression **ˆ (@* ("ab" | "bc") @*)** defines the strings which do not contains the substrings **"ab"** or **"bc"** (ˆ is the complementation operator). The expression **ˆ(@* ("ab" | "bc") @*) & ("a" @* "b" |"b" @* "c")** defines the strings which do not contains the substrings **"ab"** or **"bc"** starting by the character **"a"** and ending by the character **“b”** or starting by the character **“b”** and ending by the character **“c”** (**&** is the intersection operator).

Transducers, which are finite state machines with output have been widely used in Natural Language Processing (NLP) (see for instance [Abney, 1996], [Grefenstette, 1996], [Hobbs, 1996] or [Roche and all, 1996]). In particular, they have been used in shallow parsing and local grammar implementation. A shallow parser aims to identify phrasal constituents, such as noun phrases, and the functional role of some of the words, such as the main verb, and its direct complements [Abney, 1996]. Local grammars are used to describe local linguistic structures in the form of graphs (See [Silberztein, 1993], [Masson, 1993]).

The rules discussed in the previous chapter are implemented by finite state machines whose transitions are of two types; simple transitions and epsilon transitions. Simple transition maps an element of the input alphabet and a state to another state. Epsilon transitions maps the empty word (commonly called epsilon) and a state to another state. In addition, they produce a meta-character. These meta-characters are used to control further processing on the recognized strings. Typically these machines are union of machines which can be represented by regular expressions of the form:

<**lpat1**> **0:“(”** <**cpat1**> **0:“)”** <**rpat1**> **. . .** <**lpati**> **0:“(”** <**cpati**> **0:“)”**
<**rpati**> **. . .** <**lpatn**> **0:“(”** <**cpatn**> **0:“)”** <**rpatn**>
0:“command1” . . . 0:“commandi” . . . 0:“commandn”

where epsilon productions **0:“(”** and **0:“)”** are used to mark the beginning and the end of the strings recognized by the surrounded regular expressions <**cpati**>. The epsilon productions **0:“commandi”** specify the processing on the corresponding recognized strings. For instance the expression:

“at” <**sep**>+ **“all”** <**sep**>+
0:“(” “<Pricing>” 0:“)” “cost” 0:“(” “</Pricing>” 0:“)” <**sep**>+
0:“rep: $1” 0:“rep: $2”

recognizes the string, **“at all <Pricing>cost</Pricing>”** and produces the string **“at all cost”**. In this expression; <**sep**>= **“ ”|“\t”|“\n”|“\r”;** It means <sep> is a separator (a space, a tabulation, a new line, or a carriage return), and therefore <**sep**>+ means one or many separators.

The command **0:“rep: $1”**, means: replace the string matched by the regular expression surrounded by **0:“(”** and **0:“)”** by nothing.

The machine evaluator rewrites all characters not recognized by a machine and applies the specified processing to the recognized substrings. So these machines are transducers; the output of a transducer can be used as input to another one; such an operation is called a cascade. The evaluation is very fast because they are almost deterministic: the only possible backtracks occurring for the epsilon productions **0:“(”**. Despite their simplicity they can be used to capture many linguistic phenomena. It is important to have in mind that other commands than ‘rep’ and ‘tag’ can be easily defined to control their outputs and performing actions on the Topic Maps.

Let's see how to implement the shallow reading method previously described with this mechanism. We suppose the corpus analyzed is organized in a Topic Map having topics of type chapter, section, sentence, or whatsoever whose instances have textual occurrences. In the example of survey analysis we are interested in, these topics are instance of comment type and their occurrences are the transcripts of these comments.

Once the triggers have been collected we build a first automaton <**T1**> which is the union of expressions of the form:

0:"(" "<trigger>" 0:")" 0:"tag: $1 <isotopy>"

In these expressions <**trigger**> is the written form of a trigger and <**isotopy**> the name of the isotopy it triggers, for instance:

0:"(" "cost" 0:")" 0:"tag: $1 Pricing"

This first automaton is applied to the transcripts and produces new transcripts adorned by XML tags corresponding to the argument labels of the tag command. This same command adds dynamic associations of type '**is indexed by**' between the question topic the transcript is an occurrence of and the corresponding isotopy topic. It also adds the position of the trigger in the transcript as an occurrence of the triggering word. For instance, let's suppose the following transcript:

"Mr. Redford was happy with fisher.com commitment to reach their objectives at all cost"

is an occurrence of the question topic 'overall' in the interview "Redford Entertainments 20035," then application of the first automaton will produce:

"<Hum>Mr. Redford</Hum> was <Euph>happy</Euph> with <Comp>fisher.com<Comp> commitment to reach their objectives at all <Pricing>cost</Pricing>"

and adds an association of type '**is indexed by**' between 'overall' and 'Euphoric' just like an occurrence of 'happy' at the position 27 of the transcript, provided that 'happy' has been declared as a trigger of the isotopy 'Euphoric' in the automaton <**T1**>. It is now possible to access through the Topic Maps the texts where the isotopy 'Euphoric' occurs, in particular in the transcript given below.

"Mr. Charles is moderately happy with the service he received"

rewritten as

"<Hum>Mr. Charles</Hum> is moderately <Euph>happy</Euph> with the service he received".

The simplest way to inhibit the trigger 'happy' when it comes after 'moderately' is to add the following rule to the first automaton <**T1**>.

<be> <sep>+ 0:"(" "moderately" <sep>+ "happy" 0:")" <sep>+ 0:"tag: $2 Dysph".

A second automaton <**T2**> allows for the inhibition of certain triggers. This automaton is made of rules like:

"at" <sep>+ "all" <sep>+ 0:"(" "<pricing>" 0:")" "cost"
0:"(" "</pricing>" 0:")" <sep>+ 0:"rep:" 0:"rep:"

which recognizes the string **“at all <pricing>cost</pricing>”** produces the string, **“at all cost”** and remove the association of type ‘**is indexed by**’ between the isotopy topic ‘pricing’ and the corresponding comment. This second automaton <**T2**> cascaded with the first one <**T1**> allows the implementation of simple positive and negative rules, activating or inhibiting triggers.

The contexts analysis often suggests taking into account recurring linguistics constructions such as:

“Ms. Wilson [**declares, pointed out, thinks . . .**] **that . . . ”**

which can be captured by expressions like:

<**HumanChunk**> <**sep**>+ <**say**> <**sep**>+ **“that” 0:“(” ˆ(@* “.” @*) 0:“)” “.” 0:“mark: $1 Statement”**,

where <**HumanChunk**> is a regular expression detecting a nominal chunk containing the seme /Human/, <**say**> is a regular expression detecting a verbal chunk containing verbs introducing a statement. Hence the mark comment will surround the string recognized by **ˆ(@* “.” @*)** with the tags <**Statement**> and </**Statement**>. Expressions like this, aimed at detecting recurring discourse structures [Marcu, 1999], can be compiled in a third automaton <**T3**> which, when cascaded with the previous automata, allows for stressing on important contexts.

Experimentation

We have used this method to semi-automatically index the customer satisfaction reports of a very big global company. Basically the two generic isotopies we were interested in were:

- The pervasive issues clients were talking about; What were the topics clients were praising or complaining about?
- The feelings of the interviewed customers; Were they happy or not? What were they expecting?

The vocabulary of about 200 reports was read and categorized in relation to the semes:

- /PI/ for pervasive issues (subdivided in more specific semes like /responsiveness/, /costing/, etc. . .),
- /Euph/ for euphoric,
- /Dysph/ for dysphoric,
- /Exp/ for expectation,
- /H/ for human.

Around 600 words and expressions (out of 20000 words) were selected out of the corpus. We used a <**T0**> automaton to mark grammatical words, modal and auxiliary verbs. The <**T1**> automaton was slightly different than the one we presented in the previous chapter because in a real case application a same word potentially triggers multiple isotopies but overall

the schema was the same as the one sketched in the previous chapter.
A post XSLT [W3C, 2005] processing colored in red chunks containing dysphoric semes, in green those containing euphoric semes. Furthermore this same post processing underscored typical expressions and colored in blue pervasive issues and proper nouns.
The documents have then been indexed manually, focusing on the colored and underlined parts (See B. [Pincemin, 2001]). All these results have been combined in a web application based on XSLT applied to the XML format of the resulting topic map (XTM). The main advantage, from a business point of view, has been that for the first time transcripts of the interviews have been really used by business analysts because of the easiness to access them trough the web Topic Map interface and the pervasive issue indexing sheds a different light on the survey results than the mere scores given by the clients.
Gathering vocabulary and typical contexts took approximately three days and indexing the final 200 reports took an additional day. The main problem we faced was to translate the rules expressed by the business analyst indexer in regular expressions; the regular expressions language being too complex for non-knowledgeable people.

Concluding remarks

Our aim was to evaluate how the notion of isotopy is perceived by business people and how cascades of automaton can be used by them in order to automatically spot them. Even if the reading of a huge vocabulary is a tedious task, the proposed method and the isotopy notion has been quite well accepted. However, the usage of regular expressions and cascades of automaton is completely out of scope, requiring a lengthy learning phase .
The linguistic development environment Intex designed by Max Silberztein [Silberztein, 1993] offers an example of user friendly interface for finite state technology, it is however not designed for indexing purposes and not integrated into a standard indexing framework like Topic Maps, representing finite state machines by the mean of graphs is good and can be used for applications bound to a large public.
A very important issue not discussed in this paper deals with the global analysis of the resulting indexing. Indeed Topic Maps provides great browsing capacity but cannot encompass the global corpus structure. A preliminary work shows that demographic clustering [Johannes Grabmeier, 2002] techniques can be integrated in the Topic Maps framework in order to check the indexing consistency and discrimination power. This point is very important because it extends the structural principle of opposition to the whole corpus; what are the isotopies opposing or gathering texts in a corpus? This complementary technique is the missing retroaction loop in the shallow reading

process.
Future work will focus on graphical user interfaces (GUI) making finite state machines easier to use in the presented framework, and integration of demographic clustering techniques in the shallow reading process.

References

[Abney, 1996]Steven Abney. Partial parsing via finite-state cascades. 1996.

[Alfred V. Aho, 1986]Jeffrey D. Ullman Alfred V. Aho, Ravi Sethi. *Compilers Principles, Techniques and Tools.* Addison Wesley, 1986.

[Grefenstette, 1996]Gregory Grefenstette. Light parsing as finite-state filtering. 1996.

[Greimas, 1986]Algirdas Julien Greimas. *Sémantique Structurale.* PUF, 1986.

[Hobbs, 1996]Jerry Hobbs. Fastus: A cascaded finite-state transducer for extracting information from natural-language text. 1996.

[ISO, 2000]ISO. *ISO/IEC 13250,Information Technology – SGML Applications – Topic Maps.* ISO, Geneva, 2000.

[Johannes Grabmeier, 2002]Andreas Rudolph Johannes Grabmeier. Techniques of cluster algorithms in data mining. 2002.

[Mai, 2000]Jens-Erik Mai. The subject indexing process: an investigation of problems in knowledge representation. 2000.

[Marcu, 1999]Daniel Marcu. Discourse trees are good indicators of importance in text. 1999.

[Masson, 1993]Olivier Masson. Automatic processing of local grammar patterns. 1993.

[Pepper, 2000]Steve Pepper. The tao of topic maps, finding the way in the age of infoglut. 2000.

[Pincemin, 1999]Bénédicte Pincemin. Sémantique interprétative et analyse automatique des textes : que deviennent les sèmes ? 1999.

[Pincemin, 2001]Bénédicte Pincemin. Résoudre la surcharge informationnelle sans la décontextualiser. 2001.

[Rastier, 2002]François Rastier. Enjeux épistémologiques de la linguistique de corpus. 2002.

[Roche and all, 1996]Emmanuel Roche and all. Fastus: A cascaded finite-state transducer for extracting information from natural-language text. 1996.

[Silberztein, 1993]Max Silberztein. Dictionnaires éléctroniques et analyse automatique de textes : le système intex. 1993.

[Sonesson, 2004]Göran Sonesson. Isotopy. *Internet Semiotics Encyclopedia*, 2004.

[W3C, 2005]W3C. *XSL Transformations (XSLT) Version 2.0.* W3C, 2005.

[Wellisch, 1991]H. Wellisch. *Indexing from A to Z.* 1991.

Risk Management:
An International Regulatory Framework

Mourad Bara

IBM Business Consulting Services
Management Technologies Practice
2 av Gambetta
92066 PARIS La Défense Cedex, France
(e-mail: mourad_bara@fr.ibm.com)

Abstract. In order to limit the banks against engagements far too risky which could, for some of them, to bring them to the bankruptcy, the Bank for International Settlements (BIS) proposed in 1988 the criteria on capital adequacy necessary to cover market risks and credit risks. These recommendations constituted the regulation known as Basle I which gave rise to the Cooke ratio. The objective of this paper is to review the bank regulatory framework and to explain how the Basle II framework, proposed in 2003 with Mac Donough ratio, is different from the previous one. We shall emphasize on the method of risk calculation and capital adequacy, the supervisory process, the market discipline and communications. We shall talk about their impacts on the banks, the product pricing, their profitability and on the economic environment. We will conclude by giving a progress report on certain solutions to improve the regulation. An outline on its later evolutions will also be discussed.

Data Mining for Advanced Customer Management

Beatriz Sanz Sáiz

Ernst & Young (e-mail: beatriz.sanzsaiz@es.ey.com)

1 Position of the problem

The financial sector is under a complete process of transformation regarding strategic and tactical customer management.

The main cause of this is the use of quantitative methods (data mining) for data management. This is done in order to extrapolate behavior patterns and to forecast possible conducts to define more effective business plans.

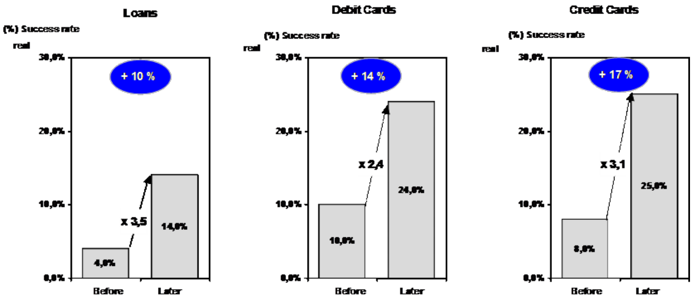
All of this is quite new when talking about marketing and business strategy, but it is not so if we focus on risks management, in which these methods have been previously applied to forecast the ability of payment or default of customers. Obviously, the existence of a regulator, like the Bank for International Settlements in Basel, and regulations, i.e. Basel I and II, has allowed this development.

If we make a closer analysis, why do we think that the current application of these techniques has been terribly underused in the business area? Moreover, why do we lately notice a significant change in trends? The following reasons might answer the questions:

- Human factor: there used to be a way of thinking in commercial areas that business knowledge was enough to identify possible customers to target for cards, funds, pension plans, etc. Strong investments in advertising and the idea that 'the more customers we contact, the more successful our campaigns will be', is something that, until today, has been followed by many organizations. These organizations continue to define their business strategy as focused on a 'product' vision, without thinking about adapting it to existing and potential customer segments.

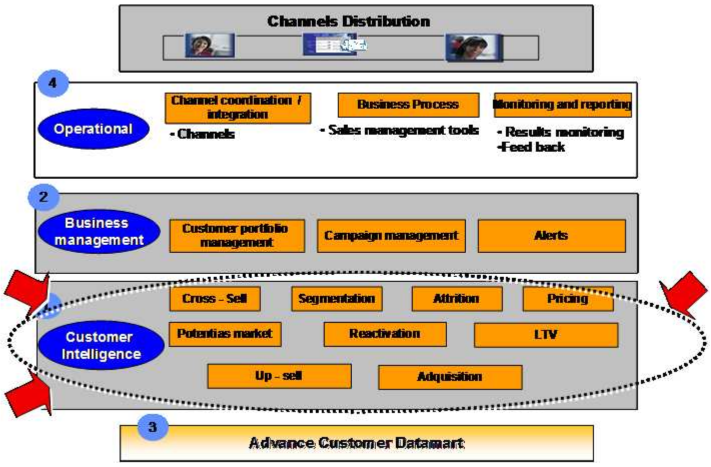
Results are the best way to convince others of the previous statements. Increases in the rates of sales obtained, thanks to the application of data mining techniques, is the trigger of this trend change.

- Distrust in unknown techniques: when a commercial or marketing director hears about algorithms of segmentation, distances, neuronal networks, etc., they may have two different reactions. On one hand, there is an admiration for the unknown, but on the other hand, they may worry about not being able to understand or explain WHY a customer has been proposed as a target of a specific business action to the sales network



The easy interpretation of results and the discovery of hidden behavior patterns are some of the keys to present data mining as the differentiating element when designing an advanced strategy in customer management.

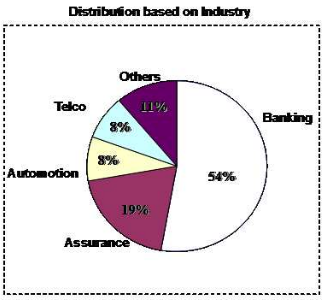
• Department initiatives: the truth is that, currently, only a few companies have a specific structure of analytic and commercial research from the organizational point of view. Up to now, several department initiatives have been developed in order to detect 'who are the potential purchasers of a product?' or 'what customers will cancel our products in the future?' However, there are few companies that have redefined an advanced commercial strategy and created more proactive and dynamic commercial models based on the systematization of the acquired knowledge to make it regularly available for commercial networks.

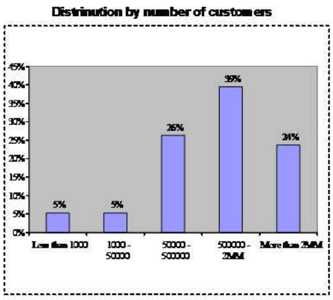


• Technology as an easy creator: up to now, any project with data analysis meant long processes of previous handling. Design and construction of advanced information data marts, with the added information of past behavior of customers, allows us to remarkably speed up these activities, reducing significantly the amount of time and resources necessary to develop a predictive model.

However, tools have been developed that significantly reduce the amount of time to compute the mathematical algorithms needed for processing millions of records.

Next, we present the results of a survey performed over 120 Spanish companies. The results are represented in a quantitative way and reflect the current Status of Companies in Advanced Customer Management Field





Finance Sector (Banking + Assurance) recruites the 73% of the Entities surveyed

63% of the Companies have more than 500.000 customers

Main Remarks

. . . There is a still long way to go . . .

Only 33% of the companies declare that they have a coordinated effort for customer management.

The financial sector (banking) is the most evolved in 'Advanced Customer Management'. There are significant differences with other sectors, such as assurance, which still stands out with a product/channel focus.

Currently, nearly 40% of the companies do have integrated customer information.

Not even 5% of the companies have calculated the following information regarding their customers: attrition risk, potential value, tendency of purchase, decision unit to which they belong, etc.

Only 13% of the surveyed companies declare to be applying business intelligence,

(data mining) as a method to define and apply strategies and specific commercial actions. 95% of this 13 % belong to the banking sector.

The concept of commercial planning is mainly spread out, but the availability of management is still very limited on the companies' side.

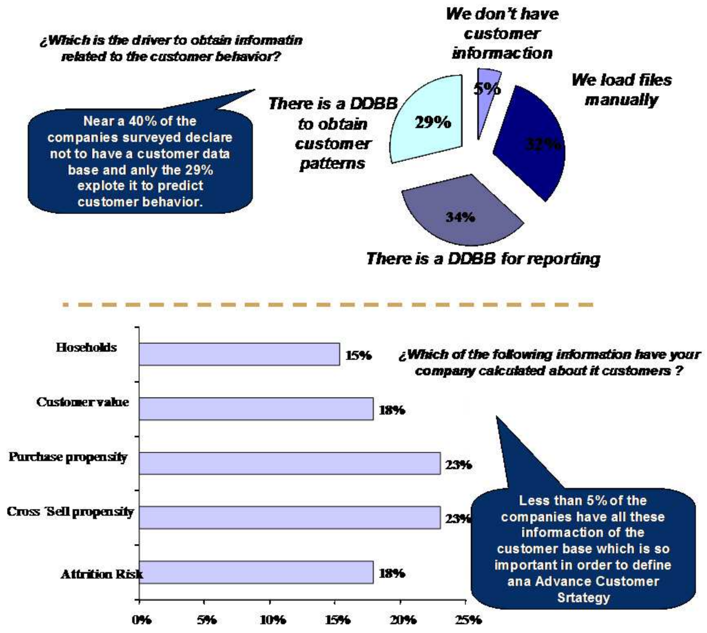
Uniquely, 8% of the companies take less than a week to design and start up a commercial campaign.

Only 34% of the surveyed companies have defined specific commercial plans for client clusters.

61% of the companies currently do not send unique selling propositions about a specific target to their commercial nets.

Their target selection systems are still based on products. Just the 3% answered that they are customer focused.

Related to customer intelligence, the answers were:



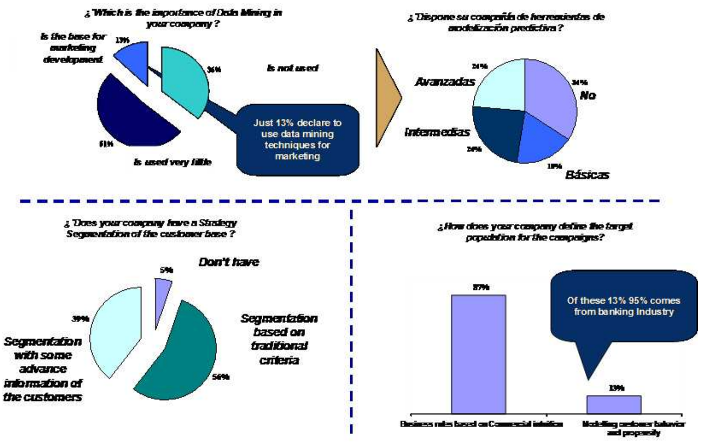
Near a 40% of the companies surveyed declare not to have a customer data base and anly the 29% explote it to predict customer behavior.

Less than 5% of the companies have all these informaction of the customer base which is so important in order to define ana Advance Customer Srtategy

In my point of view, the market conjuncture and the survey results prove that this trend is unstoppable

Advanced customer management is turning itself into an essential element in order to outline and receive continued growth in income.

Banking has historically been the industry that starts more impressive transformations in the world of business than any other. It is not different in this occasion. The banking sector is on a total revolution in reference to its own redefinitions of the commercial strategies of approach and customer management; and they have already been followed by other sectors such as telecommunications, the hotel industry, utilities, etc.



The advantages are obvious; the only requirement is an understanding of the underlying analytical techniques, and the conviction and implication of the whole company for change.

The first model to think about in order to define advanced customer management is, from my point of view, a strategic customer segmentation that allows the setting of commercial strategies attending to customer profiles instead of the company products catalog.

It is an almost definite fact that all financial entities currently have customer segmentation, but these are defined by traditional parameters such as activity or type of client (individual or company). The information that is currently known about customers allows inferring behaviors or values that are important in defining an advanced customer management strategy. This is the case in their income level or their potential value.

The next pages detail the methodology used to calculate the values and the business applications underlying segmentation.

2 Practical Case: Strategic Customer Segmentation

The segmentation proposed is compounded for magnitudes such as age, rent level, potential value, and customer connection.
These magnitudes clearly define the reasons for specialized commercial strategy.

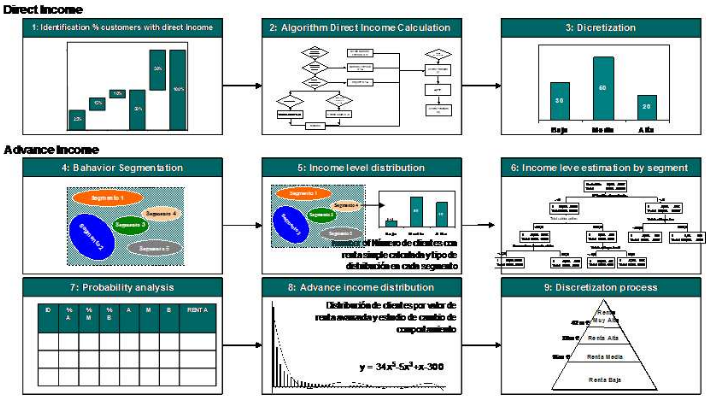
• Age: there is no doubt that age conditions different behaviors or attitudes in customers.
• Income level: the customer's income level is, in banks, one of the main parameters used to define a customer profile and what they are expecting to receive from the entity.
• Customer value: understood that future expected fluxes of profitability, incomes, etc., it fixes the investment for each type of customer.

• Client connection: this links the customer relationship and loyalty level with the entity.

We are now going to show in detail the methodological process to calculate the income level and the potential value of the previous parameters.

Income Level

The procedure to calculate a customer's income level has several steps:


Direct Income
1: Identification % customers with direct Income
2: Algorithm Direct Income Calculation
3: Dicretization
30
60
20
Baja
Media
Alta
Advance Income
4: Bahavior Segmentation
Segmento 1
Segmento 2
Segmento 3
Segmento 4
Segmento 5
5: Income level distribution
Número de clientes con renta simple calculada y tipo de distribución en cada segmento
6: Income leve estimation by segment
7: Probability analysis
ID
% A
% M
% B
A
M
B
RENTA
8: Advance income distribution
Distribución de clientes por valor de renta avanzada y estudio de cambio de comportamiento
$y = 34x^5-5x^3+x-300$
9: Discretizaton process
Renta Muy Alta
Renta Alta
Renta Media
Renta Baja

a. - Direct Income Calculation:

There are a certain percentage of customers for whom the income calculation is obtained directly from derived transformations of other variable values. For example, payroll or pension, recurring incomes, etc. . . .

On average, in the Spanish Banking Sector is able to calculate income level by this process for 40% of the population.

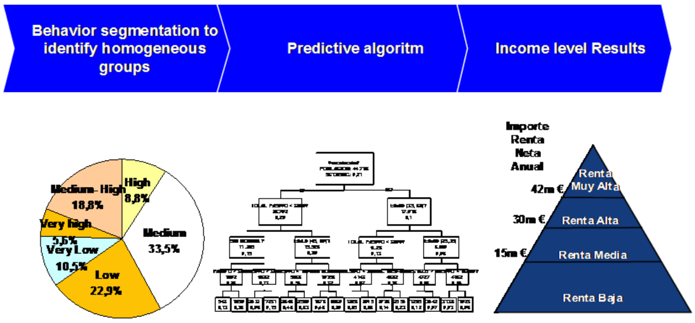
b. - Advance Income Estimation:

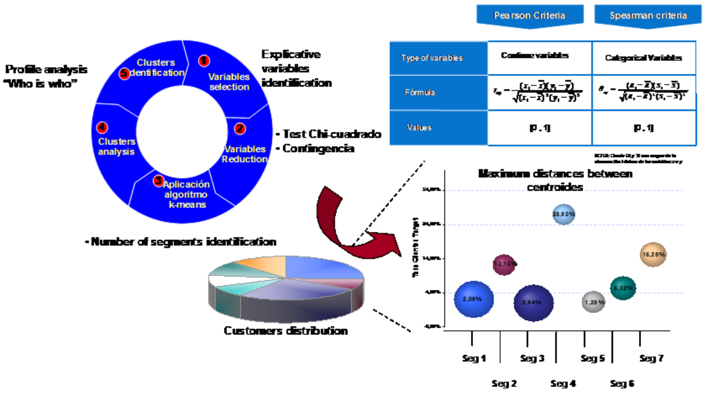
For the other 60% of the population, the income calculation is obtained by statistical inference. They assign to each customer, whose rent is unknown, the same rent interval of those others with which the distance in behavioral terms is lowest.

This is calculated with the following:

b.1. - Behavioral clustering:

By vote algorithm (condorcet) or a distance-based algorithm (mainly k-means), we are able to identify, through an iterative procedure defined in five steps, customer segments with similar behavioral. This previous step is essential in order to set a predictive algorithm to assign each customer a winning probability of having a certain income level.





b.2. - Forecasting classification

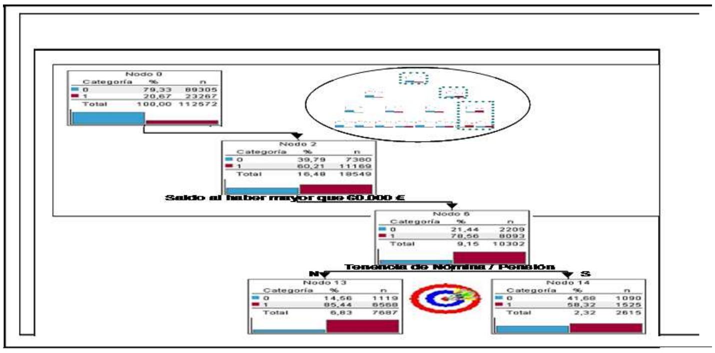
For each of the identified segments and desired levels of income forecast (i.e. low, medium, high), a forecasting algorithm is created (i.e. Chaid, RRNN) that will determine the probability that a customer belongs to each of the categories.

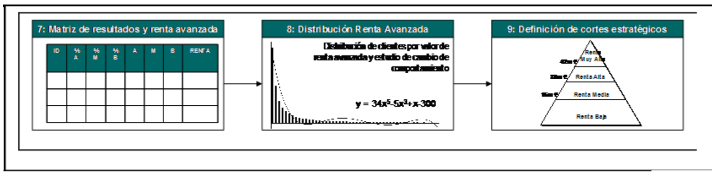
Customer niches are identified with high probabilities of belonging to a specific group.

b.3. - Distribution and transformation analysis

Once the probability of belonging to each level of income for each customer is calculated, a comparison of the income distribution over the population will be performed. This income will be calculated with direct methods, whose final value has been forecasted by data mining.

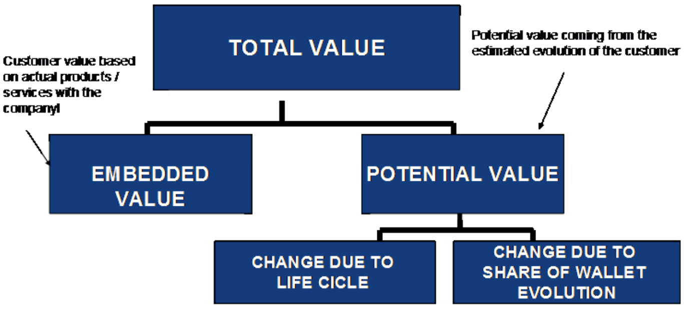
Using classification and weighting criteria, we are finally able to assign to each customer the ‘income band’ of major probability.




7: Matriz de resultados y renta avanzada
ID
% A
% M
% B
A
M
B
RENTA
8: Distribución Renta Avanzada
Distribución de clientes por valor de renta avanzada y estudio de cambio de comportamiento
y = 34x5-5x3+x-300
9: Definición de cortes estratégicos
Renta Muy Alta
Renta Alta
Renta Media
Renta Baja

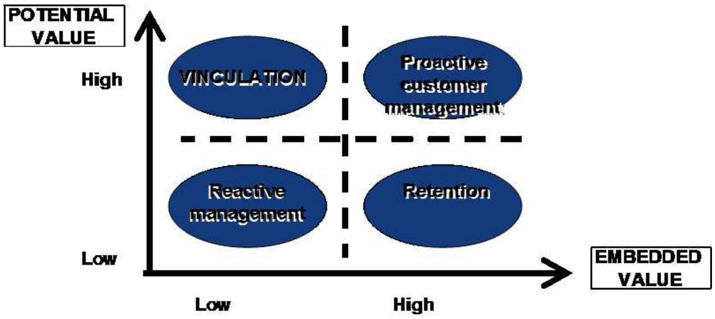
Customer value calculation

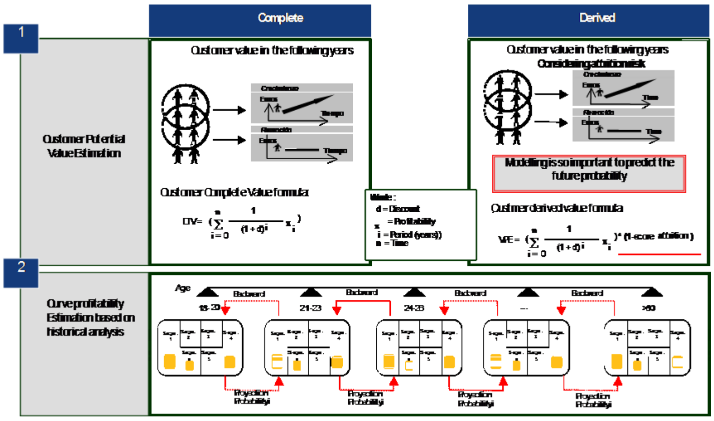
The methodological procedure of the customer value calculation is developed by taking into account the present positioning of each customer and the future projections derived of the potential increase.


TOTAL VALUE
Customer value based on actual products / services with the companyl
Potential value coming from the estimated evolution of the customer
EMBEDDED VALUE
POTENTIAL VALUE
CHANGE DUE TO LIFE CICLE
CHANGE DUE TO SHARE OF WALLET EVOLUTION

From a strategic point of view, the breakdown of the total customer value into the present and potential value will allow the definition of differential commercial strategies.

There are two variants in the potential value calculation: complete and derived, understanding that derived is the value of the customer, also considering the possible risk of attrition.




Complete
Derived
1
Customer Potential Value Estimation
Customer value in the following years
Customer Complete Value formula
Customer value in the following years Considering attrition risk
Modelling is so important to predict the future probability
Customer derived value formula
d = Discount
x = Profitability
i = Period (years)
n = Time
2
Curve profitability Estimation based on historical analysis
Age
Backward
18-20
21-23
24-29
...
>60
Proyection Probability

Each customer's micro segment, the evolution of profitability curves, should be evaluated in a quantitative way. They should be inferred based on the evolution of a customer's collective, which in the past had a similar behavior to them in the present moment. In this sense it is necessary the use of clustering algorithms as well as Markov streams, due to the probability that each status is conditioned by the previous one.

Conclusions of the strategic segmentation

Strategic segmentation is, as we have seen, the starting point for the definition of an advances customer management strategy.

We have proven with this example how data mining techniques allow for the calculation of advanced customer information, which supports a more specialized decision, and, therefore, more efficient business management. It is considered as the most necessary driver to improve business efficiency and obtaining a continuous income increase.

ARQAT:
An Exploratory Analysis Tool For Interestingness Measures

Xuan-Hiep Huynh, Fabrice Guillet, and Henri Briand

LINA CNRS FRE 2729 - Polytechnic school of Nantes University
La Chantrerie BP 50609
44306 Nantes Cedex 3, France
(e-mail: xuan-hiep.huynh@polytech.univ-nantes.fr,
fabrice.guillet@polytech.univ-nantes.fr,
henri.briand@polytech.univ-nantes.fr)

Abstract. Finding interestingness measures to evaluate association rules has become an important knowledge quality issue in KDD. Many interestingness measures may be found in the literature, and many authors have discussed and compared interestingness properties in order to help choose the best measures for a given application. As interestingness depends both on the data structure and on the decision-maker's goals, some measures may be relevant in some context, but not in others. Therefore, it is necessary to design new contextual approaches in order to help the decision-maker to select the best interestingness measures. In this paper, we present ARQAT a new tool to study the specific behavior of a set of 34 interestingness measures in the context of a specific dataset and in an exploratory data analysis perspective. The tool implements 14 graphical and complementary views structured on 5 levels of analysis: ruleset analysis, correlation and clustering analysis, best rules analysis, sensitivity analysis, and comparative analysis. The tool is described and illustrated on the mushroom dataset in order to show the interest of both the exploratory approach and the use of complementary views.
Keywords: interestingness measure, ARQAT, exploratory analysis.

1 Introduction

In the last decade, the designing of Interestingness Measure (IM) to evaluate association rules has become an important knowledge quality challenge in the context of KDD. This is because association rule [Agrawal *et al.*, 1993] is one of the few models dedicated to unsupervised discovery of rule tendencies in data. It is unfortunately confronted to a major difficulty: the user (a decision-maker or a data-analyst) must cope with a large amount of extracted rules in order to validate and select the best ones [Piatetsky-Shapiro, 1991]. One way to reduce the cost of the user's task is to help him/her with the measurement of rule interestingness adapted to both his/her goals and the dataset studied.

In initial research works [Agrawal *et al.*, 1993][Agrawal and Srikant, 1994] on association rules, these precursors have introduced the first two statistical measures: support and confidence. These measures are well adapted to

Apriori algorithm constraints, but are not sufficient to capture rule interestingness. To improve this limit, many complementary IMs have been then introduced in the research literature. As interestingness depends both on the user's goals and data characteristics, two kinds of IMs may be distinguished [Freitas, 1999]: subjective and objective. First, subjective measures depend on the user's goals and his/her knowledge or beliefs, and are combined to specific supervised algorithms in order to compare the extracted rules with what the user knows or wants [Padmanabhan and Tuzhilin, 1998][Liu *et al.*, 1999]. Hence, subjective measures allow capturing rule novelty and unexpectedness in relation to the user's knowledge or beliefs. Second, objective measures are statistical indexes that only rely on data structure and more precisely on itemset frequency. Many interesting surveys summarize their definitions and properties (see [Bayardo and Agrawal, 1999], [Hilderman and Hamilton, 2001],
[Tan *et al.*, 2002], [Tan *et al.*, 2004], [Piatetsky-Shapiro, 1991],
[Lenca *et al.*, 2004], [Guillet, 2004]). These surveys address two joint research issues, the definition of the set of principles or properties that lead to the design of a good IM, and their comparison from a data-analysis point of view to study IM behavior in order to help the user to select the best ones. In [Vaillant *et al.*, 2003] a tool HERBS is also presented.

In this paper, we present a new approach and a dedicated tool ARQAT (Association Rule Quality Analysis Tool) to study the specific behavior of a set of IMs in the context of a specific dataset and in an exploratory analysis perspective. More precisely, ARQAT is a toolbox designed to help a data-analyst to capture the best measures and as a final purpose, the best rules within a specific ruleset.

The paper is structured as follows. In section 2, we introduce the principles and the structure of ARQAT tool. In the three next sections, we describe 3 groups of ARQAT views: ruleset statistics, correlation analysis, and best rules analysis. We illustrate each view on the mushroom dataset, in order to show the interest of the exploratory approach for IM analysis.

2 Principles of ARQAT tool

ARQAT is an exploratory analysis tool that embeds 34 objective IMs studied in surveys. We complete this list of IMs with three complementary measures: Implication Intensity (II) introduced by Gras [Gras, 1996] [Guillaume *et al.*, 1998], Entropic Implication Intensity (EII) [Gras *et al.*, 2001] [Blanchard *et al.*, 2003], and the informational ratio modulated by the contra-positive (TIC)
[Blanchard *et al.*, 2004] (See Appendix 1 for a complete list of selected measures).

ARQAT (Fig. 1) implements a set of 14 complementary and graphical views structured in 5 task-oriented groups: ruleset analysis, correlation and clustering analysis, best rules analysis, sensitivity analysis, and comparative analysis.

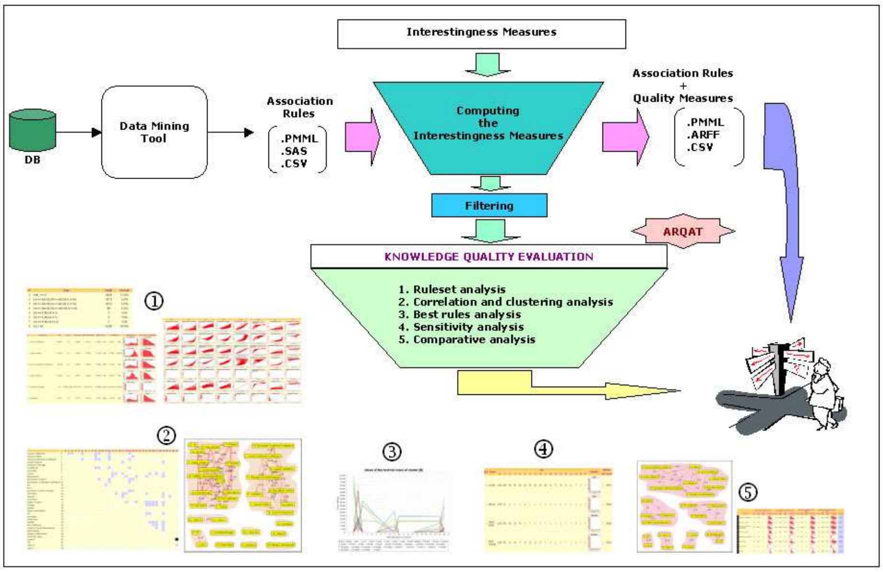


Fig. 1. ARQAT structure.

For the input, ARQAT requires an association ruleset where each association rule $a \Rightarrow b$ must be associated to 4 cardinalities $(n, n_a, n_b, n_{a\overline{b}})$. More precisely, n is the number of transactions, n_a (resp. n_b) the number of transactions satisfying the itemset a (resp. b), and $n_{a\overline{b}}$ is the number of transactions satisfying $a \wedge \overline{b}$ (negative examples).

In a first stage, the input ruleset is preprocessed in order to compute the IM values of each rule, and the correlations between all IM pairs. The results are stored in two tables: an IM table (R×I) where rows are rules and columns are IM values, and a correlation matrix (I×I) crossing IMs. At this stage, the ruleset may also be sampled in order to focus the study on a more restricted subset of rules.

In a second stage, the data-analyst can then drive the graphical exploration of results through a classical web-browser. ARQAT is structured in 5 groups of task-oriented views. The first group (1 in Fig. 1) is dedicated to ruleset and simple IM statistics to better understand the structure of the IM table (R×I). The second group (2) is oriented to the study of IM correlation in table (I×I) and IM clustering in order to select the best IMs. The third one (3) focuses on rule ordering to select the best rules. The fourth group (4) proposes to study the sensitivity of IMs. The last group (5) offers the possibility to compare the results obtained from different rulesets.

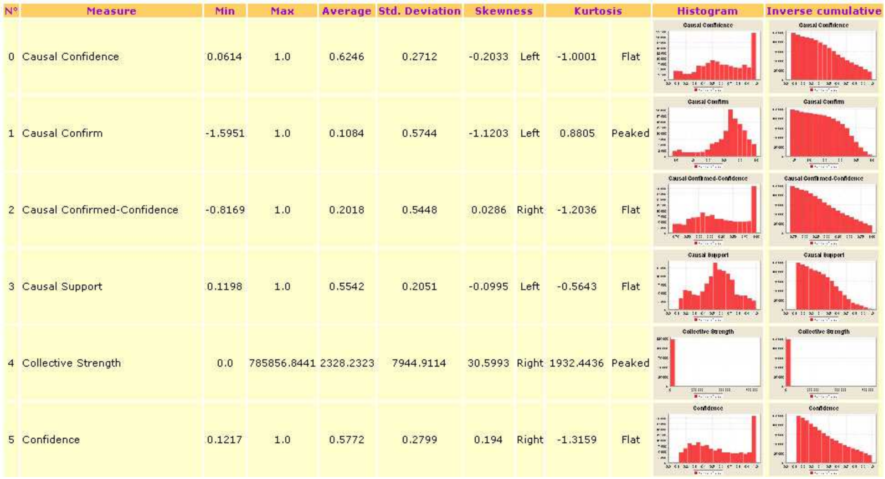
The next sections will focus on the description of the first three groups and will illustrate it with the same ruleset: 120000 association rules extracted by Apriori algorithm (support 10%) from mushroom dataset [Blake and Merz, 1998].

3 Ruleset statistics

This first group of ARQAT tools delivers 3 views summarizing some simple statistics in the ruleset structure. The first one, ruleset characteristics , shows the distributions underlying rule cardinalities, in order to detect borderline cases.

The second view, IM distribution (Fig. 2), draws the histograms for each IM. The distributions are also completed with minimum, maximum, average, standard deviation, skewness and kurtosis values. In Fig. 2, one can see that Confidence (line 5) has an irregular distribution and a great number of rules with 100% confidence, it is very different from Causal Confirm (line 1).

The third view, joint-distribution analysis (Fig. 3), shows the scatter-plot matrix of all IM pairs. This graphical matrix is very useful to see the details of the relationships between IMs. For instance, Fig. 3 shows four disagreement shapes: Rule Interest vs Yule's Q (4), Sebag & Schoenauer vs Yule's Y (5), Similarity Index vs Support (6), and Yule's Y vs Support (7) (strongly uncorrelated). On the other hand, we can notice four agreement shapes on Putative Causal Dependency vs Rule Interest (1), Putative Causal Dependency vs Similarity Index (2), Rule Interest vs Similarity Index (3), and Yule's Q vs Yule's Y (8) (strongly correlated).



N°	Measure	Min	Max	Average	Std. Deviation	Skewness		Kurtosis		Histogram	Inverse cumulative
0	Causal Confidence	0.0614	1.0	0.6246	0.2712	-0.2033	Left	-1.0001	Flat	Causal Confidence	Causal Confidence
1	Causal Confirm	-1.5951	1.0	0.1084	0.5744	-1.1203	Left	0.8805	Peaked	Causal Confirm	Causal Confirm
2	Causal Confirmed-Confidence	-0.8169	1.0	0.2018	0.5448	0.0286	Right	-1.2036	Flat	Causal Confirmed-Confidence	Causal Confirmed-Confidence
3	Causal Support	0.1198	1.0	0.5542	0.2051	-0.0995	Left	-0.5643	Flat	Causal Support	Causal Support
4	Collective Strength	0.0	785856.8441	2328.2323	7944.9114	30.5993	Right	1932.4436	Peaked	Collective Strength	Collective Strength
5	Confidence	0.1217	1.0	0.5772	0.2799	0.194	Right	-1.3159	Flat	Confidence	Confidence

Fig. 2. Distribution of some measures on mushroom dataset.

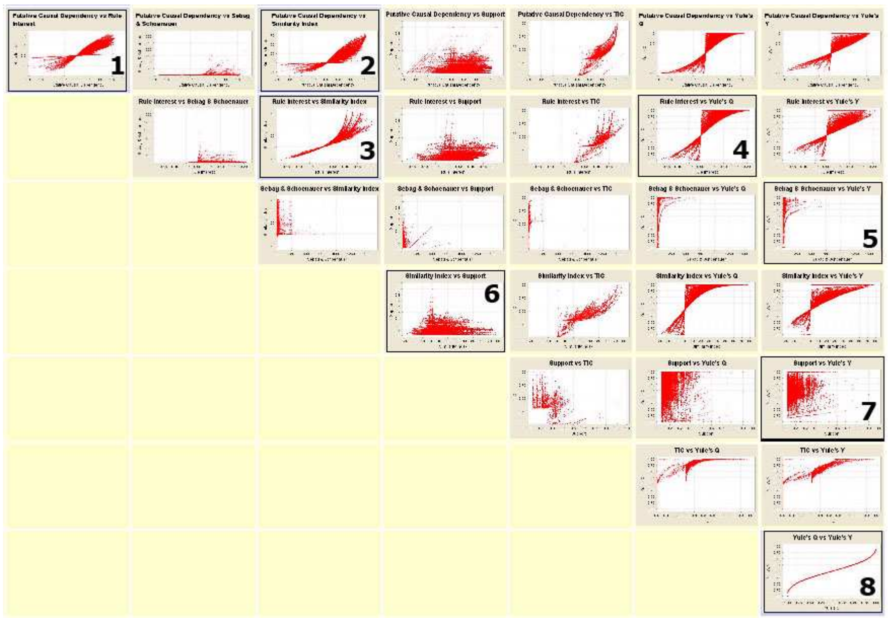


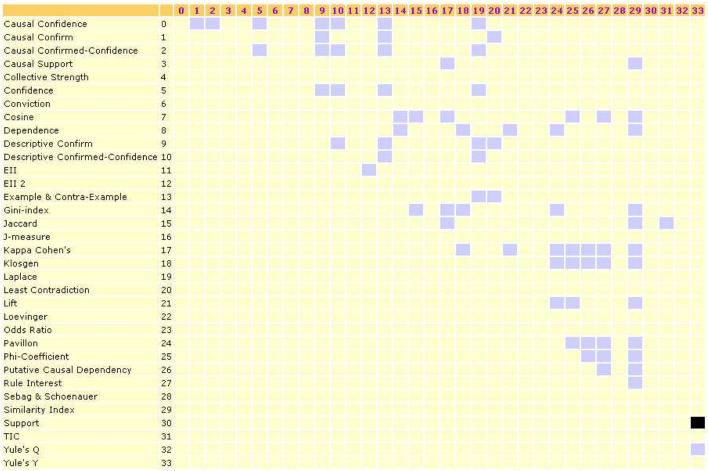
Fig. 3. Scatterplot matrix of joint-distributions on mushroom dataset.

4 Correlation analysis

This second group is dedicated to IM correlation study in order to deliver IM clustering and facilitate the choice of the subset of IMs that is the best-adapted to describe the ruleset. The correlations between IM pairs were computed in the preprocessing stage by using the Pearson's correlation coefficient and stored in the correlation matrix (I × I). The user has two visual possibilities to explore the matrix. The first one is a simple summary matrix in which each significant correlation value is visually associated to a different color (a level of gray). For instance, the only one dark cell from Fig. 4 shows a low correlation value between Yule's Y and Support. The other seventy-four gray cells correspond to high correlation values.

The second one (Fig. 5) is a graph-based view of the correlation matrix. As graphs are a good way to offer relevant graphical insights on data structure, we use the correlation matrix as the relation of an undirected and valued graph, called correlation graph. In a correlation graph, a vertex represents an IM and an edge value is the correlation value between 2 vertices/measures. We also add the possibility to set a minimal threshold τ (resp. maximal threshold θ) to retain only the edges associated to a high correlation (resp. low correlation), that deliver a partial subgraph CG+ (resp. CG0).

These two partial subgraphs can then be processed in order to extract clusters of measures. Each cluster is defined as a maximal connected subgraph. In CG+, each cluster will gather correlated or anti-correlated mea-



		0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33
Causal Confidence	0		▨	▨			▨				▨	▨			▨						▨														
Causal Confirm	1										▨				▨							▨													
Causal Confirmed-Confidence	2						▨				▨	▨			▨						▨														
Causal Support	3																		▨												▨				
Collective Strength	4																																		
Confidence	5										▨	▨			▨						▨														
Conviction	6																																		
Cosine	7															▨	▨		▨								▨		▨		▨				
Dependence	8															▨				▨			▨			▨					▨				
Descriptive Confirm	9											▨			▨						▨	▨													
Descriptive Confirmed-Confidence	10														▨						▨														
EII	11													▨																					
EII 2	12																																		
Example & Contra-Example	13																				▨	▨													
Gini-index	14																▨		▨	▨						▨					▨				
Jaccard	15																		▨												▨		▨		
J-measure	16																																		
Kappa Cohen's	17																			▨			▨			▨	▨	▨	▨		▨				
Klosgen	18																									▨	▨	▨	▨		▨				
Laplace	19																																		
Least Contradiction	20																																		
Lift	21																									▨	▨				▨				
Loevinger	22																																		
Odds Ratio	23																																		
Pavillon	24																										▨	▨	▨		▨				
Phi-Coefficient	25																											▨	▨		▨				
Putative Causal Dependency	26																												▨		▨				
Rule Interest	27																														▨				
Sebag & Schoenauer	28																																		
Similarity Index	29																																		
Support	30																																		■
TIC	31																																		
Yule's Q	32																																		▨
Yule's Y	33																																		

Fig. 4. Summary matrix of correlation on mushroom dataset.

sures that may be interpreted similarly: they deliver a close point of view on data. Moreover, in CG0 each cluster will contain uncorrelated measures: measures that deliver a different point of view.

Hence, as each graph depends on a specific ruleset, the user will use the graphs as data insight, which will graphically help him/her to select the minimal set of the measures best adapted to his/her data. For instance in Fig. 5, CG+ graph contains 11 clusters on 34 measures, the user can select in each cluster the most representative measure, and then retain it to validate the rules.

A close watch on the CG0 graph (Fig. 5) shows an uncorrelated cluster formed by Support and Yule's Y measures (also the dark cell in Fig. 4). This observation is confirmed on Fig. 3 (7). CG+ graph shows a trivial cluster where Yule's Q and Yule's Y are strongly correlated. This is also confirmed on Fig. 3 (8) showing a functional dependency between the two measures. These two examples show the interest to use the scatterplot matrix complementarily (Fig. 3) with the correlation graphs CG0, CG+ (Fig. 5) in order to evaluate the nature of the correlation links, and overcome the limits of the correlation coefficient.

5 Best rule analysis

In order to help a user to select the best rules, we have implemented two specific views. The first view (Fig. 6) collects a set of given number of best rules for each measure in one cluster, in order to answer the question ”How interesting are the rules of this cluster?”

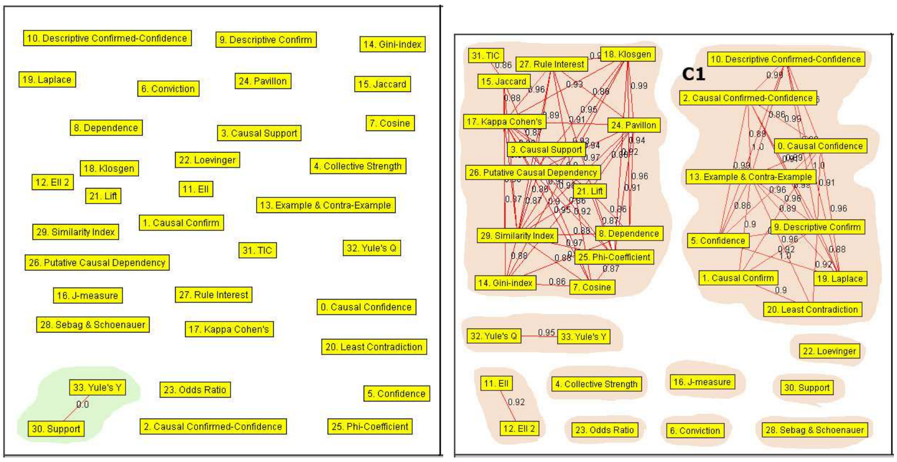
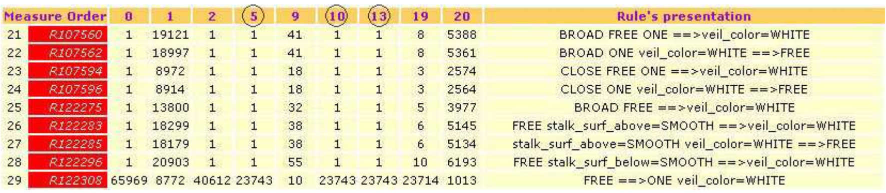


Fig. 5. CG0 and CG+ graphs on mushroom dataset (clusters are highlighted with a gray background).

The selected rules can alternatively be visualized with parallel coordinates drawing (Fig. 7). The main interest of such a drawing is to rapidly see the IM rankings of the rules, and then to facilitate their interpretation.

These two views can be used with IM values of a rule or alternatively with the rank of the value. For instance, Fig. 6 and Fig. 7 use the rank to evaluate the union of the ten best rules for each of the nine IMs in the C1 cluster (see Fig. 5). The Y-axis in Fig. 7 holds the rule rank for the corresponding measure. By observing the concentration lines on low rank values, we can obtain 3 measures: Confidence(5), Decsriptive Confirmed-Confidence(10), and Example & Contra-Example(13) (on points 1, 2, 3 respectively) that are good for a majority of best rules. This can also be retrieved from columns 5, 10, 13 of Fig. 6.



Measure Order		0	1	2	5	9	10	13	19	20	Rule's presentation
21	R107560	1	19121	1	1	41	1	1	8	5388	BROAD FREE ONE ==>veil_color=WHITE
22	R107562	1	18997	1	1	41	1	1	8	5361	BROAD ONE veil_color=WHITE ==>FREE
23	R107594	1	8972	1	1	18	1	1	3	2574	CLOSE FREE ONE ==>veil_color=WHITE
24	R107596	1	8914	1	1	18	1	1	3	2564	CLOSE ONE veil_color=WHITE ==>FREE
25	R122275	1	13800	1	1	32	1	1	5	3977	BROAD FREE ==>veil_color=WHITE
26	R122283	1	18299	1	1	38	1	1	6	5145	FREE stalk_surf_above=SMOOTH ==>veil_color=WHITE
27	R122285	1	18179	1	1	38	1	1	6	5134	stalk_surf_above=SMOOTH veil_color=WHITE ==>FREE
28	R122296	1	20903	1	1	55	1	1	10	6193	FREE stalk_surf_below=SMOOTH ==>veil_color=WHITE
29	R122308	65969	8772	40612	23743	10	23743	23743	23714	1013	FREE ==>ONE veil_color=WHITE

Fig. 6. Union of the ten best rules of the first cluster on mushroom dataset (extract).

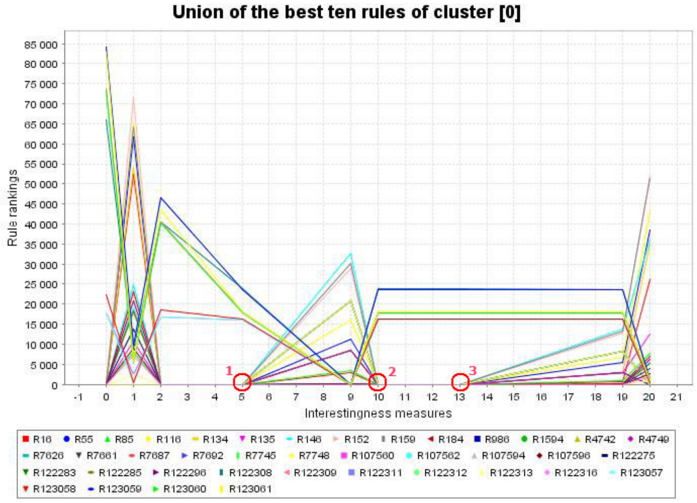


Fig. 7. Plot of the union of the ten best rules of the first cluster on mushroom dataset.

6 Conclusion

We have designed and described some features of a new tool, ARQAT, implementing an exploratory data-analysis approach for IM behavior analysis on a specific dataset.

Technically, ARQAT is written in Java and embeds a set of 14 graphical tools. For exchange facilities, three common file formats are used for importing/exporting the rulesets: PMML (XML data-mining standard), CSV (Excel and SAS) and ARFF (used by WEKA). ARQAT will be freely available at www.polytech.univ-nantes.fr/arqat.

In this paper, we have shown the interest of such an exploratory approach, where the intensive use of graphical and complementary visualizations improves and facilitates data insight for the user.

ARQAT is a first step toward a larger analysis platform in the domain of knowledge quality research. Our future research will investigate the two following directions. First, we will improve the correlation analysis by introducing a better measure than Pearson coefficient whose limits are stressed in the literature. Second, we will also improve the IM clustering analysis with IM aggregation techniques to facilitate the user's decision making from the best IMs.

References

[Agrawal and Srikant, 1994]R. Agrawal and R. Srikant. Fast algorithms for mining association rules. In *Proceedings of the 20th VLDB Conference*, pages 487–499, 1994.

[Agrawal *et al.*, 1993]R. Agrawal, T. Imielinski, and A. Swami. Mining association rules between sets of items in large databases. In *Proceedings of 1993 ACM-SIGMOD International Conference on Management of Data*, pages 207–216, 1993.

[Bayardo and Agrawal, 1999]Jr.R.J. Bayardo and R. Agrawal. Mining the most interestingness rules. In *Proceedings of the Fifth ACM SIGKDD International Confeference On Knowledge Discovery and Data Mining*, pages 145–154, 1999.

[Blake and Merz, 1998]C.L. Blake and C.J. Merz. *UCI Repository of machine learning databases, http://www.ics.uci.edu/∼mlearn/MLRepository.html.* University of California, Irvine, Dept. of Information and Computer Sciences, 1998.

[Blanchard *et al.*, 2003]J. Blanchard, P. Kuntz, F. Guillet, and Gras R. Implication intensity: from the basic statistical definition to the entropic version. In *Statistical Data Mining and Knowledge Discovery*, pages 475–493, 2003.

[Blanchard *et al.*, 2004]J. Blanchard, F. Guillet, R. Gras, and H. Briand. Mesurer la qualité des règles et de leurs contraposés avec le taux informationnel tic. In *Revue des Nouvelles Technologies de l'Information (RNTI)*, pages 287–298, 2004.

[Freitas, 1999]A.A. Freitas. On rule interestingness measures. In *Knowledge-Based Systems*, pages 309–315, 1999.

[Gras *et al.*, 2001]R. Gras, P. Kuntz, R. Couturier, and F. Guillet. Une version entropique de l'intensité d'implication pour les corpus volumineux. In *Extraction des Connaissances et Apprentissage (ECA)*, pages 69–80, 2001.

[Gras, 1996]R. Gras. *L'implication statistique - Nouvelle méthode exploratoire de données.* La pensée sauvage édition, 1996.

[Guillaume *et al.*, 1998]S. Guillaume, F. Guillet, and J. Philippé. Improving the discovery of association rules with intensity of implication. In Lecture Notes in Compuper Science, editor, *Proceedings of 2nd European Symp. on Principles of Data Mining and Knowledge Discovery, PKDD'98*, pages 318–327, 1998.

[Guillet, 2004]F. Guillet. Mesures de la qualité des connaissances en ecd. In *Actes des tutoriels, 4ème Conférence francophone Extraction et Gestion des Connaissances (EGC'2004), http://www.isima.fr/ egc2004/*, pages 1–60, 2004.

[Hilderman and Hamilton, 2001]R.J. Hilderman and H.J. Hamilton. *Knowledge Discovery and Measures of Interestingness.* Kluwer Academic Publishers, 2001.

[Lenca *et al.*, 2004]P. Lenca, P. Meyer, P. Picouet, B. Vaillant, and S. Lallich. Evaluation et analyse multi-critères des mesures de qualité des règles d'association. In *Revue des Nouvelles Technologies de l'Information - Mesures de Qualité pour la Fouille de Données, RNTI-E-1*, pages 219–246, 2004.

[Liu *et al.*, 1999]B. Liu, W. Hsu, L. Mun, and H. Lee. Finding interestingness patterns using user expectations. In *IEEE Transactions on knowledge and data mining 11(1999)*, pages 817–832, 1999.

[Padmanabhan and Tuzhilin, 1998]B. Padmanabhan and A. Tuzhilin. A belief-driven method for discovering unexpected patterns. In *Proceedings of the 4th international conference on knowledge discovery and data mining*, pages 94–100, 1998.

[Piatetsky-Shapiro, 1991]G. Piatetsky-Shapiro. Discovery, analysis and presentation of strong rules. In G. Piatetsky-Shapiro and W. Frawley, editors, *Knowledge Discovery in Databases*, pages 229–248, 1991.

[Tan *et al.*, 2002]P.N. Tan, V. Kumar, and J. Srivastava. Selecting the right interestingness measure for association patterns. In *Proc. of the Eighth ACM*

SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-2002), pages 32–41, 2002.

[Tan *et al.*, 2004]P.N. Tan, V. Kumar, and J. Srivastava. Selecting the right objective measure for association analysis. In *Information Systems 29(4)*, pages 293–313, 2004.

[Vaillant *et al.*, 2003]B. Vaillant, P. Picouet, and P. Lenca. An extensible platform for rule quality measure benchmarking. In R. Bisdorff, editor, *Human Centered Processes (HCP'2003)*, pages 187–191, 2003.

Appendix 1: IM formulas

N°	Interestingness Measure	$f(n, n_a, n_b, n_{a\overline{b}})$
0	Causal Confidence	$1 - \frac{1}{2}(\frac{1}{n_a} + \frac{1}{n_{\overline{b}}})n_{a\overline{b}}$
1	Causal Confirm	$\frac{n_a + n_{\overline{b}} - 4n_{a\overline{b}}}{n}$
2	Causal Confirmed-Confidence	$1 - \frac{1}{2}(\frac{3}{n_a} + \frac{1}{n_{\overline{b}}})n_{a\overline{b}}$
3	Causal Support	$\frac{n_a + n_{\overline{b}} - 2n_{a\overline{b}}}{n}$
4	Collective Strength	$\frac{(n_a - n_{a\overline{b}})(n_{\overline{b}} - n_{a\overline{b}})(n_a n_{\overline{b}} + n_b n_{\overline{a}})}{(n_a n_b + n_{\overline{a}} n_{\overline{b}})(n_b - n_a + 2n_{a\overline{b}})}$
5	Confidence	$1 - \frac{n_{a\overline{b}}}{n_a}$
6	Conviction	$\frac{n_a n_{\overline{b}}}{n n_{a\overline{b}}}$
7	Cosine	$\frac{n_a - n_{a\overline{b}}}{\sqrt{n_a n_b}}$
8	Dependence	$\lvert \frac{n_{\overline{b}}}{n} - \frac{n_{a\overline{b}}}{n_a} \rvert$
9	Descriptive Confirm	$\frac{n_a - 2n_{a\overline{b}}}{n}$
10	Descriptive Confirmed-Confidence	$1 - 2\frac{n_{a\overline{b}}}{n_a}$
11	EII ($\alpha = 1$)	$\sqrt{\varphi \times I^{\frac{1}{2\alpha}}}$
12	EII ($\alpha = 2$)	$\sqrt{\varphi \times I^{\frac{1}{2\alpha}}}$
13	Example & Contra-Example	$1 - \frac{n_{a\overline{b}}}{n_a - n_{a\overline{b}}}$
14	Gini-index	$\frac{(n_a - n_{a\overline{b}})^2 + n_{a\overline{b}}^2}{n n_a} + \frac{(n_b - n_a + n_{a\overline{b}})^2 + (n_{\overline{b}} - n_{a\overline{b}})^2}{n n_{\overline{a}}} - \frac{n_b^2}{n^2} - \frac{n_{\overline{b}}^2}{n^2}$
15	Jaccard	$\frac{n_a - n_{a\overline{b}}}{n_b + n_{a\overline{b}}}$
16	J-measure	$\frac{n_a - n_{a\overline{b}}}{n} log_2 \frac{n(n_a - n_{a\overline{b}})}{n_a n_b} + \frac{n_{a\overline{b}}}{n} log_2 \frac{n n_{a\overline{b}}}{n_a n_{\overline{b}}}$
17	Kappa Cohen's	$\frac{2(n_a n_{\overline{b}} - n n_{a\overline{b}})}{n_a n_{\overline{b}} + n_{\overline{a}} n_b}$
18	Klosgen	$\sqrt{\frac{n_a - n_{a\overline{b}}}{n}}(\frac{n_{\overline{b}}}{n} - \frac{n_{a\overline{b}}}{n_a})$
19	Laplace	$\frac{n_a + 1 - n_{a\overline{b}}}{n_a + 2}$
20	Least Contradiction	$\frac{n_a - 2n_{a\overline{b}}}{n_b}$
21	Lift	$\frac{n(n_a - n_{a\overline{b}})}{n_a n_b}$
22	Loevinger	$1 - \frac{n n_{a\overline{b}}}{n_a n_{\overline{b}}}$
23	Odds Ratio	$\frac{(n_a - n_{a\overline{b}})(n_{\overline{b}} - n_{a\overline{b}})}{n_{a\overline{b}}(n_b - n_a + n_{a\overline{b}})}$
24	Pavillon	$\frac{n_{\overline{b}}}{n} - \frac{n_{a\overline{b}}}{n_a}$
25	Phi-Coefficient	$\frac{n_a n_{\overline{b}} - n n_{a\overline{b}}}{\sqrt{n_a n_b n_{\overline{a}} n_{\overline{b}}}}$
26	Putative Causal Dependency	$\frac{3}{2} + \frac{4n_a - 3n_b}{2n} - (\frac{3}{2n_a} + \frac{2}{n_{\overline{b}}})n_{a\overline{b}}$
27	Rule Interest	$\frac{1}{n}(\frac{n_a n_{\overline{b}}}{n} - n_{a\overline{b}})$
28	Sebag & Schoenauer	$1 - \frac{n_{a\overline{b}}}{n_a - n_{a\overline{b}}}$
29	Similarity Index	$\frac{n_a - n_{a\overline{b}} - \frac{n_a n_b}{n}}{\sqrt{\frac{n_a n_b}{n}}}$
30	Support	$\frac{n_a - n_{a\overline{b}}}{n}$
31	TIC	$\sqrt{TI(a \rightarrow b) \times TI(\overline{b} \rightarrow a)}$
32	Yule's Q	$\frac{n_a n_{\overline{b}} - n n_{a\overline{b}}}{n_a n_{\overline{b}} + (n_b - n_{\overline{b}} - 2n_a)n_{a\overline{b}} + 2n_{a\overline{b}}^2}$
33	Yule's Y	$\frac{\sqrt{(n_a - n_{a\overline{b}})(n_{\overline{b}} - n_{a\overline{b}})} - \sqrt{n_{a\overline{b}}(n_b - n_a + n_{a\overline{b}})}}{\sqrt{(n_a - n_{a\overline{b}})(n_{\overline{b}} - n_{a\overline{b}})} + \sqrt{n_{a\overline{b}}(n_b - n_a + n_{a\overline{b}})}}$

Kernel Methods and Visualization for Interval Data Mining

Thanh-Nghi Do[1] and François Poulet[2]

[1] College of Information Technology,
Can Tho University,
1 Ly Tu Trong Street, Can Tho, VietNam
(e-mail: dtnghi@cit.ctu.edu.vn)
[2] ESIEA Pôle ECD,
BP 0339, 53003 Laval-France
(e-mail: poulet@esiea-ouest.fr)

Abstract. We propose to use kernel methods and visualization tool for mining interval data. When large datasets are aggregated into smaller data sizes we need more complex data tables e.g interval type instead of standard ones. Our investigation aims at extending kernel methods to interval data analysis and using graphical tools to explain the obtained results. The user deeply understands the models' behaviour towards data. The numerical test results are obtained on real and artificial datasets.

Keywords: Kernel methods, Support vector machines, Visualization, Interval data, Data mining, Visual data mining.

1 Introduction

In recent years, real-world databases have increased rapidly, so that the need to extract knowledge from very large databases is increasing. Data mining can be defined as the particular pattern recognition task in the knowledge discovery in databases process. It uses different algorithms for classification, regression, clustering or association. The SVM algorithms proposed by [Vapnik, 1995] are a well-known class of algorithms using the idea of kernel substitution. They have shown practical relevance for classification, regression and novelty detection tasks. The successful applications of SVM and other kernel-based methods [Cristianini and Shawe-Taylor, 2000], [Shawe-Taylor and Cristianini, 2004] have been reported for various fields.

While SVM and kernel-based methods are a powerful paradigm, they are not favourable to deal with the challenge of large datasets. The learning task is accomplished through the quadratic program possessing a global solution. Therefore, the computational cost of an kernel approach is at least square of the number of training data points and the memory requirement makes them intractable. We propose to scale up their training tasks based on the interval data concept [Bock and Diday, 1999]. We summarize the massive datasets into the interval data. We adapt the kernel algorithms to deal with this data. We construct a new RBF kernel of interval data used for classification, regression and novelty detection tasks. The numerical test results are obtained on real and artificial datasets.

Although SVM gives good results, the interpretation of these results is not so easy. The support vectors found by the algorithms provide limited information. Most of the time, the user only obtains information regarding support vectors and accuracy. He can not explain or understand why a model constructed by SVM makes a good prediction. Understanding the model obtained by the algorithm is as important as the accuracy because the user has a good comprehension of the knowledge discovered and more confidence in this knowledge. Our investigation aims at using visualization methods to try to explain the SVM results. We use interactive graphical decision tree algorithms and visualization techniques to give an insight into classification, regression and novelty detection tasks with SVM. We illustrate how to combine some strengths of different visualization methods to help the user to improve the comprehensibility of SVM results.

This paper is organized as follows. In section 2, we present a new Gaussian kernel construction to deal with interval data. In section 3, we briefly introduce classification, regression and novelty detection of interval data with SVM algorithms and other kernel-based methods. Section 4 presents a way to explain SVM results by using interactive decision tree algorithms. We propose to use an approach based on different visualization methods to try to interpret SVM results in section 5 before the conclusion and future works in section 6.

2 Non linear kernel function for interval data

Assume we have two data points x and $y \in R^n$. Here, we are interested in RBF kernel function because it is general and efficient. The RBF kernel formula in (1) of two data vectors x and y of continuous type is based on the Euclidean distance between these vectors, $d_E(x,y) = \parallel x - y \parallel$.

$$K\langle x,y\rangle = \exp\left(-\frac{\parallel x-y \parallel^2}{\gamma}\right) \tag{1}$$

For dealing with interval data, we only need to measure the distance between two vectors of interval type, then we substitute this distance measure for the Euclidean distance into RBF kernel formula. Thus the new RBF kernel can deal with interval data. The dissimilarity measure between two data vectors of interval type is the Hausdorff distance.

Suppose that we have two intervals represented by low and high values: $I_1 = [low_1, high_1]$ and $I_2 = [low_2, high_2]$, the Hausdorff distance between two intervals I_1 and I_2 is defined by (2):

$$d_H(I_1, I_2) = \max\left(|low_1 - low_2|, |high_1 - high_2|\right) \tag{2}$$

Let us consider two data vectors u, $v \in \Omega$ having n dimensions of interval type:
$u = ([u_{1,low}, u_{1,high}], [u_{2,low}, u_{2,high}], \ldots, [u_{n,low}, u_{n,high}])$
$v = ([v_{1,low}, v_{1,high}], [v_{2,low}, v_{2,high}], \ldots, [v_{n,low}, v_{n,high}])$
The Hausdorff distance between two vectors u and v is defined by (3):

$$d_H(u,v) = \sqrt{\sum_{i=1}^{n} \max\left(|u_{i,low} - v_{i,low}|^2, |u_{i,high} - v_{i,high}|^2\right)} \tag{3}$$

By substituting the Hausdorff distance measure d_H into RBF kernel formula, we obtain a new RBF kernel for dealing with interval data. This modification tremendously changes kernel algorithms for mining interval data. No algorithmic changes are required from the habitual case of continuous data other than the modification of the RBF kernel evaluation. All the benefits of the original kernel methods are kept. The kernel-based learning algorithms like Support Vector Machines (SVM [Vapnik, 1995]), Kernel Fisher's Discriminant Analysis (KFDA [Mika *et al.*, 1999]), Kernel Principal Component Analysis (KPCA [Schölkopf *et al.*, 1998]), Kernel Partial Least Squares (KPLS [Rosipal and Trejo, 2001]) can use the RBF function to build interval data models in classification, regression and novelty detection.

3 Interval data analysis with kernel methods

3.1 Support vector machines

$$\begin{gathered} \min\,(1/2)\sum_{i=1}^{m}\sum_{j=1}^{m} y_i y_j \alpha_i \alpha_j K\langle x_i, x_j\rangle - \sum_{i=1}^{m}\alpha_i \\ s.t. \sum_{i=1}^{m} y_i\alpha_i = 0 \\ C \geq \alpha_i \geq 0 \end{gathered} \tag{4}$$

where C is a positive constant used to tune the margin and the error.

Let us consider a binary linear classification task with m data points in a n-dimensional input $x_1, x_2, \ldots, x_m$ having corresponding labels $y_i = \pm 1$. SVM classification algorithm aims to find the best separating surface as being furthest from both classes. It is simultaneously to maximize the margin between the support planes for each class and minimize the error. This can be accomplished through the quadratic program (4).

From the α_i obtained by the solution of (4), we can recover the separating surface and the scalar b determined by the support vectors (for which $\alpha_i > 0$). By changing the kernel function K as a linear inner product, a polynomial, a radial basis function or a sigmoid neural network, we can get different classification model. The classification of a new data point x is based on:

$$sign(\sum_{i=1}^{\sharp SV} y_i \alpha_i K\langle x, x_i\rangle - b)$$

For one-class (novelty detection), the SVM algorithm is to find a hypersphere with a minimal radius R and center c which contains most of the data. And then novel test points lie outside the boundary of the hypersphere.

SVM can also be applied to regression problem by the introduction of an alternative loss function. By using an ϵ-insensitive loss function proposed by Vapnik, Support vector regression (SVR) aims to find a predictive function *f(x)* that has at most ϵ deviation from the actual value y_i.

These tasks can be also accomplished through the quadratic program. [Bennett and Campbell, 2000] and [Cristianini and Shawe-Taylor, 2000] provide more details about SVM and others kernel-based learning methods.

We have added a new construction kernel code to the publicly available toolkit, LibSVM (ref. http://www.csie.ntu.edu.tw/∼cjlin/libsvm). Thus, the software program is able to deal with interval data in classification, regression and novelty detection tasks. To apply the SVM algorithms to the multi-class classification problem (more than 2 classes), LibSVM uses one-against-one strategy. Assume that we have k classes, LibSVM construct *k*(k-1)/2* models. A model separates i^{th} class against j^{th} class. Then to predict the class for a new data point, LibSVM just predicts with each model and finds out which one separates the furthest into the positive region. We have used datasets from Statlog, the UCI Machine Learning Repository (ref. http://www.ics.uci.edu/∼mlearn/MLRepository.html), Regression Datasets (ref. http://www.liacc.up.pt/∼ltorgo/Regression/DataSets.html) and Delve (ref. http://www.cs.toronto.edu/∼delve). By using K-means algorithm [MacQueen, 1967], the large datasets are aggregated into smaller ones. A data point in interval datasets corresponds to a cluster, the low and high values of an interval are computed by the cluster data points. Some other methods for creating interval data can be found in [Bock and Diday, 1999]. The interval version of datasets is shown in table 1 and 2. We report the cross validation accuracy on classification results and mean squared error on regression results presented in table 1.

The results on novelty detection task are presented in table 2 with the number of outliers and significant outliers (furthest from other data points in the dataset). To the best of our knowledge, there is no other available algorithm being able to deal with interval data in both non linear classification, regression and novelty detection tasks. There is not experimental results on interval data mining provided by the others algorithms. Therefore, we only report results obtained by our approach. It is difficult to compare with the others ones.

Datasets	Points	Dims	Protocol	Accuracy	Mean squared error
Wave(3 classes)	30	21	leave-1-out	80.00%	0.462389
Iris(3 classes)	30	4	leave-1-out	100.00%	0.078389
Wine(3 classes)	36	13	leave-1-out	97.22%	0.075182
Pima(2 classes)	77	8	leave-1-out	79.22%	0.212736
Segment(7 classes)	319	19	10-fold	91.22%	1.696050
Shuttle(7 classes)	594	9	10-fold	94.78%	1.096640

Table 1. SVM classification and regression results

Datasets	Points	Dims	Nb. ourliers	Significant outliers
Shuttle	594	9	31	9
Bank8FM	450	8	12	6

Table 2. One-class SVM results

3.2 Other kernel-based methods

Many multivariate statics algorithms based on generalized eigenproblems can be also kernelized [Shawe-Taylor and Cristianini, 2004], e.g Kernel Fisher's Discriminant Analysis (KFDA), Kernel Principal Component Analysis (KPCA), Kernel Partial Least Squares (KPLS), etc. These kernel-based methods can also use the RBF function to build interval data models. We use KPCA and KFDA to visualize datasets in the embedding space where the user can intuitively see the separating boundary between the classes based on the human pattern recognition capabilities. The eigenvectors of the data can be used to detect directions of maximum variance, and thus, linear PCA is to project data onto principal components by solving a eigenproblem. By

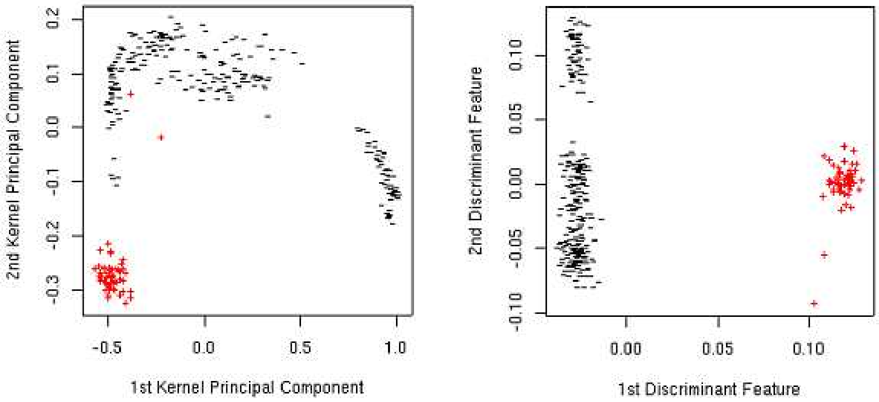


Fig. 1. Visualization of Kernel Principal Component Analysis (left) and Kernel Fisher's Discriminant Analysis (right) on the Segment dataset.

using a kernel function instead of the linear inner product in the formula, we obtain non linear PCA (KPCA). An example of the visualization of the Segment interval dataset (the class 7 against all) with KPCA using the RBF kernel function is shown in figure 1 (left).

In linear FDA, we consider projecting all the multi-dimensional data onto a generic direction w, and then separately observing the mean and the variance of the projections of the two classes. By substituting the kernel function for a linear inner product into the linear FDA formula, we have non linear FDA (KFDA). An example of the visualization of the Segment interval dataset (the class 7 against all) with KFDA using the RBF kernel function is shown in figure 1 (right).

And thus, the separating boundary between two classes is clearly represented in the embedding space.

4 Inductive rules extraction for explaining SVM results

Although SVM algorithms have shown to build accurate models, their results are very difficult to understand. Most of the time, the user only obtains information regarding the support vectors being used as "black box" to classify the data with a good accuracy. The user does not know how SVM models can work. For many data mining applications, understanding the model obtained by the algorithm is as important as the accuracy.
We propose here to use interactive decision tree algorithms [Poulet, 2003] to try to explain the SVM results. The SVM performance in classification task is deeply understood in the way of IF-THEN rules extracted intuitively from the graphical representation of the decision trees that can be easily interpreted by humans.

Figure 2 is an example of the inductive rule extraction explaining support vector classification results on the Segment interval dataset. The SVM algorithm using the RBF kernel function classifies the class 7 (considered as +1 class) against all other classes (considered as -1 class) with 100.00 % accuracy. CIAD uses 2D scatter plot matrices [Carr *et al.*, 1987] for visualizing interval data [Poulet, 2003]: the data points are displayed in all possible pair-wise combinations of dimensions in 2D scatter plot matrices. For n-dimensional data, this method visualizes *n(n-1)/2* matrices. A data point in two interval dimensions is represented by a two dimensions primitive cross and color corresponds to the class. The user interactively chooses the best separating split (parallel to an axis) to interactively construct the decision tree (based on the human pattern recognition capabilities) or with the help of automatic algorithms. The obtained decision tree having 4 leaves (corresponding to 4 rules) can explain the SVM model. One rule is created for each path from the root to a leaf, each dimension value along a path forms a conjunction and the leaf node holds the class prediction. And thus, the non linear SVM is

Input: non label dataset *SP* et a SVM classification function *f*
Output: inductive rule set *IND-RULE* explaining the SVM model

1. Classify non label dataset *SP* using SVM classification function *f*, we obtain label set *L* assigned to *SP*:

$$\{SP,\ L\} = \text{SVM\_classify}(SP,\ f)$$

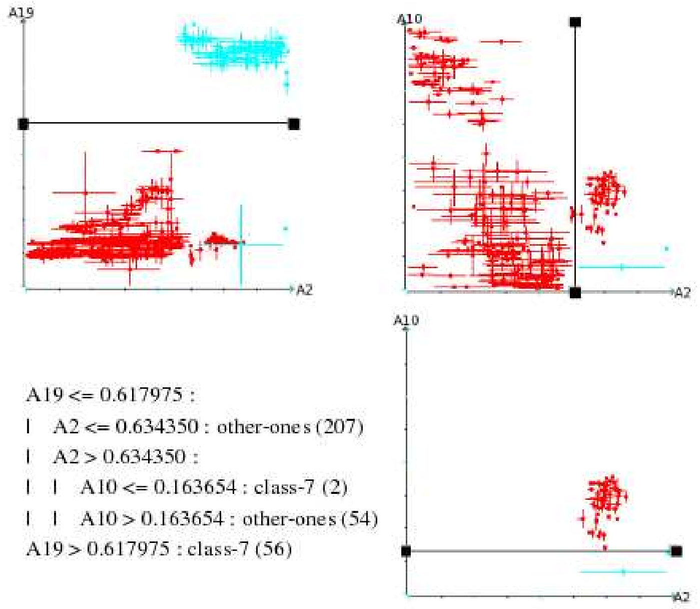
2. Interactively constructing decision tree model *DT* on dataset $\{SP,\ L\}$ using visual data mining decision tree algorithms, e.g CIAD [Poulet, 2003].

3. User extracts inductive rules *IND-RULE* from graphical representation of decision tree model *DT*:

$$\mathit{IND\text{-}RULE} = \text{HumanExtract(graphical } DT)$$

Table 3. Inductive rules extraction from SVM models

interpreted in the way of the 4 inductive rules (IF-THEN) that will be easy to understand.



```
A19 <= 0.617975 :
|   A2 <= 0.634350 : other-ones (207)
|   A2 > 0.634350 :
|   |   A10 <= 0.163654 : class-7 (2)
|   |   A10 > 0.163654 : other-ones (54)
A19 > 0.617975 : class-7 (56)
```

Fig. 2. Visualization of the decision tree explaining the SVM result on the Segment dataset.

5 Visualization tool for explaining SVM results

We have studied some ways to try to explain SVM results by using the graphical representation of high dimensional data. The information visualization methods guide the user towards the most appropriate visualizations for viewing mining results (post-processing step). There are many possibilities to visualize data by using different visualization methods, but all of them have some strengths and some weaknesses. We use the linking technique to combine different visualization methods to overcome the single one. The same information is displayed in different views with different visualization techniques providing useful information to the user. The interactive brushing technique allows the user to focus on a region (brush) in the data displayed to highlight groups of data points. And thus, the linked multiple views provide more information than the single one. We use the interactive brushing and linking techniques and the different visualization methods to try to explain SVM results. For classification tasks with SVM algorithms, understanding the margin (furthest distance between +1 class and -1 class) is one of the most important key of the support vector classification. For this, it is necessary to see the points near the separating boundary between the two classes.

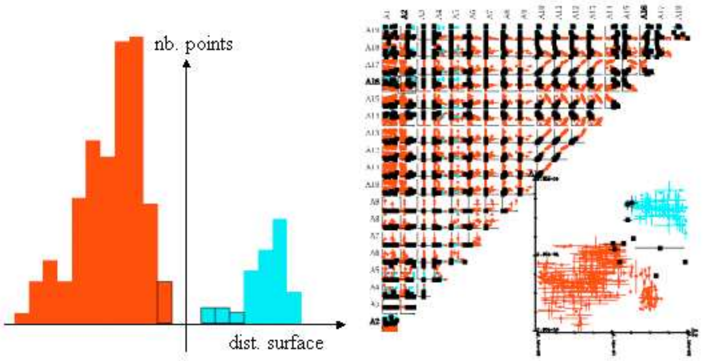


Fig. 3. Visualization of the classification result on the Segment dataset.

For achieving this goal, we propose to use the data distribution according to the distance to the separating surface. While the classification task is processed (based on the support vectors), we also compute the data distribution according to the distance to the separating surface. For each class, the positive distribution is the set of correctly classified data points and the negative distribution is the set of misclassified data points. The data points being near the frontier correspond to the bar charts near the origin. When the bar charts corresponding to the points near the frontier are selected, the data points are also selected in the other views (visualization methods) by using the brushing and linking technique. We use 2D scatter plot matrices for visualizing interval data. The user can see approximately the boundary between classes and the margin width. This helps the user to evaluate the robustness of the model obtained by support vector classification. He can also know the interesting dimensions (corresponding to the projections providing a clear boundary between the two classes) in the obtained model. Figure 3 is an example of visualizing support vector classification results on the Segment interval dataset (the class 7 against all). From data distribution according to the distance to the separating surface, the 4 bar charts near the origin are brushed, and then the corresponding points are linked and displayed in 2D scatter plot matrices. The dimensions 2 and 16 corresponding to the projection provides a clear boundary between the two classes and are interesting in the model obtained.

We have extended this idea for visualizing support vector regression results. We have also computed the data distribution according to the distance to the regression function. After that, we combine the histogram with 2D scatter plot matrices for visualization. When the user selects the data points far from the regression function, he can know how the function fits data. If the function well predicts the data points of high density region then the model obtained is interesting.

For a novelty detection task, we visualize the outliers allowing the user to valid them. The approach is based on the interactive linking and brushing of the histogram and 2D scatter plot views. The histogram displays the data distribution according to the distance to the hypersphere obtained by one class SVM. The data points far from the hypersphere are brushed in the histogram view, thus they are automatically selected in 2D scatter plot view. The user can validate the outliers. And then, the dimensions corresponding to the projection presents clearly the outliers and are interesting in the obtained model.

6 Conclusion

We have presented in this paper the interval data mining approach using kernel-based and visualization methods.

We have proposed to construct a new RBF kernel on interval data. This modification tremendously changes kernel-based algorithms. No algorithmic changes are required from the usual case of continuous data other than the modification of the RBF kernel evaluation. Thus, kernel-based algorithms can deal with interval data in classification, regression and novelty detection. It is extremely rare algorithms being able to construct non linear models on interval data for the three problems: classification, regression and novelty detection.

We have also proposed two ways to try to explain SVM results that are a well-known "black box". The first one is to use interactive decision tree algorithms for explaining the SVM results. The user can interpret the SVM performance in the way of IF-THEN rules extracted intuitively from the graphical representation of the decision trees that can be easily interpreted by the user. The second one is based on a set of different visualization techniques combined with linking and brushing techniques gives an insight into classification, regression and novelty detection tasks with SVM. The graphical representation shows the interesting dimensions in the obtained model.

A forthcoming improvement will be to extend our approach to data of taxonomic or mixture types.

References

[Bennett and Campbell, 2000]K. Bennett and C. Campbell. Support vector machines: Hype or hallelujah ?. *SIGKDD Explorations*, pages 1–13, 2000.

[Bock and Diday, 1999]H-H. Bock and E. Diday. *Analysis of Symbolic Data.* Springer-Verlag, 1999.

[Carr *et al.*, 1987]D-B. Carr, R-J. Littlefield, W-L. Nicholson, and J-S. Littlefield. Scatterplot matrix techniques for large n. *Journal of the American Statistical Association*, pages 424–436, 1987.

[Cristianini and Shawe-Taylor, 2000]N. Cristianini and J. Shawe-Taylor. *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods.* Cambridge University Press, 2000.

[MacQueen, 1967]J. MacQueen. Some methods for classification and analysis of multivariate observations. *Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability*, pages 281–297, 1967.

[Mika *et al.*, 1999]S. Mika, G. Rätsch, J. Weston, B. Schölkopf, and K-R. Müller. Fisher discriminant analysis with kernels. *Neural Networks for Signal Processing IX*, pages 41–48, 1999.

[Poulet, 2003]F. Poulet. Interactive decision tree construction for interval and taxonomical data. In *Proceedings of VDM@ICDM'03, 3nd Workshop on Visual Data Mining*, pages 183–194, 2003.

[Rosipal and Trejo, 2001]R. Rosipal and L-J. Trejo. Kernel partial least squares regression in reproducing kernel hilbert space. *Journal of Machine Learning Research*, pages 97–123, 2001.

[Schölkopf *et al.*, 1998]B. Schölkopf, A. Smola, and K-R. Müller. Nonlinear component analysis as a kernel eigenvalue problem. *Neural Computation*, pages 1299–1319, 1998.

[Shawe-Taylor and Cristianini, 2004]J. Shawe-Taylor and N. Cristianini. *Kernel Methods for Pattern Analysis.* Cambridge University Press, 2004.

[Vapnik, 1995]V. Vapnik. *The Nature of Statistical Learning Theory.* Springer-Verlag, 1995.

Contingency table with a double partition on rows and columns. Visualization and comparison of the partial and global structures

Mónica Bécue[1], Jerôme Pagès[2], and Campo-Elías Pardo[3]

[1] EIO. Universitat Politècnica de Catalunya
08028 Barcelona - Spain
(e-mail: monica.becue@upc.edu)

[2] Agrocampus Rennes
65 rue de Saint-Brieuc, CS 84215
F-35042 Rennes cedex, France
(e-mail: jerome.pages@agrocampus-rennes.fr)

[3] Departamento de Estadística. Universidad Nacional de Colombia
Bogotá, Colombia
(e-mail: cepardot@unal.edu.co)

Abstract. Internal correspondence analysis (ICA) deals with frequency tables having a double partition structure on the columns and rows, offering their representation on principal axes which reflects the inner structure of the subtables as defined by the two partitions. We enrich this global representation by the superimposed representation of the rows (respectively, the columns) as described separately by every group of columns (respectively, by every group of rows). The new aids to interpretation that we propose, give information about the common and specific structures in the subtables.

Keywords: Correspondence analysis, Internal correspondence analysis, Multiple factor analysis, Common dispersion directions, Multicanonical analysis.

1 Introduction

Some applications lead to build up contingency tables having a double partition on the columns and rows. This characteristic induces objectives such as giving an account of the global structure of the table that takes into account its specificity as well as comparing all the partial row structures (respectively, partial column structures) induced on the rows by each group of columns (respectively, each group of rows). Concerning the first objective, internal correspondence analysis (ICA) [Cazes *et al.*, 1988] offers an interesting approach. However, ICA does not provide any result relative to the partial structures. In this work, in the framework of ICA, we propose several tools in order to compare them, favoring their simultaneous visualization.

§2 introduces the notation. After reminding the basic principles of ICA (§3), we propose a methodology to compare the global and partial points of wiew in §4. The §5 shows the interest of these tools as applied to an example. Finally, we conclude with some remarks.

2 Notation

We consider a table **F** of proportions (the overall sum amounts to 1) divided into $L \times J$ sub-tables (Fig. 1). The I rows are partitioned in L groups with, respectively, $I_1, I_2, \ldots, I_L$ rows. The K columns are structured in J groups with, respectively, $K_1, K_2, \ldots, K_J$ columns. The subtable (l, j) has I_l rows and K_j columns.

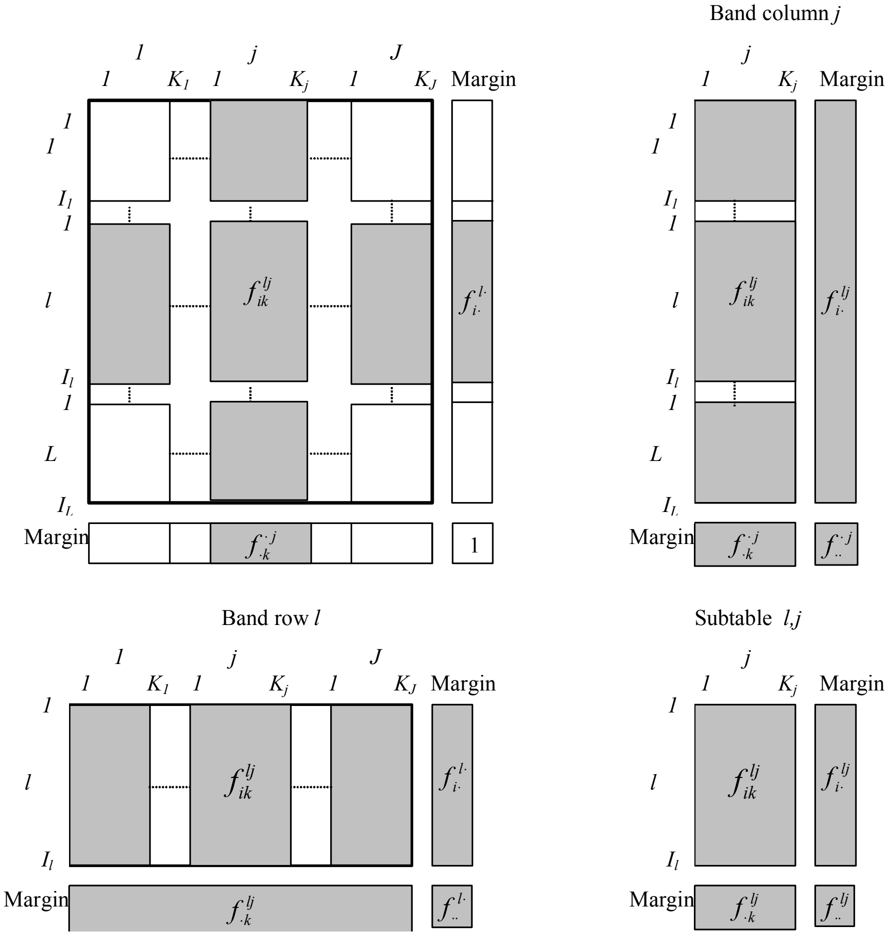


Fig. 1. The global table **F** of proportions, as partitioned into rows and columns groups

3 Internal correspondence analysis (ICA)

3.1 Correspondence analysis with respect to a model

The classical CA refers to the independence model, as given by the products of the margins. Other models can be considered. The CA of **F** with respect to any model **A** having the same margins than **F** [Escofier, 1984], is equivalent to PCA(**X**,**M**,**D**), **X** being the matrix with the general term $x_{ik}^{lj} = \frac{f_{ik}^{lj} - a_{ik}^{lj}}{f_{i\cdot}^{l\cdot} f_{\cdot k}^{\cdot j}}$, using the metric $\mathbf{M} = diag(f_{\cdot k}^{\cdot j})$ and the weights $\mathbf{D} = diag(f_{i\cdot}^{l\cdot})$ in the row space (respectively, the metric $\mathbf{D} = diag(f_{i\cdot}^{l\cdot})$ and the weights $\mathbf{M} = diag(f_{\cdot k}^{\cdot j})$ in the column space).

3.2 Within and double within-tables correspondence analysis

A particular case arises when analyzing the row-wise juxtaposition of J tables (respectively, the column-wise juxtaposition of L tables). The within-tables CA [Benzécri, 1983], [Escofier and Drouet, 1983], [Escofier and Pagès, 1998, p.229], takes as model the within-table independence in order to globally study the deviations of every columns subcloud to their own centroid. For example, if we only take into account the partition on the columns of **F** and consider this table as the row-wise juxtaposition of J tables, the general term of the within-table independence model is $a_{ik}^{lj} = \frac{f_{i\cdot}^{lj} f_{\cdot k}^{\cdot j}}{f_{\cdot\cdot}^{\cdot j}}$. This model has the same margins as table **F**.

In order to take into account both partitions on the rows and columns, [Cazes *et al.*, 1988] propose the internal correspondence analysis (ICA) that considers the model whose general term is given in (1):

$$a_{ik}^{lj} = \frac{f_{\cdot k}^{lj} f_{i\cdot}^{l\cdot}}{f_{\cdot\cdot}^{l\cdot}} + \frac{f_{i\cdot}^{lj} f_{\cdot k}^{\cdot j}}{f_{\cdot\cdot}^{\cdot j}} - \frac{f_{i\cdot}^{l\cdot} f_{\cdot k}^{\cdot j}}{f_{\cdot\cdot}^{l\cdot} f_{\cdot\cdot}^{\cdot j} / f_{\cdot\cdot}^{lj}} \tag{1}$$

This model has the same margins as table **F**. ICA can be seen as a double within correspondence analysis. The matrix **X** analysed in the PCA(**X**,**M**,**D**) corresponding to ICA inherits the double partition structure of **F**.

4 Comparison of the partial and global structures

ICA offers a representation of the global structure, of the rows and columns, on principal planes in a CA-like way. This global representation can be enriched by looking for representing the rows (resp., the columns) as described separately by every group of columns (resp., every group of rows). For that goal, we adopt a MFA-like point of view [Escofier and Pagès, 1994] by looking for a simultaneous visualization of the partial structures (of rows or of columns) on the principal planes corresponding to the global analysis. We enrich this simultaneous representation with a series of aids to interpretation.

4.1 Superimposed representation of the partial and global rows on a common referential

To each column-band matrix j of $\mathbf{X}$ (as defined in ICA applied to $\mathbf{F}$), we associate the cloud N_I^j of the rows as described by only the columns of this matrix. As this cloud lies in the subspace $\mathbf{R}^{Kj}$ of $\mathbf{R}^K$, we assimilate it to the cloud of the rows of the matrix $\tilde{\mathbf{X}}_\mathbf{j}$, having the same dimension as $\mathbf{X}$ and derived from $\mathbf{X}_\mathbf{j}$ in the following way:

$$\tilde{\mathbf{X}}_j = \begin{array}{|c|c|c|c|} \hline \mathbf{0} & \mathbf{0} & \mathbf{X}_j & \mathbf{0} \\ \hline \end{array}$$

The coordinates of the partial rows, belonging to the cloud N_I^j, on the s-axis issued from the global analysis, are $\tilde{\mathbf{F}}_s^j = \tilde{\mathbf{X}}_j\mathbf{M}\mathbf{u}_s$. To every row (l,i), we associate the $N_{(l,i)}^J$ cloud of its J partial points. In order to obtain a superimposed representation in such a way that the global point corresponds to the centroid of the subcloud $N_{(l,i)}^J$, the coordinates $\tilde{F}_s^j(l,i)$ are amplified by J and then projected on the global representation.

4.2 Aids to interpretation of superimposed representation of the partial and global rows

Quality of representation of the partial clouds: the quality of representation of every cloud N_I^j on the s-axis is measured, in a classical way, through the ratio between the projected inertia and the total inertia.

Measure of the similarity between the partial clouds: the union of the whole of the $N_{(l,i)}^J$ clouds (i.e. the cloud of all the partial row-points noted N_I^J) contains $I \times J$ partial points. These $I \times J$ partial points can be divided into I subclouds, with J points $(l,i)^j$ in every subcloud, corresponding to the same row (l,i). So, the total inertia of N_I^J can be decomposed into within-inertia (inertia within the $N_{(l,i)}^J$ subclouds) and between-inertia (inertia between the $N_{(l,i)}^J$ subclouds). The ratio [between-inertia/total-inertia], calculated axis by axis, measures the proximity of the partial points corresponding to a same row and so, the global similarity between the J partial clouds as projected on this axis. If this ratio is close to 1, the homologous points $(l,i)^j$ are close to one another and the s-axis represents a structure common to the different groups of columns.

Selection of rows and of partial rows with a high contribution to the within-inertia: the within-inertia can be decomposed into the contributions of every row, in order to detect those whose behavior varies from the different points of view represented by the groups of columns. So, the more heterogeneous (respectively, more homogeneous) rows on every axis can be identified in order to interpret the global ratios.

4.3 Superimposed representation of the partial and global columns

The superimposed representation of the partial and global columns clouds and its interpretation aids are obtained in a symmetric way.

5 Application

To illustrate the superimposed representation and their interpretation aids, we utilize the example Ardèche [Cazes *et al.*, 1988], a faunal table crossing species (43 rows) and dates×sites (35 columns, corresponding to 35 dates×sites samplings). This data set is available in [Cazes *et al.*, 1988]. The 43 species are distributed in 4 taxonomic groups (*Ephemeroptera*, *Plecoptera*, *Coleoptera*, *Trichoptera*) which induce the partition on the rows (4 groups). 6 sites (A, B, C, D, E, F) are observed at 6 dates (*jul82*, *aug82*, *nov82*, *feb83*, *apr83*, *jul83*) chosen in different seasons, but the observation of the site F at date 1 is missing. We consider the partition on the columns induced by the different dates (6 groups).

Global representation through ICA

By recentering the subclouds corresponding to a same date, ICA solves the problem of eliminating the time-associated faunal structures and allows for interpreting the spatial typology and for assessing the ability of the taxonomic groups to be used as biological descriptors.

Figure 2 shows the dates×sites on the first principal plane issued from ICA. As [Cazes *et al.*, 1988] note, ICA puts to the fore the originality of the site B, mainly contrasting with A and D. Site D presents a very specific faunal composition in winter (*D-feb83* and *D-apr83*). Mainly F, but also E present outstanding differences between winter (at the left of the first axis: *F-feb83*, *E-feb83* , *F-apr83*, and *E-apr83*) and summer (at the right of the first axis: *F-aug82* and *F-jul83*). Finally, the rise of the water in November standardizes the faunal distribution and, therefore, the subcloud *Nov82*×sites is close to the centroid.

Concerning the species, the inertia on the first axis is mainly due to the great dispersion of the trichopterans: in this group, the species with sheath are attracted by the sites presenting sand or stones with vegetation, while the free trichopterans prefer hard substratum soil (see Fig. 3). *Coleoptera* dispersion strongly contributes to the inertia of the second axis, contrasting the species depending on their preference for strong current or not.

[Cazes *et al.*, 1988] conclude that there is a summer typology, mainly defined by *Coleoptera* and a winter typology, due to *Trichoptera* and corresponding to the originality of site D and the standardization of the fauna in November.

Comparison of the partial structures

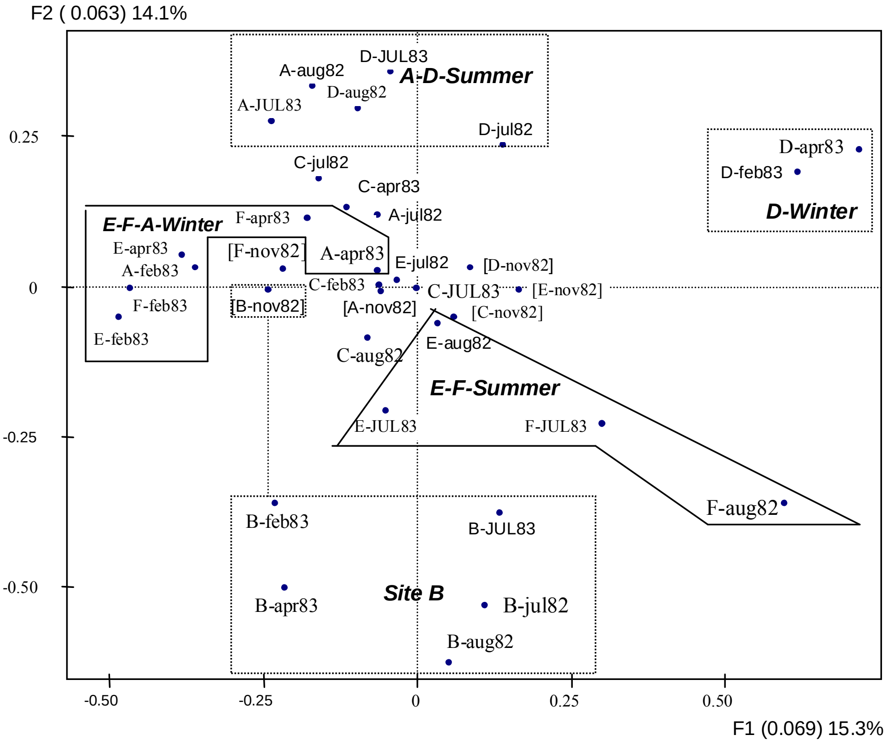


Fig. 2. The column-points on the first principal plane issued from ICA

However, the specific structure of the table, leads to other kinds of questions. For example, *D-feb83* and *D-apr83* lie in close positions from a global point of view (from the whole of the taxonomic groups), but are they also close from the point of view of every taxonomic group? In the same way, it is interesting to know, for example, if the species *Nemoura spp.* and *Eleuctra fusca*, very different from a global point of view (from the whole of the sites×dates) are alike at some date. The superimposed representations of the global and partial row-points (respectively of the global and partial column-points) will contribute to answer these questions.

5.1 Superimposed representation of the species

The global similarity between the six clouds of species, as induced by every date (partial row clouds) and as projected on the first and second axes, is measured by the ratio [between-inertia/total-inertia]. This ratio is equal to 31.9% and 38.8%, respectively. These relatively low values indicate that it exists a notable difference between the inter-species distances from one date to the other.

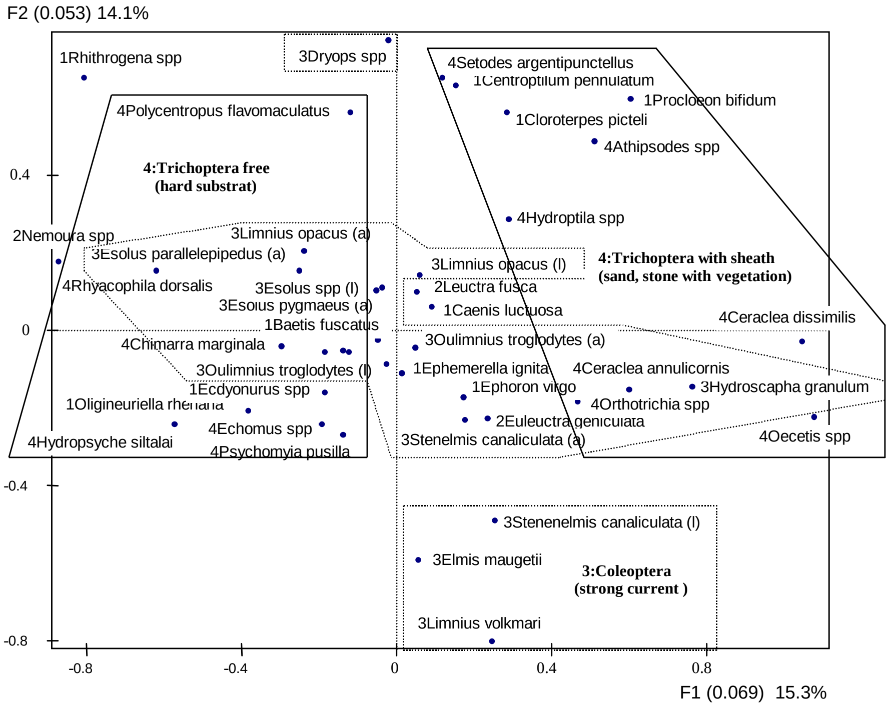


Fig. 3. The species on the first principal plane issued from ICA

In order to identify the species whose behavior varies more depending on the date, and to interpret these specific behaviors, we look for those which present the largest within-inertia on the first axes. Then, we search the partial point(s) responsible of the high dispersion of the concerned species, that are not in accordance to the other homologous partial points. For example, the species *Nemoura spp.* presents the second highest within-inertia on the first axis (equal to 10.4% of the total within-inertia on this axis, as summed up on all the rows). Furthermore, the partial point *Nemoura spp.-feb83* brings 75.9% of the within-inertia due to this species on the first axis. So, the position of this taxon (Fig. 4) suggests that it is a good indicator in February and in April, but not in summer: in fact, this species was almost never observed in summer (discarding one case). Moreover, discarding two cases, this taxon is the only plecopteran observed in February, which explains the more characteristic position of this partial point.

Regarding *Leuctra fusca* the most homogeneous of plecopteran (1.4% of the total within-inertia of the first axis), its partial points are globally close to the origin, except *feb83* (65.0% of within-inertia of this specie on first axis). In fact, this species was frequently observed, except in November and February. As not any other plecopteran was observed in November, *Leuctra fusca* is characteristic (by its absence) only in February. These examples

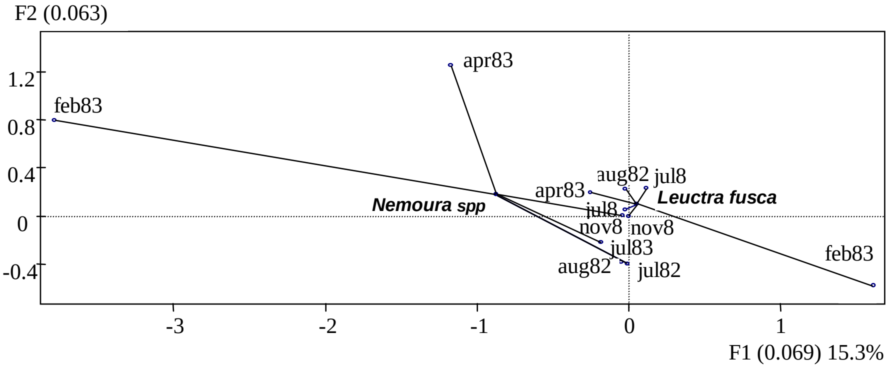


Fig. 4. Superimposed representation of two species belonging to *Plecoptera* on the first principal plane issued from ICA

5.2 Superimposed representation of the columns

The ratios [between-inertia/total-inertia] corresponding to the column clouds (the clouds of dates×sites as induced by each taxonomic group) as projected on the first and second axes are equal to 49.5% and 50.1%, respectively.

Figure 5 presents an excerpt of the superimposed representation, on the first principal plane, of the two dates×sites presenting the highest within-inertia on the first axis (see Table 1): *D-apr83* and *D-feb83* (i.e. the same site at different dates) and also of the dates×sites *C-jul82* and *D-jul82* to illustrate the case of different sites at the same date. Table 1 completes this representation with some information about dates×sites.

We can note that *D-apr83* and *D-feb83* are very similar as described globally (i.e. from all the taxonomic groups) but also as described by any taxonomic group: the partial points corresponding to the different taxonomic groups are close in every case. According to the graph, these two couple (date, site) are mainly characterized by *Trichoptera*. The two trichopterans having a high positive coordinate along axis 1 are *Caraclea dissimilis* and *Oecetis spp.* (Fig. 3). In fact, in February and April, these two taxa were observed quite only on the site *D*. From another point of view, the usual correspondence analysis applied only to the subtable (*Trichoptera*, *Feb83*) or the subtable (*Trichoptera*, *Apr83*) provides a first plane clearly showing the association between the site *D* and these two taxa. Concerning the sites *C* and *D* at July 82, we can see that, they have quite similar profiles in *Plecoptera* but different in *Coleoptera*.

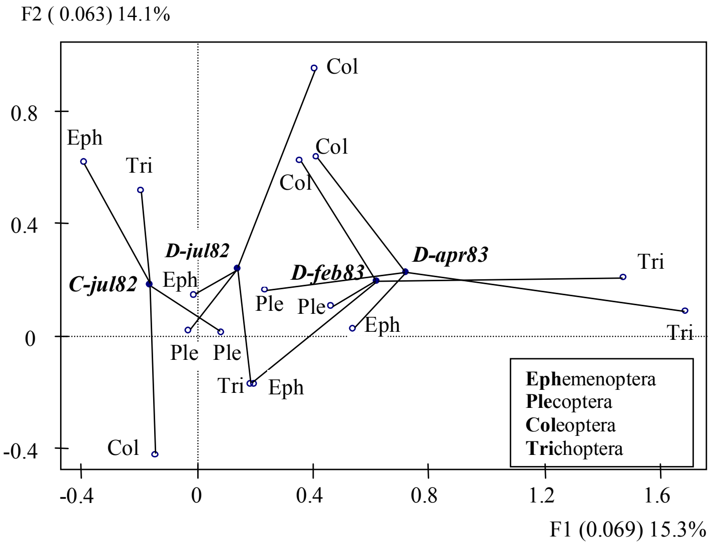


Fig. 5. Excerpt of the superimposed representation of the partial column points on the first principal plane issued from ICA

Table 1. Some interpretation aids of columns represented in Fig. 5

Column	Within Inertia x100000		Coordinates x1000		Contribution %		Weight %
	Axis-1	Axis-2	Axis-1	Axis-2	Axis-1	Axis-2	
D-apr83	309	55	720	227	28.6	30.8	3.8
D-feb83	268	88	619	190	23.7	24.2	4.2
C-jul82	19	117	-160	179	10.0	13.5	2.7
D-jul82	27	171	140	234	9.9	30.1	3.5

6 Conclusion

The comparison of the partial rows and columns structures enriches the results from ICA. This comparison induces a representation of the row-profiles (respectively, the column-profiles) not only from the global but also the partial points of view as induced by each group of columns (respectively, rows). In the case of the Ardèche example, the superimposed representation of the partial columns allows for better visualizing the taxonomic groups which are responsible of the differences observed among the sites according to the date.

Software note

The calculations are performed with ADE4 [Thioulouse *et al.*, 2004], in R environment [R Development Core Team, 2004].

References

[Benzécri, 1983]J.P. Benzécri. Analyse de l'inertie intraclasse par l'analyse d'un tableau de correspondances. *Les Cahiers de l'Analyse des Données*, 8(3):351–358, 1983.

[Cazes *et al.*, 1988]P. Cazes, D. Chessel, and S. Dolédec. L'analyse des correspondances internes d'un tableau partitionné. Son usage en hydrobiologie. *Revue de Statistique Appliquée*, 36(1):39–54, 1988. http://pbil.univ-lyon1.fr/R/articles/arti054.pdf.

[Escofier and Drouet, 1983]B. Escofier and D. Drouet. Analyse des différences entre plusieurs tableaux de fréquence. *Les Cahiers de l'Analyse des Données*, 8(4):491–499, 1983.

[Escofier and Pagès, 1994]B. Escofier and J. Pagès. Multiple factor analysis: afmult package. *Comput. Statist. Data Anal*, 18:121–140, 1994.

[Escofier and Pagès, 1998]B. Escofier and J. Pagès. *Analyses factorielles simples et multiples. Objectifs, méthodes et interprétation.* Dunod, Paris, 3 edition, 1998.

[Escofier, 1984]B. Escofier. Analyse factorielle en référence à un modèle. Application à l'analyse de tableaux d'échanges. *Revue de Statistique Appliquée*, 32(4):25–36, 1984.

[R Development Core Team, 2004]R Development Core Team. *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria, 2004. http://www.R-project.org.

[Thioulouse *et al.*, 2004]J. Thioulouse, A.B. Dufour, and D. Chessel. *ADE4: Analysis of Environmental Data : Exploratory and Euclidean method Multivariate data analysis and graphical display.* Lyon, France, November 2004. http://cran.univ-lyon1.fr/src/contrib/Descriptions/ade4.html.

Cost of Low-Quality Data over Association Rules Discovery

Laure Berti-Équille

IRISA
Campus Universitaire de Beaulieu
35042 Rennes Cedex, France
(e-mail: berti@irisa.fr)

Abstract. Quality in data mining critically depends on the preparation and on the quality of processed data sets. Indeed data mining processes and applications require various forms of data preparation (and repair) with several data formatting and cleaning techniques, because the data input to the mining algorithms is assumed to conform to nice data distributions, containing no missing, inconsistent or incorrect values. This leaves a large gap between the available dirty data and the available machinery to process and analyze the data for discovering knowledge. This paper presents a theoretical probabilistic framework for modeling the cost of low-quality data on discovered association rules.

Keywords: Data Quality, Quality of Discovered Association Rules, Minimal Cost Statistical Model.

1 Introduction

In an error-free database or datawarehouse system with perfectly clean data, knowledge discovery techniques (such as clustering, mining association rules or visualization) can be relevantly used from a decisional perspective to automatically derive new knowledge, new concepts, or knowledge patterns from numerical data. Unfortunately, most of the time, these data are neither rigorously chosen from different heterogeneous sources nor carefully controled for quality. Under the general acronym *ETL*, the Extraction-Transformation-Loading activities cover the most prominent tasks of data preparation before the warehousing and mining processes. They include [Vassiliadis *et al.*, 2003]: *i)* the identification of relevant information at the source side, *ii)* the extraction of this information, *iii)* the transformation and integration of the information coming from multiple sources into a common format and, *iv)* the cleaning and correction of the integrated data set. Data preparation and cleaning processes are complex, costly and critical despite the specialized ETL tools mainly dedicated to relational data available in the market [ETI, 2005], [MS, 2005], [DataMirror, 2005], [ArdentSoftware, 2005]. And the area raised lot of interest with research results [Dasu and Johnson, 2003], [Rahm and Do, 2000], [Winkler, 2003], [Vassiliadis *et al.*, 2003] and several academic tools (Telcordia [Caruso *et al.*, 2000], AJAX [Galhardas *et al.*, 2001], Potter's Wheel [Raman and Hellerstein, 2001], Arktos [Vassiliadis *et al.*, 2000], IntelliClean [Low *et al.*, 2001], Tailor [Elfeky *et al.*, 2002]).

In the presence of inconsistencies, errors or missing values in the data, it is nevertheless important to estimate the risk of discovering low-quality knowledge by mining low-quality data.

In this paper, our contribution is to present a probabilistic decision model that estimates the cost of discovering low-quality association rules by mining potentially polluted data.

The rest of the paper is organized as follows. Section 2 briefly provides some background information on association rules, data quality and other decision models mainly used in record linkage and data cleaning. Section 3 introduces our decision model and the notation that is used throughout this paper. Section 4 provides concluding remarks and guidelines for future extensions of this work.

2 Background

Among traditional descriptive data mining techniques, association rules discovery identifies intra-transaction patterns in a database and describes how much the presence of a set of attributes in a database's record (or transaction) implicates the presence of other distinct set of attributes in the same record (resp. transaction). The quality of association rules is commonly evaluated by looking at their support and confidence. The support of a rule measures the occurence frequency of the pattern in the rule while the confidence is the measure of the strength of implication. Association rule mining is commonly stated as follows: let $I = \{i_1, \ldots, i_n\}$ be a set of *items* and T be a set of data cases. Each data case consists of a subset of items in I. An association rule is an implication of the form $LHS \longrightarrow RHS$, where $LHS \subset I$, $RHS \subset I$, and $LHS \cap RHS = \emptyset$.

The support s of the rule $LHS \longrightarrow RHS$ is measured by the fraction of transactions that contain both LHS and RHS. More formally,

$$s = \frac{\text{number of transactions containing } LHS \cup RHS}{\text{number of transactions}} \tag{1}$$

The confidence c of the rule $LHS \longrightarrow RHS$ states that $c\%$ of transactions that contain LHS also contain RHS and it's the conditional probability of seeing RHS, given that we have seen LHS. More formally,

$$c = \frac{\text{number of transactions containing } LHS \cup RHS}{\text{number of transactions containing } LHS} \tag{2}$$

The problem of mining association rules is to generate all association rules that have support and confidence greater than the user-specified minimum support and confidence thresholds. Besides support and confidence, many other measures for knowledge evaluation have been proposed in the literature with the purpose of supplying subsidies to the user in the understanding and use of the acquired knowledge [Tan *et al.*, 2002], [Lavrac *et al.*, 1999]. Rules may have intrinsic properties (noise tolerance, asymetry, dataset size or dimensionality sensitivity, etc.) or collective properties when considering a set of rules (redundancy, transitivity, consistency, etc.). But, the main drawback of the objective and subjective interestingness measures is to neglect the initial quality of processed data. Data quality is a multidimensional, complex and morphing concept [Dasu and Johnson, 2003]. Table 1 presents some of the dimensions of data quality among more than 200 dimensions that have been proposed in the literature [Wang *et al.*, 1995], [Huang *et al.*, 1999], [Redman, 1996].

Dimension	**Definition**
Availability	Time the data is accessible based on technical equipment and statistics
Freshness	How up-to-date the information is
Accessibility	Estimation of waiting time for information retrieval processing
Security	Estimation of the number of corrupted data
Coverage	Estimation of the number of data for a specific information domain
Accuracy	Estimation of the number of data free-of-error
Completeness	Estimation of the number of missing data or null values
Credibility	User grade based on the reputation of data sources

Table 1. Some Data Quality Dimensions proposed by [Naumann, 2002]

As an illustrative example, one might legitimately wonder whether a so-called “interesting” rule $LHS \longrightarrow RHS$ is meaningful when 30% of the data describing the items of LHS are not up-to-date, 17% of RHS's data are not accurate, 14% of LHS's data come from sources that have bad credibility. In this paper, we consider that identifying interesting rules should also take into account the quality of underlying data used by the rule mining process: despite high interestingness measures, there are interesting rules discovered from dirty data, others from clean data, but they don't have the same added-value. This can be seen as a classification problem where the goal is to correctly assign cases (measurements, observations, etc.) to one of a finite number of classes. Most of the currently available algorithms for classification are designed to minimize error rate, *i.e.*, the number of incorrect predictions made. This implicity assumes that all errors are equally costly. In our context, there are many different types of cost involved on the selection of discovered rules. For instance, discovering interesting rules from inaccurate data may not have the same cost (or impact) than discovering rules from out-of-date data. In this study, we consider only the cost of misclassification error which is related to assigning different weights to different misclassification errors. Misclassification costs may be generally described by an arbitrary cost matrix C, with elements of the form c_{ij}, meaning the cost of predicting that an example belongs to the class i when in fact it belongs to j. The Bayesian decision approach is based on the assumption that the decision problem is posed in probabilistic terms, and that all the relevant probability values are known. In this paper, we propose a constant error cost Bayesian model which means that the cost of a certain type of error may be constant. In some cases, we are uncertain about the actual costs. To account for this uncertainty, we can use a probability distribution over a range of possible costs. To keep the presentation simple, we do not consider probability distributions over costs in this study. Our work is correlated to several works in data cleaning and Table 2 presents several decision models proposed in the literature mainly for record linkage. Our model is similar to the one proposed by Verykios *et al.* [Verykios *et al.*, 2003] as it minimizes the cost of making a decision rather than the probability of error in a decision of record matching. Our contribution is to adapt this model for association rule mining and for minimizing the cost of the rule selection in presence of low-quality data and of a misclassification region that can occur when erroneous data can be classified correct because they're in the range of correct values and correct data can be classified erroneous because they're in the range of erroneous values or outliers.

Model *(Tool)*	**Authors**	**Type of Model**
Error-based Model	[Fellegi and Sunter, 1969]	Probabilistic
EM-based Method	[Dempster *et al.*, 1977]	Probabilistic
Bayesian Cost-based Model	[Verykios *et al.*, 2003]	Probabilistic
Induction	[Bilenko and Mooney, 2003]	Probabilistic
Clustering for Record Linkage *(Tailor)*	[Elfeky *et al.*, 2002]	Probabilistic
1-1 matching	[Winkler, 2004]	Probabilistic
Bridging File	[Winkler, 2003]	Probabilistic
sorted-NN method	[Hernandez and Stolfo, 1995]	Empirical
XML Object Matching	[Weis and Naumann, 2004]	Empirical
Hierarchical Structure *(Delphi)*	[Ananthakrishna *et al.*, 2002]	Empirical
Matching Prediction based on clues	[Buechi *et al.*, 2003]	Knowledge-based
Functional Dependencies Inference	[Lim *et al.*, 1993]	Knowledge-based
Transformation functions *(Active Atlas)*	[Tejadaa *et al.*, 2001]	Knowledge-based
Rules and sorted-NN *(Intelliclean)*	[Low *et al.*, 2001]	Knowledge-based

Table 2. Decision Models for Record Linkage and Duplicate Identification

3 Cost-based Probabilistic Model

Let j ($j = 1, 2, \ldots, k$) be the dimensions of data quality (e.g., data freshness, credibility, accuracy, completeness, etc.). Let $x_{ij} \in [min_{ij}, max_{ij}]$ be a scoring value for the quality dimension j. The vector, that keeps the values of all quality dimensions for each data item (normalized in $[0, 1]$, is called *quality vector* q. The set of all possible vectors, is called *quality space* Q. Despite good confidence, support or other interestingness measures, selecting an association rule is a decision that designates the rule as *legitimately interesting* (noted D_1), *potentially interesting* (D_2), or *not interesting* (D_3) based on the information contained in the quality vectors of the data item sets composing the LHS and RHS parts of the rule.

3.1 Definition and Notations

Consider the item $x \in LHS \cup RHS$ of a given rule, we use $P_{CE}(x)$ to denote the probability that the item x will be classified as "erroneous" (or "polluted") *wrt* to one or more quality dimensions relevant to the application, and $P_{CC}(x)$ denotes the probability that the item x will be classified as "correct" (*i.e.*, in the range of acceptable values for each pre-selected quality dimensions). Also, $P_{AE}(x)$ represents the probability that the item x is actually erroneous (AE), and $P_{AC}(x)$ represents the probability that it is actually correct (AC). Intuitively, the item x can be an attribute whose quality dimensions are measured and aggregated from all the existing values of the attribute domain.

For an arbitrary average quality vector $\bar{q} \in Q$ on all data items in $LHS \cup RHS$ of the rule, we denote by $P(\bar{q} \in Q|CC)$ or $f_{CC}(\bar{q})$ the conditional probability of the pattern $\bar{q}$ that corresponds to the average of quality vectors of the items that are classified as correct (CC). Similarly, we denote by $P(\bar{q} \in Q|CE)$ or $f_{CE}(\bar{q})$ the conditional probability of the pattern $\bar{q}$ corresponds to the average of quality vectors of the items that are classified erroneous (CE). We denote by d the decision of the predicted class of the rule (*i.e.*, *legitimately interesting* D_1, *potentially interesting* D_2, or *not interesting* D_3), and by s the actual status of quality of the item sets upon which the rule has been computed. Let us also denote by $P(d = D_i, s = j)$ and $P(d = D_i|s = j)$ correspondingly, the joint and the conditional probability that the decision D_i is taken, when the actual status of data quality (*CC, CE, AE, AC*) is j. We also denote by c_{ij} the cost of making a decision D_i for classifying a rule with actual data quality status j of the items sets composing the parts of the rule.

3.2 Cost-based Bayesian Decision Model

Based on the example in Table 3 where we can see how the cost of different decisions could affect the result of selection among interesting rules, we need to minimize the mean cost $\bar{c}$ that results from making such a decision.

Cost	**Decision for Rule Selection**	**Actual Data Quality Status**
c_{10}	D_1	CC
c_{11}	D_1	CE
c_{12}	D_1	AE
c_{13}	D_1	AC
c_{20}	D_2	CC
c_{21}	D_2	CE
c_{22}	D_2	AE
c_{23}	D_2	AC
c_{30}	D_3	CC
c_{31}	D_3	CE
c_{32}	D_3	AE
c_{33}	D_3	AC

Table 3. Costs of various decisions for classifying interesting rules

The mean cost is written as follows:

$$\bar{c} = c_{10}.P(d = D_1, s = CC) + c_{20}.P(d = D_2, s = CC) + c_{30}.P(d = D_3, s = CC) \tag{3}$$

$$+ \quad c_{11}.P(d = D_1, s = CE) + c_{21}.P(d = D_2, s = CE) + c_{31}.P(d = D_3, s = CE) \tag{4}$$

$$+ \quad c_{12}.P(d = D_1, s = AE) + c_{22}.P(d = D_2, s = AE) + c_{32}.P(d = D_3, s = AE) \tag{5}$$

$$+ \quad c_{13}.P(d = D_1, s = AC) + c_{23}.P(d = D_2, s = AC) + c_{33}.P(d = D_3, s = AC) \tag{6}$$

From the Bayes theorem, the following is true:

$$P(d = D_i, s = j) = P(d = D_i | s = j).P(s = j) \tag{7}$$

where $i = 1, 2, 3$ and $j = CC, CE, AE, AC$. Let us also assume that q̄ is the average quality vector drawn randomly from the space of all quality vectors of items sets of the rule. The following equality holds for the conditional probability $P(d = D_i | s = j)$:

$$P(d = D_i | s = j) = \sum_{\bar{q} \in D_i} f_j(\bar{q}) \tag{8}$$

where $i = 1, 2, 3$ and $j = CC, CE, AE, AC$. f_j is the probability density of the quality vectors when the actual quality status is j. We also denote the a priori probability of CC or else $P(s = CC)$ as π^0, the a priori probability of $P(s = AC) = \pi^0_{AC}$, the a priori probability of $P(s = AE) = \pi^0_{AE}$ and the a priori probability of $P(s = CE) = 1 - \pi^0 + \pi^0_{AE} - \pi^0_{AC}$. Without misclassification region $P(s = CE)$ could be simplified as $1 - \pi^0$.

The mean cost c̄ in Eq. 3 based on Eq. 7 is written as follows:

$$\bar{c} = \quad c_{10}.P(d = D_1|s = CC).P(s = CC) \tag{9}$$

$$+ \quad c_{20}.P(d = D_2|s = CC).P(s = CC) + c_{30}.P(d = D_3|s = CC).P(s = CC) \tag{10}$$

$$+ \quad c_{11}.P(d = D_1|s = CE).P(s = CE) \tag{11}$$

$$+ \quad c_{21}.P(d = D_2|s = CE).P(s = CE) + c_{31}.P(d = D_3|s = CE).P(s = CE) \tag{12}$$

$$+ \quad c_{12}.P(d = D_1|s = AE).P(s = AE) \tag{13}$$

$$+ \quad c_{22}.P(d = D_2|s = AE).P(s = AE) + c_{32}.P(d = D_3|s = AE).P(s = AE) \tag{14}$$

$$+ \quad c_{13}.P(d = D_1|s = AC).P(s = AC) \tag{15}$$

$$+ \quad c_{23}.P(d = D_2|s = AC).P(s = AC) + c_{33}.P(d = D_3|s = AC).P(s = AC) \tag{16}$$

$$\tag{17}$$

and by using Eq. 8 and by dropping the dependent vector variable q̄, Eq. 9 becomes:

$$\bar{c} = \quad \sum_{\bar{q} \in D_1} [f_{CC}.c_{10}.\pi^0 + f_{CE}.c_{11}.(1 - \pi^0 - \pi^0_{AC} + \pi^0_{AE}) \tag{18}$$

$$+ \quad f_{AE}.c_{12}.\pi^0_{AE} + f_{AC}.c_{13}.\pi^0_{AC}] \tag{19}$$

$$+ \quad \sum_{\bar{q} \in D_2} [f_{CC}.c_{20}.\pi^0 + f_{CE}.c_{21}.(1 - \pi^0 - \pi^0_{AC} + \pi^0_{AE}) \tag{20}$$

$$+ \quad f_{AE}.c_{22}.\pi^0_{AE} + f_{AC}.c_{23}.\pi^0_{AC}] \tag{21}$$

$$+ \quad \sum_{\bar{q} \in D_3} [f_{CC}.c_{30}.\pi^0 + f_{CE}.c_{31}.(1 - \pi^0 - \pi^0_{AC} + \pi^0_{AE}) \tag{22}$$

$$+ \quad f_{AE}.c_{32}.\pi^0_{AE} + f_{AC}.c_{33}.\pi^0_{AC}] \tag{23}$$

$$\tag{24}$$

Every point q̄ in the decision space D, belongs either in partition D_1, or in D_2 or D_3 in such a way that its contribution to the mean cost is minimum. This will lead to the optimal selection for the three sets of rules which we denote by D^0_1, D^0_2, and D^0_3. Based on this observation, a point q̄ is assigned to the three optimal areas as follows:
To D^0_1 if:
$f_{CC}.c_{10}.\pi^0 + f_{CE}.c_{11}.(1 - \pi^0 - \pi^0_{AC} + \pi^0_{AE}) + f_{AE}.c_{12}.\pi^0_{AE} + f_{AC}.c_{13}.\pi^0_{AC}$
$\leq f_{CC}.c_{30}.\pi^0 + f_{CE}.c_{31}.(1 - \pi^0 - \pi^0_{AC} + \pi^0_{AE}) + f_{AE}.c_{32}.\pi^0_{AE} + f_{AC}.c_{33}.\pi^0_{AC}$
and,

$f_{CC}.c_{10}.\pi^0 + f_{CE}.c_{11}.(1 - \pi^0 - \pi^0_{AC} + \pi^0_{AE}) + f_{AE}.c_{12}.\pi^0_{AE} + f_{AC}.c_{13}.\pi^0_{AC}$
$\leq f_{CC}.c_{20}.\pi^0 + f_{CE}.c_{21}.(1 - \pi^0 - \pi^0_{AC} + \pi^0_{AE}) + f_{AE}.c_{22}.\pi^0_{AE} + f_{AC}.c_{23}.\pi^0_{AC}.$
To D_2^0 if:
$f_{CC}.c_{20}.\pi^0 + f_{CE}.c_{21}.(1 - \pi^0 - \pi^0_{AC} + \pi^0_{AE}) + f_{AE}.c_{22}.\pi^0_{AE} + f_{AC}.c_{23}.\pi^0_{AC}$
$\leq f_{CC}.c_{30}.\pi^0 + f_{CE}.c_{31}.(1 - \pi^0 - \pi^0_{AC} + \pi^0_{AE}) + f_{AE}.c_{32}.\pi^0_{AE} + f_{AC}.c_{33}.\pi^0_{AC}$
and,
$f_{CC}.c_{20}.\pi^0 + f_{CE}.c_{21}.(1 - \pi^0 - \pi^0_{AC} + \pi^0_{AE}) + f_{AE}.c_{22}.\pi^0_{AE} + f_{AC}.c_{23}.\pi^0_{AC}$
$\leq f_{CC}.c_{10}.\pi^0 + f_{CE}.c_{11}.(1 - \pi^0 - \pi^0_{AC} + \pi^0_{AE}) + f_{AE}.c_{11}.\pi^0_{AE} + f_{AC}.c_{13}.\pi^0_{AC}.$
To D_3^0 if:
$f_{CC}.c_{30}.\pi^0 + f_{CE}.c_{31}.(1 - \pi^0 - \pi^0_{AC} + \pi^0_{AE}) + f_{AE}.c_{32}.\pi^0_{AE} + f_{AC}.c_{33}.\pi^0_{AC}$
$\leq f_{CC}.c_{10}.\pi^0 + f_{CE}.c_{11}.(1 - \pi^0 - \pi^0_{AC} + \pi^0_{AE}) + f_{AE}.c_{12}.\pi^0_{AE} + f_{AC}.c_{13}.\pi^0_{AC}$
and,
$f_{CC}.c_{30}.\pi^0 + f_{CE}.c_{31}.(1 - \pi^0 - \pi^0_{AC} + \pi^0_{AE}) + f_{AE}.c_{32}.\pi^0_{AE} + f_{AC}.c_{33}.\pi^0_{AC}$
$\leq f_{CC}.c_{20}.\pi^0 + f_{CE}.c_{21}.(1 - \pi^0 - \pi^0_{AC} + \pi^0_{AE}) + f_{AE}.c_{22}.\pi^0_{AE} + f_{AC}.c_{23}.\pi^0_{AC}.$
For the sake of simplicity, let's now consider the case of the absence of the misclassification region (*i.e.*, f_{AC}, f_{AE} are null and $\pi^0_{AE} = \pi^0_{AC} = 0$, we thus can simplify the inequalities above:

$$D_1^0 = \{\bar{q} : \frac{f_{CE}}{f_{CC}} \leq \frac{\pi^0}{1-\pi^0}.\frac{c_{30}-c_{10}}{c_{11}-c_{31}} \text{ and, } \frac{f_{CE}}{f_{CC}} \leq \frac{\pi^0}{1-\pi^0}.\frac{c_{20}-c_{10}}{c_{11}-c_{21}}\} \tag{25}$$

$$D_2^0 = \{\bar{q} : \frac{f_{CE}}{f_{CC}} \geq \frac{\pi^0}{1-\pi^0}.\frac{c_{20}-c_{10}}{c_{11}-c_{21}} \text{ and, } \frac{f_{CE}}{f_{CC}} \leq \frac{\pi^0}{1-\pi^0}.\frac{c_{30}-c_{20}}{c_{21}-c_{31}}\} \tag{26}$$

$$D_3^0 = \{\bar{q} : \frac{f_{CE}}{f_{CC}} \geq \frac{\pi^0}{1-\pi^0}.\frac{c_{30}-c_{10}}{c_{11}-c_{31}} \text{ and, } \frac{f_{CE}}{f_{CC}} \geq \frac{\pi^0}{1-\pi^0}.\frac{c_{30}-c_{20}}{c_{21}-c_{31}}\} \tag{27}$$

These inequalities give rise to three different threshold values L, P and N (respectively for *legitimately, potentially and not interesting* rules) in the decision space that define concretely the decision regions based on the cost of rule selection decision such as:

$$L = \frac{\pi^0}{1-\pi^0}.\frac{c_{30}-c_{10}}{c_{11}-c_{31}} \tag{28}$$

$$P = \frac{\pi^0}{1-\pi^0}.\frac{c_{20}-c_{10}}{c_{11}-c_{21}} \tag{29}$$

$$N = \frac{\pi^0}{1-\pi^0}.\frac{c_{30}-c_{10}}{c_{11}-c_{31}} \tag{30}$$

3.3 Minimal and Maximal Quality of Association Rules for Correctly Classified Data

For categorical quality dimensions, errors or pollutions (*non-quality*) in the data sets (e.g., in LHS and RHS parts of the rules) might be measured using contingency table approach where the item subsets with actual and estimated quality for a selected sample of items. It is then possible to calculate the proportion of data items that are correctly classified and estimated the quality of the entire set based on inferential statistics. Another more sophisticated approach can utilizes a random-stratified sampling method in which the same number of samples is chosen from each item subsets. This has the advantage that minor item subsets are not under-represented in the sample, which makes it possible to calculate the average quality of individual item sets.

We present now a model in which the quality of a given rule r, $P_{CC}[\bar{Q}_r]$ is defined as the probability of data items in the left and right-hand sides data sets of the rule that are correctly classified. Given two data sets with average qualities of $P_{CC}[\bar{Q}_{LHS}]$ and $P_{CC}[\bar{Q}_{RHS}]$, the quality of the rule $P_{CC}[\bar{Q}_r]$, is given by:

$$P_{CC}[\bar{Q}_r] = P_{CC}[\bar{Q}_{LHS}].P_{CC}[\bar{Q}_{RHS}|\bar{Q}_{LHS}] \qquad (31)$$

The conditional probability $P_{CC}[\bar{Q}_{RHS}|\bar{Q}_{LHS}]$ is the probability of correctly classified data items in LHS that are also correclty classified in RHS. The equation can be expanded for situations involving more than two item sets composing the rule.

From the preceding equations, the maximum and minimum quality of a given association rule can be determined based on the average quality of the several item sets I_i composing the rule.
Maximum quality is given by:

$$P_{CC}[\bar{Q}_r^{max}] = min\{P[\bar{Q}_{I_i}]\} \text{ with } i = 1, 2, \ldots, n \qquad (32)$$

Minimum quality is given by:

$$P_{CC}[\bar{Q}_r^{min}] = max\{0, \big(1 - \sum_{i=1}^{n} P_{CE}[\bar{Q}_{I_i}]\big)\} \qquad (33)$$

where $P_{CE}[\bar{Q}_{I_i}]$ is the average quality probability of the items in the data set I_i that are classified erroneous. These formulae lead to several general conclusions about composite rule quality. Composite rule quality will at the best be equal to the quality of the least quality data set. At worst composite rule quality will be equal to one minus the sum of the probability of misclassified items on each data set (or to zero if this value is negative).

4 Conclusion

This paper presents a prospective work on a theoretical probabilistic framework for estimating the cost of low-quality data on discovered association rules. Our future plans regarding this work, are to study the optimality of our decision model, to propose error estimation and to validate the model with experiments on large data sets and discovered rules with several multidimensional quality metrics.

References

[Ananthakrishna *et al.*, 2002]R. Ananthakrishna, S. Chaudhuri, and V. Ganti. Eliminating fuzzy duplicates in data warehouses. In *Proc. of the 28th Intl. Conf. on Very Large Data Bases (*VLDB*)*, Hong Kong, China, 2002.

[ArdentSoftware, 2005]Datastage suite. Available at http://www.ardentsoftware.com/, 2005.

[Bilenko and Mooney, 2003]Mikhail Bilenko and Raymond J. Mooney. Adaptive duplicate detection using learnable string similarity measures. In *KDD*, pages 39–48, 2003.

[Buechi *et al.*, 2003]M. Buechi, A. Borthwick, A. Winkel, and A. Goldberg. Cluemaker: a language for approximate record matching. In *Proc. of the 8th Inl. Conf. on Information Quality (*ICIQ *2003)*, Boston, MA, 2003.

[Caruso *et al.*, 2000]F. Caruso, M. Cochinwala, U. Ganapathy, G. Lalk, and P. Missier. TELCORDIA's database reconciliation and data quality analysis tool. In *Proc. of the Intl. Conf. on Very Large Data Bases (*VLDB*)*, 2000.

[Dasu and Johnson, 2003]T. Dasu and T. Johnson. *Exploratory Data Mining and Data Cleaning.* Wiley, 2003.

[DataMirror, 2005]Transformation server. Available at http://www.datamirror.com/, 2005.

[Dempster *et al.*, 1977]A.P. Dempster, N.M. Laird, and D.B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *ournal of the Royal Statistical Society*, 39:1–38, 1977.

[Elfeky *et al.*, 2002]M.G. Elfeky, V.S. Verykios, and A.K. Elmagarmid. Tailor: A record linkage toolbox. In *Proc. of the Intl. Conf. on Data Engineering (*ICDE*)*, 2002.

[ETI, 2005]ETI*EXTRACT. Available at http://www.eti.com/, 2005.

[Fellegi and Sunter, 1969]I.P. Fellegi and A.B. Sunter. A theory for record linkage. *Journal of the American Statistical Association*, 64, 1969.

[Galhardas *et al.*, 2001]H. Galhardas, D. Florescu, D. Shasha, E. Simon, and C. Saita. Declarative data cleaning: Language, model and algorithms. In *Proc. of the Intl. Conf. on Very Large Data Bases (*VLDB*)*, pages 371–380, 2001.

[Hernandez and Stolfo, 1995]M. Hernandez and S. Stolfo. The merge/purge problem for large databases. In *Proc. of* ACM *Special Interest Group on Management of Data Intl. Conf. (*SIGMOD *1995)*, San Jose, California, 1995.

[Huang *et al.*, 1999]K. Huang, Y. Lee, and R. Wang. *Quality Information and Knowledge Management.* Prentice Hall, New Jersey, 1999.

[Lavrac *et al.*, 1999]Nada Lavrac, Peter A. Flach, and Blaz Zupan. Rule evaluation measures: A unifying view. In *ILP*, pages 174–185, 1999.

[Lim *et al.*, 1993]L. Lim, J. Srivastava, S. Prabhakar, and J. Richardson. Entity identification in database integration. In *Proc. of the Intl. Conf. on Data Engineering (*ICDE*)*, Wien, Austria, 1993.

[Low *et al.*, 2001]W.L. Low, M.L. Lee, and T.W. Ling. A knowledge-based approach for duplicate elimination in data cleaning. *Information System*, 26(8), 2001.

[MS, 2005]MS data transformation services. Available at http://www.microsoft.com/sql/evaluation/features/datatran.asp, 2005.

[Naumann, 2002]F. Naumann. *Quality-Driven Query Answering for Integrated Information Systems.*, volume 2261 of *Lecture Notes in Computer Science.* Springer-Verlag, 2002.

[Rahm and Do, 2000]E. Rahm and H. Do. Data cleaning: Problems and current approaches. *IEEE Data Eng. Bull.*, 23(4):3–13, 2000.

[Raman and Hellerstein, 2001]V. Raman and J. M. Hellerstein. Potter's wheel: an interactive data cleaning system. In *Proc. of the Intl. Conf. on Very Large Data Bases (*VLDB*)*, 2001.

[Redman, 1996]T.C. Redman. *Data Quality for the Information Age.* Artech House, 1996. ISBN 0-89006-8836.

[Tan *et al.*, 2002]Pang-Ning Tan, Vipin Kumar, and Jaideep Srivastava. Selecting the right interestingness measure for association patterns. In *KDD*, pages 32–41, 2002.

[Tejadaa *et al.*, 2001]S. Tejadaa, C.A. Knoblock, and S. Minton. Learning object identification rules for information integration. *Information Systems*, 26(8), 2001.

[Vassiliadis *et al.*, 2000]P. Vassiliadis, Z. Vagena, S. Skiadopoulos, and N. Karayannidis. ARKTOS: A tool for data cleaning and transformation in data warehouse environments. *IEEE Data Eng. Bull.*, 23(4):42–47, 2000.

[Vassiliadis *et al.*, 2003]P. Vassiliadis, A. Simitsis, P. Georgantas, and M. Terrovitis. A framework for the design of ETL scenarios. In *Proc. of the 15th Conf. on Advanced Information Systems Engineering (*CAiSE*'03)*, Klagenfurt, Austria, 2003.

[Verykios *et al.*, 2003]V.S. Verykios, G.V. Moustakides, and M.G. Elfeky. A bayesian decision model for cost optimal record matching. *VLBD Journal*, 12(4):28–40, 2003.

[Wang *et al.*, 1995]R. Wang, V. Storey, and C. Firth. A framework for analysis of data quality research. IEEE *Transactions on Knowledge and Data Engineering*, 7(4):623–638, 1995.

[Weis and Naumann, 2004]M. Weis and F. Naumann. Detecting duplicate objects in XML documents. In *Proc. of the 1st Intl. ACM SIGMOD Workshop on Information Quality in Information Systems*, 2004.

[Winkler, 2003]W.E. Winkler. Data cleaning methods. In *Proc. of the Intl. Conf. KDD*, 2003.

[Winkler, 2004]W.E. Winkler. Methods for evaluating and creating data quality. *Information Systems*, 29(7), 2004.

Part V

Clustering

Validation in unsupervised symbolic classification

André Hardy

Department of Mathematics
University of Namur,
8 Rempart de la Vierge,
B - 5000 Namur, Belgium
(e-mail: andre.hardy@fundp.ac.be)

Abstract. One important topic in unsupervised classification is the objective assessment of the validity of the clusters found by a clustering algorithm. The determination of the "best" number of "natural" clusters has often been presented as the central problem of cluster validation. In this paper we investigate the problem of the determination of the number of clusters for symbolic objects described by interval, multi-valued and modal variables. We consider five classical methods for the determination of the number of clusters and two hypothesis tests based on the Poisson point process, and we show how these methods can be extended to symbolic data. We present applications of these symbolic methods to real data sets.

Keywords: Validation, Number of clusters, Poisson process, Symbolic data.

1 Introduction

The aim of cluster analysis is to identify a structure within a data set. When hierarchical algorithms are used, an important problem is then to choose one solution in the nested sequence of partitions of the hierarchy. On the other hand, optimization methods for cluster analysis usually require the a priori specification of the number of classes. So most clustering procedures demand the user to fix the number of clusters, or to determine it in the final solution.

Some studies have been proposed to compare procedures for the determination of the number of clusters. For example, Milligan and Cooper [Milligan and Cooper, 1985] conducted a Monte Carlo evaluation of thirty indices for determining the number of clusters. [Hardy, 1996] compared three methods based on the Hypervolumes clustering criterion with four other methods available in the Clustan software. [Gordon, 1996] modified the five stopping rules whose performance was best in the Milligan and Cooper study in order to detect when several different, widely-separated values of c, the number of clusters, would be appropriate, that is, when a structure is detectable at several different scales.

In this paper we consider two hypothesis tests for the number of clusters based on the Hypervolumes clustering criterion: the Hypervolumes test and the Gap test. These statistical methods are based on the assumption that the points we observe are generated by a homogeneous Poisson process [Karr, 1991] in k disjoint convex sets. We consider also the five best stopping rules for the number of clusters analysed by [Milligan and Cooper, 1985]. We show how these methods can be extended in order to be applied to symbolic objects described by interval, multi-valued and modal variables [Bock and Diday, 2000].

2 The clustering problem

The clustering problem we are interested in is the following.
$E = \{x_1,\ x_2,\ ...,\ x_n\}$ is a set of objects. On each of the n objects we measure the value of p variables $Y_1,\ Y_2,\ ...,\ Y_p$. The objective is to find a "natural" partition $P = \{C_1,\ C_2,\ ...,\ C_k\}$ of the set E into k clusters.

3 Statistical models based on the Poisson process

3.1 The Hypervolumes clustering method

The Hypervolumes clustering method [Hardy and Rasson, 1982] assumes that the n p-dimensional observation points $x_1,\ x_2,\ ...,\ x_n$ are generated by a homogeneous Poisson process in a set D included in the Euclidean space R^p. The set D is supposed to be the union of k disjoint convex domains $D_1,\ D_2,\ ...,\ D_k$. We denote by $C_i \subset \{x_1,\ x_2,\ ..,\ x_n\}$ the subset of the points belonging to D_i $(1 \leq i \leq k)$. The Hypervolumes clustering criterion is deduced from that statistical model, using maximum likelihood estimation. It is defined by

$$W(P,k) := \sum_{i\,=\,1}^{k} m(H(C_i))$$

where $H(C_i)$ is the convex hull of the points belonging to C_i and $m(H(C_i))$ is the multidimensional Lebesgue measure of that convex hull. That clustering criterion has to be minimised over the set of all the partitions of the observed sample into k clusters.

3.2 The generalised Hypervolumes clustering method

The generalised Hypervolumes clustering method [Rasson and Granville, 1996] assumes that the n p-dimensional points $x_1,\ x_2,\ ...,\ x_n$ are generated by a nonhomogeneous Poisson process in a set D. D is the union of k disjoint convex domains $D_1,\ D_2,\ ...,\ D_k$. The generalised Hypervolumes clustering criterion is deduced from that statistical model, using maximum likelihood estimation. It is defined by

$$W(P,k) := \sum_{i\,=\,1}^{k} \int_{H(C_i)} q(x)m(dx)$$

where q(x) is the intensity of the nonhomogeneous Poisson process.

4 Statistical tests for the number of clusters based on the Poisson point process

4.1 The Hypervolumes test

The statistical model based on the Poisson process allows us to define a likelihood ratio test for the number of clusters [Hardy, 1996]. Let us denote by $C = \{C_1, C_2, ..., C_\ell\}$ the optimal partition of the sample into ℓ clusters and $B = \{B_1, B_2, ..., B_{\ell-1}\}$ the optimal partition into $\ell - 1$ clusters. We test the hypothesis H_0: $t = \ell$ against the alternative H_A: $t = \ell - 1$, where t represents the number of "natural" clusters ($\ell \geq 2$). The test statistics is defined by

$$S(x) := \frac{W(P,\ell)}{W(P,\ell-1)}.$$

Unfortunately the sampling distribution of the statistics S is not known. But $S(x)$ belongs to $[0,1[$. Consequently, for practical purposes, we can use the following decision rule: reject H_0 if S is close to 1. We apply the test in a sequential way: if ℓ_0 is the smallest value of $\ell \geq 2$ for which we reject H_0, we choose $\ell_0 - 1$ as the best number of "natural" clusters.

4.2 The Gap test

The Gap test [Kubushishi, 1996] [Rasson and Kubushishi, 1994] is based on the same statistical model (homogeneous Poisson process). We test H_0 : the $n = n_1 + n_2$ observed points are a realisation of a Poisson process in D against H_A: n_1 points are a realisation of a homogeneous Poisson process in D_1 and n_2 points in D_2 where $D_1 \cap D_2 = \emptyset$. The sets D, D_1, D_2 are unknown. Let us denote by C (respectively C_1 , C_2) the set of points belonging to D (respectively D_1, D_2). The test statistics is given by

$$Q(x) = \left(1 - \frac{m(\triangle)}{m(H(C))}\right)^n$$

where $\triangle = H(C) \setminus (H(C_1) \cup H(C_2))$ is the "gap space" between the clusters. The test statistics is the Lebesgue measure of the gap space between the clusters.

The decision rule is the following [Kubushishi, 1996]. We reject H_0, at level α, if (asymptotic distribution)

$$\frac{nm(\triangle)}{m(H(C))} - \log n - (p-1)\log\log n \geq -\log(-\log(1-\alpha)).$$

5 Other methods for the determination of the number of clusters

We consider the best methods from the [Milligan and Cooper, 1985] study: the Calinski and Harabasz index [Calinski and Harabasz, 1974], the Duda and Hart rule [Duda and Hart, 1973], the C index [Hubert and Levin, 1976], the γ index [Baker and Hubert, 1975] and the Beale test [Beale, 1969]. The Calinski and Harabasz, Duda and Hart, and Beale indices use various forms of sum of squares within and between clusters. The Duda and Hart rule and the Beale test are statistical hypothesis tests on the number of clusters.

6 Symbolic data analysis

Symbolic data analysis [Bock and Diday, 2000] is concerned with the extension of classical data analysis and statistical methods to complex data called symbolic data. We will consider sets of objects described by interval, multi-valued and modal variables.

6.1 Interval, multi-valued and modal variables

This paper is based on the following definitions [Bock and Diday, 2000].

A variable Y is termed set-valued with the domain $\mathcal{Y}$, if for all $x_k \in E$,

$$\begin{aligned} Y : E &\rightarrow \mathcal{B} \\ x_k &\longmapsto Y(x_k) \end{aligned}$$

where $\mathcal{B} = \mathcal{P}(\mathcal{Y}) = \{U \neq \emptyset \mid U \subseteq \mathcal{Y}\}$.

A set-valued variable is called multi-valued if its values $Y(x_k)$ are all finite subsets of the underlying domain $\mathcal{Y}$; so $|Y(x_k)| < \infty$, for all elements $x_k \in E$.

A set-valued variable Y is called **categorical multi-valued** if it has a finite range $\mathcal{Y}$ of categories and **quantitative multi-valued** if the values $Y(x_k)$ are finite sets of real numbers.

A modal variable Y on a set $E = \{x_1, ..., x_n\}$ with domain $\mathcal{Y}$ is a mapping

$$Y(x_k) = (U(x_k), \pi_k), \quad \text{for all } \mathrm{x_k \in E}$$

where π_k is, for example, a frequency distribution on the domain $\mathcal{Y}$ of possible observation values and $U(x_k) \subseteq \mathcal{Y}$ is the support of π_k in the domain $\mathcal{Y}$.

Y is an interval variable if for all $x_k \in E$,

$$Y : E \rightarrow \mathcal{B} : x_k \mapsto Y(x_k) = [\alpha_k, \beta_k] \subset \mathcal{R}$$

where $\mathcal{B}$ is the set of all closed bounded interval of $\mathcal{R}$.

7 Symbolic clustering procedures

In order to generate partitions, we consider several symbolic clustering methods. SHICLUST [Hardy, 2004] is a module containing the symbolic extensions of four well-known hierarchical clustering methods: the single link, complete link, centroid and Ward methods. SCLUST [Verde *et al.*, 2000] is a partitioning clustering method; it is a symbolic extension of the well-known Dynamic clouds clustering method [Celeux *et al.*, 1989]. DIV [Chavent, 1997] is a symbolic hierarchic monothetic divisive clustering procedure based on the extension of the within class sum-of-squares criterion. SCLASS [Pirçon, 2004] is a symbolic hierarchic monothetic divisive method based on the generalised Hypervolumes clustering criterion. The first part of HIPYR [Brito, 2000] is also a module including four hierarchical symbolic clustering methods.

8 Determination of the number of clusters

8.1 Methods based on a dissimilarity matrix

In order to apply the five best methods for the determination of the number of clusters from the Milligan and Cooper [Milligan and Cooper, 1985] study, it is necessary to define a dissimilarity matrix for symbolic objects described by interval, multi-valued and modal variables.

Let us consider the case of n objects described by p interval variables

$$Y_j : E \to \mathcal{B}_j : x_i \mapsto Y_j(x_i) = x_{ij} = [\alpha_{ij}, \beta_{ij}].$$

We first define p dissimilarity indices δ_1, ..., δ_p on the sets $\mathcal{B}_j$. Let $x_{uj} = [\alpha_{uj}, \beta_{uj}]$ and $x_{vj} = [\alpha_{vj}, \beta_{vj}]$. We consider three distances for interval variables

The Haussdorff distance:

$$\delta_j\ (x_{uj}, x_{vj}) = \max\{\ |\ \alpha_{uj} - \alpha_{vj}\ |, |\ \beta_{uj} - \beta_{vj}\ |\ \}$$

The L_1 distance:

$$\delta_j\ (x_{uj}, x_{vj}) = \ |\ \alpha_{uj} - \alpha_{vj}\ | + |\ \beta_{uj} - \beta_{vj}\ |$$

The L_2 distance:

$$\delta_j\ (x_{uj}, x_{vj}) = \ (\alpha_{uj} - \alpha_{vj})^2 + (\beta_{uj} - \beta_{vj})^2.$$

We combine the p dissimilarity indices δ_1, ..., δ_p in order to obtain a global dissimilarity measure on E.

$$d : E \times E \longrightarrow R^+ : (x_u, x_v) \longmapsto d(x_u, x_v) = \left(\sum_{j=1}^{p} \delta_j^2(x_{uj}, x_{vj}) \right)^{1/2}.$$

For multi-valued and modal variables, we define suitable L_1 and L_2 distances and we use also the de Carvalho distance [Hardy, 2004].

Concerning the four hierarchical procedures included in SHICLUST, the five indices for the determination of the number of clusters are computed at each level of the hierarchies. For SCLUST, we select the best partition into ℓ clusters, for each value of ℓ ($\ell = 1, \cdots, K$) (K is a reasonably large integer fixed by the user) and we compute the indices available for nonhierarchical classification. The analysis of theses indices should provide the "best" number of clusters.

8.2 Tests based on the Poisson point processes

The Hypervolumes test and the Gap test are now available only for classical quantitative and for interval data. These tests are not based on the existence of a dissimilarity matrix, but only on the positions of the points. For interval data, we use the following modelisation. We represent an interval by two numbers: its middle and its lenght. So each interval can be represented by a point in a two-dimensional space, and an object by a point in a $2p$-dimensional space. We first determine the best number of clusters for each interval variable. A synthesis is then made in order to precise the actual structure of the set of symbolic data.

9 Examples

9.1 Merovingian buckles - VI-VIII a.c. Century

The set of symbolic data is constituted by 58 buckles described by six symbolic multi-valued variables. These variables and the corresponding categories are presented in Table 1. The complete data set is available at http://www-rocq.inria.fr/sodas/WP6/data/data.html.

Variables	Categories
Fixation	iron nail; bronze bump; none
Damascening	bichromate; predominant veneer; dominant inlaid; silver monochrome
Contours	undulations; repeating motives; geometric frieze
Background	silver plate, hatching; geometric frame
Inlaying	filiform; hatching banner; dotted banner; wide ribbon
Plate	arabesque; large size; squared back; animal pictures; plait; circular

Table 1. Merovingian buckles: six categorical multi-valued variables

The 58 buckles have been examined by archeologists. They identified two natural clusters. SCLUST and the four hierarchical clustering methods included in SHICLUST have been applied to that data set in order to generate partitions. The true structure has been detected by most of the stopping rules.

9.2 e-Fashion stores

That data set describes the sales in a group of stores (items of clothing and accessories), belonging to six different countries. These sales concern the years 1999, 2000 and 2001. The 13 objects are the stores (Paris 6th, Lyon, Rome, Barcelona, Toulouse, Aix-Marseille, Madrid, Berlin, Milan, Brussels, Paris 15th, Paris 8th, London). Eight modal variables are recorded on each of the 13 objects, describing the items sold in these stores. For example, the variable "family product" has 13 categories (dress, sweater, T-shirt, ...). The proportion of sales in each store is associated with all these categories. The variable "month" describes the proportion of sales for each month of the year.

9.3 Fats and oils

The data set contains eight fats and oils described by four quantitative features of interval type: specific gravity, freezing point, iodine value and saponification [Ichino and Yaguchi, 1994] [Gowda and Diday, 1994].

References

[Baker and Hubert, 1975]F.B. Baker and L.J. Hubert. Measuring the power of hierarchical cluster analysis. *Journal of the American Statistical Association*, pages 31–38, 1975.

[Beale, 1969]E.M.L. Beale. Euclidean cluster analysis. *Bulletin of the International Statistical Institute*, pages 92–94, 1969.

[Bock and Diday, 2000]H.-H. Bock and E. Diday. *Analysis of Symbolic Data.* Springer Verlag, 2000.

[Brito, 2000]P. Brito. Hierarchical and pyramidal clustering with complete symbolic objects. In H.H. Bock and E. Diday, editors, *Analysis of Symbolic Data Analysis*, pages 312–323, 2000.

[Calinski and Harabasz, 1974]T. Calinski and J. Harabasz. A dendrite method for cluster analysis. *Communications in Statistics*, pages 1–27, 1974.

[Celeux *et al.*, 1989]G. Celeux, E. Diday, G. Govaert, Y. Lechevallier, and H. Ralanbondrainy. *Classification automatique des données.* Bordas, 1989.

[Chavent, 1997]M. Chavent. *Analyse des données symboliques - Une méthode divisive de classification.* Thèse. Université Paris Dauphine, 1997.

[Duda and Hart, 1973]R.O. Duda and P.E. Hart. *Classification and Scene Analysis.* Wiley, 1973.

[Gordon, 1996]A.D. Gordon. How many clusters? an investigation of five procedures for detecting nested cluster structure. In C. Hayashi, N. Ohsumi, K. Yajima, Y. Tanaka, H.H. Bock, and Y. Baba, editors, *Data Science, Classification, and Related Methods*, pages 109–116, 1996.

[Gowda and Diday, 1994]K.C. Gowda and E. Diday. Symbolic clustering algorithms using similarity and dissimilarity measures. In E. Diday, Y. Lechevallier, M. Schader, P. Bertrand, and B. Burtschy, editors, *Data Science, Classification, and Related Methods*, pages 414–422, 1994.

[Hardy and Rasson, 1982]A. Hardy and J.P. Rasson. Une nouvelle approche des problèmes de classification automatique. *Statistique et Analyse des Données*, pages 41–56, 1982.

[Hardy, 1996]A. Hardy. On the number of clusters. *Computational Statistics and Data Analysis*, pages 83–96, 1996.

[Hardy, 2004]A. Hardy. Les méthodes de classification et de détermination du nombre de classes: du classique au symbolique. In M. Chavent, O. Dordan, C. Lacomblez, M. Langlais, and B. Patouille, editors, *Comptes rendus des 11èmes Rencontres de la Société Francophone de Classification*, pages 48–55, 2004.

[Hubert and Levin, 1976]L.J. Hubert and J.R. Levin. A general statistical framework for assessing categorical clustering in free recall. *Psychological Bulletin*, pages 1073–1080, 1976.

[Ichino and Yaguchi, 1994]M. Ichino and H. Yaguchi. Generalized minkowsky metrics for mixed feature type data analysis. *IEEE Transactions System, Man and Cybernetics*, pages 698–708, 1994.

[Karr, 1991]A.F. Karr. *Point Processes and their Statistical Inference.* Marcel Dekker, 1991.

[Kubushishi, 1996]T. Kubushishi. *On some Applications of the Point Process Theory in Cluster Analysis and Pattern Recognition.* PhD Thesis, University of Namur, Belgium, 1996.

[Milligan and Cooper, 1985]G.W. Milligan and M.C. Cooper. An examination of procedures for determining the number of clusters in a data set. *Psychometrika*, pages 159–159, 1985.

[Pirçon, 2004]J.Y. Pirçon. *Le clustering et les processus de Poisson pour de nouvelles méthodes monothétiques.* PhD. thesis, University of Namur, Belgium, 2004.

[Rasson and Granville, 1996]J.P. Rasson and V. Granville. Geometrical tools in classification. *Computational Statistics and Data Analysis*, pages 105–123, 1996.

[Rasson and Kubushishi, 1994]J.P. Rasson and T. Kubushishi. The gap test: an optimal method for determining the number of natural classes in cluster analysis. In E. Diday, Y. Lechevallier, M. Schader, P. Bertrand, and Butschy B., editors, *New Approaches in Classification and Data Analysis*, pages 186–193, 1994.

[Verde *et al.*, 2000]R. Verde, F. de Carvalho, and Y. Lechevallier. A dynamical clustering algorithm for multi-nominal data. In H. Kiers, J.P. Rasson, P. Groenen, and M. Schader, editors, *Data Analysis, Classification, and Related Methods*, pages 387–393, 2000.

Attribute Selection for High Dimensional Data Clustering

Lydia Boudjeloud and François Poulet

ESIEA Recherche
38, rue des docteurs Calmette et Guérin,
Parc Universitaire de Laval-Changé,
53000 Laval-France
(e-mail: boudjeloud,poulet@esiea-ouest.fr)

Abstract. We present a new method to select an attribute subset (with few or no loss of information) for high dimensional data clustering. Most of existing clustering algorithms loose some of their efficiency in high dimensional data sets. One possible solution is to use only a subset of the whole set of dimensions. But the number of possible dimension subsets is too large to be fully parsed. We use a heuristic search for optimal attribute subset selection. For this purpose we use the best cluster validity index to first select the most appropriate cluster number and then to evaluate the clustering performed on the attribute subset. The performances of our new approach of attribute selection are evaluated on several high dimensional data sets. Furthermore, as the number of dimensions used is low, it is possible to display the data sets in order to visually evaluate and interpret the obtained results.

Keywords: Attribute Selection, Clustering, Genetic Algorithm, Visualization.

1 Introduction

Data collected in the world are so large that it becomes more and more difficult for the user to access them. Knowledge Discovery in Databases (KDD) is the non-trivial process of identifying valid, novel, potentially useful and ultimately understandable patterns in data [Fayyad *et al.*, 1996]. The KDD process is interactive and iterative, involving numerous steps. Data mining is one step of the Knowledge Discovery in Databases (KDD) process. This paper focus on clustering in high dimensional data sets, which is one of the most useful tasks in data mining for discovering groups and identifying interesting distributions and patterns in the underlying data. Thus, the goal of clustering is to partition a data set into subgroups such that objects in each particular group are similar and objects in different groups are dissimilar [Berkhin, 2002]. In real world clustering situations, with most of algorithms the user has first to choose the number of clusters. Once the algorithm has performed its computation the clustering method must be validated. To validate the clustering algorithm results we usually compare them with the results of other clustering algorithms or with the results obtained by the same algorithm while varying its own parameters. We can also validate the obtained clustering algorithms results using some validity indexes described in [Milligan and Cooper, 1985]. Some of these indexes are based on the maximization of the sum of squared distances between the clusters and the minimization of the sum of squared distances within the clusters. The objective of all clustering algorithms is to maximize the distances between the clusters and minimize the distances between every object in the group, in other words, to determine the optimal distribution of the data set. The idea treated in this paper is to use the best index (according to Milligan and Cooper, it is the Calinski index), to first select the most appropriate number of clusters and then to validate the clustering performed on a subset of attributes. For this purpose we use attribute selection methods successfully used to improve cluster quality. These algorithms find a subset of dimensions to perform clustering by removing irrelevant or redundant dimensions. In section 2, we start with a brief description of the different attribute subsets search techniques and the clustering algorithm we have chosen (without forgetting that our objective is not to obtain a better clustering algorithm but to select a pertinent attribute subset with few or no loss of information for clustering). In section 3, we describe the methodology used to find the optimal number of clusters then we describe our search strategy and the method to qualify and select the subset of attributes. In section 5, we comment the obtained results and visualize the results to try to interpret them before the conclusion.

2 Attribute subset search and clustering

Attribute subset selection problem is mainly an optimization problem which involves searching the space of possible attribute subsets to identify one that is optimal or nearly optimal with respect to f (where $f(S)$ is a performance measure used to evaluate a subset S of attributes with respect to criteria of interest) [Yang and Honavar, 1998]. Several approaches of attribute selection have been proposed [Dash and Liu, 1997], [John *et al.*, 1994], [Liu and Motoda, 1998]. Most of these methods focus on supervised classification and evaluate potential solutions in terms of predictive accuracy. Few works [Dash and Liu, 2000], [Kim *et al.*, 2002] deal with unsupervised classification (clustering) where we do not have prior information to evaluate potential solution. Attribute selection algorithms can broadly be classified into categories based on whether or not attribute selection is done independently of the learning algorithm used to construct the classifier: filter and wrapper approaches. They can also be classified into three categories according to the search strategy used: exhaustive search, heuristic search, randomized search. Genetic algorithms [Goldberg, 1989] include a class-related randomized, population-based heuristics search techniques. They are inspired by biological evolution processes. Central to such evolutionary systems is the idea of a population of potential solutions that are members of a high dimensional search space. We have seen this decade, an increasing use of this kind of methods. Related works can be found in [Yang and Honavar, 1998]. However, all tests of the different authors are performed on data sets having less than one hundred attributes. The large number of dimensions of the data set is one of the major difficulties encountered in data mining. We are interested in high dimensional data sets, our objective is to determine pertinent attribute subsets in clustering, for this purpose we use genetic algorithm population-based heuristics search techniques using validity index as fitness function to validate optimal attribute subsets. furthermore, a problem we face in clustering is to decide the optimal number of clusters that fits a data set, that is why we first use the same validity index to choose the optimal number of clusters. We apply the wrapper approach to k-means clustering [McQueen, 1967], even if the framework presented in this paper can be applied to any clustering algorithm.

3 Finding the number of clusters

When we are searching for the best attribute subset, we must choose the same number of clusters than the one used when we run clustering in the whole data set, because we want to obtain a subset of attributes having same information (ideally) on the one obtained in the whole data set. [Milligan and Cooper, 1985] have compared thirty methods for estimating the number of clusters using four hierarchical clustering methods. The criteria that performed best in these simulation studies with a high degree of error in the data is a pseudo F-statistic developed by [Calinski and Harabasz, 1974]: it is a measure of the separation between clusters and is calculated by the formula: $\frac{S_b/(k-1)}{S_w/(n-k)}$, where S_b is the sum of squares between the clusters, S_w the sum of squares within the clusters, k is the number of clusters and n is the number of observations. The higher the value of this statistic, the greater the separation between groups. We first use the described statistic (Calinski index) to find the best number of clusters for the whole data set. The method is to study the maximum value max_k of i_k (where k is the number of clusters and i_k the Calinski index value for k clusters). For this purpose, we use the k-means algorithm [McQueen, 1967] on the Colon Tumor data set (2000 attributes, 62 points) from the Kent Ridge Biomedical Data set Repository [Jinyan and Huiqing, 2002], Segmentation (19 attributes, 2310 points) and Shuttle (9 attributes, 42500 points) data sets from the UCI Machine Learning Repository [Blake and Merz, 1998]. We compute all Calinski index values where k takes values in the set (2, 3,. . . , a maximum value fixed by the user) and select the maximum value max_k of the Calinski index and the corresponding value of k. The index evolution according to the different values of k for the Shuttle data set is shown in the figure 1 (we search the maximal value of the curve). We notice that the optimal value of Calinski index is obtained effectively for k=7. We obtain k=7 for Segmentation and

Shuttle data sets and k=2 for Colon Tumor data set. The optimal values found are similar to the original number of classes. Of course, these data sets are supervised classification data sets we have removed the class information. Now we try to find an optimal combination of attribute subset with a genetic algorithm having the Calinski index as fitness function. Our objective is to find a subset of attributes that best represent the configuration of the data set and discover the same configuration of the clustering (number, contained data, ...) for each cluster. The number of cluster is the value obtained for the whole data set and we search the attribute subset that has optimal value of Calinski index. The validity indexes give a measure of the quality of the resulting partition and thus usually can be considered as a tool for the experts in order to evaluate the clustering results. Using this approach of cluster validity our goal is to evaluate the clustering results in the attribute subset selected by the genetic algorithm.

4 Genetic algorithm for attribute search

Genetic algorithms (GAs) [Goldberg, 1989] are stochastic search techniques based on the mechanism of natural selection and reproduction. We use standard genetic algorithm with usual parameters (population, mutation probability), variation of these parameters have no effect for the convergence of our genetic algorithm. Our genetic algorithm starts with a population of 60 individuals (chromosomes) and a chromosome represents a combination (subset) of dimensions. The visualization of the data set is a crucial verification of the clustering results. With large multidimensional data sets (more than some hundred dimensions) effective visualization of the data set is difficult as shown in the figure 2.

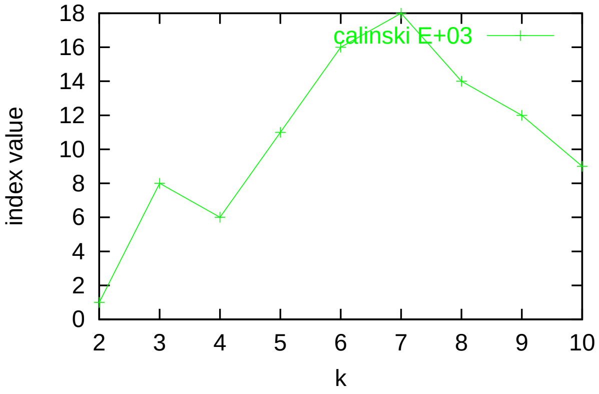


Fig. 1. Calinski index evolution for the Shuttle data set.

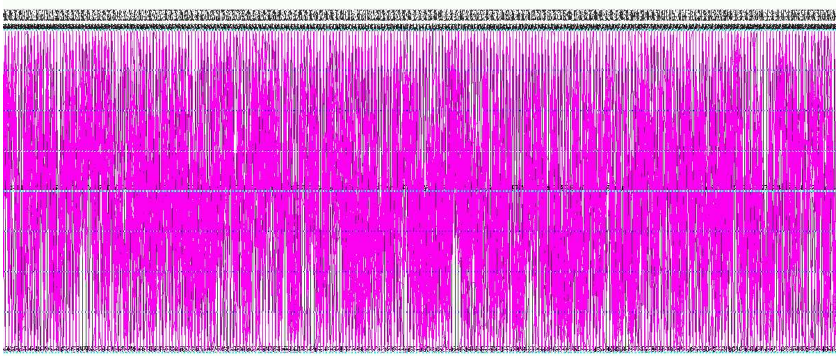


Fig. 2. Visualization of one hundred dimensions of Lung cancer data set.

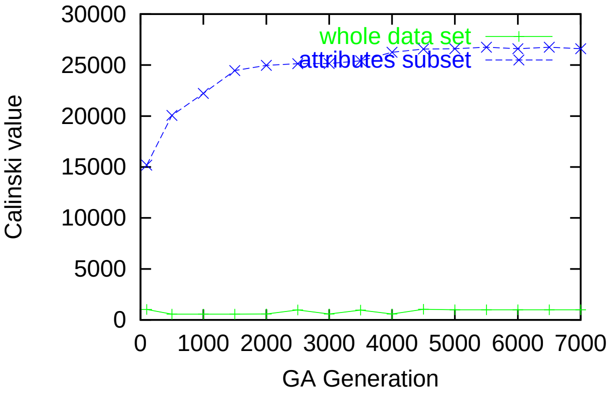


Fig. 3. Calinski index evolution for the Segmentation data set along genetic algorithm generations.

This is why the individuals (chromosomes) use only a small subset of the data set dimensions (3 or 4 attributes), we have used the same principle for outlier detection in [Boudjeloud and Poulet, 2004]. We evaluate each chromosome of the population with the Calinski index value. This procedure finds the combination of dimensions that best represents the data set with the same k as obtained for the whole data set and search attribute subset that have optimal Calinski index value. Once the whole population has been evaluated and sorted, we operate a crossover on two parents chosen randomly. Then, one of the children is muted with a probability of 0.1 and is substituted randomly for an individual of the second part of the population, under the median. The genetic algorithm ends after a maximum number of iterations. The best element will be considered as the best subset to describe the whole data, we will visualize the data set according to this most pertinent attribute subset.

5 Tests and results

We have tested GA with size 4 for the subset of attributes for the Segmentation and the Colon tumor data sets and size 3 for the Segmentation data set. Figure 3 shows the evolution of the Calinski index for all generations of the genetic algorithm for the Segmentation data set. We can see a large gap between the indexes computed with the whole data set and the indexes calculated with a subset of attributes. Our objective was to try to find the same index value for a subset of attributes as the one obtained with the whole data set. The obtained results show that the values of the indexes with the subset of attributes are better than those obtained with the whole data set. One can explain this by the fact that the data set can be noisy according to some attributes and when we select some other attributes we can get rid of the noise and therefore we obtain better results. To confirm the obtained results, we have performed tests to verify the clustering result in the different subsets of attributes that are supposed to be optimal and compared these results with the clustering obtained in the whole data set. We have used the Calinski index as reference because it is classified as the best index by Milligan and Cooper. The results with the colon Tumor data set are shown in table 1. This table describes different values obtained when we change the value of k (cluster number), we illustrate the obtained index values when k=2 and k=3, the optimal value is obtained for k=2 with 18 objects in the cluster number 1 and 44 objects in the cluster number 2. We have tested the program for a subset of 20 attributes, we describe in the third column the results obtained when we compute different index values for a subset of 20 randomly chosen attributes, after this we apply the GA to optimize the result of the index. We obtain a better Calinski index with object affectation not very different from the whole data set. We also tested our program for a subset of 4 attributes and we have obtained the optimal values described in the table (last 2 columns) for the subset of attributes: 1089, 890, 1506, 1989. We note that the cluster content for this optimal subset is similar to the cluster content in the whole data set. We presented the optimal solution of GA i.e. the subset of attributes, which has obtained the optimal values of

	Whole data set 2000 att.	Whole data set 2000 att.	Data set 20 att.	Data set 20 att. **GA opt.**	Data set 4 att.	Data set 4 att. **GA opt.**
Nbr. clusters (k)	2	3	2	2	2	2
Nbr. elemt./Cluster	18/44	10/30/22	11/51	48/14	11/51	**18/44**
Calinski	**28.91**	21.88	41.66	**56.06**	79.84	**88.50**

Table 1. GA optimization results.

all indexes. Then we visualize these results using both parallel-coordinates [Inselberg, 1985] and 2D scatter-plot matrices [Carr *et al.*, 1987], to try to explain why these attribute subsets are different from the other ones. These kinds of visualization tools allow the user to see how the data are presented in this projection. For example, figure 4 shows the visualization of clustering, with the optimal subset of attributes obtained by the GA and we can see a separation between the two clusters.

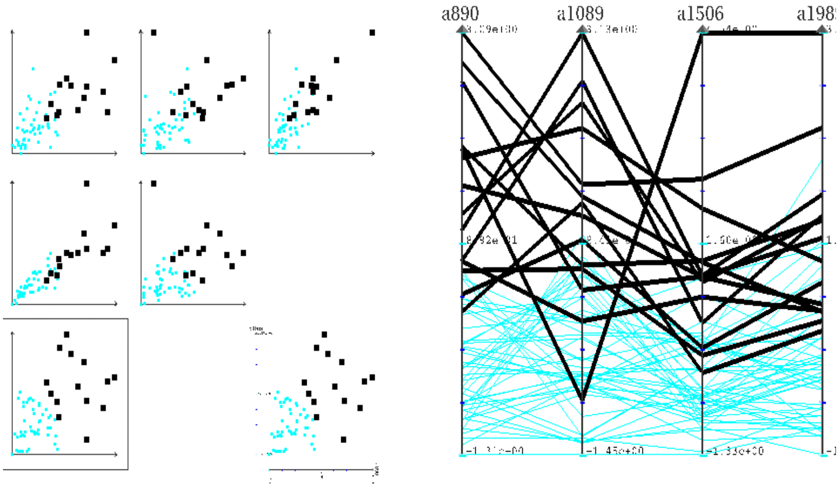


Fig. 4. Optimal subset visualization for the Colon data set.

6 Conclusion and future work

We have presented a way to select the cluster number and to evaluate a relevant subset of attributes in clustering. We used validity index of clustering algorithm not to compare clustering algorithms, but to evaluate a subset of attributes as a representative one or pertinent one for clustering results. We have used the k-means clustering algorithm, the best validity index (Calinski index) described by [Milligan and Cooper, 1985] and a genetic algorithm for the attribute selection, having the value of the validity index as fitness function. We introduced a new representation of genetic algorithm individual, our choice is fixed on small sizes of attribute subsets to facilitate visual interpretation of the results and then show the relevance of the attributes for clustering application. Nevertheless, the user is free to set up the size of the attribute subset and there is no complexity problem with the size of the population of genetic algorithm. Our first objective is to obtain subsets of attributes that best represent the configuration of the data set (number, contained data). When we tested our method by verifying clustering results we notice that the optimal subset obtained has optimal value for the index with a number of elements in the clusters similar to the ones in the whole data set and they have the same elements. Furthermore, as the number of dimensions is low, it is possible to visually evaluate and interpret the obtained results using scatter-plot matrices or/and parallel coordinates. We must keep in mind that we work with high dimensional data sets. This step is only possible because we use a subset of dimensions of the original data. This interpretation of the results would be absolutely impossible if considering all the set of dimensions (figure 2). We think to follow our objective that is to find the best attribute combination to reduce the research space without any loss in result quality. We must find a factor or a fitness function for the genetic algorithm qualifying attribute combination to optimize the algorithm and improve execution time. We think also to involve more intensively the user in the process of cluster search in data subspace [Boudjeloud and Poulet, 2005].

References

[Berkhin, 2002]P. Berkhin. Accrue software: Survey of clustering data mining techniques. In *Working paper*, 2002.

[Blake and Merz, 1998]C.L. Blake and C.J. Merz. Uci repository of machine learning databases. University of California, Irvine, Dept. of Information and Computer Sciences, 1998. http://www.ics.uci.edu/∼mlearn/MLRepository.html.

[Boudjeloud and Poulet, 2004]L. Boudjeloud and F. Poulet. A genetic approach for outlier detection in high dimensional data sets. In *Modelling, Computation and Optimization in Information Systems and Management Sciences, MCO'04*, pages 543–550. Le Thi H.A., Pham D.T. Hermes Sciences Publishing, 2004.

[Boudjeloud and Poulet, 2005]L. Boudjeloud and F. Poulet. Visual interactive evolutionary algorithm for high dimensional data clustering and outlier detection. In *to appear in proc. of The Ninth Pacific-Asia Conference on Knowledge Discovery and Data Mining*. PAKDD'05, 2005.

[Calinski and Harabasz, 1974]R.B. Calinski and J. Harabasz. A dendrite method for cluster analysis. In *Communication in statistics*, volume 3, pages 1–27, 1974.

[Carr *et al.*, 1987]D. B. Carr, R. J. Littlefield, and W. L. Nicholson. Scatterplot matrix techniques for large n. *Journal of the American Statistical Association*, 82:424–436, 1987.

[Dash and Liu, 1997]M. Dash and H. Liu. Feature selection for classification. In *Intelligent Data Analysis*, volume 1, 1997.

[Dash and Liu, 2000]M. Dash and H. Liu. Feature selection for clustering. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 110–121, 2000.

[Fayyad *et al.*, 1996]U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth. From data mining to knowledge discovery in databases. In *AI Magazine*, volume 17, pages 37–54, 1996.

[Goldberg, 1989]D.E. Goldberg. *Genetic Algorithms in Search: Optimization and Machine Learning*. Addison-Wesley, 1989.

[Inselberg, 1985]A. Inselberg. The plane with parallel coordinates. In *Special Issue on Computational Geometry*, volume 1, pages 69–97, 1985.

[Jinyan and Huiqing, 2002]L. Jinyan and L. Huiqing. Kent ridge bio-medical data set repository. 2002. http://sdmc.-lit.org.sg/GEDatasets.

[John *et al.*, 1994]G. John, R. Kohavi, and K. Pfleger. Irrelevant features and subset selection problem. In Morgan Kaufmann New Brunswick, NJ, editor, *the eleventh International Conference on Machine Learning*, pages 121–129, 1994.

[Kim *et al.*, 2002]Y. Kim, W. N. Street, and F. Menczer. Evolutionary model selection in unsupervised learning. volume 6, pages 531–556. IOS Press, 2002.

[Liu and Motoda, 1998]H. Liu and H. Motoda. Feature selection for knowledge discovery and data mining. In *Kluwer International Series in Engineering and Computer Science, Secs*, 1998.

[McQueen, 1967]J. McQueen. Some methods for classification and analysis of multivariate observations. In *Fifth Berkeley Symposium on Mathematical Statistics and Probability*, pages 281–297, 1967.

[Milligan and Cooper, 1985]G. W. Milligan and M. C. Cooper. An examination of procedures for determining the number of clusters in a data set. volume 50, pages 159–179, 1985.

[Yang and Honavar, 1998]J. Yang and V. Honavar. Feature subset selection using a genetic algorithm. In *IEEE Intelligent Systems*, volume 13, pages 44–49, 1998.

A Validation Methodology in Hierarchical Clustering

Fernanda Sousa and Jorge Tendeiro

Faculdade de Engenharia/CEC, Universidade do Porto
Rua Dr. Roberto Frias
4200-465 Porto, PORTUGAL
(e-mail: fcsousa@fe.up.pt, jorgetendeiro@net.sapo.pt)

Abstract. This paper presents a validation methodology in ascending hierarchical clustering. The objects in validation are clustering hierarchies, and simulation is used. Under certain conditions, this methodology allows us to evaluate the quality of hierarchical structures, its robustness and fiability, according to the data structure. The effect of the application of a given criterion on some kind of structures is also analyzed.

Keywords: Cluster Analysis, Hierarchical Clustering, Robustness, Validation.

1 Introduction

The use of clustering methods has progressively increased. On one hand, there are several computer programs which include these methodologies; on the other, there are big data sets which need to be studied (and summarised). Since nowadays it is quite easy and not very expensive to have big databases, it is essential to find tools in order to extract relevant information.

Generally speaking, the main goal of clustering is to define partitions or hierarchies of partitions, over a set of two-by-two comparable elements, that respect the resemblence between them in a predefined optimal manner. The elements to classify may be objects or variables of a data set.

This work belongs to the ascending hierarchical clustering (A.H.C.) field, whose usual output is a succession of partitions whose classes are partially ordered by inclusion. This methodology begins with the most refined partition (with singleton classes); in each stage the most resembled classes are gathered together according to a predefined criterion. The most common graphical result drawn out is named a classification tree or dendrogram. Choosing an element of the succession of partitions we get a division of the elements in clusters, as well as the history of the formation of each class.

Although clustering is a powerful tool in analysing data, we need to assure that the division into several clusters suggested by the algorithm does not distort the structure of the initial data. In other words, the relations between the elements to classify cannot lead to artificial clusters without real meaning. The need and importance of a next stage for the attainment of results in a clustering method is unquestionable. This stage, here named as validation, consists of questioning over the (preliminary) results obtained, in order to the achieve final results or conclusions.

Section two of this paper is dedicated to clustering validation. In section three we present a methodology based on Monte Carlo simulation which allows us to evaluate the quality and robustness of a hierarchic structure, and to give us information on the quality of adjustment to data structure. This methodology also allows us to analyse the effect that the use of a given clustering criterion can have when applied to a specific kind of structure. Section four presents an application of the proposed methodology. Finally, in section five we lay out conclusions, as well as some perspectives of developments.

2 Validation in clustering

The application of a clustering algorithm to a data set leads to a partition or a hierarchy of partitions over the set of elements in classification. After an accurate interpretation, this result will give us information on the relations between the elements in classification.

A clustering method requires choices, and it is well known that these choices affect the result of the process of clustering. In other words, different choices may lead to different classifications. This fact creates a new problem: whether to decide which choice is to result into the best clustering. We admit that each method has its own underlying structure model, which can be optimized in each situation. Also, a clustering method always produces a partition or a hierarchy of partitions, inducing a structure on the data. It seems reasonable to question the existence of structure on the initial data and, if that is the case, if there's a close relation between the initial and final structures. These and other questions really do justify the existence of a stage of validation in clustering results before its interpretation.

Several authors have studied different approaches to clustering validation. We can mention (among others) Bock's investigations [Bock, 1985], [Bock, 1996], which insert clustering models into a probabilistic context, assuming that the observed data is a sample of a structured multivariate population. [Gordon, 1994], [Gordon, 1996] and [Milligan, 1996] (and other references indicated in these works) appeal to empirical, descriptive or exploratory tools in analysing the quality of the clusters obtained.

It is usual to apply several clustering methods to a data set with the goal of choosing one of them or a new one determined from the data set. This issue includes the comparison of clustering trees or dendrograms ([Lapointe and Legendre, 1990], [Lapointe and Legendre, 1995]) and the consensus theory [Barthélemy *et al.*, 1986] and other references in [Gordon, 1999]. Another kind of issue is trying to understand the quality and stability of the results obtained from a clustering process. Here validation may also be considered under different perspectives: we may wish to validate a single cluster, a partition, or a hierarchy. The validation of a single cluster was studied (among others) by [Gordon, 1994] and [Bel Mufti, 1998]. For references about research on validation of partitions see [Hubert, 1987] and [Gordon, 1999]. Due to its complexity, hierarchies validation research has less references. Note that validation of hierarchies often appears connected to validation of partitions, since hierarchies are successions of partitions.

3 Validation methodology in A.H.C.

An A.H.C. algorithm has underlying two important choices: an index, comparison function between pairs of elements of the set to classify, and the comparison function between clusters associated to the aggregation criterion. It is assumed that the result of an A.H.C. depends on the data set as well as on these choices. Each method tends to adjust its structure model in each situation; effectiveness depends on the type and structure intensity of the data. When analysing a structure obtained by a clustering method, it is important to evaluate which part of it is due to the criterion used.

In this section we present a validation proceeding in A.H.C., whose main purpose is to help to understand some of the following questions:

- Does the data really have a clustering structure? In the affirmative case, does the hierarchy obtained reveal that structure?
- How can we choose the level of a hierarchy which gives the best partition?
- After applying several aggregation criteria or indexes to a data set, how can we decide which one is the best? Does it even exist?

The methodology here presented allows us to provide information about the problems in comparing indexes, hierarchies, indexes and hierarchies, and the effects that random perturbations have on them. This procedure intents to evaluate the quality of the final result using effectiveness and stability of a given A.H.C. method, supporting the results interpretation and helping the choices to be made. The methodology developed uses two main tools: the comparison of clustering structures and random generation of dendrograms.

3.1 Comparison of clustering structures

Let E be a set of m elements to classify, and F the set of the subsets of E with two distinct elements ($\text{card}(F) = \binom{m}{2} := M$):

$$F = \big\{\{x, y\} : x, y \in E, x \neq y\big\}.$$

Consider $\gamma : E \times E \longrightarrow \mathbb{R}_0^+$, the comparison function between pairs of elements of E, and the function $h : E \times E \longrightarrow \mathbb{R}_0^+$ which associates to each element $(x, y) \in E \times E$ the index of aggregation of the smaller cluster that contains simultaneously x and y. A function γ can be associated to a vector of dimension M that contains information about the structure of the data, and a hierarchy H can also be associated to a vector of dimension M that contains information about the clustering structure. Our goal is to compare those kind of structures. In this work we adopted an ordinal approach to do this comparison, associating preordenations to the various structures. We can define a (total) preordenation over E defining a (total) preorder over F.

The choice of a comparison function of the elements of E, γ, defines a total preorder over F; in fact, if γ is a dissimilarity we just consider

$$\forall (\{x,y\},\{z,t\}) \in F \times F : \{x,y\} \leqslant \{z,t\} \underset{\text{def.}}{\Longleftrightarrow} \gamma(x,y) \leqslant \gamma(z,t). \tag{1}$$

This total preorder over F is the total preordenation over E associated to γ. If γ is injective (which happens often in practice), this preorder is, in fact, an order.

A hierarchy H over the elements of E always defines a total preordenation over F; in fact, we have the following relation:

$$\forall (\{x,y\},\{z,t\}) \in F \times F : \{x,y\} \leqslant \{z,t\} \underset{\text{def.}}{\Longleftrightarrow} h(x,y) \leqslant h(z,t). \tag{2}$$

This total preorder over F is the total preordenation over E associated to H.

Given a partition π of E consisting of k classes $E_1, E_2, \ldots, E_k$, we can define a partition ξ of F in two classes:

- $R(\pi) = \big\{\{x,y\} \in F : x, y \in E_i \text{ for some } i = 1, 2, \ldots, k\big\}$;
- $S(\pi) = \big\{\{x,y\} \in F : x \in E_i, y \in E_j, i \neq j\big\}$.

It is easy to verify that [Lerman, 1981] ξ defines a (non total) preordenation over E. Alternatively, we can specify a total preorder over F associated to ξ as follows:

$$\forall (\{x,y\},\{z,t\}) \in F \times F : \{x,y\} \leqslant \{z,t\} \underset{\text{def.}}{\Longleftrightarrow} \begin{cases} \{x,y\},\{z,t\} \in R(\pi) \\ \underline{\text{or}} \\ \{x,y\},\{z,t\} \in S(\pi) \\ \underline{\text{or}} \\ (\{x,y\},\{z,t\}) \in R(\pi) \times S(\pi) \end{cases} \tag{3}$$

So, due to the relations (1), (2) and (3) we conclude that instead of comparing comparison functions, partitions or hierarchies of partitions we can compare the corresponding preordenations (with the same length).

There are several coefficients which allow us to compare two preordenations. The results included in this paper were obtained using the Goodman-Kruskal coefficient:

$$T_{GK} = \frac{C - D}{C + D}, \tag{4}$$

where C and D are, respectively, the number of positive and negative agreements between both preordenations. All the methodology that is going to be described can easily be applied to another coefficient.

Assintotic results on the distributions of these coefficients are not adequate in this context. The main problem is that the deduction of such distributions is based on independence between preordenations, which cannot be verified here in practice. In fact, preordenations that result from relations (1), (2) and (3) may not be independent if, for example, they come out of the application of different clustering processes over a common data set. In these situations, independence is many times what the researcher does not want, because the goal is to prove that there is information shared by the outcoming of several results. Moreover, preordenations related to clustering processes have restrictions imposed by the ultrametric property: property that verify ultrametric matrices associated to clustering structures. By this we mean that not all preordenations can be the outcome of a clustering process.

For the described reasons, it becomes necessary to deduce proper distributions for the comparison coefficients. It is not feasible to deduce the exact distribution, because the number of distinct dendrograms of order m increases very rapidly ($d(m) = \frac{m!(m-1)!}{2^{m-1}}$).

Simulation is the alternative solution, since assintotic distributions do not fit our purposes. To generate empirical distributions we need to be able to generate random clustering structures. At this stage, methods of random generation of dendrograms are extremely useful.

3.2 Random generation of dendrograms or ultrametric matrices

There are some algorithms that allow the random generation of dendrograms (or equivalent structures). Note that the point here is to generate random topologies, labels and aggregation levels; few methods attend at these three features simultaneously. We mention four methods: Double Permutation method [Lapointe and Legendre, 1990]; Uniform generation method [Sousa, 2000]; RA method [Podani, 2000]; Shape Parameter method [Sousa, 2000]. The first three methods are random *sensu* Furnas [Furnas, 1984], in other words, they can generate (for a given order) each possible dendrogram in an equiprobable manner (with probability $\frac{1}{d(m)}$). The Shape Parameter method introduces a coefficient (shape parameter) that, once settled, allows to predict (with some probability) the final shape of the generated dendrogram. This method is a very useful tool for validation in A.H.C., for it is well known that some clustering methods tend to generate particular kinds of trees.

Using one of the methods of random generation of dendrograms we can randomly generate a pair of dendrograms for a given order. By this way, it is simple to deduce empirical distributions for a chosen ordinal comparison coefficient of structures, allowing to give statistical significance to its values.

There is an alternative way in approaching the problem of random generation of clustering structures that is based on the notion of combinatorial structure ([Flajolet *et al.*, 1994] and [Van Cutsem, 1996]).

3.3 Algorithm

We now present a methodologic sequence using Monte Carlo simulation. Our goal is to supply a method that can help us answer some of the questions previously stated.

For a given topologic type of structure of data and for a fixed number of elements to classify, consider the following steps given:

1. Generate a random dendrogram; the associated ultrametric matrix, M_0, will be taken as the (initial) dissimilarity matrix.
2. For each A.H.C. criterion to study: obtain a hierarchy H_0, and compare M_0 with H_0 (comparison $\mathcal{C}^1$).
3. Disturb matrix M_0 by settling a disturbance coefficient; this creates the dissimilarity matrix M_i. Compare M_0 with M_i (comparison $\mathcal{C}^2$).
4. For each A.H.C. criterion to study: obtain a hierarchy H_i, compare M_i with H_i (comparison $\mathcal{C}^3$) and compare H_0 with H_i (comparison $\mathcal{C}^4$).
5. Repeat the steps 3. and 4. a great number of times for the same disturbance coefficient.
6. Repeat the steps 3. to 5. for different values of the disturbance coefficient.

The several comparisons, considered according to section 3.1, try to:

- $\mathcal{C}^1$: Analyse a criterion behaviour when applied to ultrametric data.
- $\mathcal{C}^2$: Control the impact of the disturbance over the associated preordenations.
- $\mathcal{C}^3$: Analyse the ability of a criterion to recover a structure after disturbance.
- $\mathcal{C}^4$: Evaluate if the hierarchical structure maintained, and try to understand what disturbance value is implied in the damaging of the structure.

4 An application

The presented methodology comprehends a diversity of choices to be made in each simulation. We now refer the several options we made in this specific application. The number of elements to classify equal 10. There were considered three types of data structures to generate: predominantly chain type trees, predominantly balanced trees (obtained with the Shape Parameter method), and also trees obtained with Uniform method. Note that both chain and balanced types are very important in classification, either for their association with well known classical methods as for their extreme characteristics. Concerning the A.H.C. methods, we tried to evaluate the performance of a set of methods belonging to classical and probabilistic approaches (the latter is known as VL approach– Validity of the Link due to I.C. Lerman) in which the aggregation criterion of clusters is based on a statistic of central tendency [Sousa, 2000]. The criteria here considered are: Single Linkage (SL), Complete Linkage (CL), Mean Linkage (HMEAN) and Median Linkage

(HMED) (classical approach), Validity of the Mean (AVM [Nicolau, 1980]), Validity of the Median (HVMED) and a method of the VL parametric family AVB proposed by [Bacelar-Nicolau, 1985] (VL approach). The disturbance was carried out adding to each element of M_0 a quantity $\delta(2x-1)$, where x comes from a uniform random variable over $]0,1[$ and δ is the disturbance factor. Values for δ were considered between 0.05 and 0.5. For the comparison of structures it was used the T_{GK} coefficient given by (4).

We now present some conclusions that illustrate how this methodology can give us information.

From $\mathcal{C}^1$ comparison we can say that classical methods recover completely the structure of an ultrametric matrix, while VL methods produce hierarchies that can be slightly different. When the dissimilarity matrix differs from the ultrametric structure ($\mathcal{C}^3$ comparison), the methods that give higher values for T_{GK} are HMEAN and HMED, followed by SL. The CL behaviour is similar to SL's for $\delta \leqslant 0.25$, but is different for greater values of δ. In general, HVMED is the VL criterion that works better, but when the data structure approaches chain type we see that AVM and HVMED are equally effective. For balanced structures, AVB seems to be the best method. $\mathcal{C}^4$ comparison allows us to conclude that the stability of the structures produced by some methods strongly depends on the type of data structure. Usually the most stable methods are HMEAN, SL and HMED, and the less stable is CL, followed by AVM. AVM is very stable when trees of chain type are considered, and for balanced trees AVB method has better T_{GK} values. The VL method less influenced by structure is HVMED.

The results obtained let us quantify some known characteristics related to the application of these criteria to real data. In fact, AVM and HVMED methods tend to produce trees of chain type, while AVB tends to produce trees with clusters of similar number of elements (balanced).

5 Conclusion and perspectives

The methodology here presented claims out to be a contribution for the A.H.C. validation subject, and it can be quite general. What was done for hierarchic clustering can easily be adjusted for partitions, too. During experiences, it was necessary to make some choices in specifing some parameters' values. This feature is considered very important. The number of possible combinations of choices is enormous, and the timing of simulation and analyse of results increases dramatically. However, a few new wise options should be tried out, particularly the application to real data.

The validity methodology here presented allows us to say that the behaviour of a clustering method strongly depends on the kind and intensity of the data structure.

The methods of central tendency of classical approach seem to have some common properties that lead to good results. The VL methods, on account of its own approach, can lead to a good performance in particular cases.

References

[Bacelar-Nicolau, 1985]H. Bacelar-Nicolau. The affinity coefficient in cluster analysis. *Methods of Operation Research*, 53:pages 507–512, 1985.

[Barthélemy *et al.*, 1986]J.-P. Barthélemy, B. Leclerc, and B. Monjardet. On the use of ordered sets in problems of comparison and consensus classifications. *Journal of Classification*, 3:pages 187–224, 1986.

[Bel Mufti, 1998]G. Bel Mufti. *Validation d'une Classe par Estimation de sa Stabilité, Ph.D. Thesis.* Université Paris IX– Dauphine, Paris, 1998.

[Bock, 1985]H.H. Bock. On some signficance tests in cluster analysis. *Journal of Classification*, 2:pages 77–108, 1985.

[Bock, 1996]H.H. Bock. Probability models and hypotheses testing in partitioning cluster analysis. P. Arabie and L.J. Hubert and G. de Soete editors, *Clustering and Classification*, pages 377–453, 1996.

[Flajolet *et al.*, 1994]P. Flajolet, P. Zimmerman, and B. Van Cutsem. A calculus for the random generation of labeled combinatorial structures. *Theoretical Computer Science*, 132:pages 1–35, 1994.

[Furnas, 1984]G.W. Furnas. The generation of random, binary unordered trees. *Journal of Classification*, 1:pages 187–233, 1984.

[Gordon, 1994]A.D. Gordon. Identifying genuine clusters in a classification. *Computational Statistics & Data Analysis*, 18:pages 561–581, 1994.

[Gordon, 1996]A.D. Gordon. Hierarchical classification. P. Arabie and L.J. Hubert and G. de Soete editors, *Clustering and Classification*, pages 65–121, 1996.

[Gordon, 1999]A.D. Gordon. *Classification.* Chapman & Hall, London, 2nd edition, 1999.

[Hubert, 1987]L.J. Hubert. *Assignment Methods in Combinatorial Data Analysis.* Marcel Dekker, New York, 1987.

[Lapointe and Legendre, 1990]F.J. Lapointe and P. Legendre. A statistical framework to test the consensus of two nested classifications. *Systematic Zoology*, 39:pages 1–13, 1990.

[Lapointe and Legendre, 1995]F.J. Lapointe and P. Legendre. Comparison tests for dendrograms: A comparative evaluation. *Journal of Classification*, 12:pages 265–282, 1995.

[Lerman, 1981]I.C. Lerman. *Classification et Analyse Ordinale des Données.* Dunod, Paris, 1981.

[Milligan, 1996]G.W. Milligan. Clustering validation: Results and implications for applied analyses. P. Arabie and L.J. Hubert and G. de Soete editors, *Clustering and Classification*, pages 341–375, 1996.

[Nicolau, 1980]F.C. Nicolau. *Critérios de Análise Classificatória Hierárquica Baseados na Função Distribuição, Ph.D. Thesis.* Faculdade de Ciências da Universidade de Lisboa, Lisboa, 1980.

[Podani, 2000]J. Podani. Simulation of random dendrograms and comparison tests: Some comments. *Journal of Classification*, 17:pages 123–142, 2000.

[Sousa, 2000]F. Sousa. *Novas Metodologias e Validação em Classificação Hierárquica Ascendente, Ph.D. Thesis.* Universidade Nova de Lisboa, Lisboa, 2000.

[Van Cutsem, 1996]B. Van Cutsem. Combinatorial structures and structures for classification. *Computational Statistics & Data Analysis*, 23:pages 169–188, 1996.

Determining the number of groups from measures of cluster stability

G. Bel Mufti[1], P. Bertrand[2] and L. El Moubarki[3]

[1] U.R. CEFI, ESSEC de Tunis, 4 rue Abou Zakaria El Hafsi, Montfleury, 1089 Tunis, Tunisie (e-mail: belmufti@yahoo.com)

[2] GET - ENST Bretagne, Dept. Lussi, CS 83818, 29238 BREST Cedex 3, France (e-mail: patrice.bertrand@enst-bretagne.fr)

[3] ISG de Tunis, 41 rue de la liberté, Cité Bouchoucha, Le Bardo, 2000 Tunis, Tunisie (e-mail: elmoubarki_lassad@yahoo.fr)

Abstract. An important line of inquiry in cluster validation involves measuring the stability of a partition with respect to perturbations of the data set. Several authors have recently suggested that the 'correct' number of clusters in a partition can be determined simply by examining the partition stability measures for different values of numbers of clusters. In this paper, we consider the clustering stability measures that were recently proposed in [Bertrand and Bel Mufti, 2005], and we present experiments that compare the method for predicting the number of clusters that is derived from these stability measures with two of the most successful methods reported in recent surveys.

Keywords: Cluster Stability, Monte Carlo Test, Cluster Isolation and Cluster Cohesion, Loevinger's measure, Number of clusters of a partition.

1 Introduction

A major challenge in cluster analysis is the validation of clusters resulting from cluster analysis algorithms. One relevant approach involves defining an index measuring the adequacy of a cluster structure to the data set and establishing how likely a given value of the index is under some null model formalizing 'no cluster structure', e.g., [Bailey and Dubes, 1982], [Jain and Dubes, 1988], [Gordon, 1994], [Milligan, 1996] and [Gordon, 1999]. Another type of approach is concerned with the estimation of the stability of clustering results. Informally speaking, cluster stability holds when membership of the clusters is not affected by small changes in the data set [Cheng and Milligan, 1996]. Several recent approaches, see for example [Tibshirani *et al.*, 2001], [Levine and Domany, 2001], [Ben-Hur *et al.*, 2002] and [Bertrand and Bel Mufti, 2005], suggest that cluster stability is a valuable way to determine the number of clusters of any partitioning of the data. Such a stability based approach aims to identify those values of the number of clusters (or any other parameter of the clustering method) for which local maxima of stability are reached.

The main contribution of this paper is to compare this stability based approach with two of the most (classical) successful methods of predicting the number of clusters. In what follows, we restrict our attention to the measures of cluster stability that were introduced by Bertrand and Bel Mufti [Bertrand and Bel Mufti, 2005]. In section 2, a summarized description of the cluster validation method introduced by Bertrand and Bel Mufti [Bertrand and Bel Mufti, 2005] is presented. This method involves the definition of stability measures both of the partition and of its clusters. Each stability measure is defined as Loevinger's measure of a rule quality, that is assessed by a probability significance which is approximated by comparing the value of the measure with values that would be obtained under a null model that specifies the absence of cluster stability. In section 3, we compare three methods for determining the number of clusters of any partitioning of a data set, on the basis of their experimental results obtained for the partitioning of two data sets. The first method is the stability based approach that is briefly mentioned here above and that is specified by the stability measures of Bertrand and Bel Mufti [Bertrand and Bel Mufti, 2005]. The other two methods are classical methods performing the best for estimating the number of clusters, according to the survey of Milligan and Cooper [Milligan and Cooper, 1985].

2 The cluster stability measures proposed by Bertrand and Bel Mufti (2005)

In this section, we briefly describe the stability based method of cluster validation that was recently introduced by Bertrand and Bel Mufti, and we refer the reader to [Bertrand and Bel Mufti, 2005] for more details.

We will denote as $\mathcal{X}$ an arbitrary data set of n objects to be clustered, and as P_k any generic k-way partitioning algorithm. The partition obtained by running P_k on the data set $\mathcal{X}$ will be denoted by $\mathcal{P}$, in other words $\mathcal{P} = \mathrm{P}_k(\mathcal{X})$. The validation method proposed in [Bertrand and Bel Mufti, 2005] is designed to estimate the stability of both the partition $\mathcal{P}$ and its clusters, with regards to both cluster isolation and cluster cohesion criteria. The perturbed data sets are (random) samples of the population $\mathcal{X}$. If all partitions into k clusters obtained from running algorithm P_k on different samples of $\mathcal{X}$ are close in structure to partition $\mathcal{P}$, then $\mathcal{P}$ can be deemed to be stable. In order to guarantee that each cluster of $\mathcal{P}$ is still represented in each random sample of $\mathcal{X}$, we use a sampling procedure, called *proportionate stratified sampling*. More precisely, given any cluster A of $\mathcal{P}$ and denoting by n_A the size of A, and by f some sampling ratio, this sampling procedure involves selecting randomly and without replacement n'_A elements in each cluster of $\mathcal{P}$, where n'_A is the integer value obtained by rounding down fn_A to the nearest integer. On the basis of experimental results presented in [Bertrand and Bel Mufti, 2005] and recommendations given in [Levine and Domany, 2001] and [Ben-Hur *et al.*, 2002], the value of f has to be chosen in the interval $[0.7, 0.9]$.

Let us focus on the single criterion of cluster isolation. Informally speaking, there is much evidence that any cluster of $\mathcal{P}$, say A, is isolated whenever the following rule holds for any sample $\mathcal{X}'$ of $\mathcal{X}$:

> (R) *Isolation rule of A.* If two objects of $\mathcal{X}'$ are not clustered together by partition $\{A, \mathcal{X} \setminus A\}$, then they are not in the same cluster of $\mathrm{P}_k(\mathcal{X}')$.

Any measure of rule quality can assess the rule (R). However, due to its specific properties and its simplicity of interpretation (see [Lenca *et al.*, 2003]), Loevinger's measure ([Loevinger, 1947]) is preferred to other measures of rule quality. Loevinger's measure of rule $E \Rightarrow F$, is defined as the expression $1 - P(E \cap \neg F)/P(E)P(\neg F)$. Denoting by $t(A,\, \mathcal{X}')$ Loevinger's measure of the quality of rule (R), we obtain:

$$t(A,\, \mathcal{X}') = 1 - \frac{n'(n'-1)m_{(\mathcal{X}';\, A, \overline{A})}}{2n'_A(n'-n'_A)\, m_{(\mathcal{X}')}}\ , \tag{1}$$

where $m_{(\mathcal{X}')}$ is the number of pairs of objects that are clustered together by $\mathrm{P}_k(\mathcal{X}')$, and where $m_{(\mathcal{X}';\, A, \overline{A})}$ is the number of pairs of sampled objects that are in the same cluster of $\mathrm{P}_k(\mathcal{X}')$ and for which exactly one of the two objects belongs to A. Taking into account only the criterion of cluster isolation, the stability measure of cluster A is defined simply as the average, denoted here by $\overline{t}_N(A)$, of the values $t(A,\, \mathcal{X}'_i)$ obtained for a large number N of samples $\mathcal{X}'_i$ $(i = 1, \ldots, N)$:

$$\overline{t}_N(A) = \frac{1}{N}\sum_{i=1}^{N} t(A,\, \mathcal{X}'_i). \tag{2}$$

It should be noted that $\overline{t}_N(A)$ is an (unbiaised) estimation of the expected value of the random variable $t(A, \mathcal{X}')$, when $\mathcal{X}'$ is considered as a random sample. This leads us to select a value of N large enough so that both the central limit theorem holds and the length of the approximate standard 95%-confidence interval is less than some maximal desired length l.

Several other stability measures were similarly defined in order to assess other characteristics of any cluster, *i.e.*, its isolation with respect to another cluster, its cohesion and its validity. In addition, the same three characteristics (isolation, cohesion and validity) of any partition were defined. Furthermore, it was proved that each stability measure of any partition that concerns isolation (resp. cohesion) is a weighted mean of the stability measures of all its clusters with respect to the criterion of isolation (resp. cohesion).

One important issue concerns the interpretation of the order of magnitude of the observed values of stability measures. This is a general problem in cluster validation: Jain and Dubes ([Jain and Dubes, 1988] p.144) noted that it is easy to propose indices of cluster validity, but that it is very difficult to fix thresholds on such indices that define when the index is large or small enough to be 'unusual'. The difficulty is solved by following the general procedure presented by Jain and Dubes [Jain and Dubes, 1988] (see also [Gordon, 1994]), since it seems reasonable to specify the absence of cluster stability by the absence of clustering structure:

> **Step 1.** Define a null model $\mathcal{M}_0$ that specifies the null hypothesis H_0 of absence of cluster stability for the data set under investigation; in the case of a data set that can be represented by n points of an euclidean space, an example of such a null model is the uniform distribution of n points in the convex hull of the data set.
>
> **Step 2.** Estimate the probability significance of the observed value of the stability measure under the null hypothesis H_0. Since the analytic expression of the distribution of the stability measure under the null model $\mathcal{M}_0$ is usually unknown, this step generally involves performing a Monte Carlo test: a large number, say M, of data sets are simulated according to the model $\mathcal{M}_0$, and each of them is partitioned and the corresponding value of stability measure is computed. The probability significance is then estimated on the basis of these M values of the stability measure.

For example, the value $\overline{t}_N(A) = 0.899$ is an indication of high stability if and only if its estimated probability significance value under H_0 is less than 5%.

3 Experimental comparison with two methods for determining the number of clusters

As previously mentioned in section 1, a method for determining the 'optimal' number of clusters in a partitioning of a data set can easily be derived from the stability measure of a partition introduced in [Bertrand and Bel Mufti, 2005]: a k-clusters partition is considered as meaningful if the value of the partitional stability measure is a local maximum when k varies. In what follows, this partitional stability measure will be denoted as $BB(k)$, when k is the number of clusters of the partition. The information provided by the stability index $BB(k)$ can be refined by considering its probability significance under the null hypothesis H_0, and also by taking into account the stability measures (concerning isolation and cohesion) of each cluster in the partition together, with their probability significances.

Otherwise, many indices that measure the adequation between the partition and the data set were proposed to determine the number of clusters. According to the survey of Milligan and Cooper [Milligan and Cooper, 1985], the index of Calinski and Harabasz [Calinski and Harabasz, 1974] and the index of Krzanowski and Lai [Krzanowski and Lai, 1985] are among the indices that perform the best (see also Tibshirani et al [Tibshirani *et al.*, 2001] for another experimental comparison). The index of Calinski and Harabasz is defined by:

$$CH(k) = \frac{B(k)/(k-1)}{W(k)/(n-k)} \tag{3}$$

where k denotes the number of clusters, and $B(k)$ and $W(k)$ denote the between and within cluster sums of squares of the partition, respectively. An optimal number of clusters is then defined as a value of k that maximizes $CH(k)$. The index of Krzanowski and Lai is defined by:

$$KL(k) = |\frac{DIFF(k)}{DIFF(k+1)}|, \tag{4}$$

where:

$$DIFF(k) = (k-1)^{2/p}W(k-1) - (k)^{2/p}W(k), \tag{5}$$

and p denotes the number of features in the data set. A value of k is optimal if it maximizes $KL(k)$.

The rest of the section is devoted to the comparison of the performance of the three indices BB, CH and KL on the basis of results obtained for two data sets: an artificial data set and the well known Iris data set.

3.1 An artificial data set

We consider the artificial data set that is represented in Figure 1. This data set is a 200 point sample of a mixture of four normal distributions.

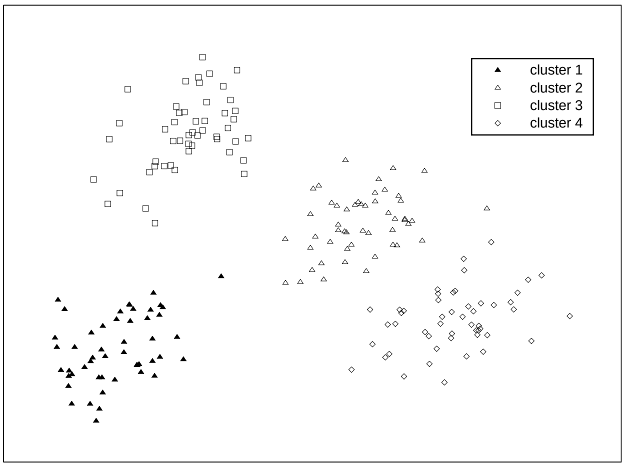


Fig. 1. Artificial data set structured into four clusters.

Each cluster is indeed a 50 point sample of one of the four normal distributions, and except for one point, the four clusters are easily identified by looking at Figure 1. The four normal distributions are centered respectively at $\mu_1 = (-1.5, -.5)$, $\mu_2 = (3, 2)$, $\mu_3 = (0, 4)$ and $\mu_4 = (4.5, 0)$ and have the same variance-covariance matrix $V = .5I$, where I denotes the identity matrix.

This data set was partitioned using the batch K-means method and the stability measures were computed with the ratio sampling $f = 0.8$. The values of the three indices are given in Table 1 for $k \in \{2, 3, 4, 5, 6\}$. The probability significances under (H_0) (p-values) suggest that the 4-partition is the most significant.

		Number of clusters (k)				
Index		2	3	4	5	6
$CH(k)$		145	414	580 *	494	446
$KL(k)$		.26	3.36	3.89	1.39	5.95 *
$BB(k)$		.779	.958	.992 *	.914	.816
Prob. sign. of $BB(k)$	(%)	48 − 61	2.4 − 6.8	0 − 1	0 − 4.5	2.5 − 9.2

Table 1. Values of the three indices for partitions of the artificial data. According to each index (row), a symbol $(^*)$ indicates the optimal numbers of clusters.

Table 2 contains all the cluster stability measures concerning the 4-partition. These values indicate that the four clusters are stable: all the stability measures are high and assessed as being significant under H_0 by low p-values.

Each stability measure in Table 3 was computed with a precision of at least 0.01, which required running the batch K-means method on $N = 140$ samples of the artificial data set. The slight lack of isolation of cluster 2 (.984), just like the slight lack of cohesion of cluster 1 (.980), suggests the presence of an outlier between these two clusters (see Fig.1). The partition into 4 clusters is also identified as optimal by the index CH, but the index KL suggests that $k = 6$ is the optimal number of clusters.

Table 3 presents the stability measures of the 5-partition. Note that with a p-value that is less than 4.5% at a 97.5%-approximate coverage probability, the global validity of the partition into 5 clusters can be deemed as significant. Each of these stability measures were computed with a precision of at least 0.02, and $N = 1500$ samples were necessary in order to obtain this precision. It turns out that the clusters 1, 2 and 3 (which coincide with clusters 3, 4

		Isolation		*Cohesion*		*Validity*	
			%		%		%
Cluster	**1**	.990	$0-1$	.980	$0-5$	.986	$0-1$
	2	.984	$0-1$	.992	$0-2$	.987	$0-1$
	3	1.	$0-1$	1.	$0-1$	1.	$0-1$
	4	.994	$0-1$	.996	$0-2$	.995	$0-1$
Partition		.992	$0-1$	.992	$0-1$	.992	$0-1$

Table 2. Stability measures for the 4-partition (prec. 0.01), and their p-values (%).

and 2 respectively in Figure 1) are clearly stable, for all cluster characteristics except the cohesion of cluster 3. Clusters 4 and 5 (obtained by splitting the cluster 1 of Figure 1 into two clusters) are assessed by low stability values (*i.e.*, .716 and .777) and by high p-values (*i.e.*, in the intervals $34-50\%$ and $22-39\%$). Therefore, their existence is clearly dubious. Stability measures for partial isolation between clusters were also computed: the extremely weak stability measure for partial isolation between cluster 4 and cluster 5 (*i.e.*, -.999) suggests that the split represents more a dissection than a real cluster structure involving separate and homogeneous clusters.

		Isolation		*Cohesion*		*Validity*	
			%		%		%
Cluster	**1**	.993	$0-1$	.939	$0-1$	.973	$0-1$
	2	.993	$0-1$	.936	$0-1$	.972	$0-1$
	3	.989	$0-5$	.873	$1-13$	.945	$0-8$
	4	.696	$32-49$	.798	$48-65$	.716	$34-50$
	5	.727	$29-47$	.980	$1-9$	.777	$22-39$
Partition		.915	$0-4.5$	.913	$0-1$	.914	$0-4.5$

Table 3. Stability measures (prec. 0.01) of the 5-partition, and their p-values (%).

3.2 Iris data

The famous Iris data set reports four characteristics of 3 species namely the iris setosa, versicolor and virginica. Each class contains 50 instances. One class (namely, the virginica) is linearly separable from the others, but the latter are not linearly separable from each other. Iris data were partitioned using the batch K-means method, taking into account only the two variables petal length and width. As in the previous subsection, we have set the value of the ratio sampling f to 0.8.

		Number of clusters (k)			
Index		2	3	4	5
$CH(k)$		756	1211	1266	1358*
$KL(k)$		4.83	6.01*	1.3	1.12
$BB(k)$		.992*	.959	.881	.900
Prob. signif. of $BB(k)$	(%)	$.3 - 3.4$	$6.7 - 11.9$	> 34	$5.2 - 9.4$

Table 4. Values of the indices on Iris data partitions. According to each index (row), a symbol (*) indicates an optimal number of clusters.

Table 4 shows the values of the three indices used for choosing the optimal number of clusters on Iris data. The 2-partition with a p-value between .3 and 3.4% is the most stable partition according to the index BB, followed by the 5-partition and the 3-partition with p-values in the intervals $5.2 - 9.4\%$ and $6.7 - 11.9\%$, respectively. Even if the p-values of the last two partitions do not differ significantly, the large p-values of the stability measures of two clusters of the 5-partition (*i.e.*, in the intervals $39 - 53\%$ and $52 - 65\%$) raise doubts about the validity of this partition (see also [Bertrand and Bel Mufti, 2005]). The stability measure BB is the only one to identify the trivial partition in two clusters, and the KL index identifies the 3-partition as the optimal one. Choosing the 5-partition, the index CH is the worst performer on the Iris data set.

4 Conclusion

The results presented in this paper confirm that measuring cluster stability can be a valuable approach to determine the 'correct' number of clusters of any partition. A real advantage of this general approach is that it does not require selecting or using any measure of adequation between the data set and the partition examined.

It can be noticed that the p-values for assessing the measures of cluster stability may be decisive when estimating the stability of clusters. For example, the p-values of Table 1 show that the stability value .915, which assesses the stability of the 5-partition, is statistically more significant under the null hypothesis of absence of structure, than the stability value .958 which assesses the stability of the 3-partition. In addition, an advantage of the stability based approach that is proposed in [Bertrand and Bel Mufti, 2005] is that a careful interpretation of the p-values of the stability measures enables one to identify not only a pertinent partition but also several sources of variation in the partitional stability, such as individual cluster isolation and cohesion.

References

[Bailey and Dubes, 1982]T. A. Bailey and R. Dubes. Cluster validity profiles. *Pattern Reconition* 15, 61–83, 1982.

[Ben-Hur *et al.*, 2002]A. Ben-Hur, A. Elisseeff and I. Guyon. A stability based method for discovering structure in clustered data. *Pacific Symposium on Biocomputing* 7, 6–17, 2002.

[Bertrand and Bel Mufti, 2005]P. Bertrand and G. Bel Mufti. Loevinger's measures of rule quality for assessing cluster stability. *Computational Statistics and Data Analysis*, 2005, to appear.

[Calinski and Harabasz, 1974]R. B. Calinski and J. Harabasz. A dendrite method for cluster analysis. *Communications in Statistics* 3, 1–27, 1974.

[Cheng and Milligan, 1996]R. Cheng and G. W. Milligan. Measuring the influence of individual data points in a cluster analysis. *J. Classification* 13, 315–335, 1996.

[Gordon, 1994]A. D. Gordon. Identifying genuine clusters in a classification. *Computational Statistics and Data Analysis* 18, 561–581, 1994.

[Gordon, 1999]A. D. Gordon. *Classification.* Chapman & Hall, 1999.

[Jain and Dubes, 1988]A. K. Jain and R. Dubes. *Algorithms for clustering data.* Prentice-Hall, Englewood Cliffs, NJ, 1988.

[Krzanowski and Lai, 1985]W. J. Krzanowski and Y. T. Lai. A criterion for determing the number of groups in a data set using sum-of-squares clustering. *Biometrics* 44, 23–34, 1985.

[Lenca *et al.*, 2003]P. Lenca, P. Meyer, B. Vaillant and S. Lallich. Critères d'évaluation des mesures de qualité en ECD. *Revue des Nouvelles Technologies de l'Information (Entreposage et Fouille de données)*, 1, 123–134, 2003.

[Levine and Domany, 2001]E. Levine and E. Domany. Resampling method for unsupervised estimation of cluster validity. *Neural Comput.* 13, 2573–2593, 2001.

[Loevinger, 1947]J. Loevinger. A systemic approach to the construction and evaluation of tests of ability. *Psychological Monographs*, 61 (4), 1947.

[Milligan and Cooper, 1985]G. W. Milligan and M. C. Cooper. An examination of procedures for determining the number of clusters in a data set. *Psychometrika* 50, 159–179, 1985.

[Milligan, 1996]G. W. Milligan. Clustering validation: results and implications for applied analyses. In P. Arabie, L. J. Hubert and G. De Soete, editors, *Clustering and Classification*. Word Scientific Publ., River Edge, NJ, pp. 341-375, 1996.

[Tibshirani *et al.*, 2001]R. Tibshirani, G. Walther, D. Botstein and P. Brown. Cluster validation by prediction strength. *Stanford Technical Report, Department of Statistics, Stanford University, USA*, 2001.

Hierarchical Classification for Seabed Characterization

Christophe Osswald and Arnaud Martin

ENSIETA - E^{3}I^2 - EA3876
2 rue François Verny,
29806 BREST Cedex 09, France
(e-mail: Christophe.Osswald@ensieta.fr, Arnaud.Martin@ensieta.fr)

Abstract. The automatic seabed characterization is a difficult problem. Most automatic characterization approaches are based on texture analysis. Indeed, the sonar seabed images present many homogeneous areas of sediment that can be interpretated as a sonar texture.

Here, we optimize the agglomerative hierarchical clustering algorithm to produce homogenous clusters of sediments images, combining known and unknow data.

Keywords: Classification, Sonar, Seabed Characterization.

1 Introduction

The problem of automatic seabed characterization is very important and difficult. The seabed characterization is important in order to make seabed maps for sedimentologists, for autonomous underwater vehicle navigation or pollution. One approach in order to characterize the seabed is the use of a sonar. The main issues with sonar images is those are particularly difficult to characterize by automatic process. The expert has never the certainty to differentiate well the sand from the silt for example: the difference between these sediments comes only from the granulometry that varies continuously.

We first expose the principle of agglomerative hierarchical classification. In section 2 we present the sonar images data and the considered texture analysis. We study the usual clustering methods applied on sonar small-images. Then in section 5 we define a hierarchy quality in order to choose better agregation functions for hierarchical classification. In section 6, we present results of the combination of known and unknow data in order to characterize the sediment of the sonar images.

2 Agglomerative Hierarchical Classification

Agglomerative hierarchical classification (AHC) is a common approach to build a clustering system from a dataset. The algorithm considers the objects of the dataset as trivial clusters of size 1: $d(\{x\},\{y\}) = d(x,y)$. Then, at each step, the algorithm merges the two nearest clusters into a new cluster, and computes the distance between the new cluster and the other ones. The index associated to the cluster $C = A \cup B$ is the dissimilarity $d(A, B)$.

The dissimilarity induced by an indexed hierarchy (*i.e.* dissimilarity between x and y is the smallest index of a cluster containing x and y) is an *ultrametric.*

The natural clustering system of a dissimilarity, in the way of Jardine and Sibson [Jardine and Sibson, 1971], is composed of the maximal cliques of its threshold graphs, indexed by the diameter of the clusters. So a set A is a cluster for the dissimilarity d with an index λ if:

(i) there exists x and y in A such that $d(x, y) = \lambda$,
(ii) u and v in A brings $d(u, v) \leqslant \lambda$,
(iii) for any z not in A there exists t in A such that $d(z, t) > \lambda$.

The indexed clustering system induced by an ultrametric is an indexed hierarchy. It is well-known that ultrametrics and indexed hierarchies are in bijection.

Let d be a dissimilarity on X and used as a dissimilarity on the singletons of X. An agglomerative hierarchical clustering (AHC) can be summarized in three steps:

1. find A and B such that $d(A, B)$ is minimal.
2. merge A and B in a cluster C.
3. for each remaining cluster D, compute $d(C, D)$.
4. go back to step 1 unless $C = X$.

Differences between algorithms are mainly the way $d(C, D)$ is computed, but steps 1 and 2 can have more than one interpretation. When more than one pair $\{A, B\}$ realize the minimum of d, the choice can be random or lexicographic, or d can be transformed such that the choice has no further consequence [Barthelémy and Guénoche, 1991]. This usually leads to clusters C larger than $A \cup B$. Due to the origin of our data, minimum of d can be considered as unique, and therefore the possible strategies for steps 1 and 2 are equivalent.

Many strategies for computing the distance between the new cluster $C = A \cup B$ and the other clusters have been explored by Lance and Williams [Lance and Williams, 1967], and formalized under the formula:

$$d^p(C, D) = \alpha_A d^p(A, D) + \alpha_B d^p(B, D) + \beta d^p(A, B) + \gamma |d^p(A, D) - d^p(B, D)|$$

Chen [Chen, 1996] restricts the form of α, β and γ in order to explicit the properties of the indexed hierarchy produced by the algorithm. They are functions of three parameters:

$$r_A = \frac{|A|}{|A \cup B|} \qquad r_B = \frac{|B|}{|A \cup B|} \qquad r_D = \frac{|D|}{|A \cup B|}$$

The parameter p is a nonzero real number.

$$\begin{aligned}d^p(C,D) = \alpha(r_A, r_D)d^p(A,D) + \alpha(r_B, r_C)d^p(B,D) + \\ \beta(r_A, r_B, r_C)d^p(A,B) + \gamma(r_C)|d^p(A,D) - d^p(B,D)|\end{aligned}$$

Most usual agglomerative hierarchical classification algorithm can be written under this formalism:

Algorithm	$\alpha(u,w)$	$\beta(u,v,w)$	$\gamma(w)$	p
single linkage	1/2	0	−1/2	1
complete linkage	1/2	0	1/2	1
Ward's method	$\frac{u+w}{1+w}$	$-\frac{w}{1+w}$	0	2

Such an algorithm is called $LW(\alpha, \beta, \gamma, p)$. One should notice that the value of p for single linkage and complete linkage, which is usually 1, can be any nonzero real number. An LW algorithm is said *space-conserving* if

$$\min\{d(A,D), d(B,D)\} \leqslant d(A \cup B, D) \leqslant \max\{d(A,D), d(B,D)\}$$

Single linkage and complete linkage are space conserving. So ultrametrics are fixed points for these algorithms. Ward method is not space-conserving, but it is *space-dilatating*: the dissimilarity produced by the algorithm is greater than the input, and can be different even if the input dissimilarity is an ultramtetric.

To produce an admissible hierarchy indexed by f, the condition $A \subseteq B \implies f(A) \leqslant f(B)$ must be respected. To achieve this goal on any dissimilarity, the LW algorithm must be *monotonic* [Dragut, 2001]:

(i) $\alpha(u,w) + \alpha(1-u,w) + \beta(u, 1-u, w) \geqslant 1$
(ii) $\alpha(u,w) \geqslant 1$
(iii) $\gamma(w) \geqslant \max\{-\alpha(u,w), -\alpha(1-u,w)\}$

Many aggregation functions cannot be written as LW functions, but can be used to produce an indexed hierarchy. It is the case for any *internal* aggregation function. A family of AHC algorithms based on median functions have been studied in [Osswald, 2003].

3 Data

The database contains 26 sonar images provided by the GESMA (Groupe d'Études Sous-Marine de l'Atlantique). Theses images were obtained with a Klein 5400 sonar with a resolution of 20 until 30 cm in azimuth and 3 cm in range. The sea-bottom deep was between 15 m and 40 m.

These 26 sonar images of different sizes (about 92 m width and 92 m to 322 m length) have been segmented in small-images with a size of 64x384 pixels (*i.e.* of approximately 1152 cm × 1152 cm). We have obtained 4003

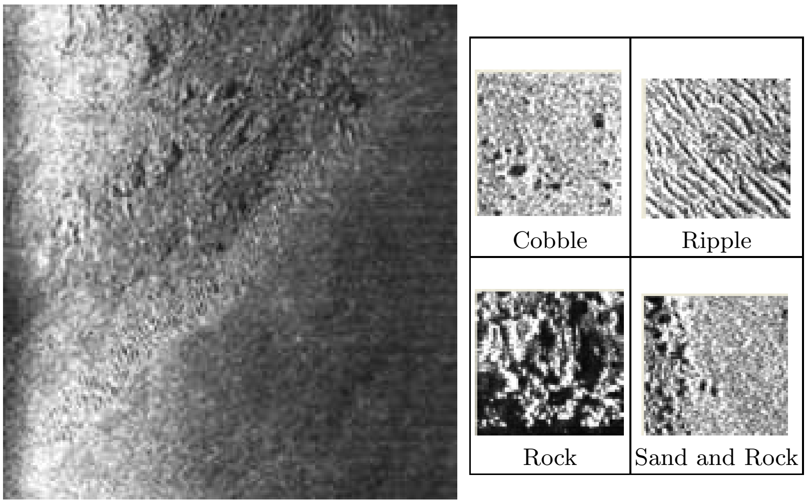


Fig. 1. Sonar image example (provided by the GESMA) and extracted and segmented small-images examples.

small-images. On table 1 we show a sonar image and a sample of these small-images represented in order to obtain a size of 64x64 pixels.

Each small-image is characterized manually by the type of sediment (rock, cobbles, sand, ripple, silt) or shadow when the information is unknown (see Table 1). Moreover the existence of more than one kind of sediment on the small-image is indicated. In this case the type of sediment affected to the small-image is the most present.

Sediment	%	code	% patchworked
Sand	56.06	s	32.00
Rock	19.91	r	43.29
Ripple	9.34	p	61.50
Shadow	8.02	o	47.66
Silt	5.85	i	35.04
Cobble	0.82	c	84.85

Table 1. Percentage and code of type of sediment

From Table 1 we note that the sand sediment is the most represented one. The cobbles sediment is particularly few represented. One of the difficulties of classification step comes from this difference.

There is 38.87% of small-image with more than one kind of sediment (named patch-worked images).

Note that such database is quite difficult to realize. Indeed, the expert has a subjective experience, and can make a mistake for some small-images.

From these small-images, we have extracted texture features. Different texture extraction methods are presented in [Martin *et al.*, 2004]. Each method allow to extract some features that can be redundant, but calculated differently. We choose here to use a wavelet transform.

Indeed, this approach can consider the translation invariance in the directions. The discrete translation invariant wavelet transform is based on the choice of the optimal translation for each decomposition level. Each decomposition level gives four new images on which three features are calculated: the energy, the entropy and a mean. We keep a decomposition level of 3 giving 63 parameters.

So, each small-image is represented in a 63-space. We have calculated the euclidean distance between each small-image: it is the initial dissimilarity used by the AHC algorithms.

4 Usual clustering methods applied on small sonar images

4.1 Some general properties of AHC algorithms

Dissimilarity induced by the single linkage algorithm has the property of being *subdominant*: it is the greatest ultrametric smaller than the original dissimilarity. This constraint often leads to more efficient algorithms [Brucker, 2001]. In the case of ultrametrics, it leads to an algorithm in $\mathcal{O}(n^2)$ operations instead of $\mathcal{O}(n^3)$ for the other LW algorithms.

The single linkage hierarchy is also known to have an *unbalancing effect*: paths from leaves to root have often very different lengths. When A and B are two non-trivial clusters, we also often have $A \subset B$ or $B \subset A$. So it is hard to separate objects into classes: partitions obtained from such a hierarchy are composed of one huge class, and many very small ones.

Other AHC algorithm are not well-defined: applying twice the complete linkage on a dataset may produce two distinct hierarchies, when the dissimilarity d between clusters admits two minimums, and choosing a random one can modify the hierarchy obtained. As our data is composed of floating numbers calculated from real sonar data, the probability of having two minimum in our dissimilarity matrix is nearly 0, so the LW algorithm we use is univocal, and produce binary hierarchies.

4.2 Exemples

Applied to our data, single linkage, complete linkage and Ward algorithm give the trees of figure 2. Index used for the representation is cluster size, for the real index does not allow us to distinguish all the clusters, and the following treatments will only use the clustering structure, not the indices.

We proceed by taking k small-images of each class (k is 4 for examples of figure 2, 12 or 15 for figure 3 data). The proportion of patchworked images, when allowed, is the same than in the original data. As there are only 5 not patchworked cobble images, we consider classes of different size when dealing with larger sets of not patchworked images.

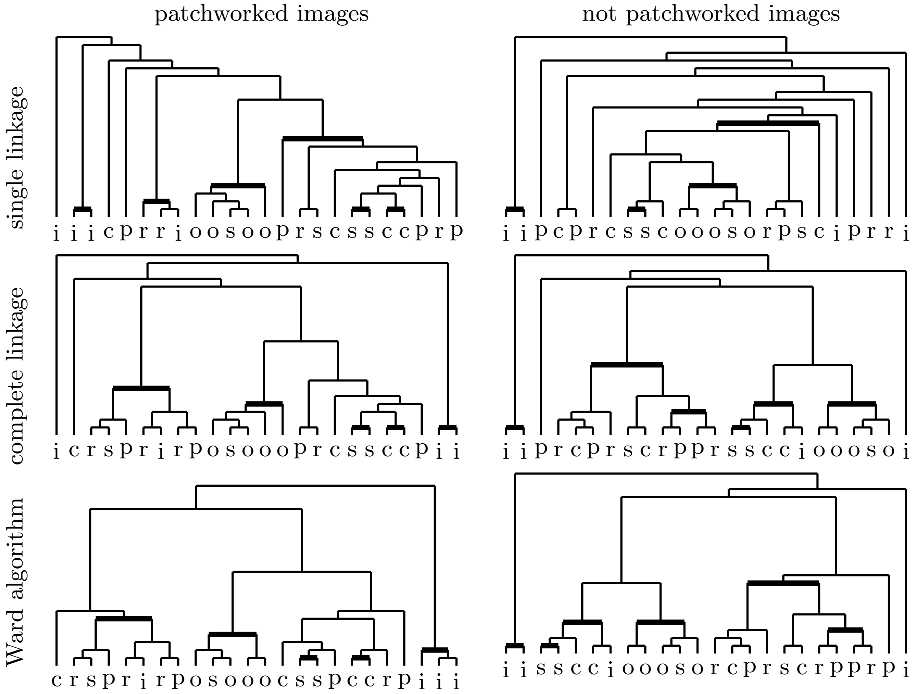

patchworked images
not patchworked images
single linkage
complete linkage
Ward algorithm
i i i c p r r i o o s o o p r s c s s c c p r p
i i p c p r c s s c o o o s o r p s c i p r r i
i c r s p r i r p o s o o o p r c s s c c p i i
i i p r c p r s c r p p r s s c c i o o o s o i
c r s p r i r p o s o o o c s s p c c r p i i i
i i s s c c i o o o s o r c p r s c r p p r p i

Fig. 2. Usual AHC algorithms applied on some small-images

5 Hierarchy quality

We consider that a hierarchy is efficient for seabed characterization if it contains clusters that are *representative* of each sediment. An expert has defined six classes $M_1, \ldots, M_6$ of small-images, partitionning our data into six sediment classes. We search in the hierarchy $\mathcal{H}$ for clusters A that maximize the quality of an association pattern $A \leftrightarrow M_i$, for i between 1 and 6.

Our concern is how the clusters of the hierarchy can be used as natural clusters for the data. We limit our qualiy measures to the shape of the hierarchy, not its index. A standard (quadratic) distance between the ultrametric induced by the AHC algorithm and the original distance would not help us to reach this goal. As we will see later, Ward's method leads to the most efficient hierarchies, but is space-dilatating. Such a measure would have favored a AHC algorithm between single linkage and complete linkage.

5.1 Measure of hierarchy quality

Tan *et al.* [Tan *et al.*, 2002] have made an exhaustive study of the measures used to measure the quality of association patterns. To obtain a simple measure, depending as little as possible on the size of the dataset, and possible to combine by multiplication, we choose the Jaccard measure, where $P(A)$ is the proportion of elements of A in the dataset:

$$\zeta(A \leftrightarrow M_i) = \frac{P(A \cap M_i)}{P(A) + P(M_i) - P(A \cap M_i)}$$

Combined on a hierarchy, we obtain the quality measure $q(\mathcal{H})$. Bold clusters on figure 2 are the clusters maximizing the ζ measure for at least one type of sediment.

$$q(\mathcal{H}) = \prod_{i=1}^{6} \max_{A \in \mathcal{H}} \zeta(A \leftrightarrow M_i)$$

What is used in the characterization step is not usually a pattern $A \leftrightarrow M_i$ but a rule $A \rightarrow M_i$. As we do not need (and often not want to have) a symmetrical measure for A and M_i, we should use an association rule measure instead of an association pattern measure.

As we want to avoid too small rules, *i.e.* $A \rightarrow M_i$ with $|A| \ll |M_i|$, our measure must take into account the unexplained examples, *i.e.* elements of M_i which are not in A. The Confidence measure $(c((A \leftrightarrow M_i) = 1 - \frac{P(A \cap \overline{M_i})}{P(A)})$ and all the other similar measures are not accurate to achieve this duty (see Vaillant *et al.*, [Vaillant *et al.*, 2004]). The Piatetsky-Shapiro measure, a non-symmetrical extension of the support measure, seem to be the most accurate: $PS(A \rightarrow M_i) = P(A)P(\overline{M_i}) - P(A, \overline{M_i})$ where $\overline{M_i}$ is the complementary of M_i.

5.2 Parameters for Lance-Williams algorithms

Lance and Williams functions associated to single linkage, complete linkage and Ward's method are given section 2.

We build a continuous family of LW algorithms containing those three usual methods. In order to guarantee that $\alpha(u, w) + \alpha(1 - u, w) + \beta(u, 1 - u, w) \geqslant 1$ and therefore that the AHC algorithm obtained is monotonic, we use an intermediary link:

$$\alpha_i(u,w) = \frac{u + w/2}{1+w} \quad \beta_i(u, 1-u, w) = 0 \quad \gamma_i(w) = 0 \quad p_i = 2$$

We use three segments of the space of admissible monotonic *LW* algorithms. The parameter x varies in $[0, 1]$.

	$\alpha(u,w)$	$\beta(u,v,w)$	$\gamma(w)$	p
Single to Complete	$1/2$	0	$x - 1/2$	1
Complete to Intermediary	$(1-x)/2+$ $x(u+w/2)/(1+w)$	0	$(x-1)/2$	2
Intermediary to Ward	$(u+(1+x)w/2)/(1+w)$	$-xw/(1+w)$	0	2

We apply this family to random restrictions of our set of small-images, composed of pure small-images or a combination of pure and patchworked small-images. We estimate the efficiency of the *LW* functions on these restrictions.

The quality measure relies on the form of the hierarchy: presence or absence of a cluster. Let $\mathcal{H}(x)$ be the hierarchy produced by the LW algorithm of parameter x. There exists reals x_1 and x_2 such that $x_1 < x < x_2$ and for each $t \in]x_1, x_2[$ we have $\mathcal{H}(t) = \mathcal{H}(x)$. Therefore the quality measure $q(\mathcal{H}(x))$ is locally constant.

On figure 3 we can note the Ward's method is the best *LW* algorithm of the family considered to classify our data. It is not possible to extend the β function joining the intermediary linkage to Ward to x greater than 1, for a value of β lesser than $-w/(1+w)$ would not respect the (i) condition of monotonicity.

5.3 Use of optimized AHC algorithm for texture identification

To use the hierarchy as a characterization tool, we first optimize the *LW* functions on a learning set. We merge this set with small-images whose class is unknown, and we build a hierarchy on this new set, with the same *LW* functions. Then we classify the unknown elements belonging to an optimal class of a sediment type.

We use a set of 72 elements for learning purpose (12 of each sediment type), allowing patchworked small-images, and we add 228 untagged elements. The procedure give us good results for silt and shadow. 100% of small-images tagged by silt are effectivly silt, and 68% for shadow. Among the 228 small-images to classify, 41 received a correct tag, 101 received one

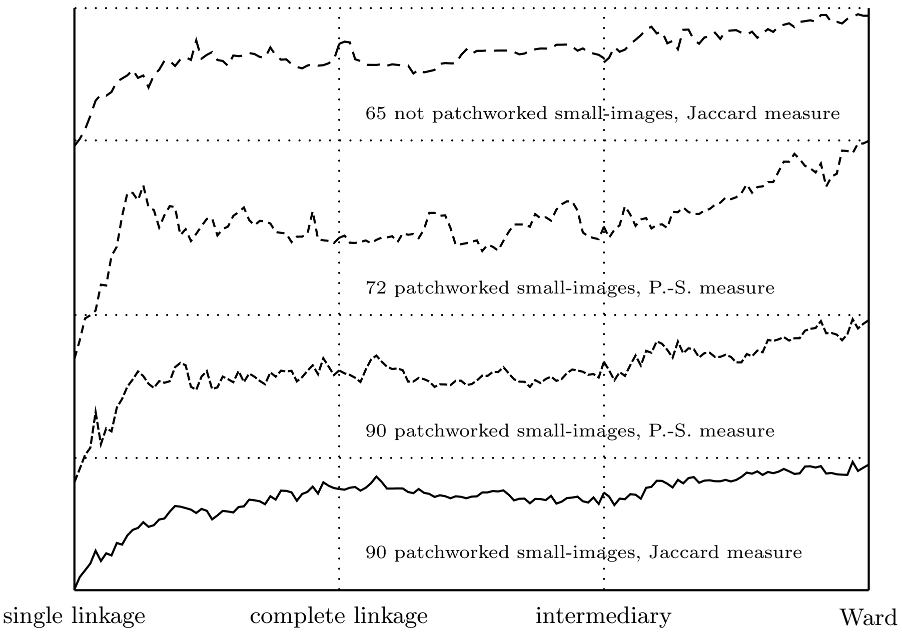


Fig. 3. Hierarchy quality

correct tag and one other tag, 75 received no correct tag and 11 received no tag at all.

Most unclassified small-images are silt (but most silt is well-classified); most ripples small-images are not correctly classified, but Martin *et al.* showed that the wavelets are not an efficient features set to discriminate ripples, as it is not rotation invariant.

6 Conclusion

This approach mixes non-supervised classification methods and supervised classification goals. The supervised context allows us to optimize the AHC parameters, and the tagging method used allow an image to receive one, zero or more than one tag. In a system were several classifiers collaborate, powerful fusion algorithms may use this information.

Here, Ward's method is the most accurate. This may be because of the way dissimilarity is calculated: inertia is closely related to euclidean model. Maybe the fact our classes are of similar size is the origin: Ward's criteria is space-dilating, so it tends to build balanced hierarchies.

References

[Barthelémy and Guénoche, 1991]J.-P. Barthelémy and A. Guénoche. *Trees and Proximity Representations.* Wiley, New York, 1991.

[Brucker, 2001]F. Brucker. *Modèles de classification en classes empiétantes.* PhD thesis, EHESS, 2001.

[Chen, 1996]Z. Chen. Space-conserving agglomerative algorithms. *Journal of classification*, 13:157–168, 1996.

[Dragut, 2001]A. Dragut. Characterization of a set of algorithms verifying the internal similarity. *Mathematical Reports*, 53(3-4):225–232, 2001.

[Jardine and Sibson, 1971]N. Jardine and R. Sibson. *Mathematical Taxonomy.* Wiley, London, 1971. part II.

[Lance and Williams, 1967]G. N. Lance and W. T. Williams. A general theory of classificatory sorting strategies. *The computer journal*, 3-4(9-10):373–380 and 271–277, 1967.

[Martin *et al.*, 2004]A. Martin, G. Sévellec, and I. Leblond. Characteristics vs decision fusion for sea-bottom characterization. In *Colloque Caracterisation in-situ des fonds marins*, Brest, France, 2004.

[Osswald, 2003]C. Osswald. Robustesse aux variations de méthode pour la classification hiérarchique. In *XXXVèmes Journées de Statistiques*, pages 751–754, Lyon, 2003. SFdS.

[Tan *et al.*, 2002]P.-T. Tan, V. Kumar, and J. Srivastava. Selecting the right interestingness measure for association patterns. In *SIGKDD '02*, Edmonton, Canada, 2002.

[Vaillant *et al.*, 2004]B. Vaillant, P. Lenca, and S. Lallich. Association rule interestingness measures: an experimental study. Technical Report 2, GET / ENST Bretagne, 2004.

On the Fitting and Consensus of Classification Systems

Bruno Leclerc

CAMS - EHESS
54 bd Raspail
75270 PARIS Cedex 06, France
(e-mail: leclerc@ehess.fr)

Abstract. Classification systems are families of subsets (classes) of a fixed set S that are closed for intersection and contain S and every single element subset of S. The main problem conidered here is that of the consensus of such systems. We first briefly mention results issued from lattice theory. Then, we consider the Adams approach for the consensus of hierarchies and point out its relation with closures, implications (as they appear in relational databases) and nestings. We show that Adams consensus correspond to the research of a particular subdominant nesting (or overhanging) relation, and generalize the corresponding fitting problem.
Keywords: Closure system, Classification system, Implication, Overhanging order, Lattice, Hierarchy.

1 Introduction

Let S be a finite set. We consider here the aggregation of a profile $\mathcal{F}^* = (\mathcal{F}_1, \mathcal{F}_2, ..., \mathcal{F}_k)$ of classifications on S into a consensus classification $\mathcal{F} = c(\mathcal{F}*)$. A classification will be here a family of subsets (classes) containing the whole set S and every one-element subset of S (singleton), and closed under intersection. Equivalently, classification systems are the closure systems of the literature that include all the singletons.

There are two main purposes for the research of such a consensus. First, the classification of a set S described by variables of different types. Each qualitative or quantitative variable v induces a partition or a quasi-order on S, which in turn induces a classification system. With such a common formalization for various structures, a set of k variables leads to a profile $\mathcal{F}^*$ of k such systems. The idea is to aggregate the elements of $\mathcal{F}^*$ into a unique system $c(\mathcal{F}*)$ that summarize the profile in some useful sense (see [Domenach and Leclerc, 2004b] for more details).

The other reason is that several consensus problems already studied in the literature are particular cases of the consensus of closure systems. The basic example is provided by hierarchies, where, frequently in a purpose of phylogenetic reconstruction, many works have followed those of [Adams III, 1972] and [Margush and McMorris, 1981] (see the survey [Leclerc, 1998]). Other usual classification models correspond, directly or after straightforward completions, to closure systems. Thus, several other classical consensus problems are also particular cases, with restricted domains or codomains (or both), of the consensus of classification systems. An example is the aggregation of partitions [Régnier, 1965], [Régnier, 1983], [Mirkin, 1975], [Barthélemy and Leclerc, 1995].

2 Classifications and closure systems

Given a finite set S, and its power set $\mathcal{P}(S)$, a *classification system* on S is a family $\mathcal{F} \subseteq \mathcal{P}(S)$ of classes (subsets) of S. A class $C \in \mathcal{F}$ may be a set of elements sharing some common properties, or close to each other in some sense. Then the following conditions, although not always required, may appear as natural ones :

(C1) $S \in \mathcal{F}$;

(C2) $C, C' \in \mathcal{F} \Rightarrow C \cap C' \in \mathcal{F}$;

(C3) for all $s \in S, \{s\} \in C$.

Then, from (C2) and (C3), we have the empty class in $\mathcal{F}$. This property, although not usual, is appropriate to obtain structural coherence. A family $\mathcal{F}$ which satisfies only (C1) and (C2) is a so-called *closure system* (or *Moore family*).

The most usual classification models correspond to such classification systems, sometimes with the addition of some trivial classes. For instance, the addition of the empty class to a hierarchy $\mathcal{H}$, or the addition of S, the empty set and the lacking singletons to a partition provide classification systems. Pyramids (or quasi-hierarchies) and weak hierarchies, in their intersection-closed variants, are further examples.

We find in the literature three notions (among many others) which all are in one-to-one correspondence with closure systems (cf. [Caspard and Monjardet, 2003]).

A *closure operator* φ on S is a mapping on $\mathcal{P}(S)$ satisfying the three properties of *isotony* (for all $A, B \subseteq S$, $A \subseteq B$ implies $\varphi(A) \subseteq \varphi(B)$), *extensivity* (for all $A \subseteq S$, $A \subseteq \varphi(A)$) and *idempotence* (for all $A \subseteq S$, $\varphi(\varphi(A)) = \varphi(A)$). The elements of the image $\mathcal{F}_\varphi = \varphi(\mathcal{P}(S))$ of $\mathcal{P}(S)$ by φ are the *closed* (by φ) *subsets* of S, and $\mathcal{F}_\varphi$ is a closure system on S. Conversely, a closure operator $\varphi_\mathcal{F}$ on S is associated to any closure system $\mathcal{F}$ on S by $\varphi_\mathcal{F}(A) = \bigcap\{F \in \mathcal{F} : A \subseteq F\}$ (i.e., from (C1) and (C2), the smallest class of $\mathcal{F}$ containing A exists and is $\varphi_\mathcal{F}(A)$.

A *complete implication system* on S, denoted by I, $\rightarrow_I$ or simply $\rightarrow$, is a binary relation on $\mathcal{P}(S)$ satisfying, for all $A, B, C, D \subseteq S$:

(I1) $B \subseteq A$ implies $A \to B$;
(I2) $A \to B$ and $B \to C$ imply $A \to C$;
(I3) $A \to B$ and $C \to D$ imply $A \cup C \to B \cup D$.

An *overhanging order* (*nesting order* in some contexts) on S is a binary relation on P(S) too, denoted as Œ and satisfying, for all $A, B, C \subseteq S$:
(O1) $A \text{ Œ } B$ implies $A \subset B$;
(O2) $A \subset B \subset C$ implies $A \text{ Œ } C \iff [A \text{ Œ } B$ or $B \text{ Œ } C]$;
(O3) $A \text{ Œ } A \cup B$ implies $A \cap B \text{ Œ } B$.

It is not difficult to see that Œ is then a (partial) order on $\mathcal{P}(S)$. The sets of all closure systems, closure operators, complete implication systems and overhanging orders on S are respectively denoted as **M**, **C**, **I** and **O**. They are in one-to-one correspondence to each other. Besides the correspondence recalled above, we give hereunder two further correspondences, the first one due to [Armstrong, 1974], and the second pointed out in [Domenach and Leclerc, 2004]: for all $A, B \subseteq S$,

$$A \to B \iff B \subseteq \varphi(A)$$
$$A \text{ Œ } B \iff A \subset B \text{ and } \varphi(A) \subset \varphi(B)$$

So, in a classification system, $A \to B$ means that every class including the subset A of S also includes B, while $A \text{ Œ } B$ means that B properly includes A and, moreover, there exists at least one classs including A and not B.

Further conditions correspond to particular classes of systems. For instance, an overhanging order corresponds to a classification system if and only if it satisfies the following condition (OS) below, and to a hierarchy if, moreover, the following condition (OT) replaces (O3) [Adams III, 1986], [Domenach and Leclerc, 2004]: for all $A, B, C \subseteq S, s \in S$,
(OS) $s \notin A$ implies $\emptyset \text{ Œ } \{s\} \text{ Œ } A \cup \{s\}$;
(OT) $A \text{ Œ } C$ and $B \text{ Œ } C$ imply $A \cup B \text{ Œ } C$ or $A \cap B = \emptyset$.

3 Consensus in the lattice of closure systems

The sets **M**, **C**, **I** and **O** are naturally ordered: **M** by set inclusion on $\mathcal{P}(\mathcal{P}(S))$, **I** and **O** by set inclusion on $\mathcal{P}(\mathcal{P}(S) \times \mathcal{P}(S)) = \mathcal{P}((\mathcal{P}(S))^2)$, **C** by the poinwise order on mappings: $\varphi \leq \varphi'$ if $\varphi(A) \subseteq \varphi'(A)$ for all $A \subseteq S$. The resulting orderings are either isomorphic or dually isomorphic: if φ, I and Œ (respectively φ', I' and Œ') are, respectively, the closure operator, complete implication system and overhanging order associated to a given closure system $\mathcal{F}$ (respectively to $\mathcal{F}'$), one has $\mathcal{F} \subseteq \mathcal{F}' \iff \varphi' \leq \varphi \iff I' \subseteq I \iff$ $\text{Œ} \subseteq \text{Œ}'$ (cf. [Caspard and Monjardet, 2003] and [Domenach and Leclerc, 2004b] for the case of overhangings).

The sets **M** and **I** are closed under set intersection in, respectively, $\mathcal{P}(\mathcal{P}(S))$ and $\mathcal{P}((\mathcal{P}(S))^2)$, and the set **O** is closed under set union in

$\mathcal{P}((\mathcal{P}(S))^2)$. The greatest elements of $\mathbf{M}$, $\mathbf{I}$ and $\mathbf{O}$ are, respectively, $\mathcal{P}(S)$, $\mathcal{P}(S))^2$ and $\{(A,B) : A,B \subseteq S, A \subset B\}$, whereas their lowest elements are, respectively, $\{S\}, \{(A,B) : A,B \subseteq S, B \subseteq A\}$ and the empty relation on $\mathcal{P}(S)$. So, $\mathbf{M}$ and $\mathbf{I}$ are themselves closure systems on, respectively, $\mathcal{P}(S)$ and $\mathcal{P}(S))^2)$.

Ordered by inclusion, any closure system $\mathcal{F}$ is a lattice $(\mathcal{F}, \vee, \cap)$, with $(F \vee F' = \varphi(F \cup F')$ for all closed subsets $F, F' \in \mathcal{F}$. The existence of such a lattice structure has important consequences for the consensus problem as described above, that is the aggregation of any profile $\mathcal{F}^* = (\mathcal{F}_1, \mathcal{F}_2, ..., \mathcal{F}_k)$ of closure systems into a closure system $\mathcal{F} = c(\mathcal{F}^*)$. Previous results on the consensus in lattice structures may be found, among others, in [Monjardet, 1990], [Barthélemy and M.F., 1991] and [Leclerc, 1994], with significant issues in particular cases like those of hierarchies ([Barthélemy *et al.*, 1986]), partitions ([Barthélemy and Leclerc, 1995]) or orders ([Leclerc, 2003]). Results for the particular case of closure systems are given in [Raderanirina, 2001] and [Monjardet and Raderanirina, 2004].

A *federation* on K is a family $\mathcal{K}$ of subsets of $K = \{1, ..., k\}$ satisfying $[L \in \mathcal{K}, L' \supseteq L] \Rightarrow [L' \in \mathcal{K}]$. We then define a *federation consensus function* $c_{\mathcal{K}}$ associated to the federation $\mathcal{K}$ by $c_{\mathcal{K}}(\mathcal{F}^*) = \bigvee_{L \in \mathcal{K}} (\bigcap_{i \in L} \mathcal{F}_i)$. Especially, K is an *oligarchic consensus function* if $K = \{L \subseteq K : L \supseteq L_0\}$ for a fixed subset L_0 of K.

Another class of consensus functions consists of the so-called *quota rules* $c_q = c_{\mathcal{K}}$, where $\mathcal{K} = \{L \in K : |L| \geq q\}$ for a given number q $(0 \leq q \leq k)$. Equivalently, $c_q(\mathcal{F}^*) = \bigvee\{A \subseteq S : |\{i \in K : A \in \mathcal{F}_i\}| \geq q\}$ is the closure system generated by those classes that are present in at least q of the $\mathcal{F}_i$'s. Especially, for $q = k$, the quota rule is the same as the oligarchie rule obtained with $L_0 = K$.

The above definition of federation consensus functions needs the set K (and, so, the integer k) to be fixed. Such a constraint is easily removed for quota rules by replacing the number q with a proportion (see [Barthélemy and M.F., 1991]). Note also that, if all the closure systems in $\mathcal{F}^*$ are classification systems, then the federation consensus system $c_{\mathcal{K}}(\mathcal{F}^*)$ is is still a classification system, for any federation $\mathcal{K}$. The same remark holds for quota rules.

An axiomatic approach (cf. [Day and McMorris, 2003]) of the consensus problem on $\mathbf{M}$ allowed to characterize oligarchic rules ([Raderanirina, 2001]), whereas a metric approach, based on the symmetric difference metric ∂ on $\mathbf{M}$ defined by $\partial(\mathcal{F}, \mathcal{F}') = |\mathcal{F} \triangle \mathcal{F}'|$ leads to the following result [Leclerc, 1994], where a median of $\mathcal{F}^*$ is a closure system $\mathcal{M} \in \mathbf{M}$ minimizing $\rho(\mathcal{M}, \mathcal{F}^*) = \sum_{1 \leq i \leq k} \partial(\mathcal{M}, \mathcal{F}_i)$.

Theorem. *For any profile $\mathcal{F}^*$ of* $\mathbf{M}$*, and any median $\mathcal{M}$ of $\mathcal{F}^*$, the inclusion $\mathcal{M} \subseteq c_{k/2}(\mathcal{F}^*)$ holds.*

In other terms, any class of a median closure system belongs to at least half of the closure systems of the profile. It is not difficult to see that this result remains valid when considering classification systems.

4 A fitting result based on implications and overhangings

Federation consensus functions $c_{\mathcal{K}}$ take only in account the presence or absence of classes in a qualified part of the elements of a profile. But it has been observed, in the case of hierarchies, that we have there a limitation which can prevent us to recognize common features in the elements of the profile, even evident ones. Moreover, there is a risk that a consensus based on presence of entire classes lacks of interest. For instance, if no untrivial class (other than the empty class, the singletons, and S), appears in at least half of the elements of a profile, the approaches evoked in the previous section lead to a consensus classification system with only the trivial classes, that is providing no information. For reasons of this type, [Adams III, 1986] developed a consensus method on hierarchies based on intersection of classes, and caracterized it in terms of the overhanging orders (called there nestings) associated to the involved hierarchies. The following result is a generalization of an Adams one. It concerns the more general problem of the fitting of an overhanging order to a given binary relation Ξ on $\mathcal{P}(S)$. The only condition on Ξ is: $(A, B) \in \Xi$ implies $A \subset B$.

For the proof of the next results, we need some further definitions on lattices, especially those of closed sets. First, given two closed sets C, C' in a closure system $\mathcal{F}$, C is *covered by* C' (denoted by $C \prec C'$) if, for any $C'' \in \mathcal{F}$, $C \subseteq C'' \subseteq C'$ implies $C'' = C$ or $C'' = C'$. A closed set C is *meet irreducible* if it is covered by a unique closed set C^+ in $\mathcal{F}$. These meet-irreducibles generate the whole closure system $\mathcal{F}$, in the sense that every $C \in \mathcal{F}$ is obtained as an intersection of such elements. Now, the covering relation of the closure system $\mathbf{M}$ is characterized as follows: for $\mathcal{F}, \mathcal{F}' \in \mathbf{M}$, $\mathcal{F} \prec \mathcal{F}'$ if and only if $\mathcal{F} = \mathcal{F}' - \{C\}$ for some meet-irreducible C of $\mathcal{F}'$ (cf. [Caspard and Monjardet, 2003]).

Consider the following two properties of a closure system $\mathcal{F}$ and its overhanging order Œ:

(AΞ1) $\Xi \subseteq$ Œ, (preservation of Ξ)

(AΞ2) for any meet-irreducible C of $\mathcal{F}$, $(C, C^+) \in \Xi$. (qualified overhangings)

Theorem. *Let $\mathcal{F}, \mathcal{F}' \in$* $\mathbf{M}$*. If both $\mathcal{F}$ and $\mathcal{F}'$ satisfy Conditions* (AΞ1) *and* (AΞ2)*, then $\mathcal{F} = \mathcal{F}'$.*

Proof. Observe first that the set S is in both $\mathcal{F}$ and $\mathcal{F}'$. If $\mathcal{F} \neq \mathcal{F}'$, the symmetric difference $\mathcal{F} \triangle \mathcal{F}'$ is not empty. Let C be a maximal class in $\mathcal{F} \triangle \mathcal{F}'$. Then, $C \neq S$ and it may be assumed without loss of generality that C belongs to $\mathcal{F}$ (and, so, C does not belong to $\mathcal{F}'$). If C was not a meet-irreducible element of $\mathcal{F}$, it would be an intersection of meet-irreducibles, all belonging to both $\mathcal{F}$ and $\mathcal{F}'$ and, so, C would belong to $\mathcal{F}'$.

Thus, C is a meet-irreducible, covered by a unique element C^+ of $\mathcal{F}$, with $C^+ \in \mathcal{F}'$. By (AΞ2), $(C, C^+) \in \Xi$ and, by (AΞ1), $C \ \text{Œ}'\ C^+$ (where $\text{Œ}'$ is the overhanging order associated to $\mathcal{F}'$). Set $C' = \varphi'(C)$ (where φ' is the closure operator associated to $\mathcal{F}'$). We have $C \subset C'$, since $C \in \mathcal{F}'$, and $C' \ \text{Œ}' C^+$, since $C' = \varphi'(C) = \varphi'(C') \subset \varphi'(C^+) = C^+$. But, according to the hypotheses, $C \subset C'$ implies $C' \in \mathcal{F}$, with $C \subset C' \subset C^+$, a contradiction with the hypothesis that C^+ covers C in $\mathcal{F}$.

In the particular case where $\mathcal{F}_1, \mathcal{F}_2, ..., \mathcal{F}_k$ are hierarchies on S, and $\Xi = \bigcap_{1 \leq i \leq k} \text{Œ}_i$ (where, for all $i = 1, ..., k$, Œ_i is the overhanging/nesting order associated with $\mathcal{F}_i$), we find a result implying the caracterization by Adams of his consensus method:

Corollary 1. *With the relation Ξ defined above, the Adams consensus hierarchy is the only closure system satisfying conditions* (AΞ1)*and* (AΞ2).

It is worth noticing that Adams results point out a case where it actually exists an overhanging order Œ satisfying conditions (AΞ1) and (AΞ2). Another case appears in [Semple and Steel, 2000] in the reseach of a "supertree". We exhibit other such cases in a work in preparation (for instance when Ξ is a relation satisfying conditions (O1) and (O2)). We end by the following result, where the solution to (AΞ1) and (AΞ2) appears, when it exists, to be actually an approximation of the given relation Ξ.

Corollary 2. *Let Ξ be a binary relation on $\mathcal{P}(S)$ and* Œ *an overhanging order satisfying conditions* (AΞ1) *and* (AΞ2). *Then, for any overhanging order* $\text{Œ}'$*, the inclusions* $\Xi \subseteq \text{Œ}' \subseteq \text{Œ}$ *imply* $\text{Œ}' = \text{Œ}$.

Proof. Assume $\Xi \subseteq \text{Œ}' \subset \text{Œ}$. Equivalently, if $\mathcal{F}'$ and $\mathcal{F}$ are the closure systems associated, respectively, to $\text{Œ}'$ and to Œ, there exists a meet irreducible C of $\mathcal{F}$ such that $\mathcal{F}' \subseteq \mathcal{F} - \{C\}$. It follows that $(C, C^+) \notin \text{Œ}'$, whereas, according to (AΞ2), $(C, C^+) \in \Xi$. This is a contradiction with the hypothesis $\Xi \subseteq \text{Œ}'$.

In the talk, we present examples where the data consist of a profile $\mathcal{F}^*$ of classification systems. In particular, profiles of hierarchies or phylogenies are considered. Now the above results prompt us to start from a relation Ξ obtained as another function of the Œ_i's than intersection. We are then able to obtain a consensus classification system which preserve more information from the profile than the Adams one, but is no longer a hierarchy.

References

[Adams III, 1972]E.N. Adams III. Consensus techniques and the comparison of taxonomic trees. *Systematic zoology*, pages 390–397, 1972.

[Adams III, 1986]E.N. Adams III. N-trees as nestings: complexity, similarity and consensus. *Journal of Classification*, pages 299–317, 1986.

[Armstrong, 1974]W.W. Armstrong. Dependency structures of data base relationships. *Information Processing*, pages 580–583, 1974.

[Barthélemy and Leclerc, 1995]J.P. Barthélemy and B. Leclerc. The median procedure for partitions. In Cox I.J., P. Hansen, and B. Julesz, editors, *Partitioning data sets*, pages 3–34, 1995.

[Barthélemy and M.F., 1991]J.P. Barthélemy and Janowitz M.F. A formal theory of consensus. *SIAM Journal on Discrete Mathematics*, pages 305–322, 1991.

[Barthélemy *et al.*, 1986]J.P. Barthélemy, B. Leclerc, and B. Monjardet. On the use of ordered sets in problems of comparison and consensus of classifications. *Journal of Classification*, pages 187–224, 1986.

[Caspard and Monjardet, 2003]N. Caspard and B. Monjardet. The lattices of moore families and closure operators on a finite set: a survey. *Discrete Applied Mathematics*, pages 241–269, 2003.

[Day and McMorris, 2003]W.H.E. Day and F.R. McMorris. *Axiomatic Consensus Theory in Group Choice and Biomathematics*. SIAM, Philadelphia, 2003.

[Domenach and Leclerc, 2004]F. Domenach and B. Leclerc. Closure systems, implicational systems, overhanging relations and the case of hierarchical classification. *Mathematical Social Sciences*, pages 349–366, 2004.

[Domenach and Leclerc, 2004b]F. Domenach and B. Leclerc. Consensus of classification systems, with adams' results revisited. In D. Banks, L. House, F.R. McMorris, P. Arabie, and W. Gaul, editors, *Classification, Clustering and Data Mining Applications*, pages 417–428, 2004b.

[Leclerc, 1994]B. Leclerc. Medians for weight metrics in the covering graphs of semilattices. *Discrete Applied Mathematics*, pages 281–297, 1994.

[Leclerc, 1998]B. Leclerc. Consensus of classifications: the case of trees. In A. Rizzi, M. Vichi, and H.-H. Bock, editors, *Advances in Data Science and Classification*, pages 81–90, 1998.

[Leclerc, 2003]B. Leclerc. The median procedure in the semilattice of orders. *Discrete Applied Mathematics*, pages 285–302, 2003.

[Margush and McMorris, 1981]T. Margush and F.R. McMorris. Consensus n-trees. *Bulletin of Mathematical Biology*, pages 239–244, 1981.

[Mirkin, 1975]B. Mirkin. On the problem of reconciling partitions. In *Quantitative Sociology, International Perspectives on mathematical and Statistical Modelling*, pages 441–449, 1975.

[Monjardet and Raderanirina, 2004]B. Monjardet and V. Raderanirina. Lattices of choice functions and consensus problems. *Social Choice and Welfare*, pages 349–382, 2004.

[Monjardet, 1990]B. Monjardet. Arrowian characterization of latticial federation consensus functions. *Mathematical Social Sciences*, pages 51–71, 1990.

[Raderanirina, 2001]V. Raderanirina. *Treillis et agrégation de familles de Moore et de fonctions de choix, Ph.D. Thesis.* Université Paris 1, Paris, 2001.

[Régnier, 1965]S. Régnier. Sur quelques aspects mathématiques des problèmes de classification automatique. *ICC Bulletin*, pages 175–191, 1965.

[Régnier, 1983]S. Régnier. Sur quelques aspects mathématiques des problèmes de classification automatique (seconde publication). *Mathématiques et Sciences humaines*, pages 13–29, 1983.

[Semple and Steel, 2000]C. Semple and M.A. Steel. A supertree method for rooted trees. *Discrete Applied Mathematics*, pages 147–158, 2000.

Comparison of distance indices between partitions

L. Denoeud[1][2], H. Garreta[3], and A. Guénoche[4]

[1] École nationale supérieure des télécommunications, 46, rue Barrault, 75634 Paris cedex 13 (e-mail: denoeud@infres.enst.fr)
[2] CERMSEM CNRS-UMR 8095, MSE, Université Paris 1 Panthéon-Sorbonne, 106-112, boulevard de l'Hôpital, 75647 Paris cedex 13
[3] Laboratoire d'Informatique Fondamentale, 163, avenue de Luminy, 13009 Marseille (e-mail: garreta@lif.univ-mrs.fr)
[4] Institut de Mathématiques de Luminy, 163, avenue de Luminy, 13009 Marseille (e-mail: guenoche@iml.univ-mrs.fr)

Abstract. In this paper, we compare several distance indices between partitions on the same set. First, we build a set $\mathcal{P}_k(P)$ of partitions close to each others by applying to an initial partition P, k transfers of one element from its class to another. Then we compare the distributions of several indices of distance between partitions of $\mathcal{P}_k(P)$.

Keywords: distance index, partition.

1 Introduction

The comparison of partitions is a central topic in clustering, as well for comparing partitioning algorithms as for classifying nominal variables. The literature abounds in indices defined by multiple authors to compare two partitions P and Q on the same set X. The most used are: the Rand index [Rand, 1971], the Jaccard index and the Rand index corrected for chance by Hubert and Arabie [Hubert and Arabie, 1985]. We also wanted to study the Wallace index [Wallace, 1983] and the normalized index of Lerman [Lerman, 1981]. The comparison of these indices is only interesting (in a practical point of view) if we consider close partitions, which differ randomly one from each others as it is mentioned by Youness and Saporta [Youness and Saporta, 2004]. They generate such partitions according to the *latent class model* [Bartholomew and Knott, 1999] adapted to an euclidian representation of the elements of X. We develop here a more general approach, independent of the representation space of X.

In 1964, Régnier proposed a distance between partitions which fits this type of study [Régnier, 1964]. It is the minimum number of transfers of one element from its class to another (eventually empty) to turn P into Q. We have recently studied this measure [Charon *et al.*, 2005] and called it the *transfer distance*. We compare the distributions of the distance indices above on partitions at k transfers from P. If k is small enough, these partitions are close to P since they represent only a small percentage α of all the partitions of X. This permits to define the value k_α of the maximum number of transfers allowed, and to build the set $\mathcal{P}_{k_\alpha}(P)$ of random partitions obtained by at most k_α transfers from P.

2 The transfer distance

Let P and Q be two partitions on the set X of n elements with respectively p and q classes ; we will admit that $p \leq q$.

$$P = \{C_1, .., C_p\} \text{ and } Q = \{C'_1, .., C'_q\}.$$

The minimum number of transfers to turn P into Q, denoted $\theta(P, Q)$, is obtained by establishing a bijection between the classes of P and those of Q keeping a maximum number of elements in matching classes, those that don't need to be moved. Consequently, we begin to add $q - p$ empty classes to P, so that P is considered as a partition with q classes.

Let Υ be the mapping from $P \times Q \longrightarrow \mathbb{N}$ which associates to one pair of classes the cardinal of their intersection. Classically, $n_{i,j} = |C_i \cap C'_j|$ and $n_i = |C_i|$ and $n'_j = |C'_j|$ denote the cardinals of the classes. Let Δ be the mapping which associates to each pair of classes (C_i, C'_j) the cardinal of their symmetrical difference, noted $\delta_{i,j}$. We have $\delta(i, j) = n_i + n'_j - 2 \times n_{i,j}$. So we consider the complete bipartite graph $K_{q,q}$ whose vertices are the classes of P and Q, with edges weighted either by Υ or by Δ.

Proposition 1 ([Day, 1981]) *The bijection minimizing the number of transfers between two partitions with q classes P and Q corresponds to a matching of maximum weight w_1 in $K_{q,q}$ weighted by Υ or, equivalently, to a matching of minimum weight w_2 in $K_{q,q}$ weighted by Δ; moreover, $\theta(P, Q) = n - w_1 = \frac{w_2}{2}$.*

Establishing the bipartite graph is in $O(n^2)$. The weighted matching problem in a complete bipartite graph can be solved by an assignment method well known in operational research [Kuhn, 1955], [Kuhn, 1956]. The algorithm has a polynomial complexity in $O(q^3)$. We won't go into further details, given for instance in [Faure *et al.*, 2000]. A computer program (in C) can be requested to the authors. We just develop an example of computation of the transfer distance.

Example 1 *We consider the two partitions $P = (1, 2, 3|4, 5, 6|7, 8)$ and $Q = (1, 3, 5, 6|2, 7|4|8)$. The two following tables correspond to the intersections and to the symmetrical differences of the classes of P and Q. Two extreme matchings are edited in bold. Each one gives $\theta(P, Q) = 4$.*

To the maximum weighted matching in the table Υ corresponds the series of 4 transfers: $(1, 2, 3|4, 5, 6|7, 8) \rightarrow (1, 3|4, 5, 6|2, 7, 8) \rightarrow (1, 3, 5|4, 6|2, 7, 8) \rightarrow (1, 3, 5, 6|4|2, 7, 8) \rightarrow (1, 3, 5, 6|4|2, 7|8)$.

Υ	1,3,5,6	2,7	4	8	Δ	1,3,5,6	2,7	4	8
1,2,3	**2**	1	0	0		3	**3**	4	4
4,5,6	2	0	**1**	0		**3**	5	2	4
7,8	0	**1**	0	1		6	2	3	**1**
$\emptyset$	0	0	0	**0**		4	2	**1**	1

To the minimum weighted matching in the table Delta corresponds another optimal series: $(1,2,3|4,5,6|7,8) \rightarrow (1,2,3,7|4,5,6|8) \rightarrow (2,3,7|1,4,5,6|8) \rightarrow (2,7|1,3,4,5,6|8) \rightarrow (2,7|1,3,5,6|8|4)$.

3 Close partitions in terms of transfers

We note $\mathcal{P}_n$ the set of partitions on a set of n elements and $\mathcal{P}_k(P)$ the set of partitions at k transfers from P and $\mathcal{P}_{\leq k}(P)$ the set of partitions at at most k transfers from P.

$$\mathcal{P}_k(P) = \{Q \in \mathcal{P}_n \text{ such that } \theta(P,Q) = k\}$$

$$\mathcal{P}_{\leq k}(P) = \{Q \in \mathcal{P}_n \text{ such that } \theta(P,Q) \leq k\} = \bigcup_{0 \leq i \leq k} \mathcal{P}_i(P)$$

Statistically, we consider that a partition Q is close to P at threshold α if the probability of observing a partition closer to P than $\theta(P,Q)$ is lower than or equal to α. The matter is then to know how many partitions are within a k radius from P. For $k = 0$, there is just one partition, P itself, otherwise θ would'nt be a distance. We can easily enumerate $\mathcal{P}_1(P)$, but for larger k it becomes difficult. We call *critical value* of the partition P, at threshold α, the greatest number of transfers k_α such as

$$\frac{|\mathcal{P}_{\leq k_\alpha}(P)|}{|\mathcal{P}_n|} \leq \alpha.$$

While $n \leq 12$, we can enumerate all the partitions in $\mathcal{P}_n$ and we compute $|\mathcal{P}_k(P)|$. For that, we use the procedure NexEqu in [Nijenhuis and Wilf, 1978]. Each partition is coded by the vector of the class number to which each element belongs. The algorithm builds the next partition for the lexicographic order on this code, starting from the partition with a single class.

For $n > 12$, there are too many partitions to realize an exhaustive enumeration. Then we select at random a large number of partitions, to be compared to P to estimate $|\mathcal{P}_{\leq k}(P)|/|\mathcal{P}_n|$. To obtain a correct result, the partitions must be equiprobable; the book of Nijenhuis and Wilf provides also such a procedure (RandEqu).

Thus we measure a frequency f in order to estimate a proportion p. We want to approximate $p = 0.1$ for a risk ρ fixed ($\rho = 5\%$) and a gap δ between

f and p judged as acceptable ($\delta = 0.01$). For these values, we can establish the size of the sample E by the classical formula:

$$t(\rho)\sqrt{\frac{f(1-f)}{|E|}} \leq \delta$$

in which $t(\rho)$ is given by the normal distribution of Gauss [Brown *et al.*, 2002]. We obtain that 3600 trials should be carried out, which are quite feasible. We can notice that this number decreases with p (when $p < 0.5$) and it is independent of n.

Example 2 *For* $n = 12$*, there are* $|\mathcal{P}_{12}| = 4213597$ *partitions that can be compared to* P *in order to establish the distribution of* $|\mathcal{P}_k(P)|$ *according to* k*. For* $P = \{1,2,3,4|5,6,7|8,9|10,11|12\}$*, as for all the partitions with classes having the same cardinality, the number of partitions at* $0,\ldots,8$ *transfers from* P *are respectively* 1*,* 57*,* 1429*,* 20275*,* 171736*,* 825558*,* 1871661*,* 1262358*,* 60522 *and 0 beyond. The cumulated proportions in % are respectively* 0.0*,* 0.0*,* 0.0*,* 0.5*,* 4.6*,* 24.2*,* 68.6*,* 99.6*,and* 100*. For* $\alpha = .1$ *the critical value is* 4*; indeed there are just* 4.6% *of the partitions that are at most at* 4 *transfers from* P*, while for* 5 *transfers, there are* 24.2%*. The cumulated frequencies computed from* P *and 5000 random partitions are:* $0.0, 0.0, 0.1, 0.5, 4.4, 23.9, 68.7, 98.3$ *and* 100*. Thus the critical value computed by sampling is also equal to* 4*.*

4 Indices of proximity between partitions

The comparison of partitions is based on the pairs of elements of X. Two elements x and y can be joined together or separated in P and Q. The two partitions agree on (x,y) if these elements are simultaneously joined or separated in P and Q. On the other hand there is a disagreement if x and y are joined in one of them and separated in the other. Let r be the number of pairs simultaneously joined together, s the number of pairs simultaneously separated, an u (resp. v) the number of pairs joined (resp. separated) in P and separated (resp. joined) in Q.

According to the previous notations, we have $r = \sum_{i,j} \frac{n_{i,j}(n_{i,j}-1)}{2}$. Equivalent formulas for s, u and v appear in several papers. We will note $\pi(P)$ the set of joined pairs in P, that is to say $|\pi(P)| = \sum_{i=1,p} \frac{n_i(n_i-1)}{2}$.

4.1 The Rand index

The Rand index [Rand, 1971], noted R, is simply the percentage of pairs for which there is an agreement. It belongs to $[0,1]$ and $1 - R(P,Q)$ is the symmetrical difference distance between $\pi(P)$ and $\pi(Q)$.

$$R(P,Q) = \frac{r+s}{n(n-1)/2}$$

4.2 The Jaccard index

In the Rand index, the pairs simultaneously joined or separated are counted in the same way. However, partitions are often interpreted as classes of joined elements, the separations being the consequences of this clustering. We use then the Jaccard index (1908), noted J, which does not take into account the s simultaneous separations:

$$J(P, Q) = \frac{r}{r + u + v}$$

4.3 The corrected Rand index

In their paper of 1985 [Hubert and Arabie, 1985], they noticed that the Rand index is not *corrected for chance* that is equal to zero for random partitions having the same number of objects in each class. They introduced the corrected Rand index, whose expectation is equal to zero, noted here HA, in homage to the authors.

The corrected Rand index is based on three values: the number r of common joined pairs in P and Q, the expected value $Exp(r)$ and the maximum value $Max(r)$ of this index, among the partitions of the same type as P and Q. It leads to the formula

$$HA(P, Q) = \frac{r - Exp(r)}{Max(r) - Exp(r)}$$

with $Exp(r) = \frac{|\pi(P)| \times |\pi(Q)|}{n(n-1)/2}$ and $Max(r) = \frac{1}{2}(|\pi(P)| + |\pi(Q)|)$. This maximum value is questionable since the number of common joined pairs is necessarily bounded by $\inf\{|\pi(P)|, |\pi(Q)|\}$, but $Max(r)$ insures that the maximum value of HA is 1 when the two partitions are identical. On the other hand this index can take negative values.

4.4 The Wallace index

This index is very natural, it's the number of joined pairs common to P and Q divided by the number of possible pairs [Wallace, 1983]. This last quantity depends on the partition of reference and, if we don't want to favour neither P nor Q, the geometrical average is used.

$$W(P, Q) = \frac{r}{sqrt(|\pi(P)| \times |\pi(Q)|)}$$

.

4.5 The normalized Lerman index

The Lerman index(denoted ICL) is the difference between the number of simultaneously joined pairs and its expectation, divided by its standard deviation [Lerman, 1988].

$$ICL(P,Q) = \frac{r - Exp(r)}{\sqrt{Var(r)}}$$

These two values are computed on the set of pairs of partitions having the same types as P and Q; they are defined according to the cardinals of the classes. The expected value of r already appears in the formula given by Hubert and Arabie and its variance $Var(r)$ is given by:

$$\frac{V_1(P)V_1(Q)}{2n(n-1)} + \frac{V_2(P)V_2(Q)}{n(n-1)(n-2)} + \frac{V_3(P)V_3(Q)}{4n(n-1)(n-2)(n-3)} - [\frac{V_1(P)V_1(Q)}{2n(n-1)}]^2$$

where $V_1(P) = \sum_{i=1,p} n_i(n_i - 1)$, $V_2(P) = \sum_{i=1,p} n_i(n_i - 1)(n_i - 2)$ and

$$V_3(P) = [\sum_{i=1,p} n_i(n_i - 1)]^2 - 2\sum_{i=1,p} n_i(n_i - 1)(2n_i - 3)],$$

with similar expressions for $V_1(Q)$, $V_2(Q)$ and $V_3(Q)$, in which the sums are computed on q classes and the n_i are replaced by n'_i.

The index value is not defined when $Var(r) = 0$, that is when one of the partitions has a single class or n singletons. As for the HA index, it can be negative, but it is not upper bounded. Finally, Lerman proposes a normalized index defined as a correlation coefficient given by the formula:

$$ILN(P,Q) = \frac{ICL(P,Q)}{\sqrt{ICL(P,P) \times ICL(Q,Q)}}$$

5 Comparison of indices

Let P be a partition on X with p classes, defined by its type, that is to say by the cardinal of its classes. When $n = |X| \leq 12$, we enumerate the sets $\mathcal{P}_k(P)$, then we evaluate the minimum and maximum values of each index above between P and any Q belonging to $\mathcal{P}_k(P)$. The table 1 contains the results for $P = (1,2,3,4,5|6,7,8,9,10)$. The partitions being at at most 3 transfers represent 1.7% of the 115975 partitions on 10 elements.

One can observe that, for each index, the maximum value obtained for partitions at 5 transfers are greater than the minimum value obtained for 2 transfers. Moreover the minimum values at 3 transfers are very small and don't reflect the closeness of these partitions and P. Finally, for the normalized Lerman index, the maximum values do not decrease with k and the closest partition from P is at 4 transfers.

Nb. of transfers	1	2	3	4	5	6	7	8
Nb. of partitions	20	225	1720	9112	31361	54490	17500	1546
J min	.64	.43	.32	.22	.15	.08	.04	0.0
J max	.80	.70	.60	.50	.44	.21	.10	0.0
R min	.80	.64	.53	.47	.44	.44	.44	.44
R max	.91	.87	.82	.78	.69	.64	.60	.56
HA min	.60	.28	.06	-.08	-.12	-.17	-.19	-.22
HA max	.82	.72	.63	.53	.32	.22	.11	0.0
W min	.78	.60	.49	.37	.28	.16	.09	0.0
W max	.89	.84	.77	.71	.67	.45	.32	0.0
ILN min	.61	.28	.06	-.08	-.20	-.20	-.23	-.32
ILN max	.86	.84	.95	1.15	.67	.39	.25	-.14

Table 1. Distribution of the number of partitions at k transfers from P and extreme values of the distance indices

.

In the case $n > 12$, we cannot enumerate $\mathcal{P}_n$ anymore. Then, in order to compare very close partitions in the neighborhood of a given partition P,

- we compute by sampling the critical number of transfers $k_{5\%}$;
- we build a set $\mathcal{Q}_k(P)$ of 100 partitions Q randomly selected such as $\theta(P, Q) \leq k$, with $k \leq k_{5\%}$;
- we compare all the partitions of $\mathcal{Q}_k(P)$ two by two and measure the average value and the standard deviation of each studied index.

The partitions close to P are obtained by selecting recursively at random one element; if this element is not alone in its class, its new class number is selected between 1 and $p+1$, and the number of classes is updated. Here, we restrict our study at the single partition of 100 elements spread in 5 balanced classes of 20 elements each. The critical value at 5% is 83, that is to say that only 5% of the partitions with 100 elements are at less at 83 transfers from the balanced partition with 5 classes.

The figure 1 represents the computed averages and standard deviations of each index for $k \in [5; k_\alpha]$, with a step of 5.

We can see that the indices decrease when k increases since the partitions are less close to each other. The indices of Jaccard, corrected Rand , Wallace, and Lerman have approximately the same behavior: they are high when k is small and decrease near to 0 when $k = k_\alpha$. But they reflect the closeness of partitions only when k is very small. Among these indices the Jaccard index seems to be the most accurate since it has the lowest standard deviation. The Rand index has a different behavior: its values stays above 0,8 whatever is k. Two pairs of partitions at 40 and 90 transfers from each others can have the same value.

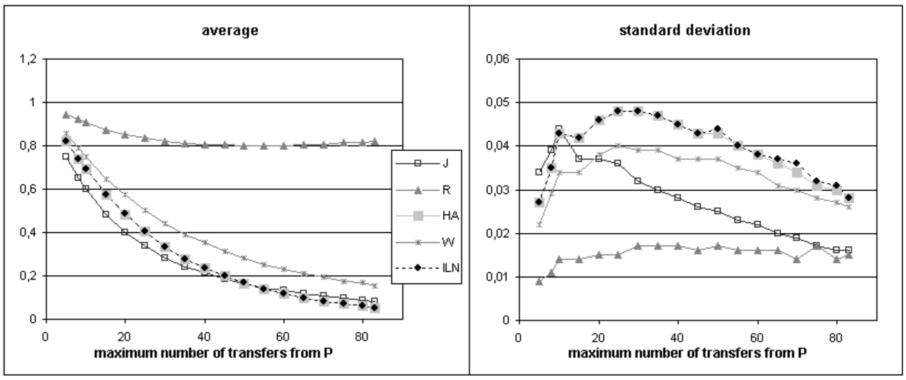


Fig. 1. Average and standard deviation of the distance indices between partitions of $\mathcal{Q}$

We have obtained the same kind of results for other initial partitions, balanced or not. Our conclusion is that the Rand index isn't very satisfying for the comparison of close partitions. Among the others, the Jaccard index seems the best, followed by the Wallace index, because they have the lowest standard deviation. The corrected Rand index and the normalized Lerman index share similar average values but the extreme values of the normalized Lerman index make it less satisfying.

Acknowledgements This work is supported by the CNRS ACI IMP-Bio. We also want to thank B. Fichet (University of Aix-Marseille II) and I.C. Lerman (University of Rennes I) for their help and advices.

References

[Bartholomew and Knott, 1999]D. Bartholomew and M. Knott. *Latent Variables Models and Factor Analysis.* Arnold, London, 1999.

[Brown *et al.*, 2002]L. Brown, T. Cai, and A. DasGupta. Confidence intervals for a binomial proportion and asymptotic expansions. *Ann. Statist.*, pages 160–201, 2002.

[Charon *et al.*, 2005]I. Charon, L. Denoeud, A. Guénoche, and Hudry O. Comparing partitions by element transfers. *submitted*, 2005.

[Day, 1981]W. Day. The complexity of computing metric distances between partitions. *Mathematical Social Sciences*, pages 269–287, 1981.

[Faure *et al.*, 2000]R. Faure, B. Lemaire, and C. Picouleau. Précis de recherche opérationnelle. *Mathematical Social Sciences*, pages 134–137, 2000.

[Hubert and Arabie, 1985]L. Hubert and P. Arabie. Comparing partitions. *J. of Classification*, pages 193–218, 1985.

[Kuhn, 1955]H.W. Kuhn. The hungarian method for the assignment problem. *Naval Res. Logist. Quart.*, pages 83–97, 1955.

[Kuhn, 1956]H.W. Kuhn. Variants on the hungarian method for the assignment problem. *Naval Res. Logist. Quart.*, pages 253–258, 1956.

[Lerman, 1981]I.C. Lerman. *Classification et analyse ordinale des données.* Dunod, Paris, 1981.

[Lerman, 1988]I.C. Lerman. Comparing partitions (mathematical and statistical aspects). In H.H Bock, editor, *Classification and Related Methods of Data Analysis*, pages 121–131, 1988.

[Nijenhuis and Wilf, 1978]A. Nijenhuis and H. Wilf. *Combinatorial algorithms.* Academic Press, New-York, 1978.

[Rand, 1971]W.M. Rand. Objective criteria for the evaluation of clustering methods. *J. of the Am. Stat. Association*, pages 846–850, 1971.

[Régnier, 1964]Régnier. Sur quelques aspects mathématiques des problèmes de classification automatique. *ICC Bulletin*, 1964.

[Wallace, 1983]D.L. Wallace. Comment. *J. of the Am. Stat. Association*, pages 569–579, 1983.

[Youness and Saporta, 2004]G. Youness and G. Saporta. Une méthodologie pour la comparaison des partitions. *Revue de Statistique Appliquée*, pages 97–120, 2004.

Kernel Logistic PLS: a new tool for complex classification

Arthur Tenenhaus[1,2], Alain Giron[1], Gilbert Saporta[3], and Bernard Fertil[1]

[1] INSERM U678, CHU Pitié-Salpêtrière, Paris, France
(e-mail: arthur.tenenhaus@imed.jussieu.fr,
alain.giron@imed.jussieu.fr,
bernard.fertil@imed.jussieu.fr)
[2] KXEN research, Suresnes, France
[3] CNAM – Conservatoire National des Arts et Métiers, France
(e-mail: saporta@cnam.fr)

Abstract. "Kernel Logistic PLS" (KL-PLS), a new tool for classification with performances similar to the most powerful statistical methods is described in this paper. KL-PLS is based on the principles of PLS generalized regression and learning via kernel. The successions of simple regressions, simple logistic regression and multiple logistic regressions on a small number of uncorrelated variables that are computed within KL-PLS algorithm are convenient for the management of very high dimensional data. The algorithm was applied to a variety of benchmark data sets for classification and in all cases, KL-PLS demonstrates its competitiveness with other state-of-art classification method. Furthermore, leaning on statistical tests related to the logistic regression, KL-PLS allows the systematic detection of data points close to "support vectors" of SVM and thus reduces the computational charges of the SVM training algorithm without significant loss of accuracy.

Keywords: Classification, Kernel, PLS Generalized Regression.

1 Introduction

Given a set of labeled experiments $\{(x_i, y_i)\}_{i=1,\ldots,n}$, $x_i \in \mathbb{R}^{p\times 1}$ and $y_i \in \{-1, 1\}$, we would like to build a prediction rule which, based on the observations, allows a prediction of the label y_{new} of a new point x_{new}. The following notation is used throughout this paper: each data point x_i (respectively each response y_i) represents the i^{th} row of the data matrix X (respectively the i^{th} row of the column vector Y). In order to handle the "generally" high dimensionality of the input space, we propose to exploit principles of the Partial Least Square regression (PLS) [Wold *et al.*, 1982, Tenenhaus, 1998]. PLS regression creates a set of orthogonal latent variables (PLS component) $t_1, t_2, \ldots, t_m$, linear combinations of the original variables but, contrary to principal component analysis (PCA), use the target Y for their determination. The PLS components t_h is obtained from the following constraints (Tucker criteria):

$$\max_{t_h} \operatorname{cov}^2(t_h, Y) \;=\; \max_{w_h} \operatorname{cov}^2(X_{h-1}w_h, Y)$$

$$\text{such that } \|w_h\| = 1 \text{ and } t_h \text{ is orthogonal to } t_1, \ldots, t_{h-1},$$

where $X_0 = X$ and X_{h-1} is the residual of the regression of X on $t_1, \ldots, t_{h-1}$.

A least square regression is then performed to relate Y to the PLS components.

But PLS was not originally designed as a tool for classification. Thus, based on the algorithmic structure of PLS regression, the PLS logistic regression was proposed for classification task [Tenenhaus, 2002, Bastien *et al.*, 2004].

PLS regression is designed to operate with input data that are high-dimensional and highly correlated (PLS is very popular in the chemometrics field), such a situation encountered by the use of kernel function [Schölkopf and Smola, 2002]. Based on kernel techniques, Rosipal and Trejo have proposed a nonlinear extension of PLS regression, the Kernel PLS regression (KPLS regression) [Rosipal and Trejo, 2001]. The approach was subsequently extended to the kernel orthonormalized PLS for classification problems [Rosipal *et al.*, 2003] using Barker and Rayens approach [Barker and Rayens, 2003].

In this paper we present a non linear extension via kernel of the PLS logistic regression: Kernel Logistic PLS (KL-PLS). Following Bennet and Embrechts who demonstrated interest of directly exploit kernel within the framework of PLS regression [Bennett and Embrechts, 2003] and noting the close connection between KPLS and PLS regression of Y on the kernel K, [Appendix 1], we propose an algorithm directly based on the factorization of the kernel matrix.

Furthermore, thanks to the statistical tests related to logistic regression, KL-PLS allows detecting points close to "support vectors" (points used by the Support Vector Machines (SVM) to compute the decision boundary). It is therefore possible to select a subset of the training set that is sufficient to derive the SVM decision boundary.

2 Kernel Logistic PLS (KL-PLS)

2.1 Algorithm

Principle of KL-PLS is to compute orthogonal latent variables in the space induced by the kernel matrix before performing logistic regression in the derived feature space. Therefore, KL-PLS is a 3-step algorithm:

1. **Computation of the kernel matrix**
   Let X be the matrix comprising the p explanatory variables x_k, $k = 1, \ldots, p$ and Y a binary variable (the target) observed on n samples. Let K be the kernel matrix associated to X. A usual kernel is given below:

$$\text{Gaussian kernel: } K(x_i, x_j) = \exp\left(- \frac{\|x_i - x_j\|^2}{2\sigma^2} \right)$$

The dimension of the kernel matrix is $n \times n$. Each cell k_{ij} is a measure of similarity between the individuals i and j.

2. **Computation of the KL-PLS components**

   2.1 ***Computation of the first KL-PLS component*** t_1

   Step 1: Compute the regression coefficient a_{1j} of k_j in the logistic regression of Y on k_j, $j = 1, \dots, n$
   Step 2: Normalize the column vector a_1 made by a_{1j}'s: $w_1 = a_1/\|a_1\|$
   Step 3: Compute the first KL-PLS component as $t_1 = Kw_1$

   2.2 ***Computation of the*** h^{th} ***KL-PLS component*** t_h

   Let assume that in the previous steps, the KL-PLS components $t_1, \dots, t_{h-1}$ have been yielded. This block is designed to get variables which, in addition to - and orthogonally to - $t_1, \dots, t_{h-1}$, hold residual information on Y. The h^{th} KL-PLS component is subsequently computed from the residual of the regression of k_j, $j = 1, \dots, n$ on $t_1, \dots, t_{h-1}$.

   Step 1: Compute the residual $e_{h1}, \dots, e_{hn}$ from the multiple regression of k_j, $j = 1, \dots, n$ on $t_1, \dots, t_{h-1}$. Let K_{h-1} be the matrix comprising $\epsilon_{h1}, \dots, \epsilon_{hn}$.
   Step 2: Compute the coefficients a_{hj} of e_{hj} in the logistic regression of Y on $t_1, \dots, t_{h-1}$ and e_{hj}.
   Step 3: Normalize the column vector a_h made by a_{hj}'s: $w_h = a_h/\|a_h\|$.
   Step 4: Compute the h^{th} PLS component: $t_h = K_{h-1}w_h$.
   Step 5: Express the component t_h in terms of K as $t_h = Kw_h^*$.

3. ***Logistic regression of*** Y ***on the*** m ***retained KL-PLS components***

$$P(Y = 1|K = k) \; = \; \frac{e^{\alpha_0 + \sum_{h=1}^{m} \alpha_h t_h}}{1 + e^{\alpha_0 + \sum_{h=1}^{m} \alpha_h t_h}} \, .$$

2.2 Remarks

2.3 Selection of the number of useful KL-PLS components

Computation of the KL-PLS component t_h may be simplified by setting non-significant regression coefficients a_{hj} to 0. Only variables that are significantly related to Y contribute to the computation of t_h. The number m of KL-PLS components to be retained may be chosen by cross-validation or by observing that the component t_{m+1} is not significant because none of the coefficients $a_{(m+1)j}$ is significantly different from 0.

2.4 Expression of KL-PLS component in term of original variables

Expression of PLS components in terms of original variables is a fundamental step to analyze new data. Indeed, let $Ktest$ be the new dataset. The matrix product $Ttest = Ktest \times W^*$ allows to compute the values of the KL-PLS components for the new dataset.

2.4.1 Computation of w_h^*

a. The first KL-PLS component is already expressed in terms of original variables : $t_1 = Kw_1$ and $w_1^* = w_1$.

b. The second KL-PLS component is expressed in terms of the residuals in the regression of the original variables on t_1. From $K = t_1p_1' + K_1$ and $t_2 = K_1w_2$ we get:

$$t_2 = K_1w_2 = (K - t_1p_1') = (K - Kw_1p_1')w_2 = K\underbrace{(I - w_1p_1')w_2}_{w_2^*} = Kw_2^*.$$

c. In a similar way, it can be shown that t_h is expressed in terms of the original variables as:

$$\begin{aligned} t_h = K_{h-1}w_h &= \left(K - \sum_{i=1}^{h-1} t_ip_i\right) \cdot w_h = \left(K - \sum_{i=1}^{h-1} Kw_i^*p_i'\right) \cdot w_h \\ &= K\underbrace{\left(I - \sum_{i=1}^{h-1} w_i^*p_i'\right) \cdot w_h}_{w_h^*} = Kw_h^*. \end{aligned}$$

3 Kernel Logistic PLS and detection of support vectors

3.1 Preliminary considerations

SVM was designed to find the "optimal separating hyperplane" i.e. the hyperplane whose minimal distance to the training examples is maximum (fig. 1) [Vapnik, 1998]. The optimal hyperplane is defined by a vector β and a scalar β_0 through the equation:

$$\arg\max_{\beta,\beta_0} \min\Big\{\|x - x_i\| : x \in \mathbb{R}^n,\ (x^t\beta + \beta_0) = 0,\ i = 1, \ldots, n\Big\}.$$

Points which "support" hyperplanes H_1 and H_2 are the "support vectors". Only support vectors take part in the construction of the SVM decision boundary. We propose an approach which is able to detect points, called "ambiguous points" thereafter, close to support vectors. This procedure is achieved by removing a subset of training examples with minimal impact on the SVM decision boundary position.

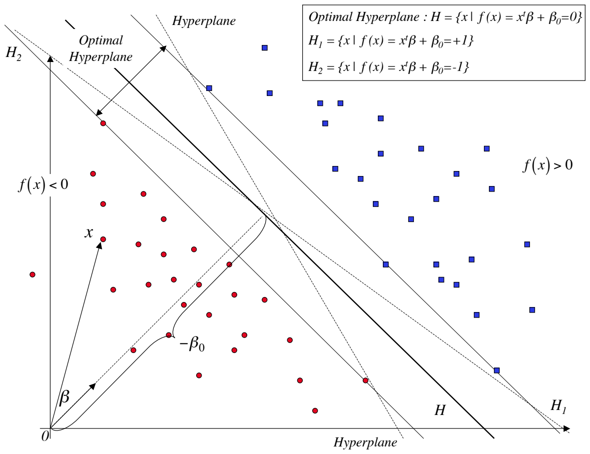

Optimal Hyperplane : $H = \{x \mid f(x) = x^t\beta + \beta_0 = 0\}$
$H_1 = \{x \mid f(x) = x^t\beta + \beta_0 = +1\}$
$H_2 = \{x \mid f(x) = x^t\beta + \beta_0 = -1\}$
Hyperplane
Optimal Hyperplane
H_2
$f(x) < 0$
$f(x) > 0$
x
$-\beta_0$
β
0
H
H_1
Hyperplane

Fig. 1. Optimal separating hyperplane.

3.2 Detection of ambiguous points

During the construction of the first KL-PLS component, coefficients a_{1j} of k_j in the logistic regression of Y on each k_j, $j = 1, \ldots, n$ are computed. If a point j is, on the average, closer to the points belonging to its own group than to the points belonging to the other group, then k_j has, on the average, a larger value (in the case of Gaussian kernel) for the individuals belonging to the group containing j than for the other individuals. We can expect the regression coefficient a_{1j} to be highly significant in this situation. Consequently, it is proposed to label points associated to non-significant a_{1j} to the risk α (Wald test) as ambiguous.

The number of ambiguous points can, subsequently be controlled by increasing the risk α.

4 Results

4.1 Banana data projection onto the two first components found by KL-PLS

Banana data is a 2D dataset (two classes). 400 × 2 training set is associated to a 4,600 × 2 testing set. Figure 2 depicts projection of the original training and testing data onto the two first components found by KL-PLS (training data). A nice linear separation of the two classes can be seen in the feature space and logistic regression is adequate to achieve an efficient classification.

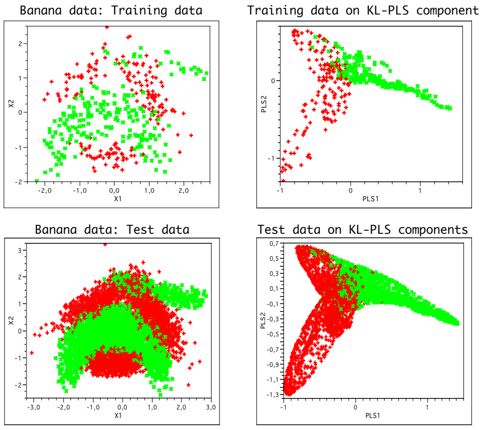


Fig. 2. Banana data depict onto the two first components found by kernel logistic PLS.

4.2 Benchmarks

The usefulness of KL-PLS was tested on several benchmark data sets (two-class classification) used in [Mika *et al.*, 1999] and [Rätsch *et al.*, 2001]. These datasets are available at http://ida.first.gmd.de/raetsch/data/benchmarks.htm. Each dataset consists of 100 different training and testing partitions. Several methods (KFD, SVM, KPLS-SVC) have already been used and results are presented in table 1. Baudat and Anouar have proposed a nonlinear extension of the Fisher Discriminant Analysis via "Kernel Trick": the Kernel Fisher Discriminant analysis (KFD) [Baudat and Anouar, 2000]. The kernel orthonormalized PLS + SVC (KPLS-SVC) is based on the kernel orthonormalized PLS method for dimensionality reduction followed by SVM on retained PLS components for classification [Rosipal *et al.*, 2003]. In all cases the Gaussian kernel was used. KL-PLS efficiency relies on the value of width of the Gaussian and the number of retained KL-PLS components Those values are selected based on the minimum classification error observed after five-fold cross validation on the first five training sets. Results of logistic regression (LR) are also presented. Results achieved for the 11 benchmarks demonstrate the efficiency of KL-PLS and its competitiveness with other state-of-the-art classification methods.

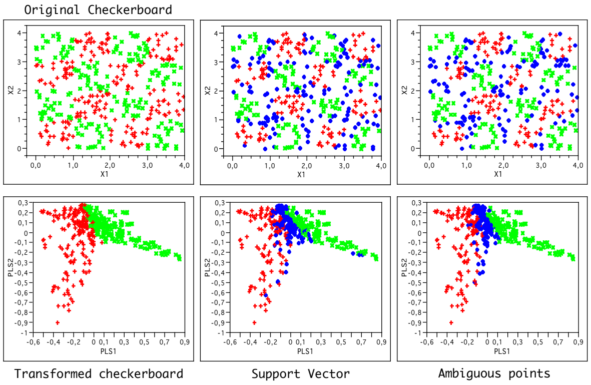
4.3 Ambigous points and support vectors

4.3.1 Simulated checkerboard A 4 × 4 checkerboard is represented in fig 3. Twenty-five uniformly points labeled according to checkerboard pattern

Table 1. Comparison of the mean and standard deviation classification errors (test set) for KFD [Mika *et al.*, 1999], SVM [Rätsch *et al.*, 2001], Kernel PLS-SVC [Rosipal *et al.*, 2003], Logistic Regression (LR) and KL-PLS. The last column provides the width of the Gaussian kernel and the number of retained KL-PLS components.

Data set	**KFD**	**SVM**	**KPLS-SVC**	**LR**	**KL-PLS**	**KL-PLS parameters**
Banana	10.8 ± 0.5	11.5 ± 0.5	10.5 ± 0.4	47.0 ± 4.48	10.7 ± 0.5	$(0.9, 10)$
B. Cancer	25.8 ± 4.6	26.0 ± 4.7	25.1 ± 4.5	27.5 ± 4.7	25.8 ± 4.4	$(50, 7)$
Diabetis	23.2 ± 1.6	23.5 ± 1.7	23.0 ± 1.7	23.3 ± 1.8	23.0 ± 1.7	$(60, 4)$
German	23.7 ± 2.2	23.6 ± 2.1	23.5 ± 1.6	24.0 ± 2.1	23.2 ± 2.1	$(20, 2)$
Heart	16.1 ± 3.4	16.0 ± 3.3	16.5 ± 3.6	16.9 ± 2.9	16.0 ± 3.2	$(20, 3)$
Ringnorm	1.49 ± 0.12	1.66 ± 0.12	1.43 ± 0.10	25.3 ± 0.8	1.44 ± 0.09	$(200, 2)$
F. Solar	33.2 ± 1.7	32.4 ± 1.8	32.4 ± 1.8	34.6 ± 3.7	32.7 ± 1.8	$(12, 1)$
Thyroid	4.20 ± 2.07	4.80 ± 2.19	4.39 ± 2.1	10.3 ± 2.7	4.35 ± 1.99	$(15, 6)$
Titanic	23.2 ± 2.06	22.4 ± 1.0	22.4 ± 1.1	22.7 ± 1.1	22.4 ± 0.04	$(300, 2)$
Twonorm	2.61 ± 0.15	2.96 ± 0.23	2.34 ± 0.11	3.81 ± 0.53	2.37 ± 0.10	$(40, 1)$
Waveform	9.86 ± 0.44	9.88 ± 0.43	9.58 ± 0.36	13.48 ± 0.7	9.74 ± 0.46	$(15, 4)$

was generated within each square. Fig. 3 depicts the projection of the 4 × 4 checkerboard from both classes onto the two first components found by KL-PLS. A nice separation of the two classes can be seen. Note that support vectors and ambiguous (blue circles) are pretty close.


Original Checkerboard

Transformed checkerboard

Support Vector

Ambiguous points

Fig. 3. Comparison between Support Vectors and Ambiguous Points (blue circles).

4.3.2 Selection of ambiguous points (banana data)

The SVM decision boundary only depends on the support vectors. In order to evaluate

Table 2. Confusion matrix between Support Vectors and ambiguous points.

	Ambiguous points	**Non ambiguous points**	**Total**
Support vector	112	36	148
Non support vector	23	229	252
Total	135	265	400

proximity between ambiguous points and support vectors, SVM was trained on "α-selected" ambiguous points. Results were compared to those obtain by SVM (full training set).

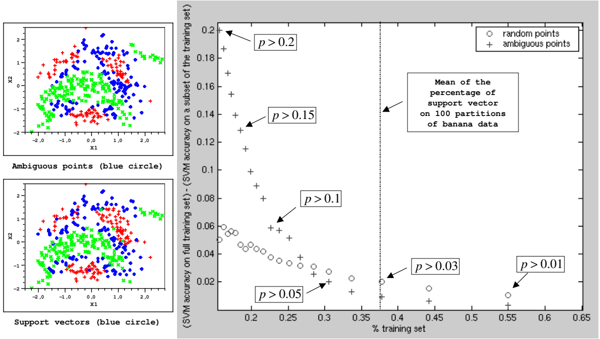


Fig. 4. Efficiency of SVM classification as function of the size of the training set. o - randomly selected points for the training set. + - points selected with respect to their significant level (p).

The following operations were carried out:

i. We compute the mean test set classification error based on SVM trained on full training set on the 100 partitions of banana data.
ii. For each $\alpha = \{0.01, 0.02, \ldots, 0.2\}$, KL-PLS was trained on the 100 partitions of banana data. It allows detection of ambiguous points for each partition. Then, SVM is trained on ambiguous points for each partition. We compute the mean test set classification error.
iii. For each $\alpha = \{0.01, 0.02, \ldots, 0.2\}$, SVM was trained on randomly selected points in the same proportion as the ambiguous points related to this value of α for each partition. We compute the mean test set classification error.

SVM train on ambiguous point gives performances similar to SVM train on full training set when the number of ambiguous points is close to the number of support vectors. Syed *et al.* have shown that the discarding of even a small proportion of the support vectors can lead to a severe reduction in generalization performance [Syed *et al.*, 1999]. They stated that this implies that the support vector set chosen by SVM is a minimal set; this can explain the behavior of the (blue - cross) curve (fig. 4) when considering low numbers of ambiguous points.

5 Discussion and conclusion

Performances of KL-PLS are equivalent to the most powerful classification methods such as SVM, KPLS-SVC or KFD. This algorithm is very simple to implement since it is solely composed of ordinary least square and logistic regressions. Furthermore, it is possible to compute KL-PLS components only by considering individual column vectors of the kernel matrix. These properties make possible to highlight 3 interests of KL-PLS:

a. KL-PLS does not require the full kernel matrix in memory but the columns of the kernel individually.
b. Inversions of small dimension matrices (number of KL-PLS components +1) take place in the algorithm.
c. The introduction of intercept when constructing the latent variables, avoid the kernel centering method proposed by Wu *et al.* [Wu *et al.*, 1997].

⇒ KL-PLS allows management of very high dimensional data.

Furthermore, direct factorization of the kernel matrix offers 2 advantages:

a. K does not need to be square
b. K does not need defining a dot product in the feature space induced by the "kernel trick". The Mercer's conditions (positive definite) are subsequently not required.

⇒ K just need to contain similarity measures.

Moreover, Kernel-PCA is often used as a preliminary step for dimensional reduction prior classification [Schölkopf *et al.*, 1998]. A more powerful goal-driven preprocessing is built in KL-PLS.

Lastly, leaning on Wald tests related to the logistic regression, it is possible to detect "ambiguous points" close to support vectors. This approach specifically selects examples from the training set close to support vectors. SVM computational charges are consequently reduced without jeopardizing classification.

Works in progress comprise the extension of KL-PLS approach to the multi-class classification problems, the study of the relationship between "ambiguous points" and "Support Vectors" and the extension of Kernel Logistic PLS to the kernel generalized PLS via generalized linear model.

Acknowledgment This work was supported by KXEN, a global analytic software company.

References

[Barker and Rayens, 2003]M. Barker and W. S Rayens. Partial least square for discrimination. *Journal of Chemometrics*, pages 166–173, 2003.

[Bastien *et al.*, 2004]P. Bastien, V. E. Vinzi, and M Tenenhaus. Pls generalized linear regression. *Computational Statistics & data analysis*, 2004.

[Baudat and Anouar, 2000]G. Baudat and F Anouar. Generalized discriminant analysis using a kernel approach. *Neural Computation*, pages 2385–2404, 2000.

[Bennett and Embrechts, 2003]K. P. Bennett and M. J Embrechts. An optimization perspective on kernel partial least squares regression, advances in learning theory: Methods, models and applications. *NATO Sciences Series III: Computer & Systems Sciences*, pages 227–250, 2003.

[Höskuldsson, 1988]A Höskuldsson. Pls regression methods. *Journal of Chemometrics*, pages 211–228, 1988.

[Mika *et al.*, 1999]S. Mika, G. Rätsch, J. Weston, Schölkopf B., and K. R Muller. Fischer discriminant analysis with kernels. *Neural Networks for Signal Processing IX*, pages 41–48, 1999.

[Rätsch *et al.*, 2001]G. Rätsch, T. Onoda, and K.R Muller. Soft margin for adaboost. *Machine Learning*, pages 287–320, 2001.

[Rosipal and Trejo, 2001]R. Rosipal and L.J Trejo. Kernel partial least squares regression in reproducing kernel hilbert space. *Journal of Machine Learning Research*, 2001.

[Rosipal *et al.*, 2003]R. Rosipal, L.J. Trejo, and B Matthews. Kernel pls-svc for linear and nonlinear classification. In *Proceeding of the twentieth international conference on machine learning (ICML-2003)*, 2003.

[Schölkopf and Smola, 2002]B. Schölkopf and A. J Smola. *Learning with kernel - Support Vector Machines Regularization, Optimization and Beyond*. The MIT Press, 2002.

[Schölkopf *et al.*, 1998]B. Schölkopf, A.J. Smola, and K.R Müller. Nonlinear component analysis as a kernel eigenvalue problem. *Neural Computation*, pages 1299–1319, 1998.

[Syed *et al.*, 1999]N.A. Syed, H. Liu, and K. K Sung. Incremental learning with support vector machines. In *Proceeding of Workshop on Support Vector Machines at International Joint Conference on Artificial Intelligence*, 1999.

[Tenenhaus, 1998]M Tenenhaus. *La Régression PLS*. Éditions Technip, 1998.

[Tenenhaus, 2002]A Tenenhaus. La régression logistique pls validée par bootstrap. In *Mémoire de DEA de Statistique, Université Pierre et Marie Curie*, 2002.

[Vapnik, 1998]V Vapnik. *Statistical Learning Theory*. Wiley, 1998.

[Wold *et al.*, 1982]S. Wold, L. Martens, and H Wold. The multivariate calibration problem in chemistry solved by the pls method. In *Conf. Matrix Pencils, Ruhe A. & Kåstrøm B, Lecture Notes in Mathematics*, pages 286–293. Springer Verlag, 1982.

[Wu *et al.*, 1997]W. Wu, D.L. Massart, and S de Jong. The kernel pca algorithms for wide data – part ii: Fast cross validation and application in classification of nir data. *Chemometrics and Intelligent Laboratory Systems*, pages 271–280, 1997.

6 Appendix: KPLS and PLS regression of $\boldsymbol{Y}$ on $\boldsymbol{K}$

6.1 Kernel PLS (KPLS)

Höskuldsson shows that the weights vector w_1^{PLS} corresponds to the eigenvector associated to the greatest eigenvalue of the matrix $X'YY'X$ [Höskuldsson, 1988]. The first PLS component is then $t_1^{PLS} = Xw_1^{PLS}$.

$\Rightarrow X'YY'Xw_1^{PLS} = \lambda w_1^{PLS}$

$\Leftrightarrow XX'YY'\underbrace{Xw_1^{PLS}}_{t_1^{PLS}} = \lambda\underbrace{Xw_1^{PLS}}_{t_1^{PLS}}$

The first PLS component is the eigenvector associated to the greatest eigenvalue of $XX'YY'$.

Within the framework of PLS 1: $Y \in \mathbb{R}^n$.

$\Rightarrow Y'Xw_1^{PLS}$ is a scalar

$\Rightarrow t_1^{PLS}$ is proportional to $XX'Y$ and thus, we can rigorously be reduced to the framework of the kernel trick, giving arise to Kernel PLS; and write that $t_1^{KPLS} = KY$.

6.2 PLS regression of $\boldsymbol{Y}$ on $\boldsymbol{K}$ (DK-PLS)

In a similar way, the weight vector w_1^{DK-PLS} corresponds to the eigenvector associated to the greatest eigenvalue of the matrix $K'YY'K$. The first DK-PLS component is then $t_1^{DK-PLS} = Kw_1^{DK-PLS}$.

$\Rightarrow K'YY'Kw_1^{DK-PLS} = \lambda w_1^{DK-PLS}$

$\Leftrightarrow KK'YY'\underbrace{Kw_1^{DK-PLS}}_{t_1^{DK-PLS}} = \lambda\underbrace{Kw_1^{DK-PLS}}_{t_1^{DK-PLS}}.$

The first DK-PLS component is the eigenvector associated to the greatest eigenvalue of $KK'YY'$.

Within the framework of PLS 1: $Y \in \mathbb{R}^n$.

$\Rightarrow Y'Kw_1^{DK-PLS}$ is a scalar

$\Rightarrow t_1^{DK-PLS}$ is proportional to $KK'Y$ being, by construction, symmetric

$\Rightarrow t_1^{DK-PLS}$ is proportional to K^2Y.

In a similar way, $t_h^{K-PLS} = K_{h-1}Y$ and $t_h^{DK-PLS} = K_{h-1}^2 Y = K_{h-1}w_h^{DK-PLS}$ where $w_h^{DK-PLS} = K_{h-1}Y$ and K_{h-1} is the matrix comprising the p residual vector $e_{h1}, \ldots, e_{hp}$ of the ordinary least square of $k_j,\ j = 1, \ldots, p$ on $t_1, \ldots, t_{h-1}$. Let us notice that $t_h^{K-PLS} = w_h^{DK-PLS}$.

A finite time stochastic clustering algorithm

Andreea B. Dragut[1] and Codrin M. Nichitiu[2]

[1] LIF, Univ. Aix-Marseille II, Aix-en-Provence, France
[2] Long Island High Tech. Incubator, Stony Brook Univ., NY, USA

Abstract. We present a finite time local search $(1 + \delta)$-approximation method finding the optimal solution with probability almost one with respect to a general measure of within group-dissimilarity. The algorithm is based on a finite-time Markov model of the simulated annealing. A dynamic cooling schedule, allows the control of the convergence. The algorithm uses as measure of within group dissimilarity a new generalized Ward index based on a set of well-scattered representative points, which deals with the major weaknesses of partitioning algorithms regarding the hyperspherical shaped clusters and the noise. We compare it with other clustering algorithms, such as CLIQUE and DBSCAN.
Keywords: Clustering, Finite-time Simulated Annealing, Approximation Scheme, Generalized Ward Index.

1 Introduction

It is generally acknowledged that there are two main families of clustering (unsupervised classification) methods: hierarchical and partitioning. The former ones create a tree structure splitting (reuniting) the initial set of objects in smaller and smaller subsets, all the way to singletons (and reverse), while the latter ones construct a partition of the initial set of objects into a certain number of classes, with the target number usually part of the input, along with the objects themselves. Most partitioning methods proposed for data mining [Jain *et al.*, 1999], [Gosh, 2003] can be divided into: discriminative (or similarity-based) approaches and generative (or model-based) approaches. In similarity-based approaches, one optimizes an objective function involving the pairwise data similarities, aiming to maximize the average similarities within clusters and minimize the average similarities between clusters. A fundamentally different approach is the model based approach which attempts to optimize the fit (global likelihood optimization) between the data and some mathematical model, and most researchers do not consider them as clustering methods. Similarity-based partitioning clustering is also closely related to a number of operations research problems such as facility location problems, which minimize some empirical loss function (performance measure). There are no efficient exact solutions known to any of these problems for general number of clusters m, and some formulations are NP-hard. Given the difficulty of exact solving, it is natural to consider approximation, either through polynomial-time approximation algorithms, which provide guarantees on the quality of their results, or heuristics, which make no guarantees. One of the most popular heuristics for the similarity-based partitioning problem is Lloyd's algorithm, often called the m-means algorithm. Define the neighborhood of a center point to be the set of data points for which this center is the closest. Thus, one can easily see that any locally minimal solution must be centroidal (i.e. each center lies at the centroid of its neighborhood). Unfortunately, m-means algorithm may converges to a local minimum that is arbitrarily bad compared to the optimal solution. Other heuristics with no proven approximation bounds are based on branch-and-bound searching, gradient descent, simulated annealing, nested partitioning, ant colony optimization, and genetic algorithms.

It is desirable to have some bounds on the quality of a heuristic. Given a constant $\delta \geq 0$, a $(1+\delta)$-approximation algorithm (for a minimization problem) produces a solution that is at most a factor $(1+\delta)$ larger than the optimal solution. With a tradeoff between approximation factors and running times, some clustering algorithms are able to produce solutions that are arbitrarily close to optimal. This includes $(1+\delta)$-approximation algorithms for the Euclidean m-median problem by [Kolliopoulos and Rao, 1999] with a running time of $O(2^{1/\delta^s} n \log n \log m)$, assuming that the dimension s is fixed. Another one is the $(1+\delta)$-approximation algorithm for the Euclidean m-center problem given by [Agarwal and Procopiuc, 1998], which runs in $O(n \log m) + (m/\delta)^{O(m^{1-1/s})}$.

Another common approach in approximation algorithms is to develop much more practical, efficient algorithms having weaker, but still constant, approximation factors. These algorithms are based on local search, that is, by incrementally improving a feasible solution by swapping a small number of points in and out of the solution set. This includes the work of [Mettu and Plaxton, 2002] on the use of successive swapping for the metric m-means problem.

Unfortunately it is well known that m-means/medians/centers partitioning clustering algorithms have a tendency to partition the data into hyperspherical shaped clusters and do not adequately deal with outliers and noise.

The algorithm presented here is a local search $(1+\delta)$-approximation method finding the optimal solution with probability almost one with respect to any general measure of within group-dissimilarity. It is actually a cooling schedule, obtained by stopping a simulated annealing algorithm in finite time, and it belongs to a family of approximation clustering algorithms of type m-median and m-means,

The algorithm addresses the weaknesses of partitioning algorithms in the way in which it constructs what we shall define as "critical" clusters, that are to be further expanded by the cooling schedule. As a measure of within group dissimilarity we introduce a new generalized Ward index based not on a single cluster representative i.e. centroid or median, but on a set of well-scattered representative points, which are shrunk toward the centroid. The idea of joining together points close to a set of representatives was introduced by [Guha *et al.*, 1998] to obtain a measure of inter-group dissimilarity in hierarchical clustering. Moreover, due to the particular choices of generation probabilities for the system of neighborhoods, the more dense a cluster is, the smaller the probability to have its elements reassigned to other clusters will be while trying to transform the current classification.

The rest of the paper is organized as follows. In Sections 2. and 3. we present the clustering problem as a combinatorial optimization problem and the general asymptotic convergence conditions for it Sections 4 and 5 describe and compare our algorithm with other clustering algorithms. Finally in Section 6. we present the conclusions and give directions for future research.

2 Setting

The general form of clustering problems considered is "given a set $X = \{1, 2, .., n\}$ of n entities, to classify these entities means to partition the linear subspace X into a number $m \leq n$ of clusters such that the $m-$partitioning is optimal according to a certain chosen criterion function defined on the set Π_m of all $m-$partitions of the set X". Each element i from the set X has an input information vector $Y(i)$. There exists also a distance d as a dissimilarity measure for every pairwise combination of entities to be clustered, and a function $\tau : P(X) \rightarrow R_+$ as a measure of within-group dissimilarities with the property that $\tau(A) = 0 \longleftrightarrow |A| = 1$. Let us consider the function $f : \Pi_m \rightarrow R$, $f(\pi_m) = \sum_{i=1}^{m} \tau(A_i)$, where $\pi_m = (A_1, A_2, ..., A_m) \in \Pi_m$.

The class of clustering problems considered is (PC) $\min_{\pi_m \in \Pi_m.} f(\pi_m)$. (PC) is a combinatorial optimization problem (see [Aarts *et al.*, 1997]) with a very large state space since the $|P(X)|$ given by the Bell number grows extremely rapidly with n; e.g., $B_{40} = 1.6 \times 10^{35}$ and $B_{100} = 4.8 \times 10^{115}$.

A first contribution of this work is the development of a stochastic search algorithm for finding $(1 + \delta)$-optimal partitions with a probability close to one. The basic idea is to construct a Metropolis-Hastings Markov chain via the simulated annealing algorithm.

A neighborhood function is a mapping $N : \Pi_m \rightarrow 2^{\Pi_m}$, which, for each classification $i \in \Pi_m$, defines a set $N(i) \subseteq \Pi_m$ of classifications that can be reached from i by a single perturbation. At the beginning, an initial classification is given. The simulated annealing algorithm starts with it, and continuously tries to transform the current classification into one of its neighbors by applying the generation mechanism and an acceptance criterion. Better-cost neighbors are always accepted. To avoid being trapped in a local minimum, worst-cost neighbors are also accepted, but with a probability that is gradually decreased in the course of the algorithm execution. The lowering of the acceptance probability is controlled by a set of parameters whose values are determined by a cooling schedule.

As we mentioned in the introduction, the algorithm solves the (PC) problem for a general measure of within-group dissimilarities $\tau : P(X) \rightarrow R_+$ such that $\tau(A) = 0 \longleftrightarrow |A| = 1$. However to the best of our bibliographical knowledge, the already existent measures of within group dissimilarity constructed by the extension of a distance do not deal with arbitrarily shaped clusters, and are very sensitive to outliers. Among those indices, the most known are: Wilks index: $\tau(A) = \frac{1}{2|A|} \sum\limits_{x,y \in A} d^2(x,y)$, and Ward index $\tau(A) = \sum\limits_{x,y \in A} d^2(x, x_A)$, where x_A is the centroid of A. The first index does not require X to be a linear space and treats any point of the cluster as a cluster representative, which gives too much unfiltered information about the shape of the set to the clustering algorithm. Also, the (PC) optimization problem with this index leads to long shaped clusters. The second index treats the centroid as the unique cluster representative. This choice gives no information about the shape of the cluster and leads to the well known squared sum of errors criterion with his already discussed problems.

The new index we propose generalizes the Ward one considering multiple representatives for a cluster. We define the representatives index to be $\tau(A) = \sum\limits_{x \in A} \min\limits_{x_r \in R} d^2(x, x_r)$, where R is the set of representatives. The idea of multiple representatives was introduced in hierarchical clustering by [Guha *et al.*, 1998]. They must be well spread across the whole cluster, and are thus obtained through an iterative selection: initially the farthest point from the centroid is picked, and then, up to $|R|$ (fixed in advance), the farthest point from the ones already picked is added. The distance from a candidate point to the set of already picked is the min of the pointwise distances from that point to each already picked. These representatives capture the geometry of the cluster, and upon a shrinking towards the centroid by a fixed factor, done after building R, the outliers get much closer to the centroid (moving more than average representatives within the bulk of the cluster).

3 The asymptotic convergence for the (PC) problem

Notation 1. S : the set of solutions for the considered combinatorial optimization problem (here $S = \Pi_m$), and S^* : the set of optimal solutions.

The simulated annealing can be mathematically modeled as a sequence of Markov chains. Each Markov chain has transition probabilities defined as

$$\forall\, i,j \in S : \quad P_{ij}(k) = \begin{cases} G_{ij}(c_k)\, A_{ij}(c_k) & \text{if } i \neq j \\ 1 - \sum\limits_{l \in S,\, l \neq i} G_{il}(c_k)\, A_{il}(c_k) & \text{if } i = j \end{cases} \qquad (1)$$

where $G_{ij}(c_k)$ denotes the probability of generating a solution j from a solution i, and $A_{ij}(c_k)$ the probability of accepting a solution j that is generated from a solution i.

The matrix P of equation (1) is stochastic and $G_{ij}\left(c_k\right)$ and $A_{ij}\left(c_k\right)$ are conditional probabilities. In the original version of simulated annealing, the acceptance probability is defined by:

$$\forall\, i,j \in S: \quad A_{ij}\left(c_k\right) = \exp\Big(-\ \left(f(j) - f(i)\right)^{+}/c_k\Big) \quad (2)$$

Theorem 1 ([Aarts *et al.*, 1988]) *Let* $(S,\, f)$ *be an instance of a combinatorial optimization problem,* N *a neighborhood function, and* $P\,(k)$ *the transition matrix defined by* (1)*, with* $c_k = c,\ \forall\ k = 0, 1, ...$ *If we have (G1)* $\forall c > 0,\ \forall i,j \in S,\ \exists p \geq 1,\ \exists l_0, l_1, ..., l_p \in S$ *with* $l_0 = i,\ l_p = j$ *and* $G_{l_k\ l_{k+1}}\left(c\right) > 0,\ k = 0, 1, ..., p-1$; *(G2)* $\forall c > 0,\ \forall i,j \in S:\ G_{ij}\left(c\right) = G_{ji}\left(c\right);$ *(A1)* $\forall c > 0,\ \forall i,j \in S:\ A_{ij}\left(c\right) = 1$ *if* $f(i) \geq f(j)$*, and* $A_{ij}\left(c\right) \in (0,1)$ *if* $f(i) < f(j)$; *(A2)* $\forall c > 0,\ \forall i,j,k \in S:\ A_{ij}\left(c\right) A_{jk}\left(c\right) A_{ki}\left(c\right) = A_{ik}\left(c\right) A_{kj}\left(c\right) A_{ji}\left(c\right);$ *(A3)* $\forall i,j \in S$ *with* $f(i) < f(j)$ $\lim_{c\to 0} A_{ij}\left(c\right) = 0.$ *then the Markov chain has a unique stationary distribution* $q(c)$*, with*

$$q_i(c) = 1/\sum_{j\in S}\left(A_{ij}\left(c\right)/A_{ji}\left(c\right)\right)\ ,\ \forall i \in S, \quad (3)$$

Remark 1 *For the following choice of the generation probabilities*

$$G_{ij} = \chi_{(N(i))}\left(j\right)/\left|N\left(i\right)\right|, \quad \forall i,j \in S, \quad (4)$$

condition (G2) is no longer needed to guarantee asymptotic convergence, and the components of the stationary distribution are given by

$$q_i(c) = \left|N\left(i\right)\right|/\sum_{j\in S}\left[\left(\left|N\left(j\right)\right| A_{ij}\left(c\right)\right)/A_{ji}\left(c\right)\right]\ \textit{for all}\ \forall i \in S, \quad (5)$$

We will consider this choice for the generation probability in order to solve the (PC) problem.

Definition 1 *A cluster* A *from* $\omega \in \Pi_m$ *is called critical for* ω *if*

$$\tau\left(A\right) = \max_{A_i\ \textit{cluster of}\ \omega} \tau\left(A_i\right).$$

Notation 2. $N'\left(\pi\right) = \{\omega = \left(A'_1, ..., A'_m\right) \in \Pi_m |\ A'_i$ are obtained from A_i by a reassignment of up to k elements from a critical cluster A, where $k = |A|\}$, for $\pi \in \Pi_m$. We say that $\left(N'\left(\pi\right)\right)_{\pi\in\Pi_m}$ is the set of critical neighborhoods.

Proposition 1 *For the (PC) problem, the set of neighborhoods defined by* $N'\left(\pi\right)$ *satisfies the (G1) condition.*

Proof. It is a fact that $\forall i,\ j \in S,\ \exists\ p \geq 1$, and $l_0, l_1, ..., l_p \in S$ with $l_0 = i$, $l_p = j$ such that for any k ,l_k and l_{k+1} are neighbors through a n-reassign system of neighborhoods. We shall prove that there also exists a path from $i \in S$ to $j \in S$ through a critical system of neighborhoods. Suppose that $u = 0, ..., p-1$ is the first step at which $l_u \in S$ and $l_{u+1} \notin N'\left(l_u\right)$. Let A_u be a cluster in l_u which has a maximal value for the within-group dissimilarity function τ. Let B_u be the cluster in l_u from which t elements are reassigned to some other clusters for obtaining l_{u+1}. Since $l_{u+1} \notin N'(l_u)$ then $\tau(A_u) > \tau(B_u)$. To get a path from $i \in S$ to $j \in S$ through a critical system of neighborhoods we will add a finite number of elements $l_u^{\setminus} \in N'(l_u)$ to the initial path. The procedure is the following: (1) We assign $k-1$ elements from A_u to B_u, where $k = |A_u|$. The new classification $l_u^{\setminus}$ has only two modified clusters $A_u^{\setminus}$, $B_u^{\setminus}$, and $\tau\left(A_u^{\setminus}\right) = 0$ since $\left|A_u^{\setminus}\right| = 1$. (2) If $\tau\left(B_u^{\setminus}\right)$ has not the maximal value then $\exists\, A_{1u}$ such that $\tau(A_{1u})$ is maximal, and we will proceed as in the case of A_u starting the construction of some $l_u^{\setminus\setminus} \in N'\left(l_u^{\setminus}\right)$. Since $1...n$ is a finite set after repeating for a finite number of times the procedure $\tau\left(B_u^{\setminus}\right)$ will be maximal. Now we construct a new classification $l_{u+1}^{\setminus} \in N'\left(l_u^{\setminus}\right)$ in the following way: from $B_u^{\setminus}$ the t elements to other clusters as in the construction step from l_u to l_{u+1}, and the elements belonging to the clusters A_u, $A_{1u}, ...$ are reassigned back to their clusters. We proceed in a similar way for all the steps q at which $l_q \in S$ and $l_{q+1} \notin N'(l_q)$ preserving the other steps.

4 Finite-time model of simulated annealing

In practical applications, asymptoticity is never attained and thus convergence to an optimal solution is no longer guaranteed. Then we shall use the simulated annealing as an approximation algorithm, implementing a cooling schedule. The general idea of a cooling schedule is the following: start with an initial value c_0 for the control parameter and repeatedly generate a finite Markov chain for a finite number of decreasing values of c until $c \backsimeq 0$. The parameters determining the cooling schedule are: the start value c_0 of the control parameter; the decreasing rule of the control parameter; the length L_k of the individual Markov chains; the stop criterion of the algorithm. We will discuss the choice of those parameters for our problem such that the convergence towards near-optimal solutions will be ensured. Our cooling schedule follows the general ideas of the statistical cooling algorithm developed in [Aarts *et al.*, 1988] and designed for symmetric generation probabilities which lead to less complicate formulas for the stationary distributions.

4.1 The start value of the control parameter

This value should be large enough to ensure that initially all configurations occur with rather equal probabilities since $\lim_{c\to\infty} q_i(c) = |N_i'| / \sum\limits_{j\in S} \left|N_j'\right|$.

We distinguish two cases. In the first one, in which the set of system configurations corresponds to values of the cost function distributed over a number of distinct intervals whose mutual distances are large compared to their size, c_0 will be computed in the classical way as $\theta \cdot \max_{i,\, j \in S} \left[f\left(j\right) - f\left(i\right) \right]$, where $\theta \gg 1$. In the second case, the values for the cost function are sufficiently uniformly distributed. Thus, we can observe the behavior of the system before the actual optimization process takes place, and adjust the value of the control parameter such that the ratio χ of the system perturbations accepted over the total number of perturbations generated is kept close to the one given by $\lim_{c\to\infty} q_i\left(c\right)$. The initial value c_0 will be the final value of c updated $m_1 + m_2$ times according to the relation:

$c = \underset{\Delta C_{i\,j} > 0}{Average} \Delta f_{ij} / \ln\left[m_2 / \left(m_2 \chi - \left(1 - \chi\right) m_1 \right) \right]$, where $\Delta f_{ij} = f\left(j\right) - f\left(i\right)$, and m_2, m_1 the numbers of rearrangements with $\Delta f_{ij} \leq 0, > 0$.

4.2 The decreasing rule of the control parameter

In the frame of the homogeneous Markov model for simulated annealing algorithm, the decreasing rule of the control parameter, as well as the lengths L_k of the Markov chains are constructed in order to satisfy the following quasi-equilibrium condition: "$a\left(L_k, c_k\right)$ is close to $q\left(c_k\right)$", where $a\left(l, c_k\right)$ denotes the probability distribution of the classifications after l transitions of the k-th Markov chain. The time behavior of the cooling schedule usually depends on the mathematical formulation of this condition. It is clear from an intuitive point of view that we will have larger differences between $q\left(c_k\right)$ and $q\left(c_{k+1}\right)$ if the decreasing rule of the control parameter allows large decrements of c_k, where we suppose we have reached the quasi-equilibrium. In this case it will be necessary to attempt more transitions at the new value c_{k+1}, for restoring the quasi-equilibrium. Thus, there is a trade-off between fast decrement of c_k and small values for L_k. We will proceed as in [Aarts *et al.*, 1988] using small decrements in c_k in order to avoid extremely long chains, and imposing for ε, δ small positive numbers: $\left\| q\left(c_k\right) - q\left(c_{k+1}\right) \right\| < \varepsilon \quad \approx \forall i \in S \quad 1/\left(1+\delta\right) < q_i\left(c_k\right) / q_i\left(c_{k+1}\right) < \left(1+\delta\right)$

Remark 2 *For the components of the stationary distribution function from* (5) *we get* $q_i\left(c\right) = \left|N_i'\right| \cdot q_0\left(c\right) \cdot A_{i_0 i}\left(c\right)$, *where* $q_0\left(c\right) = \left[\sum_{j\in S} \left|N_j'\right| \cdot A_{i_0 j}\left(c\right) \right]^{-1}$, *and* $i_0 \in S^*$.

Proof. Let $i_0 \in S^* \implies f\left(j\right), f\left(i\right) \geq f\left(i_0\right)$. For $f\left(j\right) > f\left(i\right)$ we have $A_{ji} = 1$, and $A_{ij}\left(c\right) = \exp\left(-\Delta f_{ij}/c\right) = \exp\left(-\Delta f_{i_0\,j}/c\right) \cdot \exp\left(-\Delta f_{i\,i_0}/c\right) = A_{i_0\,j}\left(c\right) \cdot \exp\left(-\Delta f_{i\,i_0}/c\right)$. For $f(j) < f(i)$ we have $A_{ij}\left(c\right) = 1$. From the (A2) property of Theorem 1 we have $A_{i_0\,j}(c) \cdot A_{ji}(c) = A_{i_0 i}\left(c\right) = \exp\left(-\Delta f_{i_0\,i}/c\right) \Rightarrow A_{ji}\left(c\right) = \exp\left(-\Delta f_{i_0\,i}/c\right) / A_{i_0\,j}\left(c\right)$. So we get $A_{ij}(c)/A_{ji}\left(c\right) = A_{i_0\,j}(c)/A_{i_0\,i}(c)$. Then we have that $q_i\left(c\right) \overset{def}{=} \left|N_i'\right| / \left[\sum_{j\in S} \left|N_j'\right| \cdot A_{ij}\left(c\right) / A_{ji}\left(c\right) \right] = \left|N_i'\right| \cdot A_{i_0\,i}\left(c\right) / \left[\sum_{j\in S} \left|N_j'\right| \cdot A_{i_0\,j}\left(c\right) \right]$.

Proposition 2 *If* $\forall i \in S, \forall k \in \mathbf{N}^* \ c_k < c_{k+1}$*, and* $A_{i_0\,i}(c_k)/A_{i_0\,i}(c_{k+1}) < 1+\delta$*, where* $i_0 \in S^*$ *then the following inequalities are satisfied:* $1/(1+\delta) < q_i(c_k)/q_i(c_{k+1}) < (1+\delta)$*.*

Proof. Obviously $\sum\limits_{j\in S} A_{i_0 j}(c_{k+1}) < \sum\limits_{j\in S} A_{i_0 j}(c_k) < (1+\delta)\sum\limits_{j\in S} A_{i_0 j}(c_{k+1})$. Then we derive that $q_0(c_{k+1})/(1+\delta) < q_0(c_k) < q_0(c_{k+1})$, relation from which using the form of $q_i(c)$'s given by the previous remark we can obtain the desired inequality. Thus, using the hypothesis the second part of the desired inequality follows directly. The first part of the desired inequality is a result of introducing in the first part of the $q_0(c)$'s inequality, the $q_i(c)$'s expression, and the obvious relation: $A_{i_0\,i}(c_k) > A_{i_0\,i}(c_{k+1})$.

Remark 3 *The relation given in the hypothesis of the previous proposition can be reformulated as:* $\forall i \in S, \forall k \in \mathbf{N}^* c_{k+1} > c_k/\left[1 + c_k \cdot \ln(1+\delta)/\Delta f_{i_0\,i}\right]$ *which is in fact a decreasing rule of the control parameter.*

To simplify the decreasing rule, we shall make an assumption often made in the literature, and supported by computational evidence (see [Aarts *et al.*, 1988] and [White, 1984]). What we really do is to restrict the decreasing rule to a set S_{c_k} of configurations that occur with a greater probability during the generation of the k-th Markov chain. We will record the cost values of the classifications $X(1), ..., X(L_k) \in S = \Pi_m$ that occur during the generation of the k-th Markov chain, and we will assume that they are normally distributed with mean $\mu_k = \mu(c_k) = \left[\sum\limits_{j=1}^{L_k} f(X(j))\right]/L_k$, and variance $\sigma_k^2 {=} \sigma^2(c_k) {=} \left[\sum\limits_{j=1}^{L_k} f^2(X(j))\right]/L_k - \mu_k^2$. Thus, $\Pr\{\Delta f_{i_0\,i} \le \mu_k - f^* + 3\sigma_k\} \backsimeq 0.99$, where f^* is the optimal value of the problem. Finally, we define $S_{c_k} = \{i{\in}S|\Delta f_{i_0\,i} {\le} \mu_k {-} f^* {+} 3\sigma_k\}$. Then $\Pr\{i \in S_{c_k}\} \backsimeq 0.99$, and we can replace the previous decreasing rule with a simpler one: $\forall i \in S_{c_k}, \forall k \in \mathbf{N}^* \ c_{k+1} > c_k/\left[1 + c_k \cdot \ln(1+\delta)/\mu_k - f^* + 3\sigma_k\right]$. For us f^* is not known but $\mu_k - f^* \ge 0$. Thus, the final decreasing rule of the control parameter is:

$$\forall i \in S_{c_k}, \forall k \in \mathbf{N}^* c_{k+1} > c_k/\left[1 + c_k \cdot \ln(1+\delta)/3\sigma_k\right] \quad (6).$$

4.3 The length $L_k$of the individual Markov chains

The length of a Markov chain is usually determined such that at each value c_k a minimum number of transitions is accepted. Since transitions are accepted with decreasing probability, one would obtain $L_k \to \infty$ for $c_k \downarrow 0$. Therefore, L_k is usually bounded by some constant $L_{\max}$ to avoid extremely long chains for small values of c_k. We take $L_{\max} = |X| - m \ge \max_{i\in S}|N'(i)|$.

4.4 The final value of the control parameter

This choice determines in fact the stopping criterion. We will follow the general idea of most of the dynamic cooling schedules (see [Aarts *et al.*, 1997]). Thus, the algorithm will stop at the c_k value for which the cost function of the classification obtained in the last trial of a Markov chain remains unchanged for a number of ρ consecutive chains. Schematically we have:

```
Compute($L_{\max}, c_0$) ; $c := c_0$; $\overline{f}[k] = MaxInt \; \forall k \in 0, ..., \rho$
repeat
for $i := 1$ to $L_{\max}$ do
   Generate($j \in N'(i)$)
   if $\Delta f_{i\,j} \leq 0$ then Accept($j$) =true
   else if $\exp(-\Delta f_{ij}/c)$ >randomize$[0, 1)$ then Accept($j$) =true;
   if Accept($j$) =true then $i := j$;
Compute($\sigma^2(c)$);   Update($\overline{f}[0, ..., \rho]$) ; $c := \lceil c/[1 + c \cdot \ln(1+\delta)/3\sigma(c)] \rceil$ ;
until $\overline{f}[k_1] = \overline{f}[k_2] \; \forall k_1, k_2 \in 0, ..., \rho$
```

5 Comparison with other algorithms

The study is done comparing the speed and also the quality of the output classification, and using synthetic data generated in a setting constructed and acknowledged by several researchers, such as [Agrawal *et al.*, 1998] and [Zait and Messatfa, 1997]. In generating the data several parameters have been varied, such as size of the classes, their mutual distances, overlap factor, and also their local dimension, smaller than the one of the whole space where points where selected.

Our algorithm was compared to CLIQUE [Agrawal *et al.*, 1998] and DB-SCAN, the latter being much less performant. For the algorithm presented here, we have noted a behavior of similar quality to the one of CLIQUE. However, CLIQUE reports overlapping classes in many cases (it has an approach based on density, varying the local dimensions in which it performs the search), and lower density zones in clusters are discarded as being outliers. Finally, CLIQUE requests the user to find appropriate values for some mandatory parameters controlling its behavior, which is a very difficult task in general. Finally, while both CLIQUE and our algorithm can end up making quite a number of passes over the data, the time required by our algorithm also depends on how fast the within-group dissimilarity τ can be computed, linear ones leading to faster algorithms. The building of the representative set R takes $O(n|R|^2)$: $|R|$ steps, when each point of the current cluster is considered, and for each one, the minimum of the pointwise distance to each member of the increasing R, so another factor of $|R|$.

6 Conclusion

We have presented a finite time stochastic approximation clustering algorithm, which finds optimal solutions with probability almost one, and performs as well as good heuristic clustering algorithms, with a mathematical assessment of its properties, within the framework of the Markov chain analysis of simulated annealing. We have also introduced a new measure of within cluster dissimilarity improving the recognition of arbitrary shaped clusters and reducing the outliers effects.

Concerning outliers, CURE random sampling can filter out a majority of them. Chernoff bounds [Motwani and Raghavan, 1995] provide equations to analytically derive the random sample size required to have a low probability of missing clusters. Also for large databases making several passes over the whole database is undesirable, and clustering the random sample dramatically improves time complexity. Afterwards, the initial non-selected points are each assigned to the cluster of the closest among a fraction of randomly selected representatives for each cluster.

References

[Aarts *et al.*, 1988]E. H. L. Aarts, J. H. M. Korst, and P. J. M. van Laarhoven. A quantitative analysis of the simulated annealing algorithm: A case study for the travelling salesman problem. *Journal of Stat. Phys.*, 50:187–206, 1988.

[Aarts *et al.*, 1997]E. H. L. Aarts, J. H. M. Korst, and P. J. M. van Laarhoven. Simulated annealing, Local search. In E. H. L. Aarts and J. K. Lenstra, editors, *Combinatorial Optimization.* John Wiley Interscience Series, New York, 1997.

[Agarwal and Procopiuc, 1998]P. K. Agarwal and C. M. Procopiuc. Exact and approximation algorithms for clustering. In *Procs. of the 9th AnnlACM-SIAM SODA*, pages 658–667, 1998.

[Agrawal *et al.*, 1998]R. Agrawal, J. Gehrke, D. Gunopulos, and P. Raghavan. Automatic subspace clustering of high dimensional data for data mining applications. In *Procs. of the 1998 ACM-SIGMOD Int'l Conf. on Management of Data*, pages 94–105, 1998.

[Gosh, 2003]J. Gosh. *Handbook of Data Mining*, chapter Scalable Clustering Methods for Data Mining. Lawrence Erlbaum Assoc, 2003.

[Guha *et al.*, 1998]S. Guha, R. Rastogi, and K. Shim. CURE: an efficient clustering algorithm for large databases. In *ACM SIGMOD Int'l Conf. on Manag. of Data*, pages 73–84, 1998.

[Jain *et al.*, 1999]A. K. Jain, M. N. Murty, and P. J. Flynn. Data clustering: a review. *ACM Computing Surveys (CSUR)*, 31(3):264–323, 1999.

[Kolliopoulos and Rao, 1999]S. Kolliopoulos and S. Rao. A nearly linear-time approximation scheme for the euclidean k-median problem. In J. Nesetril, editor, *Procs. of the 7th Annl. Euro. Symp. on Algs.*, volume 1643, pages 362–371. Springer Verlag, 1999.

[Mettu and Plaxton, 2002]R. R. Mettu and C. G. Plaxton. Optimal time bounds for approximate clustering. In *Procs. of the 8th Conf. on Uncertainty in Artif. Intell.*, pages 339–348, 2002.

[Motwani and Raghavan, 1995]R. Motwani and P. Raghavan. *Randomized Algorithms*. Cambridge University Press, 1995.

[White, 1984]S. R. White. Concepts of scale in simulated annealing. In *Proceedings IEEE of the Int'l Conf. on Computer Design*, pages 646–651, 1984.

[Zait and Messatfa, 1997]M. Zait and H. Messatfa. A comparative study of clustering methods. *Future Generation Computer Systems*, 13:149–159, 1997.

Block clustering with mixture model: comparison of different approaches

Mohamed Nadif[1] and Gérard Govaert[2]

[1] LITA - Université de Metz
Ile du Saulcy,
57045 Metz, France
(e-mail: nadif@iut.univ-metz.fr)
[2] Heudiasyc, UMR CNRS 6599
Université de technologie de Compiègne
BP 20529,
60205 Compiègne, France
(e-mail: gerard.govaert@utc.fr)

Abstract. When the data consists of a set of objects described by a set of variables, we have recently proposed a new mixture model which takes into account the block clustering problem on the both sets. In considering this problem under the maximum likelihood and classification maximum likelihood approaches, one can wonder about the performances of the algorithm obtained by block EM, block CEM or by simple uses of the EM and CEM algorithms applied on the both sets separately. The main objective of this paper is to compare these algorithms.
Keywords: Block clustering, Mixture model, EM and CEM algorithms.

1 Introduction

Cluster analysis is an important tool in a variety of scientific areas such as pattern recognition, information retrieval, microarray, data mining, and so forth. Although many clustering procedures such as hierarchical clustering, k-means or self-organizing maps, aim to construct an optimal partition of objects or, sometimes, of variables, there are other methods, called block clustering methods, which consider simultaneously the two sets and organize the data into homogeneous blocks. If $\mathbf{x}$ denotes a $n \times r$ data matrix defined by $\mathbf{x} = \{(x_i^j); i \in I \text{ and } j \in J\}$, where I is a set of n objects (rows, observations, cases) and J is a set of r variables (columns, attributes), the basic idea of these methods consists in making permutations of objects and variables in order to draw a correspondence structure on $I \times J$. These last years, block clustering (also called biclustering) has become an important challenge in data mining context. In the text mining field, [Dhillon, 2001] has proposed a spectral block clustering method by exploiting the clear duality between rows (documents) and columns (words). In the analysis of microarray data where data are often presented as matrices of expression levels of genes under different conditions, block clustering of genes and conditions has permitted to overcome the problem of the choice of similarity on the both sets found in conventional clustering methods [Cheng and Church, 2000].

The mixture model is undoubtedly one of the greatest contributions to clustering. It offers a great flexibility and solutions to the problem of the number of clusters. To take into account the block clustering situation, we have defined in [Govaert and Nadif, 2003] a block mixture model and, setting the clustering problem in the classification maximum likelihood (CML) approach [Symons, 1981], we have developed an algorithm called block CEM which is based on the alternated application of classical CEM on intermediate data matrices. More recently, setting the clustering problem in the maximum likelihood (ML) approach, we have proposed [Govaert and Nadif, 2005] a generalized EM algorithm (GEM) [Dempster *et al.*, 1977] which maximizes a variational approximation of the likelihood using an iterative algorithm whose steps are carried out by the application of the EM algorithm on intermediate mixture models. In estimation context, we have shown that this approach gives good results on simulated data.

This paper focuses on the clustering context. It deals to compare five algorithms: block CEM, block EM with two variants, two-way EM and two-way CEM, i.e. EM and CEM applied separately on I and J. Results on simulated data are given, confirming that block EM gives much better performance than the other algorithms.

In the following, for convenience, we represent a partition $\mathbf{z}$ into g clusters of the sample I by the vector $(z_1, \ldots, z_n)$, where $z_i \in \{1, \ldots, g\}$ indicates the component of the observation i or by the classification matrix $(z_{ik}, i = 1, \ldots, n, k = 1, \ldots, g)$ where $z_{ik} = 1$ if i belongs to cluster k and 0 otherwise. We will use similar notation for a partition $\mathbf{w}$ into m clusters of the set J. Moreover, to simplify the notation, the sums and the products relating to categories, row clusters will be subscripted respectively by letters i, j and k without indicating the limits of variation which will be thus implicit. Thus, for example, the sum $\sum_i$ stands for $\sum_{i=1}^n$ or $\sum_{i,j,k,\ell}$ stands for $\sum_{i=1}^n \sum_{j=1}^r \sum_{k=1}^g \sum_{\ell=1}^m$.

2 Block Mixture Model

For the classical mixture model, the probability density function of a mixture sample $\mathbf{x}$ is defined by $f(\mathbf{x}; \boldsymbol{\theta}) = \prod_i \sum_k p_k \varphi(\mathbf{x}_i; \boldsymbol{\alpha}_k)$ where the p_k's are the mixing proportions, the $\varphi(\mathbf{x}_i; \boldsymbol{\alpha}_k)$'s are the densities of each component k, and $\boldsymbol{\theta} = (p_1, \ldots, p_g, \boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_g)$. We have shown [Govaert and Nadif, 2003] that $f(\mathbf{x}; \boldsymbol{\theta})$ can be written as

$$f(\mathbf{x}; \boldsymbol{\theta}) = \sum_{\mathbf{z} \in \mathcal{Z}} p(\mathbf{z}; \boldsymbol{\theta}) f(\mathbf{x}|\mathbf{z}; \boldsymbol{\theta}), \tag{1}$$

where $\mathcal{Z}$ denotes the set of all possible partitions of I in g clusters, $p(\mathbf{z}; \boldsymbol{\theta}) = \prod_i p_{z_i}$ and $f(\mathbf{x}|\mathbf{z}; \boldsymbol{\theta}) = \prod_i \varphi(\mathbf{x}_i; \boldsymbol{\alpha}_{z_i})$. With this formulation, the data matrix $\mathbf{x}$ is assumed to be a sample of size 1 from a random (n, r) matrix.

To study the block clustering problem, we have extended the formulation (1) to propose a block mixture model defined by the following probability density function $f(\mathbf{x};\boldsymbol{\theta}) = \sum_{\mathbf{u}\in U} p(\mathbf{u};\boldsymbol{\theta})f(\mathbf{x}|\mathbf{u};\boldsymbol{\theta})$ where U denotes the set of all possible partitions of $I \times J$ and $\boldsymbol{\theta}$ is the parameter of this mixture model. In restricting this model to a set of partitions of $I \times J$ defined by a product of partitions of I and J, which will be supposed to be independent, we obtain the following decomposition

$$f(\mathbf{x};\boldsymbol{\theta}) = \sum_{(\mathbf{z},\mathbf{w})\in\mathcal{Z}\times\mathcal{W}} p(\mathbf{z};\boldsymbol{\theta})p(\mathbf{w};\boldsymbol{\theta})f(\mathbf{x}|\mathbf{z},\mathbf{w};\boldsymbol{\theta})$$

where $\mathcal{Z}$ and $\mathcal{W}$ denote the sets of all possible partitions $\mathbf{z}$ of I and $\mathbf{w}$ of J.

Now, extending the latent class principle of local independence to our block model, the x_i^j will be supposed to be independent once $\mathbf{z}_i$ and $\mathbf{w}_j$ are fixed; then, we have $f(\mathbf{x}|\mathbf{z},\mathbf{w};\boldsymbol{\theta}) = \prod_{i,j} \varphi(x_i^j;\boldsymbol{\alpha}_{z_i w_j})$ where $\varphi(x,\boldsymbol{\alpha}_{k\ell})$ is a probability density function defined on the real set $\mathbb{R}$. Denoting $\boldsymbol{\theta} = (\mathbf{p},\mathbf{q},\boldsymbol{\alpha}_{11},\ldots,\boldsymbol{\alpha}_{gm})$ where $\mathbf{p} = (p_1,\ldots,p_g)$ and $\mathbf{q} = (q_1,\ldots,q_m)$ are the vectors of probabilities p_k and q_ℓ that a row and a column belong to the kth component and to the ℓth component respectively, we obtain a block mixture model with the following probability density function

$$f(\mathbf{x};\boldsymbol{\theta}) = \sum_{(\mathbf{z},\mathbf{w})\in\mathcal{Z}\times\mathcal{W}} \prod_i p_{z_i} \prod_j q_{w_j} \prod_{i,j} \varphi(x_i^j;\boldsymbol{\alpha}_{z_i w_j}).$$

3 Various approaches

To tackle the block clustering problem, we have used the block mixture model and have considered the ML and CML approaches.

3.1 ML approach and block EM algorithm

For the ML approach, to estimate the parameters of the block mixture model, we proposed to maximize the log-likelihood $L(\boldsymbol{\theta};\mathbf{x}) = \log(f(\mathbf{x};\boldsymbol{\theta}))$ by using the EM algorithm. To describe this algorithm, we must define the complete log-likelihood, also called classification log-likelihood $L_C(\mathbf{z},\mathbf{w};\boldsymbol{\theta}) = L(\boldsymbol{\theta};\mathbf{x},\mathbf{z},\mathbf{w}) = \log f(\mathbf{x},\mathbf{z},\mathbf{w};\boldsymbol{\theta})$ which can be written

$$L_C(\mathbf{z},\mathbf{w};\boldsymbol{\theta}) = \sum_{i,k} z_{ik}\log p_k + \sum_{j,\ell} w_{j\ell}\log q_\ell + \sum_{i,j,k,\ell} z_{ik}w_{j\ell}\log\varphi(x_i^j;\boldsymbol{\alpha}_{k\ell}).$$

The EM algorithm maximizes $L(\boldsymbol{\theta};\mathbf{x})$ iteratively by maximizing the conditional expectation of the complete log-likelihood given a previous current estimate $\boldsymbol{\theta}^{(c)}$ and $\mathbf{x}$:

$$\begin{aligned} Q(\boldsymbol{\theta},\boldsymbol{\theta}^{(c)}) &= \sum_{i,k} P(z_{ik}=1|\mathbf{x},\boldsymbol{\theta}^{(c)})\log p_k + \sum_{j,\ell} P(w_{j\ell}=1|\mathbf{x},\boldsymbol{\theta}^{(c)})\log q_\ell \\ &\quad + \sum_{i,j,k,\ell} P(z_{ik}w_{j\ell}=1|\mathbf{x},\boldsymbol{\theta}^{(c)})\log\varphi(x_i^j;\boldsymbol{\alpha}_{k\ell}). \end{aligned}$$

Unfortunately, difficulties arise due to the dependence structure in the model, and more precisely, to the determination of $P(z_{ik}w_{j\ell} = 1|\mathbf{x}, \boldsymbol{\theta}^{(c)})$ and approximations are required to make the algorithm tractable. Using a variational approximation

$$P(z_{ik}w_{j\ell} = 1|\mathbf{x}, \boldsymbol{\theta}^{(c)}) \approx P(z_{ik} = 1|\mathbf{x}, \boldsymbol{\theta}^{(c)})P(w_{j\ell} = 1|\mathbf{x}, \boldsymbol{\theta}^{(c)}),$$

we proposed [Govaert and Nadif, 2005] to maximize alternatively two conditional expectations of the complete-data log-likelihood $Q'(\boldsymbol{\theta}, \boldsymbol{\theta}^{(c)}|\mathbf{d})$ and $Q'(\boldsymbol{\theta}, \boldsymbol{\theta}^{(c)}|\mathbf{c})$ where $\mathbf{c}$ and $\mathbf{d}$ are the matrices defined by the c_{ik}'s and the $d_{j\ell}$'s. We shown that these conditional expectations are associated respectively to classical mixture models

$$\sum_k p_k \psi_k(\mathbf{u}_i; \boldsymbol{\theta}, \mathbf{d}) \text{ and } \sum_\ell q_\ell \psi_\ell(\mathbf{v}^j; \boldsymbol{\theta}, \mathbf{c})$$

where $\mathbf{u}_i = (u_i^1, \ldots, u_i^m)$ and $\mathbf{v}^j = (v_1^j, \ldots, v_g^j)$ are vectors of sufficient statistics and ψ_k and ψ^ℓ are the probability density functions of the sufficient statistics. So, these maximizations can be carried out by the EM algorithm and we obtain the two following versions, called block EM(1) and block EM(2). The different steps of the first one are

1. Start from $\mathbf{c}^{(0)}$, $\mathbf{d}^{(0)}$ and $\boldsymbol{\theta}^{(0)}$.
2. Compute $(\mathbf{c}^{(c+1)}, \mathbf{d}^{(c+1)}, \boldsymbol{\theta}^{(c+1)})$ starting from $(\mathbf{c}^{(c)}, \mathbf{d}^{(c)}, \boldsymbol{\theta}^{(c)})$:
   (a) Compute $\mathbf{c}^{(c+1)}, \mathbf{p}^{(c+1)}, \alpha^{(c+\frac{1}{2})}$ by using on the data $(\mathbf{u}_1, \ldots, \mathbf{u}_n)$ the EM algorithm starting from $\mathbf{c}^{(c)}, \mathbf{p}^{(c)}, \alpha^{(c)}$.
   (b) Compute $\mathbf{d}^{(c+1)}, \mathbf{q}^{(c+1)}, \alpha^{(c+1)}$ by using on the data $(\mathbf{v}^1, \ldots, \mathbf{v}^r)$ the EM algorithm starting from $\mathbf{d}^{(c)}, \mathbf{q}^{(c)}, \alpha^{(c+\frac{1}{2})}$.
3. Iterate the steps 2 until convergence.

The different steps of the second version are

1. Start from $\mathbf{c}^{(0)}$, $\mathbf{d}^{(0)}$ and $\boldsymbol{\theta}^{(0)}$, initial values of $\mathbf{c}$, $\mathbf{d}$ and $\boldsymbol{\theta}$.
2. Compute $(\mathbf{c}^{(c+1)}, \mathbf{d}^{(c+1)})$ starting from $\boldsymbol{\theta}^{(c)}$ by iterating the following two steps (a) and (b) until convergence:
   (a) Compute $\mathbf{c}^{(c+1)}$ by using on the data $(\mathbf{u}_1, \ldots, \mathbf{u}_n)$ the E-step starting from $\mathbf{d}^{(c)}, \mathbf{p}^{(c)}, \alpha^{(c)}$.
   (b) Compute $\mathbf{d}^{(c+1)}$ by using on the data $(\mathbf{v}^1, \ldots, \mathbf{v}^r)$ the E-step starting from $\mathbf{c}^{(c)}, \mathbf{q}^{(c)}, \alpha^{(c)}$.
3. Compute $\boldsymbol{\theta}^{(c+1)} = (\mathbf{p}^{(c+1)}, \mathbf{q}^{(c+1)}, \alpha^{(c+1)})$
4. Repeat the steps (2) and (3) until convergence.

After we fit the mixture model to estimate $\boldsymbol{\theta}$, we can give an outright or hard clustering of this data by assigning each observation to the component of the mixture to which it has the highest posterior of probability of belonging. As the calculus of posterior probabilities starting form the parameter is not tractable, a simple solution is to use the probabilities c_{ik} and $d_{j\ell}$ obtained at the end of the block EM algorithm. This procedure, which assigns a partition to a value of the parameter $\boldsymbol{\theta}$, will be named “ C-step ” in the following.

3.2 CML approach and block CEM algorithm

With the CML approach, the partition is added to the parameters to be estimated. In [Govaert and Nadif, 2003], we have proposed the block CEM algorithm that is a variant of block EM. In each of the phases 2(a) and 2(b), it is sufficient to add a C-step which converts the c_{ik}'s and $d_{j\ell}$'s to a discrete classification before performing the M-step by assigning each object and each variable to cluster which has the highest posterior probability of belonging.

3.3 2EM and 2CEM algorithms

Obviously, we can also use the both classical versions EM and CEM on I and J separately (noted 2EM and 2CEM) but unfortunately it is unaware of the correspondence between I and J. It will be seen later that this process is ineffective to detect homogeneous blocs. In addition, the use of two models on the both sets is not parsimonious. Indeed, our proposed block mixture model has fewer parameters than a standard "one-dimensional" clustering: for example, with $n = 1000$ and $r = 500$ and equal proportions of mixture components, if we need to cluster binary data matrix into 4 clusters of rows and 3 clusters of columns, this leads to estimate 12 parameters with Bernoulli block mixture model instead of $5000 = 4 \times 500 + 3 \times 1000$ parameters with two Bernoulli mixture models, i.e., applied on I and J separately.

4 Numerical experiments

In this section, to illustrate the behaviors of our algorithms (2EM, 2CEM, block EM(1), block EM(2), block CEM) and to compare them, we studied their performances for the Bernoulli block mixture model where

$$\varphi(x; \alpha_{k\ell}) = (\alpha_{k\ell})^x (1 - \alpha_{k\ell})^{1-x} \text{ with } \alpha_{k\ell} \in]0, 1[.$$

With block EM(1) and block EM(2), we have two levels of convergence. The first is local; see the phases 2a) and 2b) for block EM(1) and the phase 2) for block EM(2) and the second convergence is global; see the phase (3) for block EM(1) and the phase (4) for block EM(2). In order to accelerate both algorithms, we decided to carry out less iterations locally and more at the global level. After intensive simulations, we chose to carry out only one iteration locally and considered that the global convergence is reached when $|1 - L^{(c)}/L^{(c-1)}| < \varepsilon$ where $L^{(c)}$ denotes the observed log-likelihood at c-th iteration and ε represents a threshold value which chosen on a pragmatic ground, here we took $\varepsilon = 10^{-7}$. This strategy, kept in the following, is fast and gives better results that when one chooses to carry out less iterations globally ($\varepsilon = 10^{-6}$) and more locally (This comparison is not reported here).

In our experiments, we selected twelve kinds of data arising from 3×2-component mixture model corresponding to three degrees of overlap (well separated (+), moderately separated (++) or ill-separated (+++)) of the clusters and four sizes of the data ($n \times r = 50 \times 30, 100 \times 60, 200 \times 120, 300 \times 180$). The concept of cluster separation is difficult to visualize for Bernoulli-mixture models, but the degree of overlap can be measured by the Bayes error corresponding to the block mixture model. As its computation is being theoretically difficult, we used Monte Carlo simulations and evaluated the error rate by comparing the partitions simulated and those we obtained by applying a C-step. But, this step is not direct as in classical situation of mixture model and, in these simulations, we used a modified version of the block Classification EM algorithm in which the parameter $\boldsymbol{\theta}$ is fixed to the true value $\boldsymbol{\theta}^*$. Parameters have been chosen to obtain error rates respectively in $[0.01, 0.05]$ for the well-separated, in $[0.12, 0.17]$ for the moderately and in $[0.20, 0.24]$ for the ill-separated situations. For each of these twelve data structures, we generated 30 samples and for each sample, we have run five algorithms 20 times starting from the same random situations and selected the best solution for each method. We compared 2EM, 2CEM, block CEM, block EM(1) and block EM(2) with $(g, m) = (3, 2)$.

Firstly, we focused on the comparison between block EM(1) and block EM(2). To summarize the behavior of these algorithms, we computed the mean error rate and the mean running time for each simulation. From our results of experiments (Table 1), incontestably the both versions of block EM almost always give the same results and their performance increases with the size of data and especially for block EM(1) (with 300×180 and the situation +++ the error rate is equal to 0.22 for block EM(1) versus 0.28 for block EM(2)). We can also note that block EM(1) is faster and therefore a regular update of $\boldsymbol{\theta}$ is more advantageous. For the continuation, we kept only block EM(1).

The comparisons between 2EM, 2CEM, block CEM and block EM(1) are summarized in Table 2. The first one displays the mean error rate for each situation and in Table 3, the mean running time. From these experiments, the main point arising are the following.

- The versions 2EM and 2CEM working on the two sets separately are suitably effective only when the clusters are well separated. This shows the risk of the use of such methods to obtain homogeneous blocks.
- The block CEM algorithm, even if it is faster and better than 2CEM and 2EM does not give encouraging results when the clusters are not well separated. Moreover, when the size of data increases, it has some difficulties to detect the pattern into 3×2 blocks.
- Not surprisingly, the versions 2CEM and 2EM are slower than block CEM and block EM(1).

In our comparisons we chose to use the percentage of misclassified like an approximation of the Bayes error. This choice is justified because the number of obtained clusters and simulated ones were the same ones. Furthermore, we have extended these comparisons to the cases where the numbers of clusters

Size	Degree of overlap	Error rates		Times	
		block EM(1)	block EM(2)	block EM(1)	block EM(2)
	+	.02(.02)	.02(.02)	0.11(0.07)	0.33(0.15)
(50,30)	++	.24(.08)	.23(.09)	0.53(0.36)	1.71(1.26)
	+++	.31(.14)	.31(.13)	0.48(0.32)	2.04(1.53)
	+	.02(.02)	.02(.02)	0.23(0.16)	0.77(0.70)
(100,60)	++	.14(.03)	.14(.03)	0.28(0.13)	0.93(0.24)
	+++	.28(.11)	.28(.10)	0.69(0.51)	2.13(0.95)
	+	.02(.01)	.02(.01)	0.42(0.08)	1.26(0.17)
(200,120)	++	.14(.02)	.14(.02)	1.03(0.36)	3.72(0.89)
	+++	.28(.09)	.28(.09)	2.54(1.56)	9.86(4.09)
	+	.03(.01)	.03(.01)	0.98(0.15)	3.43(0.30)
(300,180)	++	.15(.02)	.15(.02)	3.11(2.66)	10.38(2.98)
	+++	.22(.06)	.28(.06)	3.90(1.72)	14.77(4.41)

Table 1. Means and standard errors (in parentheses) of error rates and times recorded from the 20 same random situations by block EM(1) and EM(2).

are different from $(3, 2)$ and used the Rand index in comparing the agreement between the both partitions (simulated and obtained). Note that this measure is not restricted to comparing partitions with the same number of clusters. The results of experiments have confirmed the performance of block EM(1).

Size	Degree of overlap	Error rates			
		2CEM(1)	2EM(2)	block CEM	block EM(1)
	+	.09(.09)	.04(.06)	.02(.02)	.02(.02)
(50,30)	++	.38(.08)	.31(.11)	.29(.11)	.24(.08)
	+++	.51(.13)	.46(.13)	.35(.12)	.31(.14)
	+	.08(.06)	.07(.04)	.03(.02)	.02(.02)
(100,60)	++	.31(.08)	.24(.09)	.16(.08)	.14(.03)
	+++	.53(.07)	.49(.10)	.35(.11)	.28(.11)
	+	.03(.02)	.02(.01)	.02(.01)	.02(.01)
(200,120)	++	.41(.10)	.29(.09)	.16(.08)	.14(.02)
	+++	.61(.07)	.50(.08)	.46(.10)	.28(.09)
	+	.06(.02)	.05(.01)	.03(.01)	.03(.01)
(300,180)	++	.50(.06)	.31(.06)	.15(.03)	.15(.02)
	+++	.58(.07)	.39(.08)	.37(.09)	.22(.06)

Table 2. Comparison between 2CEM, 2EM, block CEM, block EM(1): means and standard errors (in parentheses) of error rates.

Size	Degree of overlap	Times			
		2CEM	2EM	block CEM	block EM(1)
	+	2.29(2.61)	0.53(0.12)	0.03(0.01)	0.11(0.07)
(50,30)	++	0.23(0.02)	0.87(0.12)	0.10(0.21)	0.53(0.36)
	+++	0.37(0.83)	0.91(0.12)	0.07(0.12)	0.48(0.32)
	+	2.19(0.48)	5.29(1.38)	0.39(0.25)	0.23(0.16)
(100,60)	++	1.60(0.45)	6.97(0.99)	0.15(0.24)	0.28(0.13)
	+++	1.16(0.09)	7.71(1.09)	0.07(0.03)	0.69(0.51)
	+	10.21(1.08)	26.49(9.14)	0.08(0.05)	0.42(0.08)
(200,120)	++	10.12(0.73)	73.03(8.40)	0.19(0.10)	1.03(0.36)
	+++	8.97(0.80)	89.79(12.12)	0.21(0.12)	2.54(1.56)
	+	37.31(2.77)	111.64(30.26)	0.27(0.27)	0.98(0.15)
(300,180)	++	33.76(2.21)	291.01(31.84)	0.13(0.09)	3.11(2.66)
	+++	35.90(6.78)	449.28(407.16)	0.23(0.17)	3.90(1.72)

Table 3. Comparison between 2CEM, 2EM, block CEM, block EM(1): means and standard errors (in parentheses) of times recorded from the 20 same random situations.

5 Conclusion

Setting the problem of block clustering under the ML and CML approaches, we have compared three block clustering algorithms (block EM(1), block EM(2), block CEM) and two classical methods applied separately on the sets of rows and columns (2EM and 2CEM). Even if the both versions of block EM do not maximize exactly the likelihood, as in the classical mixture model situation but only an approximation of the likelihood of the block mixture model, they give encouraging results on simulated binary data and are better than the other methods. Furthermore, we note, that the first version block EM(1) appears slightly better than EM(2) when the clusters are ill-separated and it is faster. It would be now necessary to apply this algorithm to real situations and to extend this approach to other types of data, such as continuous data by using Gaussian densities for example.

References

[Cheng and Church, 2000]Y. Cheng and G. Church. Biclustering of expression data. In *Proceedings of ISMB*, pages 93–103, 2000.

[Dempster *et al.*, 1977]A.P. Dempster, N.M. Laird, and D.B. Rubin. Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society*, pages 1–38, 1977.

[Dhillon, 2001]I. S. Dhillon. Co-clustering documents and words using bipartite spectral graph partitioning. In *Proceedings of the Seventh ACM SIGKDD Conference, San Francisco, California, USA*, pages 269–274, 2001.

[Govaert and Nadif, 2003]G. Govaert and M. Nadif. Clustering with block mixture models. *Pattern Recognition*, pages 463–473, 2003.

[Govaert and Nadif, 2005]G. Govaert and M. Nadif. An EM algorithm for the block mixture model. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 643–647, 2005.

[Symons, 1981]M.J. Symons. Clustering criteria and multivariate normal mixture. *Biometrics*, pages 387–397, 1981.

Un algorithme de normalisation des données à l'aide de graphes pour le traitement non-linéaire des données : Application à l'optimisation des cartes de Kohonen

Catherine Aaron

SAMOS-MATISSE- Université Paris I
90 rue de Tolbiac,
75013 PARIS, France
(e-mail: catherine_aaron@hotmail.com)

Abstract. Dans le cas où la structure des données à étudier est fortement non-linéaire les méthodes de normalisation "classiques" sont inefficaces pour rendre compte de l'organisation des données. Pour pallier ce problème on propose un algorithme de normalisation des données reposant sur le choix d'un graphe et visant à rendre les voisinages des points sphériques. La version "exhaustive" d'un tel algorithme étant coûteuse en temps de calcul, on en présentera, aussi, sa version stochastique.

En illustration de cette méthode de normalisation, nous proposerons un indicateur permettant de choisir le nombre de lignes et de colonnes à demander en entrée d'une carte de Kohonen.

Keywords: Normalisation, Graphes, Distance Curviligne, Cartes de Kohonen.

1 Normalisation et Analyse des données non linéaire

1.1 La distance curviligne

Dans le cas de l'analyse des données linéaire, l'hypothèse topologique sous-jacente est la convexité des données qui permet de lier les points par des segments en restant dans l'ensemble considéré. En revanche, dans le cas de l'analyse des données non-linéaire, la seule hypothèse topologique est la connexité, qui ne garantit que l'existence d'un chemin continu liant les points deux à deux.

Ainsi dans le cas où la structure des données serait non linéaire, la mesure de distance entre les points représentant le mieux l'organisation des données est la distance curviligne (ou géodésique) qui, en résumé, représente la longueur minimum d'un chemin continu, liant les points, au sein de l'ensemble considéré.

Cette distance (curviligne) est utilisée dans de nouvelles méthodes d'analyse des données non linéaires telles qu' "ISOMAP" ou "curvilinear distance analysis" et, on verra, dans la dernière partie, en quoi son étude peut aider au paramétrage des cartes de Kohonen.

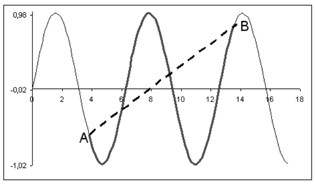


Fig. 1. distance curviligne vs distance euclidienne dans le cas d'un ensemble non convexe.

1.2 Impact de la normalisation sur la distance curviligne

Dans la pratique on approche la distance curviligne en deux étapes : dans un premier temps on détermine un graphe sur les points ($k-$ plus proches voisins, $\varepsilon-$ voisins, MST...) qui lie les points si on peut considérer qu'ils sont ”suffisamment proches”. Puis l'algorithme de Disjkstra permet de rechercher le plus court chemin liant les points et d'en donner sa longueur et d'obtenir ainsi une approximation de la distance curviligne).

Le problème est que les le graphe des liaisons est très sensible aux changements d'échelles. En illustration le graphique si dessous montre le Minimum spanning tree d'un même tirage sinusoïdal pour différentes échelles sur l'axe horizontal.

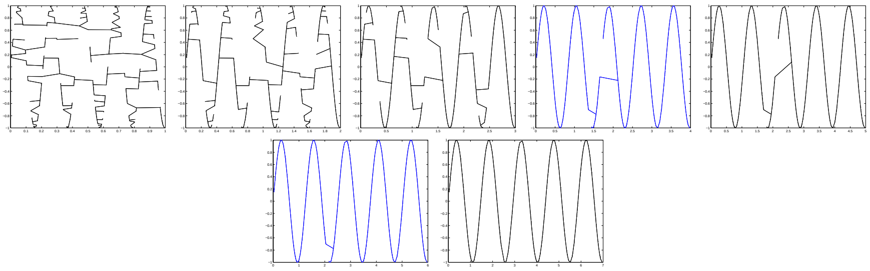


Fig. 2. MST pour différentes échelles horizontales (de 1 à 7) d'un même tirage sinusoïdal.

Le chapitre suivant présente un algorithme qui recherche les transformations à effectuer sur les données pour construire un graphe ”correct” c'est a dire résumant l'information topologique des données.

2 Algorithme de normalisation proposé

2.1 Principe

Travailler sur des données fortement non linéaire implique de s'intéresser localement aux données. Les méthodes ”classiques” reposant sur la dispersion générale autour d'un indicateur central seront ici inefficaces. Pour illustrer ce propos, dans tous les exemples proposés le point de départ des algorithmes suivant sera le résultat des données centrées réduites de manière classique (division par l'écart type).

La méthode de normalisation proposée a pour principe de rendre, en moyenne, les voisinages de forme sphérique et de rayon moyen égal à 1. Autrement dit, on veut rendre les voisinages isotropes en moyenne.

2.2 Version exhaustive

On se fixe un type de graphe ($k-$plus proches voisins, Minimum spanning tree...) qui sert à construire les voisinages que l'on veut rendre isotropes. Puis on itère l'algorithme suivant qui effectue à chaque étape :

- (1) Calcul du graphe G
- (2) Stockage de Y matrice de toutes les vecteurs liaisons
- (3) Effectue une ACP sur Y (les résultat sont une isométrie P et $Y := YP$)
- (4) Application de l'isométrie à X : $X := XP$ (comme P est une isométrie le graphe ne change pas $G(X) = G(XP)$)
- (5) Effectue $Z = |Y|$ vecteur constitué des valeurs absolu des liaisons dans toutes les (nouvelles) directions.
- (6) $pds = mean(Y)$ longueur moyenne des liaisons dans toutes les directions.
- (7) Pour tout j tel que $pds(j) \neq 0$ on effectue $X(:,j) = X(:,j)/pds(j)$

Les points (2) à (4) visent essentiellement à faire tourner les axes de manière à rendre toutes les directions de liaisons possibles. Les points (5) à (7) visent à rendre les liaisons de tailles équivalentes sur tous les axes significatifs.

Remarque : S'il existe un système de $d' < d$ axes linéaires permettant résumant complètement l'information celui ci est obtenu par l'algorithme. On obtient dans ce cas là les même résultats qu'ISOMAP mais avec un temps de calcul largement plus court car seul les graphes sont calculés et non toutes les distances curvilignes.

2.3 Version Stochastique

La version stochastique a, uniquement, comme but d'accélérer le temps de calcul du au calcul du graphe (en $O(N^2)$ pour les $k-$plus proches voisins et en $O(N^2log(N))$ pour le MST) pour cela, à chaque étape on tire (sans remise) $N' < N$ points sur lesquels on calcul le graphe, la rotation et la pondération des axes qu'on applique sur toutes les données.

2.4 Quelques résultats

Les exemples suivants présentent tous les résultats de l'algorithme exposé ci-dessus pour des données sinusoïdales en dimension 2. Le graphe de référence le MST. Les graphiques résultats se lisent verticalement :

- (1) Graphe et données dans le cas de la normalisation standard.
- (2) Pourcentage d'inertie expliquée (cumulée) par axe.
- (3) Angle de la plus grande rotation de l'isométrie.
- (4) Poids de chaque axe.
- (5) Graphe pour les données après normalisation.

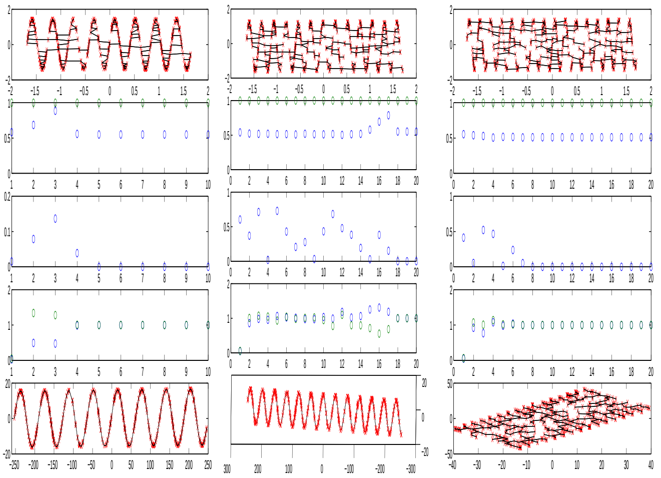


Fig. 3. 500 points avec $X(:,1) = unifrnd(0,1)$ et $X(:,2) = sin(\omega X(:,1))$ avec $\omega \in \{50, 80, 100\}$

On voit que l'algorithme donne de relativement bons résultats jusqu'à ce qu'il y ait un effet "saturation" pour des fréquences trop élevées.

Les données suivantes ont été tiré de la manière suivante : $X(:,1)$ suit une loi uniforme sur $[0,1]$ et $X(:,2) = sin(\omega X(i,1)) + \sigma\varepsilon$ avec ε suivant une loi normale centrée réduite. Puis on a appliqué une rotation d'angle $pi/4$ aux données. On observe dans ce cas que la saturation est plus rapide.

2.5 Conclusion et perspective

Les résultats sont encourageants mais nous ne sommes pas encore parvenus à quantifier la performance de l'algorithme. On sait aujourd'hui qu'il n'existe

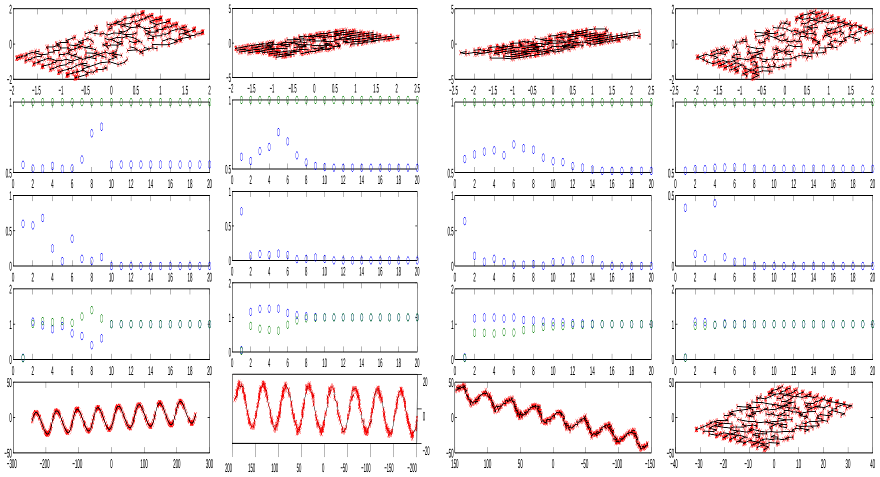


Fig. 4. exemples avec rotation : (1) $\omega = 50, \sigma = 0$,(2) $\omega = 50, \sigma = 0.1$, (3) $\omega = 50, \sigma = 0.2$,(4) $\omega = 70, \sigma = 0$

pas forcément une unique solution au problème de l'existence d'une transformation rendant les voisinages (exactement) isotropes en moyenne (mais la probabilité d'existence tend vers 1 lorsque le nombre d'individus augmente). On aimerait surtout quantifier le fait que la distance curviligne estimée à l'issue de la normalisation corresponde "au mieux" à la "vraie" distance curviligne par une vraie démonstration et, non, uniquement par des simulations.

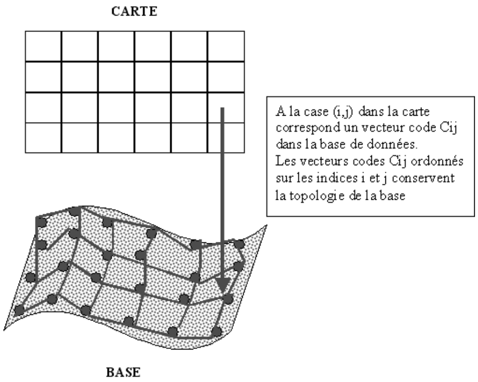
3 Paramétrage d'une carte de Kohonen

3.1 Les cartes de Kohonen

L'algorithme des cartes de Kohonen vise à projeter les données sur une "carte" i.e. une structure de voisinage fixé a l'avance. Plusieurs types d'utilisations en sont faite. Essentiellement en représentation des données en plus petite dimension (pendant non linéaire à l'ACP) et en classification. Nous nous intéresserons ici plus particulièrement à l'aspect "projection" et représentation des données et non a l'aspect "classification".

Brièvement, pour projeter des données sur une carte de Kohonen, on se fixe une topologie c'est a dire un nombre de cellules et leurs voisinages associés.

A chaque case (i, j) dans la carte correspond donc un vecteur code $\boldsymbol{C_{i,j}}$ dans l'espace des données. L'algorithme de Kohonen repose sur une conservation de la topologie c'est a dire sur le fait que la topologie induite par une



A la case (i,j) dans la carte correspond un vecteur code Cij dans la base de données.
Les vecteurs codes Cij ordonnés sur les indices i et j conservent la topologie de la base

Fig. 5. cartes de Kohonen

distance sur Z^2 (distance entre les cases) respecte la topologie induite par une distance sur $\Re^d$ (distance entre les vecteurs codes).

3.2 Un indicateur de respect de la topologie

Le propos précédent nous permet de définir un indicateur de respect de la topologie. On choisit comme distance sur Z^2 la distance euclidienne : $d_1((i,j),(i',j')) = \sqrt{(i-i')^2 + (j-j')^2}$. Le choix et la construction d'une distance sur les vecteurs codes est légèrement plus délicat. Etant donné que les données peuvent être dans un espace non linéaire on va choisir la distance curviligne mais, comme le nombre de vecteurs codes est relativement faible la détermination de la distance curviligne se fait en s'autorisant des chemins qui passent par des points de la base de donnée.

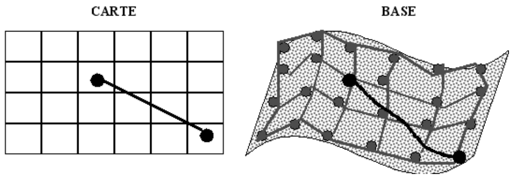


Fig. 6. Distance sur la grille et distance entre les vecteurs codes

Les données sont, en préliminaire, normées par l'algorithme décrit dans la section précédente.

Les résultats seront alors présentés de la manière suivante : Pour une base de donnée, et pour une carte de Kohonen (ici un nombre de ligne et un nombre de colonne), on va tracer les nuages de points liant la distance entre les cases et distances entre les vecteurs codes, et indiquer leur corrélation.

3.3 Résultats

Un premier exemple est constitué d'un tirage uniforme de 200 points sur $[0,1]^2$. On note alors que, comme escompté ce sont les cartes "carrées" (i.e. comportant autant de lignes que de colonnes) qui reconstituent au mieux la topologie de l'espace.

Le premier graphique donne les nuages de points entre les deux distances : première lignes pour des cartes $(1,3)$ jusqu'à $(1,10)$ deuxième lignes pour des cartes $(2,2)$ jusqu'à $(2,10)$... Le second graphique présente les coefficients de corrélation entre les distances pour toutes ces des cartes

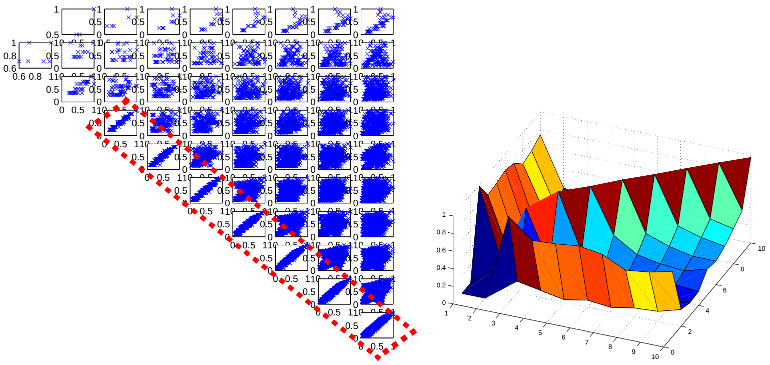


Fig. 7. Distance sur la grille et distance entre les vecteurs codes et correlation entre ces dernières dans le cas d'un tirage uniforme

Un deuxième exemple est constitué d'un tirage de 200 points avec $X(:,1)$ suivant une loi uniforme $[0,1]$ et $X(:,2) = sin(15X(:,1))$. Là aussi, comme prévu, on observe les meilleurs résultats pour une carte de Kohonen de largeur 1, c'est a dire pour une "ficelle", ce qui correspond au fait que la dimension intrinsèque des données est 1 (voir figure 8).

Pour finir, on s'est placé en dimension 3 avec $[X(:,1)X(:,2)]$ uniformément tirés dans $[0,1]^2$ et $X(:,3) = sin(15X(:,1))$. Le résultat obtenu qui semblait étonnant a première vue (préconisation d'une ficelle) est confirmé par la représentation du nuage de point et des vecteurs codes associés a une ficelle (15 cellules) voir figure 9.

References

[De Bodt, *et al.*, 2002]De Bodt,E. , Cottrel, M., Verleysen, M.: : Statistical tools to asses the reliability of self-organizing maps In: Neural Networks 15 (2002) 967-978

[Dijkstra, 1951]E.W. Dijkstra, A note on two problems in connection with graphs, *Mathematics*, (1):269-271, 1951.

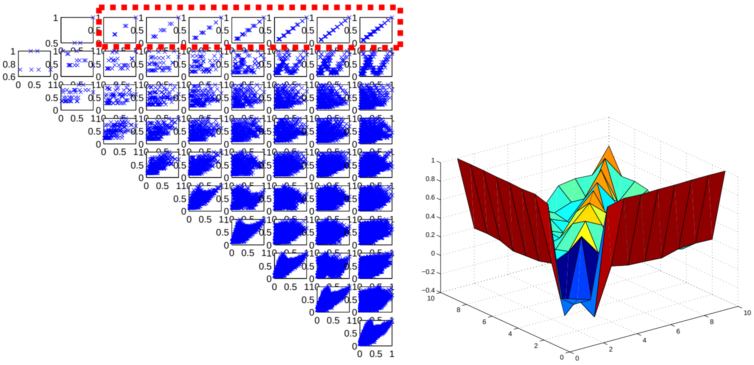


Fig. 8. Distance sur la grille et distance entre les vecteurs codes et correlation entre ces dernières dans le cas d'un tirage sinusoïdal

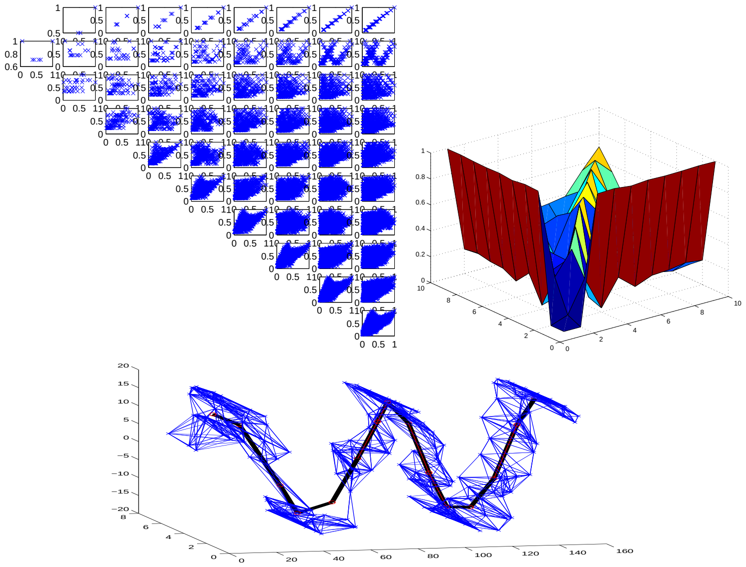


Fig. 9. Distance sur la grille et distance entre les vecteurs codes et correlation entre ces dernières dans le cas d'un tirage sinusoïdal en dimension 3

[Tenenbaum and de Silva, 2000]J.B. Tenenbaum, V. de Silva, A global geometric framework for no n-linear dimensionality reduction, *Science*, (290):2319-2323, December 2000.

[Lee *et al.*, 2000]J.A. Lee, A. Lendasse and M. Verleysen, Curvilinear Distance analysis versus isomap In M. Verleysen, editor, *proceedings of the* 8^{th} *European Symposium on Artificial Neural Networks* (ESANN 2000), d-side pub., pages 13-20, April, Bruges (Belgium), 2000.

Adaptation de l'algorithme SOM à l'analyse de données temporelles et spatiales : Application à l'étude de l'évolution des performances européennes en matière d'emploi

Catherine Aaron, Corinne Perraudin, and Joseph Rynkiewicz

SAMOS - MATISSE
Université de Paris 1
90 rue de Tolbiac,
75013 Paris, France
(e-mail: catherine_aaron@hotmail.com, Corinne.Perraudin@univ-paris1.fr, Joseph.Rynkiewicz@univ-paris1.fr)

Abstract. Cet article étudie l'évolution des performances européennes en matière d'emploi depuis le début des années quatre vingt dix, en utilisant l'algorithme SOM adapté au traitement de données qui sont à la fois temporelles et spatiales. La carte de Kohonen ainsi obtenue permet d'établir une classification des pays de l'Union Européenne qui tient compte simultanément de l'ordonnancement temporel et spatial des données, et permet alors de comparer les trajectoires des différents pays dans le temps. Nous comparons les résultats obtenus par cette méthode à ceux reposant sur une carte de Kohonen traditionnelle.

Keywords: Classification, Algorithme SOM, Emploi, Union Européenne.

1 Introduction

Malgré les modestes performances européennes en matière d'emploi très souvent dénoncées, la question de l'emploi n'a véritablement été abordée au niveau européen qu'en 1993 avec le fameux "Livre blanc Delors" sur la croissance, la compétitivité et l'emploi. Inspirés de ce document, les Conseils européens qui suivent vont intégrer progressivement l'objectif d'atteindre un niveau élevé d'emploi parmi les objectifs clés de l'Union Européenne (UE ensuite), conduisant à lui donner le même niveau d'importance que les objectifs macroéconomiques de croissance et de stabilité. Lors du sommet sur l'emploi de 1997 à Luxembourg, la Stratégie Européenne pour l'Emploi (SEE) est lancée. Elle est conçue comme un instrument de coordination et d'orientation des priorités de la politique de l'emploi des Etats membres de l'UE. Le Conseil européen de Lisbonne, en mars 2000, lui donne une impulsion supplémentaire en insérant la stratégie dans l'agenda global de l'UE en matière économique et sociale. Il vise à faire de l'UE l'économie "la plus compétitive et la plus dynamique du monde" et établit des cibles à atteindre à un horizon de dix ans.

A mi-chemin de la date butoir, il importe d'étudier la situation en matière d'emploi des pays de l'UE, et de regarder plus particulièrement si le fait d'avoir mis l'emploi au cœur des préoccupations européennes a permis une amélioration des performances en matière d'emploi. Cet article étudie l'évolution des performances des pays de l'UE en matière d'emploi depuis le début des années quatre vingt dix. Il s'agit de tenir compte de l'évolution temporelle de différents indicateurs relatifs au marché du travail afin d'étudier les trajectoires suivies par les pays de l'UE.

Nous étudions la situation des 25 pays actuellement dans l'UE, ainsi que celles de la Roumanie et de la Bulgarie, qui vont rejoindre l'Union Européenne en 2007, de la Turquie, dont les négociations devraient débuter fin 2005 et de la Croatie, candidate à l'adhésion. Nous retenons, dans une première analyse, les trois indicateurs définissant les objectifs d'emploi fixés par la Stratégie de Lisbonne à atteindre pour 2010, à savoir un taux d'emploi total de 70%, un taux d'emploi des femmes de 60% et un taux d'emploi des travailleurs âgés (55-64 ans) de 50%. Nous étendons ensuite cette analyse à un ensemble plus large d'indicateurs relatifs à l'emploi et au chômage, afin de caractériser de manière plus fine les performances des marchés du travail européens.

Pour mener à bien cette comparaison des trajectoires des pays européens selon leurs performances, nous proposons d'utiliser une adaptation de l'algorithme SOM, permettant de traiter des données qui sont la fois temporelles et spatiales. La section suivante présente l'adaptation de l'algorithme SOM. La troisième section présente les résultats de l'analyse des taux d'emploi. La quatrième section étend cette analyse à une gamme plus large d'indicateurs du marché du travail. La dernière section conclut.

2 Carte de Kohonen adaptée à l'analyse de données temporelles

L'algorithme SOM (*Self-Organizing Map*), encore appelé algorithme de Kohonen, est un algorithme stochastique de classement des données, qui respecte la topologie de l'espace des observations en intégrant une notion de voisinage entre les classes (voir [Kohonen, 1995])[1].

Afin de classifier les données européennes étudiées dans cet article, qui sont temporelles ($t = 1, \ldots, T$) et spatiales ($j = 1, \ldots, N$), on peut considérer qu'un pays j pour une année t est une observation et on peut alors classer les $T \times N$ observations sur une carte de Kohonen[2]. Cependant, cette manière de faire ne permet pas d'observer les trajectoires des individus et de les comparer facilement.

[1] Une présentation des méthodes de classification et de diverses applications construites à partir de cet algorithme est fournie par [Cottrell *et al.*, 2003].

[2] Voir par exemple [Akarçay-Gürbüz and Perraudin, 2004] pour une classification des pays européens selon leurs performances économiques dans le temps

Afin de réellement prendre en compte la dimension temporelle des données, on pourrait alors construire T cartes de Kohonen et classifier les N observations selon les variables retenues par année. Le problème avec ce type de méthode est que la classification ainsi obtenue année par année est très instable.

Ainsi, nous proposons une méthode qui permet de tenir compte de l'aspect temporel des données[3]. Plus précisément, nous adaptons l'algorithme SOM afin qu'il tienne compte de l'ordonnancement temporel en plus de l'ordonnancement spatial des données. Pour cela, nous construisons une carte dont la longueur est égale au nombre d'années T de la période d'observations, et de largeur égale au nombre de représentants par année choisi a priori (noté k). Le calcul des vecteurs codes de cette carte de Kohonen, dite généralisée dans la suite, s'effectue selon l'algorithme suivant :

- L'initialisation de l'algorithme SOM correspond à un tirage aléatoire de k pays dans l'ensemble des données. A l'unité (i, t) de la carte, on affecte les valeurs des variables[4] du pays i pour l'année t.
- A chaque itération, un pays i_0 et une année t_0 sont tirés aléatoirement dans l'ensemble de données. Ensuite, pour tout $i \in [1, k]$, on cherche l'unité (i, t_0) qui est la plus proche de l'observation sélectionnée.
- On met à jour l'unité gagnante et les unités voisines. Le voisinage décroît dans la dimension ligne durant les itérations de r à 0. Pour forcer l'organisation temporelle, pour un voisinage ligne donné r, la taille du voisinage temporel décroît de r à 0 (voir figure 1).
- Finalement, afin de garantir la convergence, on finit à 0 voisin sur les deux derniers tiers des itérations.

Une fois que l'algorithme a convergé, on positionne les pays sur la carte afin d'identifier leur position.

Un premier avantage de cette carte de Kohonen généralisée au classement de données temporelles est qu'elle permet d'observer une continuité à la fois dans la dimension temporelle et dans la dimension ligne.

Le nombre de colonnes de cette carte étant choisi a priori (arbitrairement grand), nous mettons ensuite en œuvre une classification ascendante hiérarchique à l'ensemble des vecteurs codes de la carte afin de réduire le nombre de classes. Ainsi, le nombre de super-classes par année n'est pas contraint à être le même au cours du temps. Cela nous permet d'étudier si les super-classes homogènes se forment, dans le temps ou alors pour une date donnée, et si le nombre de super-classes par année se réduit, indiquant alors une convergence des pays.

[3] Voir [Aaron *et al.*, 2003], qui proposent deux adaptations de l'algorithme de Kohonen afin de classifier les trajectoires des 15 pays de l'UE dans le temps, vers les normes fixées par les critères de Maastricht. La première méthode proposée est reprise dans cet article.

[4] Les données sont centrées et réduites par variable sur la période entière.

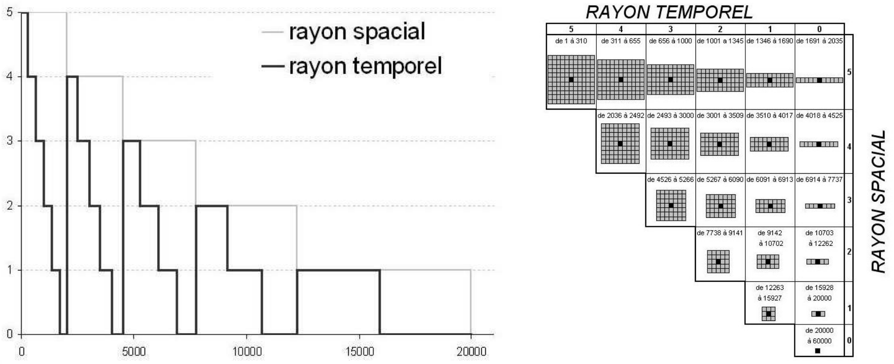


Fig. 1. Evolution du rayon dans l'algorithme généralisé.

Un deuxième avantage de cette carte est de permettre de visualiser très facilement les trajectoires des pays à travers les différentes super-classes.

3 L'évolution des taux d'emploi

Nous étudions tout d'abord les performances des pays européens sur la base du taux d'emploi total (TxEmp), du taux d'emploi des femmes (TxEmpF) et du taux d'emploi des travailleurs âgés (TxEmpTA)[5].

Le taux d'emploi total moyen sur tous les pays disponibles et sur la période 1992-2003 s'élève à 62% (respectivement 53% pour les femmes et 37,5% pour les travailleurs âgés). Il passe de 62,4% en 1992 (respectivement 52% et 39%) à 62% en 2003 (respectivement 54,3% et 39,9%)[6]. Cependant, cette apparente constance des taux d'emploi cache des disparités importantes entre les pays, que ce soit en termes de niveaux ou de trajectoires, comme nous allons le voir.

Nous avons fixé à 12 le nombre de classes par année. La carte de Kohonen obtenue (voir figure 2) comprend alors 12 lignes (correspondant aux 12

[5] Le taux d'emploi est égal au nombre de personnes en emploi rapporté à la population concernée en âge de travailler. Les données sont issues de [Commission, 2004] et du site *http://europa.eu.int*. Les variables sont disponibles de 1992 à 2003 pour l'Allemagne (DE), la Belgique (BE), le Danemark (DK), l'Espagne (ES), la Finlande (FI), la France (FR), la Grèce (EL), l'Irlande (IE), le Luxembourg (LU), les Pays-Bas (NL), le Portugal (PT), le Royaume-Uni (UK), la Suède (SE). Les données ne sont disponibles qu'à partir de 1993 pour l'Italie (IT); de 1994 pour l'Autriche (AT); de 1996 pour la Hongrie (HU), la Slovénie (SI); de 1997 pour la Pologne (PL), la Roumanie (RO); de 1998 pour la République tchèque (CZ), l'Estonie (EE), la Lettonie (LV), la Lituanie (LT), la Slovaquie (SK); de 2000 pour la Bulgarie (BG), Chypre (CY), Malte (MT), la Turquie (TR); et seulement pour 2003 pour la Croatie (CR).

[6] La moyenne étant effectuée sur 13 pays en 1992 et sur 29 en 2003.

années de la période 1992-2003) et 12 colonnes. On observe une continuité dans la répartition des trois indicateurs à la fois dans la dimension ligne et dans la dimension colonne. Chaque année, les pays sont classés selon leurs performances en matière de taux d'emploi. On constate que la carte oppose les pays ayant les meilleures performances en termes de taux d'emploi, que ce soit total, des femmes ou des travailleurs âgés (le coté gauche de la carte) et les pays ayant de moins bonnes performances (coté droit de la carte). On remarque que le coin supérieur droit de la carte correspond aux taux d'emploi total et féminin les plus faibles, mais à des taux d'emploi moyens pour les travailleurs âgés. On note que les différences de performances se sont estompées dans le temps, essentiellement parce que les moins bonnes performances en matière de taux d'emploi se sont améliorées.

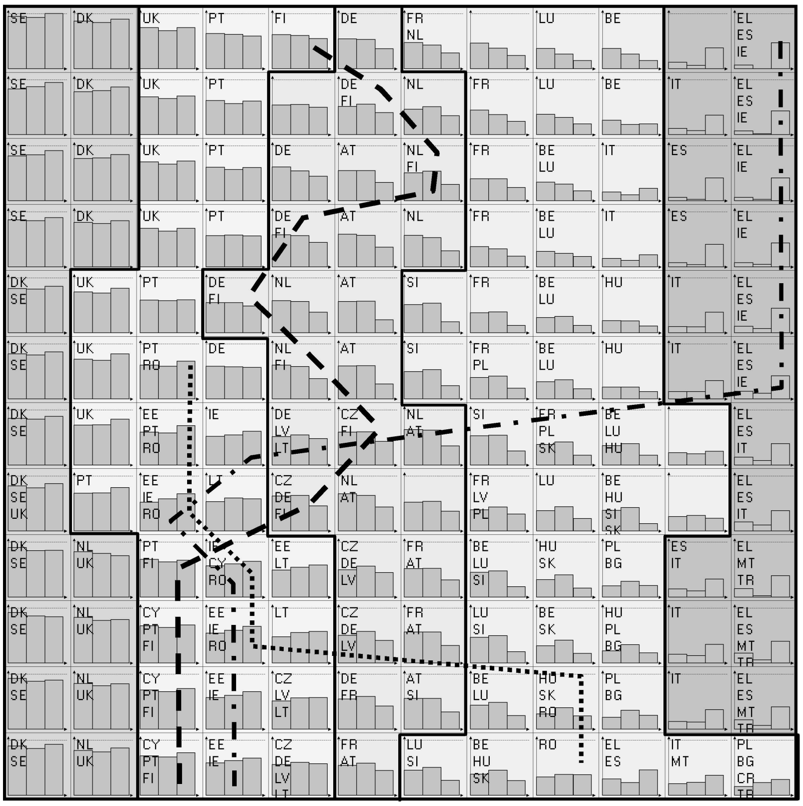


Fig. 2. Carte de Kohonen généralisée.

Note : La première ligne correspond à 1992 et la dernière à 2003. Dans chaque case de la carte sont représentées de gauche à droite les répartitions de TxEmp, TxEmpF, TxEmpTA. La ligne en pointillés (respectivement discontinue et alternée) représente la trajectoire de la Roumanie (respectivement de la Finlande et de l'Irlande.

La classification ascendante hiérarchique menée sur l'ensemble des vecteurs codes de la carte conduit à retenir 5 super-classes. Elles correspondent à des performances de moins en moins bonnes en allant de la gauche vers la droite. On remarque que les 5 super-classes coexistent sur l'ensemble de la période étudiée (sauf celle regroupant les pays aux performances médiocres qui disparaît la dernière année). La super-classe qui se trouve sur le coté gauche de la carte est la seule à enregistrer des taux d'emploi supérieurs aux cibles fixées pour 2010. Il s'agit des pays nordiques (Suède, Danemark, ainsi que Pays-Bas depuis 2000) ainsi que le Royaume-Uni. La super-classe à droite de la carte correspond aux moins bonnes performances, et on y retrouve les pays du Sud (Espagne, Grèce, Italie, ainsi que Malte, Turquie, Bulgarie, Croatie). La super-classe du milieu correspond à des performances moyennes et regroupe la France, l'Allemagne, l'Autriche, ainsi que les autres pays récemment entrés dans l'UE.

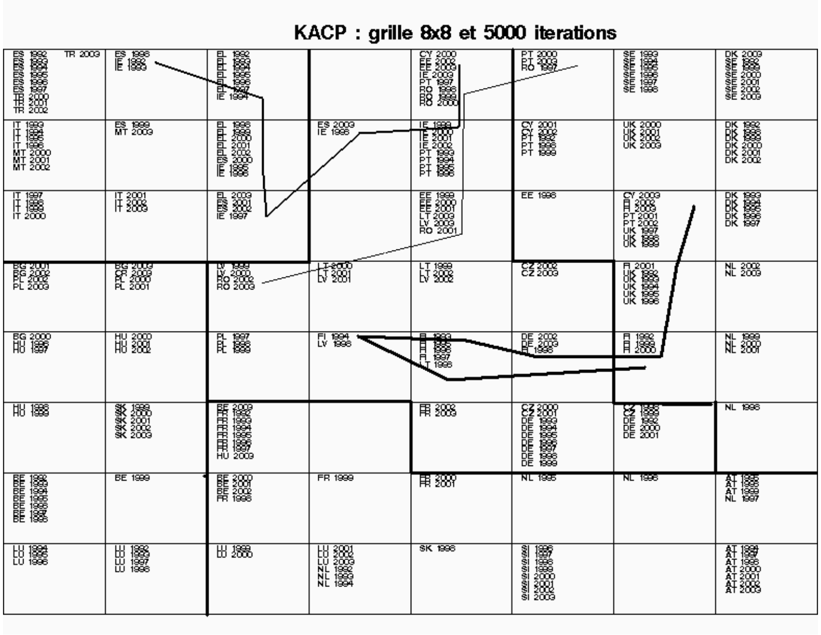
On peut très facilement observer la trajectoire suivie par chaque pays à travers la carte, et donc à travers les super-classes décrivant des performances différentes, et cela grâce à l'ordonnancement temporel imposé dans cette carte. Certains pays restent tout au long de la période dans la même super-classe. C'est le cas de la Suède et du Danemark, qui enregistrent les meilleures performances, ou de l'Espagne, l'Italie et la Grèce qui restent dans la super-classes des moins bonnes performances. Quelques pays suivent des trajectoires qui traversent différentes super-classes, indiquant une évolution, soit vers de meilleures performances (Irlande ou Finlande, dont les trajectoires sont dessinées sur la carte) soit vers des performances en déclin (cas de la Roumanie, dont la trajectoire est aussi représentée sur la carte).

Afin de souligner l'avantage de cette carte relativement à une carte de Kohonen non contrainte par l'ordonnancement temporel (dite traditionnelle dans la suite)[7], nous comparons ces résultats avec ceux obtenus en considérant un pays pour une année comme une observation.

Dans ce cas, nous obtenons une classification (voir carte figure 3) qui associe essentiellement un même pays pour différentes années. Comme dans les résultats précédents, cette carte permet de séparer très clairement les pays qui enregistrent de bonnes performances en matière de taux d'emploi (en haut à droite de la carte figure 4) aux pays ayant les moins bonnes performances (à gauche de la carte). Parmi ces derniers, on retrouve la distinction entre ceux qui ont des taux d'emploi des travailleurs âgés les plus faibles et ceux qui ont des taux moyens.

Le positionnement des pays, ainsi que les super-classes obtenues par une classification hiérarchique ascendante (en 5 super-classes), gardent le même type d'interprétation que dans les résultats précédents, mais il est moins aisé d'observer le suivi des trajectoires et de les comparer entre elles (voir carte figure 3). A titre d'illustration, nous avons reproduit sur la carte 3 la trajectoire des trois pays (Roumanie, Finlande, Irlande) mentionnés précédemment.

[7] Voir [Letrémy, 2000] pour une présentation des programmes en SAS-IML.



ES 1992 ES 1993 ES 1994 ES 1995 ES 1996 ES 1997 TR 2000 TR 2001 TR 2002 TR 2003	ES 1998 IE 1992 IE 1993	EL 1992 EL 1993 EL 1994 EL 1995 EL 1996 EL 1997 IE 1994		CY 2000 EE 2002 EE 2003 IE 2003 PT 1997 RO 1998 RO 1999 RO 2000	PT 2000 PT 2003 RO 1997	SE 1993 SE 1994 SE 1995 SE 1996 SE 1997 SE 1998	DK 2003 SE 1992 SE 1999 SE 2000 SE 2001 SE 2002 SE 2003
IT 1993 IT 1994 IT 1995 IT 1996 MT 2000 MT 2001 MT 2002	ES 1999 MT 2003	EL 1998 EL 1999 EL 2000 EL 2001 EL 2002 ES 2000 IE 1995 IE 1996	ES 2003 IE 1998	IE 1999 IE 2000 IE 2001 IE 2002 PT 1993 PT 1994 PT 1995 PT 1996	CY 2001 CY 2002 PT 1992 PT 1998 PT 1999	UK 2000 UK 2001 UK 2002 UK 2003	DK 1992 DK 1998 DK 1999 DK 2000 DK 2001 DK 2002
IT 1997 IT 1998 IT 1999 IT 2000	IT 2001 IT 2002 IT 2003	EL 2003 ES 2001 ES 2002 IE 1997		EE 1999 EE 2000 EE 2001 LT 2003 LV 2003 RO 2001	EE 1998	CY 2003 FI 2002 FI 2003 PT 2001 PT 2002 UK 1997 UK 1998 UK 1999	DK 1993 DK 1994 DK 1995 DK 1996 DK 1997
BG 2001 BG 2002 PL 2002 PL 2003	BG 2003 CR 2003 PL 2000 PL 2001	LV 1999 LV 2000 RO 2002 RO 2003	LT 2000 LT 2001 LV 2001	LT 1999 LT 2002 LV 2002	CZ 2002 CZ 2003	FI 2001 UK 1992 UK 1993 UK 1994 UK 1995 UK 1996	NL 2002 NL 2003
BG 2000 HU 1996 HU 1997	HU 2000 HU 2001 HU 2002	PL 1997 PL 1998 PL 1999	FI 1994 LV 1998	FI 1993 FI 1995 FI 1996 FI 1997 LT 1998	DE 2002 DE 2003 FI 1998	FI 1992 FI 1999 FI 2000	NL 1999 NL 2000 NL 2001
HU 1998 HU 1999	SK 1999 SK 2000 SK 2001 SK 2002 SK 2003	BE 2003 FR 1992 FR 1993 FR 1994 FR 1995 FR 1996 FR 1997 HU 2003		FR 2002 FR 2003	CZ 2000 CZ 2001 DE 1993 DE 1994 DE 1995 DE 1996 DE 1997 DE 1998 DE 1999	CZ 1998 CZ 1999 DE 1992 DE 2000 DE 2001	NL 1998
BE 1992 BE 1993 BE 1994 BE 1995 BE 1996 BE 1997 BE 1998	BE 1999	BE 2000 BE 2001 BE 2002 FR 1998	FR 1999	FR 2000 FR 2001	NL 1995	NL 1996	AT 1995 AT 1996 AT 1999 NL 1997
LU 1994 LU 1995 LU 1996	LU 1992 LU 1993 LU 1997 LU 1998	LU 1999 LU 2000	LU 2001 LU 2002 LU 2003 NL 1992 NL 1993 NL 1994	SK 1998	SI 1996 SI 1997 SI 1998 SI 1999 SI 2000 SI 2001 SI 2002 SI 2003		AT 1994 AT 1997 AT 1998 AT 2000 AT 2001 AT 2002 AT 2003

Fig. 3. Carte de Kohonen traditionnelle : répartition des pays.

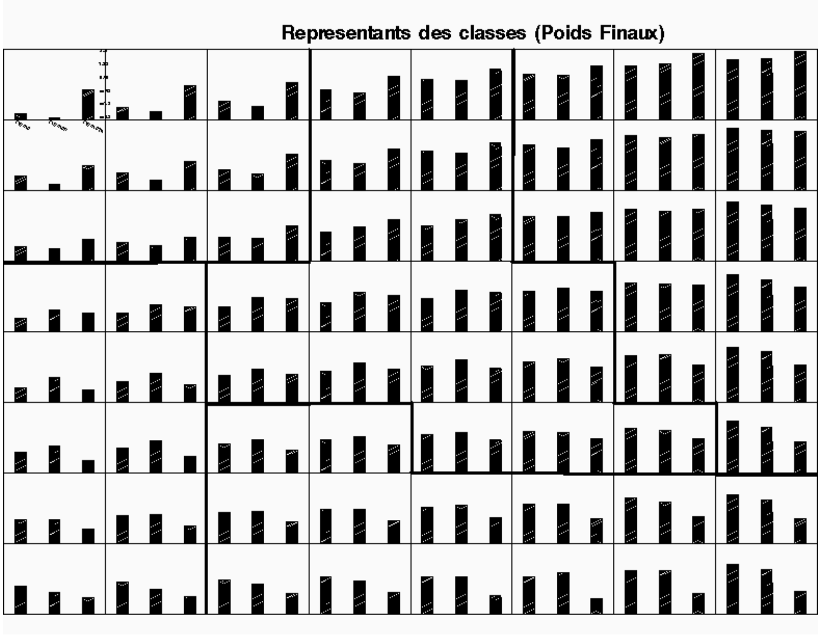


Fig. 4. Carte de Kohonen traditionnelle : répartition des variables.

4 L'évolution des taux d'emploi et d'autres indicateurs

Nous étendons dans cette section l'analyse menée sur les taux d'emploi à un ensemble plus large d'indicateurs relatifs à l'emploi et au chômage, afin de caractériser de manière plus fine les performances des marchés du travail européens. Nous considérons en plus des taux d'emploi total, des femmes et des travailleurs âgés, le taux d'emploi des jeunes (15-24 ans), le taux de croissance de l'emploi, le taux de chômage, le taux de chômage de longue durée (plus de 12 mois) et le taux de contrats à durée déterminée (CDD). La carte obtenue (figure 5) indique que la méthode proposée dans cet article est robuste à un ensemble plus important de variables.

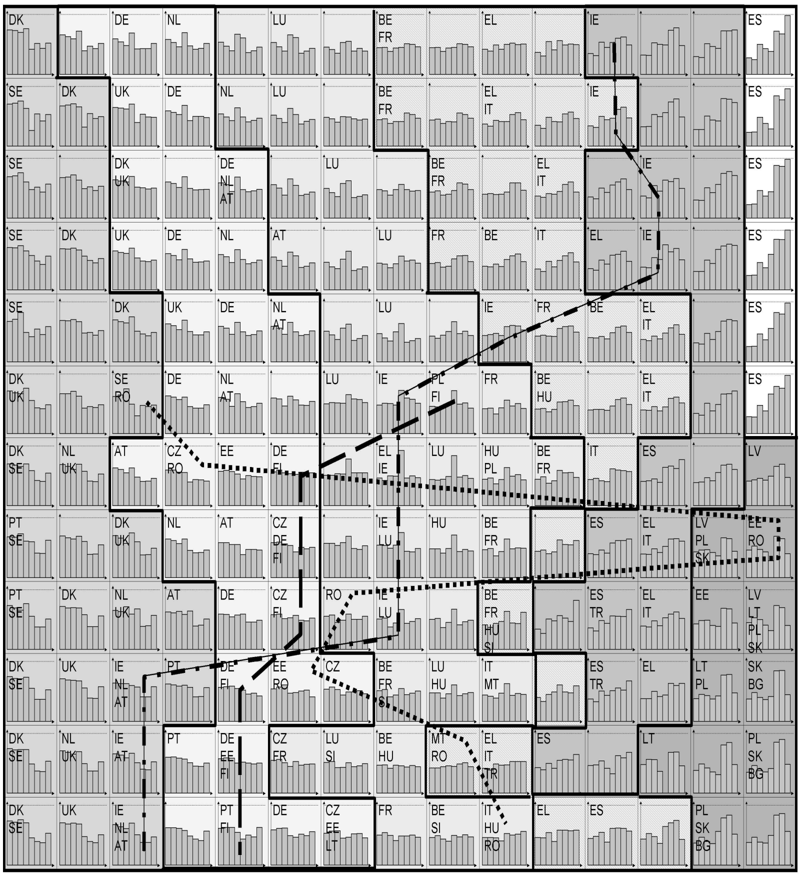


Fig. 5. Carte de Kohonen généralisée.

Note : La première ligne correspond à 1992 et la dernière à 2003. Dans chaque case de la carte sont représentées de gauche à droite les répartitions de TxEmp, TxEmpF, TxEmpTA, TxEmpJ, CroissEmp, Cho, ChoLD, CDD.

La classification hiérarchique ascendante conduit à retenir 7 super-classes. La super-classe se trouvant à gauche de la carte correspond encore aux pays enregistrant les meilleures performances en matière de taux d'emploi, de croissance de l'emploi et les plus faibles taux de chômage, combinés avec des taux moyens de CDD. On retrouve les pays nordiques (Suède, Danemark, Pays-Bas ainsi que Royaume-Uni). Les deux super-classes à coté correspondent à des performances moyennes : on trouve l'Allemagne, l'Autriche, la Finlande à la fin de la période. La super-classe en haut à droite de la carte, correspondant à des taux d'emploi très faibles et des taux de chômage très élevés (Espagne), disparaît en 1998. La classe en bas à droite est principalement caractérisée par des taux de chômage total et de longue durée très élevés, on y trouve des pays entrés récemment dans l'UE (Pologne, Slovaquie, Lituanie, ainsi que la Bulgarie). La super-classe aux performances médiocres où se trouve l'Espagne comprend aussi la Turquie dans les années 2000. On note les meilleures performances de l'Estonie, de la République tchèque ou même de la Hongrie.

5 Conclusion

A travers l'étude des performances en matière d'emploi des pays européens et des pays candidats à l'adhésion, cet article a illustré les avantages d'une carte de Kohonen adaptée afin de prendre en compte la dimension temporelle de données. Cette méthode permet notamment de suivre très facilement les trajectoires suivies par les pays.

References

[Aaron *et al.*, 2003]C. Aaron, C. Perraudin, and J. Rynkiewicz. Curves based kohonen map and adaptative classification : an application to the convergence of the european countries. Prepub SAMOS 190, 2003.

[Akarçay-Gürbüz and Perraudin, 2004]A. Akarçay-Gürbüz and C. Perraudin. How to situate the turkish economy among the european union economies? an exploratory analysis. *European Journal of Economics and Social Systems*, pages 41–62, 2004.

[Commission, 2004]European Commission. *Employment in Europe 2004*. European Communities, 2004.

[Cottrell *et al.*, 2003]M. Cottrell, S. Ibbou, P. Letrémy, and P. Rousset. Cartes auto-organisées pour l'analyse exploratoire des données et la visualisation. *Journal de la société française de statistique*, pages 67–106, 2003.

[Kohonen, 1995]T. Kohonen. *Self-Organizing Maps.* Springer, 1995.

[Letrémy, 2000]P. Letrémy. Notice d'installation et d'utilisation de programmes basés sur l'algorithme de kohonen et dédiés à l'analyse des données. Prepub SAMOS 121, 2000.

Missing values: processing with the Kohonen algorithm

Marie Cottrell and Patrick Letrémy

SAMOS-MATISSE
Université Paris 1
90, rue de Tolbiac, 75634 Paris Cedex 13, France
(e-mail: `cottrell@univ-paris1.fr, pley@univ-paris1.fr`)

Abstract. We show how it is possible to use the Kohonen self-organizing algorithm to deal with data with missing values and estimate them. After a methodological reminder, we illustrate our purpose with three applications to real-world data.

Nous montrons comment il est possible d'utiliser l'algorithme d'auto-organisation de Kohonen pour traiter des données avec valeurs manquantes et estimer ces dernières. Après un rappel méthodologique, nous illustrons notre propos à partir de trois applications à des données réelles.

Keywords: Data Analysis, Kohonen maps, Missing Values.

1 Introduction

The processing of data which contain missing values is a complicated and always awkward problem, when the data come from real-world contexts. In applications, we are very often in front of observations for which all the values are not available, and this can occur for many reasons: typing errors, fields left unanswered in surveys, etc.

Most of the statistical software (as SAS for example) simply suppresses incomplete observations. It has no practical consequence when the data are very numerous. But if the number of remaining data is too small, it can remove all significance to the results.

To avoid suppressing data in that way, it is possible to replace a missing value with the mean value of the corresponding variable, but this approximation can be very bad when the variable has a large variance.

So it is very worthwhile seeing that the Kohonen algorithm (as well as the Forgy algorithm) perfectly deals with data with missing values, without having to estimate them beforehand. We are particularly interested in the Kohonen algorithm for its visualization properties.

In Smaïl Ibbou's PHD thesis, one can find a chapter about this question, but it has not been published yet. The examples are run with the software written by Patrick Letrémy in IML-SAS and available on the SAMOS WEB page (http://samos.univ-paris1.fr).

2 Adaptation of the Kohonen algorithm to data with missing values

We do not remind of the definition of the Kohonen algorithm here, see for example Kohonen [Kohonen, 1995], or [Cottrell *et al.*, 2003].

Let us assume that the observations are real-valued p-dimensional vectors, that we intend to cluster into n classes.

When the input is an incomplete vector x, we first define the set M_x of the numbers of the missing components. M_x is a sub-set of $\{1, 2, \dots, p\}$. If $C = (C_1, C_2, ..., C_n)$ is the set of code-vectors at this stage, the winning code-vector $C_{i_0(x,C)}$ related to x is computed as by setting

$$i_0(x, C) = \arg\min_i \|x - C_i\|,$$

where the distance $\|x - C_i\|^2 = \sum_{k \notin M_x} (x_k - C_{i,k})^2$ is computed with the components present in vector x.

One can use incomplete data in two ways:

a) If we want to use them during the construction of the code-vectors, at each stage, the update of the code-vectors (the winning one and its neighbors) only concerns the components present in the observation. Let us denote $C^t = (C_1^t, C_2^t, ..., C_n^t)$ the code-vectors at time t and if a randomly chosen observation x^{t+1} is drawn, the code-vectors are updated by setting:

$$C_{i,k}^{t+1} = C_{i,k}^t + \epsilon(t)(x_k^{t+1} - C_{i,k}^t)$$

for $k \notin M_x$ and j neighbor of $i_0(x^{t+1}, C^t)$. Otherwise,

$$C_{i,k}^{t+1} = C_{i,k}^t.$$

The sequence $\epsilon(t)$ is [0,1]-valued with $\epsilon(0) \simeq 0.5$ and converges to 0 as $1/t$. After convergence, the classes are defined by the nearest neighbor method.

b) If the data are numerous enough to avoid using the incomplete vectors to build the map, one can content oneself with classifying them after the map is built, as supplementary data, by allocating them to the class with the code-vector which is the nearest for the distance restricted to non-missing components.

This method yields excellent results, provided a variable is not totally or almost totally missing, and also provided the variables are correlated enough, which is the case for most real data bases. Several examples can be encountered in Smaïl Ibbou's PHD thesis [Ibbou, 1998] and also in Gaubert, Ibbou and Tutin [Gaubert *et al.*, 1996].

3 Estimation of missing values, computation of membership probabilities

Whatever the method used to deal with missing values, one of the most interesting properties of the algorithm is that it allows an a posteriori estimation of these missing values.

Let us denote by $C = (C_1, C_2, \dots, C_n)$ the code-vectors after building the Kohonen map. If M_x is the set of missing component numbers for the observation x, and if x is classified in class i, for each index k in M_x, one estimates x_k by:

$$\hat{x}_k = C_{i,k}.$$

Because in the end of the learning the Kohonen algorithm uses no more neighbor (0 neighbor algorithm), we know that the code-vectors are asymptotically near the mean values of their classes. This estimation method therefore consists in estimating the missing values of a variable by the mean value of its class.

It is clear that this estimation is all the more precise as the classes built by the algorithm are homogeneous and well separated. Numerous simulations have shown as well for artificial data as for real ones, that when the variables are sufficiently correlated, the precision of these estimations is remarkable, [Ibbou, 1998].

It is also possible to use a probabilistic classification rule, by computing the membership probabilities for the supplementary observations (be they complete or incomplete), by putting:

$$Prob(x \in \text{Class } i) = \frac{\exp(-\|x - C_i\|^2)}{\sum_{k=1}^{n} \exp(-\|x - C_k\|^2)}.$$

These probabilities also give confirmation of the quality of the organization in the Kohonen map, since significant probabilities have to correspond to neighboring classes.

Moreover, to estimate the missing values, one can compute the weighted mean value of the corresponding components. The weights are the membership probabilities. If x is an incomplete observation, and for each index k in M_x, one estimates x_k by:

$$\hat{x}_k = \sum Prob(x \in \text{Class } i)\ C_{i,k}.$$

These probabilities also provide confidence intervals, etc. In the following sections, we present three examples extracted from real data.

4 Socio-economic data

The first example is classical. The database contains seven ratios measured in 1996 on the macroeconomic situation of 182 countries. This data set was first used by F. Blayo and P. Demartines [Blayo and Demartines, 1991] in the context of data analysis by SOMs.

The measured variables are: annual population growth (ANCRX), mortality rate (TXMORT), illiteracy rate (TXANAL), population proportion in high school (SCOL2), GDP per head (PNBH), unemployment rate (CHOMAG), inflation rate (INFLAT).

Among the set of 182 countries, only 115 have no missing values, 51 have only one missing value, while 16 have 2 or more than 2 missing values.

Therefore we use the 115 + 51 = 166 complete or almost complete countries to build the Kohonen map, and we then classify the 16 remaining countries. The data are centered and reduced as classically. We take a Kohonen map with 7 by 7 units, that is 49 classes. Figure 1 shows the contents of the classes. The 166 countries that were used for computing the code-vectors are in normal font, the 16 others in underlined italics.

Fig. 1. The 182 countries (166 + 16) on a 7 by 7 map, 1500 iterations

We can see that rich countries are in the top left hand corner, very poor ones are displayed in the top right hand corner. Ex-socialist countries are not very far from the richest, etc. As for the 16 countries which are classified after the learning as supplementary observations, we observe that the logic is respected. Monaco and Vatican are displayed with rich countries, and Guinea with very poor countries, etc.

From these computations, it is possible to calculate the membership probabilities of each supplementary observation of each of the 49 classes.

For example, the probabilities that Cuba belongs to class i are greater than 0.03 for classes $i = (1,1),(2,1),(3,1),(4,1),(5,1),(6,1),(7,1),(1,2),(2,2),$ $(3,2),(4,2),(5,2),(6,2),(7,2),(3,3),(4,3),(6,3),(7,3),$ the maximum (0.06) being reached for class $(5,2)$. We can notice (figure 1) that they are neighboring classes. From these probabilities, it is possible to estimate the distribution of the estimators of the missing values. For Cuba, the variables in question are GDP, Unemployment and Inflation.

From these results, it is possible (as it will be shown in the talk) to build super-classes by using an ascending hierarchical classification of the code-vectors and then to cross this classification with other exogenous classifications, etc.

5 Study of the property market in Ile-de-France

The second example is extracted from a study commissioned by the direction of Housing in the Regional Direction of Equipment in Ile-de-France (DHV/DREIF). This was achieved in 1993 by Paris 1 METIS and SAMOS laboratories, by Gaubert, Tutin and Ibbou, [Gaubert *et al.*, 1996].

For 205 towns in Ile-de-France considered in 1988, we have property data (housing rents and prices, old and new, collective or individual, standard or luxurious, office rents and prices, old and new). Structurally, some of the data are missing, for example the office market can be nonexistent in some towns.

This is a case where some data are structurally missing, and where the number of towns is dramatically reduced if one suppresses those which are incomplete: only 5 out of 205 would be kept! So for the learning, we use 150 towns which have less than 12 missing values out of 15. After that, the 55 towns which have more than 12 missing values out of 15 are classified as supplementary observations.

Figure 2 displays the 205 towns (with and without missing values) classified on a 7 by 7 Kohonen map. Note that there are about 63% of missing values on the data set.

In this example which is practically impossible to deal with using classical software, we see that the Kohonen algorithm nevertheless allows to classify extremely sparse data, without introducing any rough error. The results

Fig. 2. The 205 towns in Ile-de-France, in underlined italics the 55 towns which have more than 12 missing values out of 15 variables

are perfectly coherent, even though the data are seriously incomplete. The districts of Paris, Boulogne and Neuilly sur Seine are in the bottom left hand corner. On a diagonal stripe, one finds the towns of the inner suburbs (petite couronne), further right there are the towns of the outer suburbs (grande couronne). Arcueil is classified together with l'Haÿ-les-Roses (class (2,3)), Villejuif with Kremlin Bicêtre (4,6), etc.

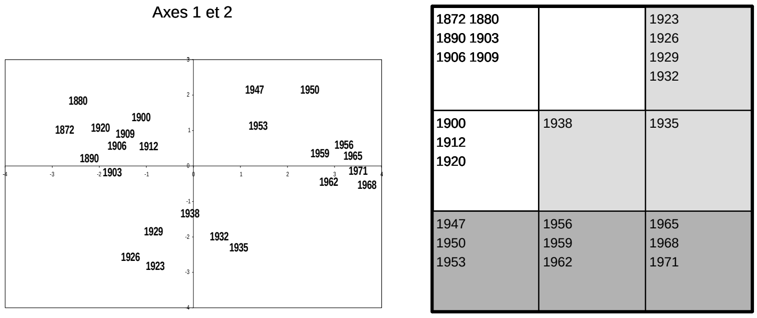
Of course, these good results can be explained by the fact that the 15 measured variables are well correlated and that the present values contain information about missing values. The examination of the correlation matrix (that SAS computes even in case of missing values) shows that 76 coefficients out of 105 are greater than 0.8, none of them being less than 0.65.

6 Structures of Government Spending from 1872 to 1971

The third example is a very classical one in data analysis, taken from the book “Que-sais-je ?” by Bouroche and Saporta , “L’analyse des données” [Saporta, 1981]. The problem is to study the government spending, measured over 24 years between 1872 and 1971, by a 11-dimensional vector: Public Authorities (Pouvoirs publics), Agriculture (Agriculture), Trade and Industry (Commerce et industrie), Transports (Transports), Housing and Regional Development (Logement et aménagement du territoire), Education and Culture (Education et culture), Social Welfare (Action sociale), Veterans (Anciens combattants), Defense (Défense), Debt (Dette), Miscellaneous (Divers). It is a very small example, with 24 observations of dimension 11, without any missing values.

A Principal Component Analysis provides an excellent representation in two dimensions with 64% of explained variance. See figure 3.

Fig. 3. On the left, the projections on the first two principal axes; on the right, the Kohonen map with 9 classes and 3 super-classes



1872 1880
1890 1903
1906 1909
		1923
		1926
		1929
		1932

1900 1912 1920	1938	1935
1947 1950 1953	1956 1959 1962	1965 1968 1971

On this projection, the years split up into three groups, which correspond to three clearly identified periods (before the First World War, between the two World Wars, after the Second World War). Only the year 1920, the first year when an expenditure item for Veterans appears, is set inside the first group, while it belongs to the second one. On the Kohonen map, the three super-classes (identical to the ones just defined) are identified by an ascending hierarchical classification of the code-vectors.

In this example, we have artificially suppressed randomly chosen values which were present in the original data, from 1 value out of 11 to 8 values out of 11, in order to study the clustering stability and compute the accuracy of the estimations that we get by taking the corresponding values of the code-vectors.

One can observe that the three super-classes remain perfectly stable as long as one does not suppress more than 3 values a year, that is 27% of the values.

Then we estimate the suppressed values in each case. The next table shows the evolution of the mean quadratic error according to the number of suppressed values.

Number of missing values	1	2	3	4	5	6	7	8
Percentage of missing values	9%	18%	27%	36%	45%	55%	64%	73%
	0.39	0.54	0.73	1.11	1.31	1.30	1.27	1.39

Table 1. Mean Quadratic Error according to the number of suppressed values

We notice that the error remains small as long as we do not suppress more than 3 values a year.

7 Conclusion

Through these three examples, we have thus shown how it is possible and desirable to use Kohonen maps when the available observations have missing values. Of course, the estimations and the classes that we get are all the more relevant since the variables are well correlated.

Example 2 shows that it can be the only possible method when the data are extremely sparse. Example 3 shows how this method allows to estimate the absent values with good accuracy. The completed data can then be dealt with using any classical treatment.

References

[Blayo and Demartines, 1991]F. Blayo and P. Demartines. Data analysis : How to compare kohonen neural networks to other techniques ? In A. Prieto, editor, *Proceedings of IWANN'91*, pages 469–476, 1991.

[Bouroche and Saporta, 1980]J.-M. Bouroche and G. Saporta. *L'analyse des Données.* PUF, Paris, 1980.

[Cottrell *et al.*, 2003]M. Cottrell, S. Ibbou, P. Letrémy, and P. Rousset. Cartes auto-organisées pour l'analyse exploratoire de données et la visualisation. *Journal de la Société Française de Statistique*, pages 67–106, 2003.

[Gaubert *et al.*, 1996]P. Gaubert, S. Ibbou, and C. Tutin. Segmented real estate markets and price mechanisms : the case of paris. *International Journal of Urban and Regional Research*, pages 270–298, 1996.

[Ibbou, 1998]S. Ibbou. *Thèse : Classification, analyse des correspondances et méthodes neuronales.* Université Paris 1 Panthéon-Sorbonne, 1998.

[Kohonen, 1995]T. Kohonen. *Self-Organizing Maps.* Springer, 1995.

Part VI

Discriminant analysis and learning

Classification via kernel regression based on univariate product density estimators

Bezza Hafidi[1], Abdelkarim Merbouha[2], and Abdallah Mkhadri[1]*

[1] Department of Mathematics,
Cadi Ayyad University, BP 3290 Marrakech, Moroco
(e-mail: b.hafidi@ucam.ac.ma, mkhadri@ucam.ac.ma)

[2] Department of Mathematics
FST-Beni-Mellal, Morocco
(e-mail: merbouhak@yahoo.fr)

Abstract. We propose a nonparametric discrimination method based on a nonparametric Nadaray-Watson kernel regression type-estimator of the posterior probability that an incoming observed vector is a given class. To overcome the curse of dimensionality of the multivariate kernel density estimate, we introduce a variance stabilizing approach which constructs independent predictor variables. Then, the multivariate kernel estimator is replaced by the univariate kernel product estimators. The new procedure is illustrated in simulated data sets and real example, confirming the usefulness of our approach.

Keywords: Classification, Density estimation, Kernel regression, Component Principal Analysis.

1 Introduction

The basic problem in classification is to assign an unknown subject to one of K groups $G_1, \dots, G_K$ on the basis of a multivariate observation $\mathbf{x} = (x_1, \dots, x_p)^t$, where p represents the number of variables and t denotes the transpose operation. However, in practice, the form of class-conditional densities is seldom known. To overcome this problem, one can consider a nonparametric classification method, which uses a nonparametric multivariate kernel density estimates instead of the parametric densities.

Indeed, recently much attention has been given to the application of nonparametric procedures in the classification problem, which have been shown to exhibit superior performance over standard parametric methods such as linear discriminant analysis (LDA) or quadratic discriminant analysis (QDA) in a wide variety of problems. The recent book of [Hastie *et al.*, 2001] presents an excellent overview of nonparametric classification methods. A disadvantage of such models may be a lack of parsimony in the final model and a sensitivity to the "curse of dimensionality" when the dimension p is large and the sample sizes are moderate.

* The third author is supported by TWAS Research grant 01-159 RG/MATHS/AF/AC.

Two semiparametric alternative models for classification, which are a generalization of the model assumed by LDA and QDA, are recently proposed by [Cooley and MacEachern, 1998] and [Amato *et al.*, 2003]. This generalization relies upon a transformation of the data based on pseudo-independent variables. Then, the multivariate kernel density estimates are replaced by the univariate product kernel estimators. [Cooley and MacEachern, 1998] used principal component analysis (PCA) to obtain a transformation matrix, while [Amato *et al.*, 2003] considered independent component analysis (ICA) (cf. [Comon, 1994]).

In this paper, we propose a nonparametric discrimination method based on a nonparametric Nadaray-Watson kernel regression type-estimator of the posterior probability that an incoming observed vector is a given class. To overcome the curse of dimensionality we introduce a Cooley and MacEachern's variance stabilizing approach which constructs independent predictor variables. Then, the multivariate kernel density estimates is replaced by product of univariate kernel estimators. Some theoretical result on Bayes risk consistence is discussed.

This article is organized as follows. In Section 2, we briefly review the nonparametric classification rules which product indirect estimation of the conditional group probability (or a posteriori probability). We also recall the classification approach based on univariate product density estimators which is an alternative interpretation of LDA and QDA. Section 3 is devoted to our new variance stabilizing kernel regression classification approach. Some theoretical asymptotic result on Bayes consistency is discussed in the same section. In Section 4, we apply our new classification rule to some simulations data sets and a real example, confirming the usefulness of our approach. Section 5 ends with some conclusions.

2 Nonparametric classification rules

The multiple classification problem is well studied in statistics. Typically, there is a qualitative random variable Y that takes on a finite number K of values which we refer to as groups: $G_1, \ldots, G_K$. To assign an individual to one of K distinct groups, we must build an allocation rule from the training sample $(\mathbf{x_1}, y_1), \ldots, (\mathbf{x_n}, y_n)$, where $\mathbf{x_i} \in I\!R^p$ is the observation vector and $y_i \in \{1, \ldots, K\}$ indicates the a priori group membership of $\mathbf{x_i}$.

As is well known, the optimal classification rule $d(\mathbf{x})$ allocates an observed p-variate vector $\mathbf{x}$ via

$$d(\mathbf{x}) = \text{argmax}_{j=1,\ldots,K} I\!P(Y = y_j|\mathbf{x}), \tag{1}$$

where

$$I\!P(Y = y_j|\mathbf{x}) = \frac{\pi_j f_j(\mathbf{x})}{\sum_{i=1}^{K} \pi_i f_i(\mathbf{x})} \tag{2}$$

is the a posteriori probability of group j (or the conditional group probability), with π_j and $f_j(.)$ the a priori probability and group-conditional density of group j, respectively. In practice classification rules are constructed either by combining (1) with (2) and estimating the group densities f_j or by estimating directly the *a posteriori* probability $P(Y = y_j|\mathbf{x})$ from the given data. The first approach is called generative method, while the latter approach is called discriminative method.

2.1 Generative nonparametric rules

Most important parametric and nonparametric generative classification rules based on the direct estimation of group densities are Gaussian discriminant analysis (GDA) and kernel density classification, respectively. In kernel density classification the group-conditional densities are estimated with multivariate kernel density estimators which have the form

$$\hat{f}_j(\mathbf{x}) = \frac{1}{n_j} \sum_{\ell=1}^{n_j} \mathcal{K}(\mathbf{x} - \mathbf{x}_{j\ell}, H_j),$$

where $n_j = \#\{i : y_i = j\}, \{\mathbf{x}_{j1}, \ldots, \mathbf{x}_{jn_j}\}$ is the training sample of group j, $\mathcal{K}(., H_j)$ denotes a multivariate kernel function from $I\!R^p$ to $I\!R$, and H_j is a usually a p-dimensional vector of smoothing parameters that governs the degree of smoothness of the estimate (cf. [Scott, 1992]). The recent book of [Hastie *et al.*, 2001] presents an excellent overview of new nonparametric classification methods. A disadvantage of such models may be a lack of parsimony in the final model and a sensitivity to the "curse of dimensionality" when the dimension p is large and the sample sizes are moderate.

2.2 Kernel univariate product estimators

In order to avoid the biased tail estimation and the curse of dimensionality common to multivariate kernel density estimation, [Cooley and MacEachern, 1998] (see also [Amato *et al.*, 2003]) present an alternative view of QDA and LDA which allows them to extend the nonparametric classification problem. In this alternative rotations of the coordinate axes are employed to obtain an assumed mutual independence among the components of the rotated data. Then, the conditional density of the kth sample group can be written as the product of univariate Gaussian density on the transformed sample, i. e.

$$f_k(\mathbf{x}) = f(H_k\mathbf{x}) = \prod_{j=1}^{p} \frac{1}{\sigma_{jk}} \phi\left(\frac{(H_k\mathbf{x})_j - (H_k\mu_k)_j}{\sigma_{jk}}\right), \tag{3}$$

where $\phi(.)$ denotes the density of a standard normal variable and H_k is the transform matrix obtained from the spectral decomposition of the covariance matrix Σ_k. Then, a natural generalization of LDA and QDA is to replace the univariate Gaussian densities with univariate kernel density, we called kernel product estimator (KPE) the resulting estimator. From the latter algorithm, QDA and LDA are therefore just affected by the way the H_k are estimated. [Cooley and MacEachern, 1998] consider the principal component analysis (PCA) to estimate H_k, while [Amato *et al.*, 2003] propose to use independent component analysis (ICA).

3 Classification via kernel regression estimator

There are several compelling reasons for using discriminative rather than generative classifiers. The first one is that in real world problems the assumed generative model is rarely exact, and asymptotically a discriminative model should typically be preferred (cf. [Vapnick, 1998]). Moreover, there are many problems in which direct classification does not suffice, and where the precise estimation of the conditional group probabilities is most important. Multiple logistic regression (polychotomous regression) has been used for a long time (cf. [Hosmer and Lemeshow, 1989]) to obtain a direct estimate of all the conditional group probabilities.

On the other hand, little is known about nonparametric kernel discriminative method. One early direct kernel approach was proposed by [Lauder, 1983], which is analogue to kernel density estimation. [Hoti and Holmström, 1999] proposed an analogue Nadaray-Watson type-estimator defined by

$$\hat{r}_n^{(k)}(\mathbf{x}) = \frac{\sum_{i=1}^n T_i^{(k)} K((\mathbf{x} - \mathbf{x}_i)/h_n)}{\sum_{i=1}^n K((\mathbf{x} - \mathbf{x}_i)/h_n)} \tag{4}$$

where $T_i^{(k)} = 1$ if $Y_i = k$ and 0 elsewhere, $\mathcal{K}()$ is a multivariate kernel and $k = 1, \ldots, K$. They further improve the flexibility of the estimator by replacing the constants Y_i^j with locally fitted polynomial functions.

3.1 Regression kernel classification method (RKCM)

We attack the problem of curse of dimensionality of the kernel regression classification method, defined via the Nadaray-Watson type-estimator (4), by adapting the Cooley and MacEachern's variance stabilizing approach to KRCM. It consists to replace in (4) the multivariate kernel density estimator by the product of univariate kernel density estimators, which leads to the new estimator

$$\tilde{r}_n^{(k)}(\mathbf{x}) = \frac{\sum_{i=1}^n T_i^{(k)} \prod_{j=1}^p \hat{f}_{kj}^* \{(\hat{H}_k \mathbf{x})_j - (\hat{H}_k \mathbf{x_i})_j\}}{\sum_{i=1}^n \prod_{j=1}^p \hat{f}_{kj}^* \{(\hat{H}_k \mathbf{x})_j - (\hat{H}_k \mathbf{x_i})_j\}}, \tag{5}$$

where $\hat{f}_{kj}^*(z) = \sum_{\ell : y_\ell = k}^n \mathcal{K}\{(z - (\hat{H}_k \mathbf{X}_{k\ell})_j)/h_{kj}\}/h_{kj} n_k$ is the univariate kernel density estimate in the jth dimension of the transformed space for group

k. We allow the pooling of sample covariance information across K groups to obtain $\hat{H}_1 = \ldots = \hat{H}_K = \hat{H}$. Then the common transformation matrix $\hat{H}$ is estimated via the application of PCA on the pooling sample covariance matrix.

Since many kernel functions are highly efficient, we adopt the Gaussian kernels which are widely used (cf. [Farhmeir and Tutz, 1994], pp. 156-157). For simplicity, we can assume that the smoothing parameter in direction j for group k is constant and equal h. Then, the classical cross-validation of the average squared error criterion is often used for the selection of the smoothing parameter h. But, the cross-validation of the misclassification error rate is more convenient in our context, since it is related to discriminant problem. However, in our experimental study, we fix $h_{kj} = 0.9\sigma_{kj}n^{-1/(p+4)}$ as in [Cooley and MacEachern, 1998] ($k = 1, \ldots, K; j = 1, \ldots, p$). A robust estimation of σ_{kj} can be taken equal to the smaller of the sample standard deviation and (1/1.34) x sample interquartile range. This choice is mainly related to density estimation, but it is simple to compute and seems to work well in our numerical study.

3.2 Consistence and convergence rate

[Cooley and MacEachern, 1998] showed that the rule based on KPE of the density of the kth group is consistent on the set $I\!\!R^p - \mathcal{N}_k$, where $\mathcal{N}_k$ is a set of Lebesgue measure 0 ($k = 1, \ldots, K$). Moreover, they established that the rate of convergence of the mean integrated squared error to 0 is O($n^{-4/5}$), regardless of the dimensionality p.

For our KRCM, we have established that the rule based on the regression kernel product estimator $g_n(.)$ is Bayes risk universally consistent, i.e. $lim_{n \longrightarrow \infty} I\!\!P\{g_n(\mathbf{X}|D_n)\} - L^* = 0$ for any distribution of the pair $(\mathbf{X}, Y)$, where L^* is the optimal Bayes error probability and D_n denote the training sample of size n. The proof is based on the verification of the three conditions of the general Stone's theorem (cf. [Devroye *et al.*, 1996], Theorem 6.3 in page 98). For saving space, this proof is not included in this note.

4 Numerical experiments

In this section, we report on some case studies for analyzing the practical behavior of KRCM relative to LDA, QDA and KPE on the basis of training and test error rate, respectively. For purposes of comparison, the smoothing parameter in direction j and group k is fixed equal to $h_{kj} = 0.9\sigma_{kj}n^{-1/(4+p)}$ for KRCM and KPE, and a priori probabilities were taken to be equal. As indicated in Section 3, σ_{kj} is estimated by the smaller of the sample standard deviation and (1/1.34) x sample interquartile range. We first present some Monte Carlo numerical experiments on simulated data sets, then we present numerical experiment on real data set.

4.1 Monte Carlo numerical experiments

The first simulated example was also considered by [Cooley and MacEachern, 1998], and has two groups and two predictors. The final predictors are combination of two initial predictors, generated from the normal mixture for the first initial predictor and the standard normal for the second one. The difference between the groups lies in the means of the normals in the mixture distribution of the first predictor (cf.[Cooley and MacEachern, 1998]).

Two hundred and fifty sets for the training and test samples consisted of 100 and 900 observations, respectively, were run from an equal mixture of the two distributions. Table 1 shows the averaged success rates for the training data set and the test data set over 200 simulations, with the standard error of the average in the parentheses. It appears that KRCM performs well than KPE, QDA and LDA, in both training and test sample respectively.

In the second example the optimal boundaries separating the group are non-additive functions of the predictors. The observations of the two groups are described by 6 predictors, the last four of which are random $\mathcal{N}(0, 1)$ noise variables for both groups. The first two predictors of group 1 are independent Uniform$[-5, 5]$ random variables, whereas the first two variables of group 2 form bivariate normal vectors with means 0, variance 1 and correlation coefficient 1/2. Similar example appears in [Cooley and MacEachern, 1998], where all relevant discriminatory information is contained in a relatively small dimension.

We select a training sample of size 500 and a test sample of size 3000, both from an equal mixture of the two populations. For both the training data set and the test data set, the averaged success rates and their standard errors over 75 replicates are summarized in Table 1. The behavior of KRCM is similar to that in the first example, where the difference with KPE is more important (15% on test data).

The third example is a well-known *waveform* problem composed of three groups with 21 predictors. The predictors are defined by

$$\begin{aligned} x_i &= uh_1(i) + (1-u)h_2(i) + \epsilon_i \qquad \text{Group1} \\ x_i &= uh_1(i) + (1-u)h_3(i) + \epsilon_i \qquad \text{Group2} \\ x_i &= uh_2(i) + (1-u)h_3(i) + \epsilon_i \qquad \text{Group3,} \end{aligned}$$

where $i = 1, \ldots, 21$, u is uniform on $[0, 1]$, $\epsilon_i \sim \mathcal{N}(0, 1)$ and the h_i are the shifted triangular forms defined by: $h_1(i) = \max(6 - [i - 11], 0), h_2(i) = h_1(i - 4)$ and $h_3(i) = h_1(i + 4)$.

The training and test sets consisted of 500 and 300 observations, respectively, are selected and their averaged success rates are shown in Table 1 where equal prior are used. Again, KRCM is better than QDA, LDA and KPE.

	Mixture data		Nonadditive boundary		Waveform	
Method	Train	Test	Train	Test	Train	Test
LDA	62.92(.050)	59.23(.015)	84.32(.018)	50.62(.012)	97.72(.005)	97.44(.008)
QDA	61.73(.046))	59.22(.013)	85.13(.021)	74.72(.054)	97.95(.007)	96.25(.017)
KPE	78.14(.043)	76.23(.015)	85.39(.023)	84.75(.012)	91.22(.002)	93.89(.003)
KRCM	83.17(.034)	77.31(.015)	99.92(.001)	99.04(.002)	100(.000)	100(.000)

Table 1. Average success rates and standard deviation in parentheses.

4.2 Real data example

The real data set considered is the Diabetes in Pima Indian Women. It is described for instance in [Ripley, 1996]. It concerns a population of $n = 532$ women who were at least 21 years old, of Pima Indian heritage and living near Phoenix, Arizona, was tested for diabetes according to World Health Organization criteria. This women described by 7 predictors and two groups. The data were collected by the US National Institute of Diabetes and Digestive and Kidney Diseases. The training set contains a randomly selected set of 200 subjects, and the sample test set contains the remaining 332 subjects.

For both the training data set and the test data set, the success rates is summarized in Table 2. Here again, the behavior of KRCM is better than all the other methods.

	Pima	
Method	Train	Test
LDA	76.000	77.108
QDA	76.500	69.879
KPE	85.000	81.626
KRCM	99.000	99.698

Table 2. Success rates corresponding to Pima data set.

5 Discussion

In this paper, we propose a nonparametric discrimination method based on a nonparametric Nadaray-Watson kernel regression type-estimator of the posterior probability that an incoming observed vector is a given class. To overcome the curse of dimensionality we introduce a Cooley and MacEachern's variance stabilizing approach which constructs independent predictor variables. Then, the multivariate kernel density estimates is replaced by product of univariate kernel estimators.

Summarizing results experiments, performance of KRCM is very good compared with KPE, LDA and QDA. Consequently, our study confirms that using discriminative rather than generative classifiers is preferred.

References

[Amato *et al.*, 2003]U. Amato, A. Antoniadis, and Gregoire. Independent component discriminant analysis. *Int. Math. J.*, pages 727–734, 2003.

[Comon, 1994]P. Comon. Independent component analysis, a new concept. *Signal Processing*, pages 187–314, 1994.

[Cooley and MacEachern, 1998]C. A. Cooley and MacEachern. Classification via kernel product estimators. *Biometrika*, pages 823–833, 1998.

[Devroye *et al.*, 1996]L. Devroye, L. Gyorfi, and G. Lugosi. *A probabilistic theory of pattern recognition.* Springer-Verlag, New-York, 1996.

[Farhmeir and Tutz, 1994]L. Farhmeir and G. Tutz. *Multivariate statistical modelling based generalized linear models.* Springer-Verlag, New-York, 1994.

[Hastie *et al.*, 2001]T. Hastie, R. Tibshirani, and J. Friedman. *The elements of statistical learning.* Springer-Verlag, New-York, 2001.

[Hosmer and Lemeshow, 1989]D.W. Hosmer and S. Lemeshow. *Applied logistic regression.* Wiley, New-York, 1989.

[Hoti and Holmström, 1999]F. Hoti and L. Holmström. Reduced kernel regression for fast classification. *P. Brenner, L. Arkeryd, J. Rergh and R. Pettersson*, pages 405–412, 1999.

[Lauder, 1983]I. J. Lauder. Direct kernel assessment of diagnostic probabilities. *Biometrika*, pages 254–256, 1983.

[Ripley, 1996]B.D. Ripley. *Pattern Recognition and Neural Networks.* Cambridge University Press, Cambridge, 1996.

[Scott, 1992]D.W. Scott. *Multivariate density estimation: theory, practice and visualization.* Wiley, New-York, 1992.

[Vapnick, 1998]V. N. Vapnick. *Statistical learning theory.* John Wiley & Sons, 1998.

Model Selection for Multi-class SVMs

Yann Guermeur[1], Myriam Maumy[2], and Frédéric Sur[1]

[1] LORIA-CNRS
Campus Scientifique, BP 239,
54506 Vandœuvre-lès-Nancy Cedex, France
(e-mail: `Yann.Guermeur@loria.fr, Frederic.Sur@loria.fr`)

[2] IRMA-ULP
7 rue René Descartes
67084 Strasbourg Cedex, France
(e-mail: `mmaumy@math.u-strasbg.fr`)

Abstract. In the framework of statistical learning, fitting a model to a given problem is usually done in two steps. First, model selection is performed, to set the values of the hyperparameters. Second, training results in the selection, for this set of values, of a function performing satisfactorily on the problem. Choosing the values of the hyperparameters remains a difficult task, which has only been addressed so far in the case of bi-class SVMs. We derive here a solution dedicated to M-SVMs. It is based on a new bound on the risk of large margin classifiers.
Keywords: Multi-class SVMs, hyperparameters, soft margin parameter.

1 Introduction

When support vector machines (SVMs) [Vapnik, 1998] were introduced in the early nineties, they were seen by some as off-the-shelf tools. This idealistic picture soon proved too optimistic. Not only does their training raise technical difficulties, but the tuning of the kernel parameters and the soft margin parameter C also remains a difficult task. In literature, this question is addressed for (two-class) pattern recognition and function estimation SVMs. The methods proposed often rest on estimates of the true risk of the machine [Chapelle *et al.*, 2002]. The case of multi-class discriminant analysis was only considered in the framework of decomposition schemes [Passerini *et al.*, 2004]. The case of multi-class SVMs (M-SVMs) calls for specific solutions. Indeed, the implementation of the structural risk minimization (SRM) inductive principle [Vapnik, 1982] utterly rests on the availability of tight error bounds and the standard uniform convergence results do not carry over nicely to the case of multi-category large margin classifiers. In this paper, we derive a new bound on the generalization performance of M-SVMs in terms of constraints on the hyperplanes. This bound, interesting in its own right, makes central use of a result relating covering problems and the degree of compactness of operators. It serves as an objective function to tune the value of the soft margin parameter. This way, the value of C and the dual variables α can be determined simultaneously, at a cost of the same order of magnitude as the one of a standard training. The organization of the paper is as follows. Section 2 is devoted to the description of the bound on which the study is based. In Section 3, the measure of capacity involved is bounded in terms of the entropy numbers of a linear operator. The resulting objective function is used in Section 4, to derive the algorithm tuning C and the parameters α. A first assessment of this algorithm on a toy problem is described in Section 5. Due to lack of space, proofs are omitted.

2 Bound on the risk of large margin classifiers

We consider the case of a Q-category pattern recognition problem, with $Q \geq 3$ to exclude the degenerate case of dichotomies. Let $\mathcal{X}$ be the space of description and $\mathcal{C} = \{C_1, \dots, C_k, \dots, C_Q\}$ the set of categories. We make the assumption that there is a joint probability measure μ, fixed but unknown, on $(\mathcal{X} \times \mathcal{C}, \mathcal{B})$, where $\mathcal{B}$ is a σ-algebra on $\mathcal{X} \times \mathcal{C}$. This measure utterly characterizes the problem of interest. Our goal is to find, in a given set $\mathcal{H}$ of functions from $\mathcal{X}$ into $\mathbb{R}^Q$, a function with the lowest "error rate" on this problem. The "error rate" of a function h in $\mathcal{H}$ with component functions h_k, $(1 \leq k \leq Q)$, is the *expected risk* of the corresponding discrimination function, obtained by assigning each pattern x to the category C_k in $\mathcal{C}$ satisfying: $h_k(x) = \max_l h_l(x)$. The patterns for which this assignation is ambiguous are assigned to a dummy category, so that they contribute to the computation of the different risks considered below. Hereafter, $C(x)$ will denote indifferently the category of the (labelled) pattern x, or the index of this category. To simplify notations, when no confusion is possible, the labels of the categories will be identified with their indices, i.e. k could be used in place of C_k. First of all, we define the functional that is to be minimized, the expected risk.

Definition 1 (Expected risk) *The* expected risk *of a function f from $\mathcal{X}$ into $\mathcal{C}$ is the probability that $f(x) \neq C(x)$ for a labelled example $(x, C(x))$ chosen randomly according to μ, i.e.:*

$$R(f) = \mu \left\{ (x,k) : f(x) \neq k \right\} = \int_{\mathcal{X} \times \mathcal{C}} 1\!\!1_{\{f(x) \neq k\}}(x,k) d\mu(x,k) \tag{1}$$

where $1\!\!1_{\{f(x) \neq k\}}$ is the indicator function of the set $\{(x,k) \in \mathcal{X} \times \mathcal{C} : f(x) \neq k\}$.

In the framework of large margin multi-category pattern recognition, the class of functions of interest is not $\mathcal{H}$ itself, but rather its image by an adequately chosen operator. Basically, this is due to the fact that the two central elements to assign a pattern to a category and to derive a level of confidence in this assignation are respectively the index of the highest output and the difference between this output and the second highest one. The operator used here was introduced in previous works.

Definition 2 ($\boldsymbol{\Delta}$ operator) *Define Δ as an operator on $\mathcal{H}$ such that:*

$$\begin{aligned}&\Delta : \mathcal{H} \longrightarrow \Delta\mathcal{H}\\ &\quad h = (h_k)_{1 \leq k \leq Q} \mapsto \Delta h = \left(\Delta h_k : x \mapsto \tfrac{1}{2}\left\{h_k(x) - \max_{l \neq k} h_l(x)\right\}\right)_{1 \leq k \leq Q}.\end{aligned}$$

Let s_m be a m-sample of examples independently drawn from μ. The empirical margin risk is defined as follows:

Definition 3 (Empirical margin risk) *The empirical risk with margin $\gamma > 0$ of h on a set s_m is*

$$R_{\gamma,s_m}(h) = \frac{1}{m} \cdot \#\left\{(x_i, C(x_i)) \in s_m : \ \Delta h_{C(x_i)}(x_i) < \gamma\right\}, \tag{2}$$

where # returns the cardinality of the set to which it is applied.

For technical reasons, it is useful to bound the values taken by the functions Δh_k in $[-\gamma, \gamma]$, the smallest interval such that this change has no incidence on the empirical margin risk. This is achieved by application of the π_γ operator.

Definition 4 ($\boldsymbol{\pi_\gamma}$ operator [Bartlett, 1998]) *Let $\mathcal{G}$ be a set of functions from $\mathcal{X}$ into $\mathbb{R}^Q$. For $\gamma > 0$, let $\pi_\gamma : g = (g_k)_{1 \leq k \leq Q} \mapsto \pi_\gamma(g) = (\pi_\gamma(g_k))_{1 \leq k \leq Q}$ be the piecewise-linear squashing operator defined as:*

$$\forall x \in \mathcal{X}, \ \pi_\gamma(g_k)(x) = \begin{cases} \gamma.sign\,(g_k(x)) & \textit{if } |g_k(x)| \geq \gamma \\ g_k(x) & \textit{otherwise} \end{cases}. \tag{3}$$

Let Δ_γ denote $\pi_\gamma \circ \Delta$ and $\Delta_\gamma\mathcal{H}$ be defined as the set of functions $\Delta_\gamma h$. Our guaranteed risk is made up of two terms, the empirical margin risk given above and a "confidence interval" involving a covering number of $\Delta_\gamma\mathcal{H}$.

Definition 5 ($\boldsymbol{\epsilon}$-cover, $\boldsymbol{\epsilon}$-net and covering numbers) *Let (E, ρ) be a pseudo-metric space and E' be a subset of E. An ϵ-*cover *of E' is a coverage of E' with balls of radius ϵ the centers of which belong to E. These centers form an ϵ-*net *of E'*[1]*. If E' has an ϵ-cover of finite cardinality, then its* covering number *$\mathcal{N}(\epsilon, E', \rho)$ is the smallest cardinality of its ϵ-covers. If there is no such finite cover, then the covering number is defined to be ∞.*

The covering number of interest uses the following pseudo-metric:

Definition 6 *Let $\mathcal{G}$ be a set of functions from $\mathcal{X}$ into $\mathbb{R}^Q$. For a set s of points in $\mathcal{X}$ of finite cardinality, define the pseudo-metric d_s on $\mathcal{G}$ as:*

$$\forall (g, g') \in \mathcal{G}^2, \ d_s(g, g') = \max_{x \in s} \|g(x) - g'(x)\|_\infty. \tag{4}$$

[1] Hereafter, we will only consider a restricted case in which the ϵ-nets of E' will be supposed to be subsets of E' itself.

Let $\mathcal{N}_{\infty,\infty}(\epsilon, \varDelta_\gamma \mathcal{H}, m) = \sup_{s_m \in \mathcal{X}^m} \mathcal{N}(\epsilon, \varDelta_\gamma \mathcal{H}, d_{s_m})$. These definitions being given, we can formulate the following theorem, which extends to the multi-class case Corollary 9 in [Bartlett, 1998].

Theorem 1 (Theorem 1 in [Guermeur, 2004]) *Let s_m be a m-sample of examples independently drawn from μ. With probability at least $1 - \delta$, for every value of γ in $(0, 1]$, the risk of any function h in the class $\mathcal{H}$ of functions computed by a Q-class large margin classifier is bounded from above by:*

$$R(h) \leq R_{\gamma,s_m}(h) + \sqrt{\frac{2}{m}\left(\ln(2\mathcal{N}_{\infty,\infty}(\gamma/4, \varDelta_\gamma \mathcal{H}, 2m)) + \ln\left(\frac{2}{\gamma\delta}\right)\right)} + \frac{1}{m}. \quad (5)$$

The practical interest of such a bound utterly rests on the possibility to derive a tight bound on the covering number appearing in the "confidence interval". To that end, a preliminary simplification is useful.

Proposition 1 $\forall(\gamma, \epsilon) : 0 < \epsilon \leq \gamma \leq 1, \quad \mathcal{N}_{\infty,\infty}(\epsilon, \varDelta_\gamma \mathcal{H}, m) \leq \mathcal{N}_{\infty,\infty}(\epsilon, \mathcal{H}, m)$.

Theorem 1 and Proposition 1 imply that deriving a guaranteed risk for $\mathcal{H}$ can boil down to deriving a bound on $\mathcal{N}_{\infty,\infty}(\epsilon, \mathcal{H}, m)$. In [Guermeur, 2004], to bound the covering number appearing in (5), we investigated a standard pathway, consisting in relating this capacity measure to a generalized VC dimension [Vapnik, 1998] through an extension of Sauer's lemma [Sauer, 1972]. It appeared then that in the multivariate case, establishing the connection between the separation of functions (with respect to the selected pseudo-metric) and their shattering capacity is no longer trivial. Taking our inspiration from [Carl and Stephani, 1990, Williamson *et al.*, 2000], we assess here a more direct approach: relating the covering numbers of $\mathcal{H}$ to the entropy numbers of a linear operator.

3 Bound on the covering numbers of M-SVMs

SVMs [Cortes and Vapnik, 1995] are learning systems introduced by Vapnik and co-workers as a nonlinear extension of the maximal margin hyperplane [Vapnik, 1982]. Originally, they were designed to compute dichotomies. In this context, the principle on which they are based can be outlined very simply. First, the examples are mapped into a high-dimensional Hilbert space thanks to a nonlinear transform. Second, the maximal margin hyperplane is computed in that space, to separate the two categories. Initially, the extension to perform multi-class discriminant analysis utterly rested on decomposition schemes. The M-SVMs are globally more recent (see [Guermeur, 2004] for references). The family $\mathcal{H}$ of functions $h = (h_k)_{1 \leq k \leq Q}$ computed by these machines can be defined by:

$$\forall k \in \{1, \ldots, Q\}, \ h_k(x) = \langle w_k, \varPhi(x) \rangle + b_k, \quad (6)$$

where Φ is some mapping from $\mathcal{X}$ into a Reproducing Kernel Hilbert Space (RKHS) [Aronszajn, 1950] $\left(E_{\Phi(\mathcal{X})}, \langle .,. \rangle\right)$, derived from a symmetric positive kernel κ. The vectors w_k belong to $E_{\Phi(\mathcal{X})}$, whereas the b_k are real numbers. As in the case of all kernel machines, Φ does not appear explicitly in the computations. Thanks to the "kernel trick", which rests on the equation:

$$\forall (x, x') \in \mathcal{X}^2, \; \kappa(x, x') = \langle \Phi(x), \Phi(x') \rangle, \tag{7}$$

all what is needed to perform training or testing are the values taken by the kernel κ. To ensure the finiteness of the capacity measures, we make the additional assumption that $\Phi(\mathcal{X})$ is included in the closed ball of radius $\Lambda_{\Phi(\mathcal{X})}$ in $E_{\Phi(\mathcal{X})}$, that is: $\forall x \in \mathcal{X}, \;\; \|\Phi(x)\| = \sqrt{\kappa(x,x)} \leq \Lambda_{\Phi(\mathcal{X})}$. To upperbound $\mathcal{N}_{\infty,\infty}(\epsilon, \mathcal{H}, m)$ when $\mathcal{H}$ is a M-SVM, we use a result regarding linear operators on Banach spaces. This implies that the covering numbers of $\mathcal{H}$ could be bounded in terms of the covering numbers of its linear counterpart.

Proposition 2 *Let $\mathcal{H}$ be the class of functions implemented by a Q-category SVM under the constraint that $b = (b_k) \in [-\beta, \beta]^Q$. Let $\tilde{\mathcal{H}}$ be the subset of $\mathcal{H}$ made up of the functions for which $b = 0$. Then, for all $\epsilon > 0$,*

$$\mathcal{N}_{\infty,\infty}(\epsilon, \mathcal{H}, m) \leq \left(2 \left\lceil \frac{\beta}{\epsilon} \right\rceil + 1\right)^Q \mathcal{N}_{\infty,\infty}(\epsilon/2, \tilde{\mathcal{H}}, m). \tag{8}$$

A function $\tilde{h}$ in $\tilde{\mathcal{H}}$ is characterized by the vector $\mathbf{w} = (w_k)_{1 \leq k \leq Q}$ in $E^Q_{\Phi(\mathcal{X})}$. This space is endowed with a Hilbertian structure. Its dot product is given by: $\forall (\mathbf{w}, \mathbf{w}') \in \left(E^Q_{\Phi(\mathcal{X})}\right)^2, \; \langle \mathbf{w}, \mathbf{w}' \rangle = \sum_{k=1}^Q \langle w_k, w'_k \rangle$. Its norm is the one derived from $\langle .,. \rangle$. Since the additional hypothesis $\|\mathbf{w}\| \leq 1$ will also be used, we introduce another proposition.

Proposition 3 *Let $\tilde{\mathcal{H}}$ be defined as above, under the additional constraint that $\|\mathbf{w}\| \leq \Lambda_w$. Let $\mathcal{U}$ be its restriction to the functions satisfying $\|\mathbf{w}\| \leq 1$.*

$$\forall \epsilon > 0, \;\; \mathcal{N}_{\infty,\infty}(\Lambda_w \epsilon, \tilde{\mathcal{H}}, m) \leq \mathcal{N}_{\infty,\infty}(\epsilon, \mathcal{U}, m). \tag{9}$$

Definition 7 (entropy numbers) *Let (E, ρ) be a pseudo-metric space. Let E' be a subset of E. The nth* entropy number *of E', $\epsilon_n(E')$, is defined as the smallest real ϵ such that there exists an ϵ-cover of E' of cardinality at most n. Let E and F be two Banach spaces. $\mathfrak{L}(E, F)$ denotes the Banach space of all (bounded linear) operators from E into F equipped with the usual norm. Let U_E be the closed unit ball of E. The nth entropy number of $S \in \mathfrak{L}(E, F)$ is defined as*

$$\epsilon_n(S) = \epsilon_n\left(S(U_E)\right). \tag{10}$$

By ℓ_p^n we denote the vector space of n-tuples equipped with the norm $\|.\|_p$.

Definition 8 (Evaluation operator) *Let s_m be any element of $\mathcal{X}^m$. We define S_{s_m} as the linear operator given by:*

$$\begin{array}{rcl} S_{s_m} : E_{\Phi(\mathcal{X})}^{Q} & \longrightarrow & \ell_{\infty}^{Qm} \\ \mathbf{w} & \mapsto & S_{s_m}(\mathbf{w}) = \left(\langle w_k, \Phi(x_i) \rangle \right)_{1 \leq k \leq Q,\ 1 \leq i \leq m} \end{array}$$

The connection between $\mathcal{N}_{\infty,\infty}(\epsilon, \mathcal{U}, m)$ and the entropy numbers of S_{s_m} is given by the following proposition.

Proposition 4 *If for all $s_m \in \mathcal{X}^m$, $\epsilon_n(S_{s_m}) \leq \epsilon$, then $\mathcal{N}_{\infty,\infty}(\epsilon, \mathcal{U}, m) \leq n$.*

To bound $\epsilon_n(S_{s_m})$, we use a result due to Maurey and Carl.

Lemma 1 (Lemma 6.4.1 in [Carl and Stephani, 1990]) *Let H be a Hilbert space, m a positive integer and $S \in \mathfrak{L}(H, \ell_{\infty}^{m})$. Then, for $1 \leq n \leq m$,*

$$\epsilon_{2^{n-1}}(S) \leq c \|S\| \left(\frac{1}{n} \log \left(1 + \frac{m}{n} \right) \right)^{1/2}, \tag{11}$$

where c is a universal constant and by log *we denote the logarithm to base* 2.

Lemma 1 still holds without the hypothesis $n \leq m$. Gathering the results from Propositions 1 to 4 together with this lemma (applied on S_{s_m}) produces a handy bound on the covering number of interest.

Theorem 2 *Let $\mathcal{H}$ be the class of functions computed by a Q-category M-SVM under the hypothesis that $\Phi(\mathcal{X})$ is included in the closed ball of radius $\Lambda_{\Phi(\mathcal{X})}$ in $E_{\Phi(\mathcal{X})}$ and the constraints that $\|\mathbf{w}\| \leq \Lambda_w$ and $b \in [-\beta, \beta]^Q$. For every value of γ in $(0, 1]$,*

$$\mathcal{N}_{\infty,\infty}(\gamma/4, \Delta_{\gamma}\mathcal{H}, 2m) \leq \left(2 \left\lceil \frac{4\beta}{\gamma} \right\rceil + 1 \right)^{Q} \cdot 2^{\frac{8c\Lambda_w \Lambda_{\Phi(\mathcal{X})}}{\gamma} \sqrt{\frac{2Qm}{\ln(2)}} - 1}. \tag{12}$$

4 Tuning the soft margin parameter

To tune C thanks to the guaranteed risk derived above, we propose a simple line search. Although it is compatible with any of the training algorithms published, for the sake of simplicity, we focus here on the case of the most common machine, introduced in [Weston and Watkins, 1998]. Training it amounts to solving the following quadratic programming (QP) problem:

Problem 1 (Primal).

$$\min_{(\mathbf{w},b)} \left\{ \frac{1}{2} \sum_{k=1}^{Q} \|w_k\|^2 + C \sum_{i=1}^{m} \sum_{k=1}^{Q} \xi_{ik} \right\}$$

$$\text{s.t.} \begin{cases} h_{C(x_i)}(x_i) - h_k(x_i) \geq 1 - \xi_{ik}, & (1 \leq i \leq m),\ (1 \leq k \neq C(x_i) \leq Q) \\ \xi_{ik} \geq 0, & (1 \leq i \leq m),\ (1 \leq k \neq C(x_i) \leq Q) \end{cases}.$$

In the objective function, the sum of slack variables is used in place of the empirical margin risk, whereas the penalty term $\frac{1}{2}\sum_{k=1}^{Q}\|w_k\|^2$ is added to perform capacity control. To the best of our knowledge, Theorem 2 offers the first justification for this choice. By setting the soft margin parameter C, one specifies a compromise between training accuracy and complexity. If the objective function itself cannot be used in that purpose, since it is only distantly related to a guaranteed risk, performing n-fold cross-validation is a sensible possibility. However, it implies training the machine n times for each value of C considered, which can be prohibitive in terms of `cpu` time requirements. Furthermore, this no longer corresponds to the implementation of the SRM principle. In that respect, our solution should prove more satisfactory. To detail it, we first introduce the formulation in which Problem 1 is solved, its Wolfe dual. Let α_{ik} be the Lagrange multiplier associated with the constraint $\langle w_{C(x_i)} - w_k, \Phi(x_i)\rangle + b_{C(x_i)} - b_k - 1 + \xi_{ik} \geq 0$. Let

$$J(\alpha) = \frac{1}{2}\left\{\sum_{i \simeq j}\sum_{k=1}^{Q}\sum_{l=1}^{Q}\alpha_{ik}\alpha_{jl}\kappa(x_i, x_j) - 2\sum_{i=1}^{m}\sum_{j=1}^{m}\sum_{k=1}^{Q}\alpha_{ik}\alpha_{jC(x_i)}\kappa(x_i, x_j)\right.$$

$$\left. + \sum_{i=1}^{m}\sum_{j=1}^{m}\sum_{k=1}^{Q}\alpha_{ik}\alpha_{jk}\kappa(x_i, x_j)\right\} - \sum_{i=1}^{m}\sum_{k=1}^{Q}\alpha_{ik},$$

with $i \simeq j$ meaning that x_i and x_j belong to the same category.

Problem 2 (Dual).

$$\min_{\alpha} J(\alpha)$$

$$\text{s.t.} \begin{cases} \sum_{x_i \in C_k}\sum_{l=1}^{Q}\alpha_{il} - \sum_{i=1}^{m}\alpha_{ik} = 0 \ (1 \leq k \leq Q-1) \\ 0 \leq \alpha_{ik} \leq C \qquad (1 \leq i \leq m), \ (1 \leq k \neq C(x_i) \leq Q) \end{cases} .$$

Based on this dual formulation, our algorithm can be expressed as follows:

```
/* Initialization */
  $C_0 := C^{(0)}$, $\alpha^{(0)} := 0_{Qm}$;
/* Main loop */
  For $i := 1$ to nb_iter do
        train_SVM($C_{i-1}$, $s_m$, $\alpha^{(i-1)}$) $\longrightarrow \alpha^{(i)}$;
        $C_i := C_{i-1} + \epsilon$;
  done
/* Termination */
  $i_0 :=$ Argmin$_{1 \leq i \leq \text{nb\_iter}}$ { compute_bound($C_{i-1}$, $s_m$, $\alpha^{(i)}$) };
  $C := C_{i_0}$;
```

In words, this algorithm consists in training the M-SVM a given number of times (calls of the function `train_SVM`) for increasing values of C, and checking each time the value of the guaranteed risk (calls of the function `compute_bound`). Eventually, the value retained is the one corresponding to the "argmin", C_{i_0}. The benefit in terms of `cpu` time springs from the fact that the initial feasible solution used for the $i+1$-th training is the optimal solution of the i-th training, $\alpha^{(i)}$. Note that this is possible since we are working with increasing values of C. As a consequence, each training procedure converges more quickly than if the starting feasible solution was simply the null vector. Obviously, this exploration of the regularization path could also benefit from the implementation of a multi-class extension of the algorithm proposed in [Hastie *et al.*, 2004].

5 Experimental results

The bound provided by the conjunction of Theorem 1 and Theorem 2 can be applied to any M-SVM, whatever the kernel is. This is not a trivial property indeed, since it means that the feature space can be infinite dimensional, as in the case of a Gaussian kernel. In this section, for the sake of simplicity, we restrict to the case of a linear machine, i.e. a machine where the kernel is the Euclidean dot product. In that case, we can make use of a simpler result than Lemma 1 to bound from above the covering numbers of interest.

Proposition 5 (Proposition 1.3.1 in [Carl and Stephani, 1990]) *Let E and F be Banach spaces and $S \in \mathfrak{L}(E, F)$. If S is of rank r, then for $n \geq 1$,*

$$\epsilon_n(S) \leq 4\|S\|n^{-1/r}. \tag{13}$$

The bound resulting from this proposition is the following.

Theorem 3 *Let $\mathcal{H}$ be the class of functions computed by a Q-category M-SVM under the hypothesis that $\Phi(\mathcal{X})$ is included in the closed ball of radius $\Lambda_{\Phi(\mathcal{X})}$ in $E_{\Phi(\mathcal{X})}$ and the constraints that $\|\mathbf{w}\| \leq \Lambda_w$ and $b \in [-\beta, \beta]^Q$. Suppose further that the dimensionality of $E_{\Phi(\mathcal{X})}$ is finite and equal to d. For every value of γ in $(0, 1]$,*

$$\mathcal{N}_{\infty,\infty}(\gamma/4, \Delta_\gamma \mathcal{H}, 2m) \leq \left(2\left\lceil\frac{4\beta}{\gamma}\right\rceil + 1\right)^Q \cdot \left(\frac{32\Lambda_w\Lambda_{\Phi(\mathcal{X})}}{\gamma}\right)^{Qd}. \tag{14}$$

The derivation of this bound rests on the fact that under the hypothesis $\dim\left(E_{\Phi(\mathcal{X})}\right) = d$, the rank of S_{s_m} (or $S_{s_{2m}}$) is bounded from above by the dimensionality of its domain, Qd. Otherwise, the sole bound on the rank available would be Qm (resp. $2Qm$), which would not meet our purpose (the guaranteed risk would not tend to the margin risk as m tends to infinity).

The algorithm of Section 4 is evaluated on a toy problem: the discrimination between three categories corresponding to isotropic Gaussian distributions in the plane with respective means and variances $\left((2.5 \cdot \sqrt{3}, -2.5), 1\right)$, $((0,5),4)$ and $\left((-2.5 \cdot \sqrt{3}, -2.5), 16\right)$. The priors on the categories are equal. The training set is made up of 3000 points, 1000 for each category. This problem is illustrated on Figure 1. The optimal separating surfaces, implementing Bayes' classifier, are two circles. The smaller one, at the bottom right of the right subfigure, corresponds to the boundary of the first category, the other one corresponding to the boundary of the second category. For this classifier, a Monte-Carlo method provides us with an estimate of the expected risk equal to 5.27%. With the same method, the estimates of the risks of the optimal linear separator and the M-SVM specified by the algorithm of Section 4 are respectively 5.85% and 6.30%. Thus, the estimation error is slightly inferior to the approximation error. Obviously, the significance of these initial results is limited, since they were obtained with a linear model, for which overfitting seldom happens. Additional experiments are currently being performed with a polynomial kernel in place of the Euclidean dot product.

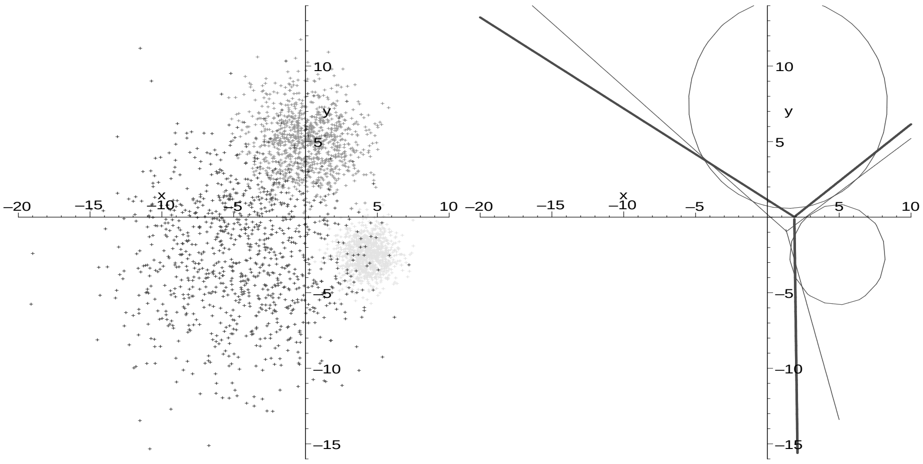


Fig. 1. Separating 3 Gaussian-distributed categories in $\mathbb{R}^2$. **Left:** training set. **Right:** Bayes' classifier (circles), optimal linear classifier and boundaries computed by the linear M-SVM for the (estimated) optimal value of C (thick lines).

6 Conclusions and future work

In this paper, a bound on the covering numbers of M-SVMs in terms of constraints on the parameters of their hyperplanes has been established. When plugged into the guaranteed risk derived in [Guermeur, 2004], it provides us with an objective function which can be used to implement the SRM inductive principle, and especially to tune the hyperparameters. An experimental validation on real-world data is underway, in protein secondary structure prediction, with the aim to improve the accuracy of the classifier introduced in [Guermeur *et al.*, 2004].

Acknowledgements

This study was inspired by an old discussion with A. Smola and a past collaboration with A. Elisseeff. The work of YG and FS is supported by the ACI "Masses de Données".

References

[Aronszajn, 1950]N. Aronszajn. Theory of reproducing kernels. *Trans. Amer. Math. Soc.*, 68:337–404, 1950.

[Bartlett, 1998]P.L. Bartlett. The sample complexity of pattern classification with neural networks: the size of the weights is more important than the size of the network. *IEEE Transactions on Information Theory*, 44(2):525–536, 1998.

[Carl and Stephani, 1990]B. Carl and I. Stephani. *Entropy, compactness, and the approximation of operators.* Cambridge University Press, Cambridge, UK, 1990.

[Chapelle *et al.*, 2002]O. Chapelle, V.N. Vapnik, O. Bousquet, and S. Mukherjee. Choosing Multiple Parameters for Support Vector Machines. *Machine Learning*, 46(1):131–159, 2002.

[Cortes and Vapnik, 1995]C. Cortes and V.N. Vapnik. Support-Vector Networks. *Machine Learning*, 20:273–297, 1995.

[Guermeur *et al.*, 2004]Y. Guermeur, A. Lifchitz, and R. Vert. A kernel for protein secondary structure prediction. In B. Schölkopf, K. Tsuda, and J.-P. Vert, editors, *Kernel Methods in Computational Biology*, pages 193–206. The MIT Press, 2004.

[Guermeur, 2004]Y. Guermeur. Large margin multi-category discriminant models and scale-sensitive Ψ-dimensions. Technical Report RR-5314, INRIA, 2004.

[Hastie *et al.*, 2004]T. Hastie, S. Rosset, R. Tibshirani, and J. Zhu. The entire regularization path for the support vector machine. *Journal of Machine Learning Research*, 5:1391–1415, 2004.

[Passerini *et al.*, 2004]A. Passerini, M. Pontil, and P. Frasconi. New results on error correcting output codes of kernel machines. *IEEE Transactions on Neural Networks*, 15(1):45–54, 2004.

[Sauer, 1972]N. Sauer. On the density of families of sets. *Journal of Combinatorial Theory (A)*, 13:145–147, 1972.

[Vapnik, 1982]V.N. Vapnik. *Estimation of Dependences Based on Empirical Data.* Springer-Verlag, N.Y, 1982.

[Vapnik, 1998]V.N. Vapnik. *Statistical learning theory.* John Wiley & Sons, Inc., N.Y., 1998.

[Weston and Watkins, 1998]J. Weston and C. Watkins. Multi-class Support Vector Machines. Technical Report CSD-TR-98-04, Royal Holloway, University of London, Department of Computer Science, 1998.

[Williamson *et al.*, 2000]R.C. Williamson, A.J. Smola, and B. Schölkopf. Entropy numbers of linear function classes. In *COLT'00*, pages 309–319, 2000.

Model Selection in Classification: the Swapping Method

Jean-Jacques Daudin and Tristan Mary-Huard

INA-PG (dépt OMIP) / INRA (dépt MIA)
16 rue Claude Bernard, Paris Cedex 05, France
(e-mail: daudin@inapg.fr, maryhuar@inapg.fr)

Abstract. In this article, the bias of the empirical error rate in supervised classification is studied. The exact formula and a robust estimator of the bias are given. From these results, we propose a new penalized criterion to perform model selection in classification. Applications to simulated and real data are presented.
Keywords: Classification, Model Selection, Covariance Penalty.

1 Introduction

The aim of supervised classification is to predict the unknown label Y of an observation (here $Y = 0$ or 1), according to some collected information X. A classifier $\phi_n^* : x \mapsto \phi_n^*(x) = \hat{y}$ is constructed on the basis of a collection of i.i.d. examples $(X_i, Y_i),\ i = 1, ..., n$ for which both the label and the information are known. An important problem is to estimate the conditional error rate (CER)

$$L_x(\phi_n^*) = \frac{1}{n}\sum_{i=1}^{n} P(\phi_n^*(x_i) \neq Y)$$

of the constructed classifier, where the x_i were observed on the training set. A natural estimator of $L_x(\Phi_n^*)$ is the empirical error rate (EER)

$$L_n(\phi_n^*) = \frac{1}{n}\sum_{i=1}^{n} I_{\{\phi_n^*(X_i)\neq Y_i\}} \quad ,$$

but this estimator is known to be optimistically biased, and we would like to gain insight into the bias of the EER estimator.

In this paper, we study the behavior of the random variable $B(\Phi_n^*) = L_x(\Phi_n^*) - L_n(\Phi_n^*)$, where Φ_n^* is constructed on the basis of an independent copy of the Y_i's, and with the same x_i's as in the initial dataset. We give an exact formula for the bias

$$E_Y\left(B(\Phi_n^*)\right) = E_Y\left(L_x(\Phi_n^*) - L_n(\Phi_n^*)\right) \quad , \tag{1}$$

along with an estimator S_n of $E_Y\left(B(\Phi_n^*)\right)$.

An important motivation for estimating (1) is to perform complexity regularization in pattern recognition. When the CER is close to the true error rate $P(\phi_n^*(X) \neq Y)$ (TER), it should be relevant to minimize the criterion

$$C(\Phi_n^*) = L_n(\Phi_n^*) + S_n \qquad (2)$$

to find a classifier with good generalization performance. We call the minimization of criterion (2) the swapping method (designated by (S)). We analyse the empirical behavior of (S) on a theoretical example, and we compare (S) with cross-validation (CV). We then present the adaption of (S) to the popular k-nearest neighbors algorithm (kNN), where (S) is used to select k. Applications to experimental data are presented to assess the performance of (S).

2 Bias estimation in classification

Let (X_i, Y_i), $i = 1, ..., n$ be n i.i.d. random vectors with distribution P. We note $p_x = P(Y = 1|X = x)$. We define ϕ_n^* as a fixed classification function obtained from a given sample. The "*" indicates that the function was found by optimization of some criterion. We also define Φ_n^* as the corresponding random classification rule obtained for any sample with the same x_is and random Y_is. In practice we would like to obtain some mathematical properties about ϕ_n^* which is the classification function we will use for prediction. However, these properties are difficult to obtain, and we must use Φ_n^* as an intermediate trick.

The following theorem gives the exact form for the bias of the EER in the general classification case:

Theorem 1 *For any classification rule Φ_n^* we have:*

$$E_Y(B(\Phi_n^*)) = \frac{2}{n} \sum_{i=1}^{n} p_{x_i}(1 - p_{x_i}) E_Y[\Phi_n^*(x_i|Y_i = 1) - \Phi_n^*(x_i|Y_i = 0)] \ , \qquad (3)$$

where $\Phi_n^*(.\ |Y_i = 1)$ is the decision rule computed from the learning dataset with Y_i set to 1.

The proof is not given here. It is worthwhile to interpret this result. The label of each observation is swapped alternatively and the consequence on the decision rule is observed. If the swap does not change the decision for the observation under concern, its contribution to the bias estimate is null. Conversely, if the decision is changed, the contribution is equal to $2p_x(1-p_x)$ with a sign - or +, usually +. Thus if a decision rule is "too versatile" the bias of the EER is high.

From Theorem 1 we can derive an unbiased estimator for the bias of any classification method:

Corollary 1 *With the notations of Theorem 1, an unbiased estimator of* $E_Y(B(\Phi_n^*))$ *is*

$$S_n = \frac{2}{n}\sum_{i=1}^{n} p_{x_i}(1-p_{x_i})[\phi_n^*(x_i|Y_i=1)-\phi_n^*(x_i|Y_i=0)] \enspace .$$

Of course this estimator is theoretical, since p_x is unknown. Many classification methods provide estimations of the posterior probabilities $\hat{p}_x$ that could be used in place of p_x in Lemma 1. But this method leads to an inconsistent estimation of the bias. We propose a robust version of the plug-in estimator:

$$\hat{p}_{x,B} = \frac{n_x\hat{p}_x + n_0 \times (1/2)}{n_x+n_0} \enspace , \tag{4}$$

where $\hat{p}_x$ is the plug-in estimator, n_x is the number of points used to compute $\hat{p}_x$ and n_0 is a fixed integer. The "B" index stands for "Bayesian". If $n_0 = 0$, then $\hat{p}_B = \hat{p}_x$ and we find the plug-in estimator. Inversely, if $n_0 = \infty$, then $\hat{p}_B = 1/2$ which corresponds to the worst case in classification
The behavior of the swapping estimate may be closely related to the value of n_0. For high levels of noise in the data and rich classes of classification functions, n_0 should be large. Conversely for low level of noise and poor classes of functions, n_0 should be small. In the following, n_0 is fixed to 10, that seems to be an omnibus compromise (see section 3).

3 Model selection by swapping

3.1 Model selection

Classification aims at finding a classifier ϕ_n^* in a class of functions $\mathcal{C}$ on the basis of data $((X_1,Y_1),...,(X_n,Y_n))$. Of course, we want the TER of ϕ_n^* to be close to the Bayes error rate, i.e. the error rate L^* of the Bayes classifier

$$\Phi^*(x) = \{ \begin{array}{l} 1 \text{ if } \mathbf{P}\{Y=1|X=x\} >1/2 \\ 0 \text{ otherwise .} \end{array}$$

In practice, ϕ_n^* is selected by empirical risk minimization on $\mathcal{C}$. Since we do not know how to choose $\mathcal{C}$, we consider many classes $\mathcal{C}_k$ with different complexities. In the classical complexity regularization framework, the EER minimizer $\phi_{n,k}^*$ is computed for each class. Then among all the candidate classifiers we choose the one that minimizes a given penalized criterion, which usually is an upper bound of the TER.
We propose to use the swapping method (S) to perform model selection. The selection among all the candidate classifiers is performed by minimizing:

$$\begin{aligned} C(\phi_{n,k}^*) &= L_n(\phi_{n,k}^*) + S_n \\ &= L_n(\phi_{n,k}^*) + \frac{2}{n}\sum_{i=1}^{n}\hat{p}_{x_i}(1-\hat{p}_{x_i})[\phi_n^*(x_i|Y_i=1)-\phi_n^*(x_i|Y_i=0)](5) \end{aligned}$$

While this strategy is also based on the minimization of a penalized criterion, the difference with the preceding strategy is the meaning of the criterion. In (5), the criterion is an estimator of the conditional error risk, while in the regularization framework the criterion is an upper bound for the true error rate. The (S) strategy can be justified with the following break-down:

$$\begin{aligned} L(\phi_n^*) &= L_n(\phi_n^*) + [L_x(\phi_n^*) - L_n(\phi_n^*)] + [L(\phi_n^*) - L_x(\phi_n^*)] \\ &= L_n(\phi_n^*) + B(\phi_n^*) + A(\phi_n^*) \ , \end{aligned}$$

where $A(\phi_n^*) = L(\phi_n^*) - L_x(\phi_n^*)$. In this paper we make the assumption that $A(\phi_n^*)$ does not strongly depend on the complexity of ϕ_n^*, and therefore can be neglected for model selection.

3.2 The Kearn's example

[Kearns *et al.*, 1997] proposed the following model for comparison of model selection methods. The interval $[0, 1]$ is divided into d equal subintervals, alternatively labelled 0 and 1. Let $((X_1, Y_1), ...(X_n, Y_n))$ be an i.i.d. sample, where X_i and Y_i are the position and label of observation i, respectively. The X_i's are drawn from the uniform distribution on $[0, 1]$. Y_i equals the label of the interval to which X_i belongs with probability $1 - \eta$, and the alternative label with probability η. η denotes the noise level of the problem.

We performed simulations according to this model with $d = 10$, $\eta = 0.1$, 0.2, 0.3 and 0.4 and $n = 20, 100, 500$. Simulations performed with $d = 100$ lead to similar findings (not shown).

Figure 1 (left) shows the EER, CER and TER averaged on 100 trials, for $\eta = 0.2$ and $n = 100$, displayed along the number of intervals k. One can see that the curves of the conditional and true error rates are nearly parallel for $k \geq d$. This behavior is observed for any value of η, d and n (data not shown). Therefore the basic condition on $A(\phi_n^*)$ assumed in this paper is satisfied for the Kearns example.

Figure 1 (right) shows the behavior of the estimate of the conditional bias given by the swapping method (S). In this example with $\eta = 0.2$ the bias is overestimated. This overestimation is higher for $\eta = 0.1$, vanishes for $\eta = 0.3$ and becomes an underestimation for $\eta = 0.4$ (not shown).

Figure 2 (left) gives the behavior of the empirical and (S) error rates (y-axis) according to the number of intervals k (x-axis), for 3 trials with $\eta = 0.2$ and $n = 100$. One can see that the empirical error rate decreases to zero. Conversely (S) estimate of the error rate decreases till $k \simeq 10$ and then grows for $k > 10$. Figure 2 (right) shows the mean values of 100 trials of the two error rate estimates with the same parameters as above.

3.3 Comparison between cross validation and swapping

We compared the swapping method selection with $n_0 = 10$ (S) to its natural competitors, the $(n-1, 1)$ cross validation (CV) and the best possible classifi-

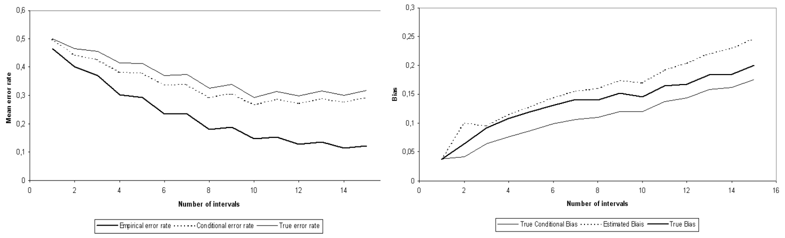


Fig. 1. Left: EER (bold line), CER (dotted line) and TER (solid line) along the number of intervals k. Average on 100 trials, with $\eta = 0.2$ and $n = 100$. **Right:** Estimated bias (dotted line), conditional bias (solid line) and true bias (bold line) along the number of intervals k.

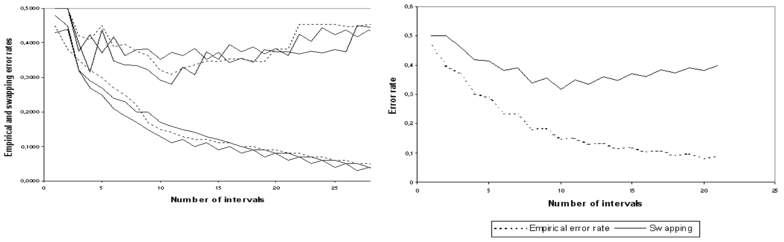


Fig. 2. Left: Empirical and (S) error rates along the number of intervals k for 3 trials. **Right:** Empirical and (S) error rates along the number of intervals k, averaged on 100 trials.

cation function "oracle" (O). (O) is the classification function that minimizes the true error rate for each sample. Figure 3 shows the results for $\eta = 0.2$. Considering Figure 3 and the results obtained for other values of η (not shown here), we draw the following conclusions:

• (S) outperforms (CV) for $\eta \leq 0.3$. The relative gain $(100(L_{CV} - L_S)/(L_{CV} - L_O)$ of (S) on (CV) for $\eta \leq 0.3$ lies between 20% and 80% (not shown here). When $\eta = 0.4$ the gain exists but is tiny.

• The (S) 95% quantile of $L_S - L_O$ is always lower than the (CV) 95% quantile of $L_{CV} - L_O$.

• The empirical error rate penalized by the (S) method gives a better estimate of the true error rate of the selected classification function. This estimate is optimistic for $\eta \geq 0.2$ and pessimistic for $\eta \leq 0.1$. (CV) systematically gives an optimistic view of the true error rate of the selected classification function.

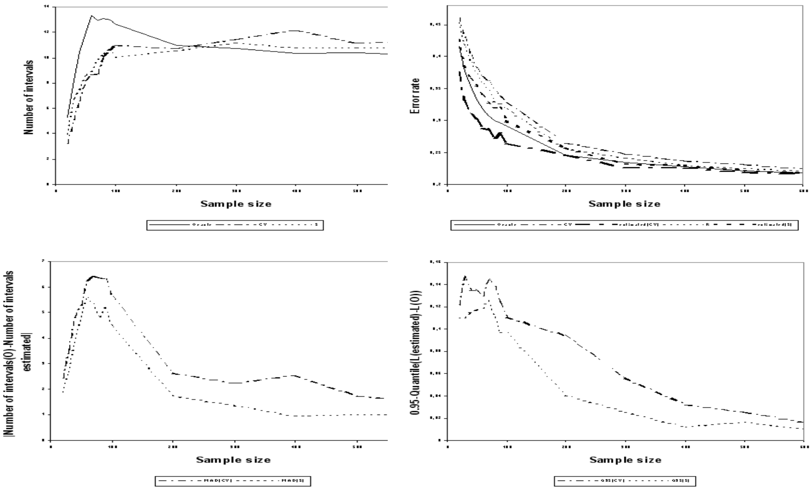


Fig. 3. Results of the (S) model selection for $\eta = 0.2$. **Top left:** Mean number of intervals k_O, k_{CV} and k_S obtained by (O), (CV) and (S), respectively. **Top right:** Mean of the true error rate of the classifiers obtained by (O) selection (solid line), (CV) selection (dashed line) and (S) selection (dotted line), respectively. Dashed lines correspond to (CV) TER estimated by (CV), and (S) TER estimated by (S). **Bottom left:** Mean of $|k_O - k_{CV}|$ and $|k_O - k_S|$. **Bottom right:** 95% quantile of $L_{CV} - L_O$ and $L_S - L_O$.

4 Application to k-nearest-neighbors

We present a simple computational trick to efficiently apply the (S) method to kNN. We then compare the performance of (CV) and (S) on a benchmarking microarray dataset.

4.1 Computation of (S) for kNN

To avoid any concern about the parity of k, in the following we consider only odd values for k, as proposed in [Fort and Lambert-Lacroix, 2004]. For a given k, we need to compute for each observation x_i the quantity $p_{x_i}(1 - p_{x_i})[\phi_n^*(x_i, 1) - \phi_n^*(x_i, 0)]$. The posterior probability p_{x_i} can be estimated according to the Bayes method presented in section 2. In this case, the Bayes estimator of p_{x_i} for the kNN is:

$$\hat{p}_{x_i,B} = \frac{k \times (m/k) + n_0 \times 1/2}{k + n_0} = \frac{m + n_0/2}{k + n_0} \quad , \tag{6}$$

where m is the number of 1 among the k neighbors of point x_i. Clearly, this posterior probability can be obtained from the kNN classifier without

additional computational time.
The difference $\phi_n^*(x_i, 1) - \phi_n^*(x_i, 0)$ can also be easily obtained from the kNN classifier considering the following argument: when the label of point x_i is swapped, its classification is not changed except in the case where x_i belongs to the majority and the majority is "short", i.e. $m = (k-1)/2$ or $m = (k+1)/2$ (remember that k is odd). Hence, the difference $\phi_n^*(x_i, 1) - \phi_n^*(x_i, 0)$ will be 1 if $m = (k-1)/2$ or $m = (k+1)/2$, and 0 otherwise. So this difference is easily obtained from the kNN algorithm.
This shows that (S) is a competing method from a computational point of view. In practice, for samples of size $n \sim 100$ and a number of variables as big as 2000, the minimization of the penalized empirical risk to select k is performed within a few seconds.

4.2 Microarray data

We consider the Colon microarray dataset, described in [Alon *et al.*, 1999]. It contains 62 tissue samples for which 2,000 genes were observed. Among the 62 observations, 40 of them are tumor tissues and 22 are normal. For comparison with other published studies, the data normalization, the preliminary gene selection, and the re-randomization study to assess the performance of (S) and (CV) were performed according to the procedures described in [Fort and Lambert-Lacroix, 2004]. It should be noticed that the high level of noise in the data along with the high number of variables considered (with possibly many of them irrelevant) should be in favor of (CV). We display the average performance of the classification rules obtained with (S) and (CV) selection methods.

Table 1 shows that (S) outperforms (CV) for three gene selections: $g = 100, 500, 2000$. As for simulations, both methods are far from the oracle results, even for the simpler case where the number of genes is 100 (which corresponds to the low level of noise case). We conclude that the (S) method for kNN is competing on simulated and real data.

	Oracle		Swapping		Cross-Valid.	
Nb. Genes	N	R	N	R	N	R
2000	6.0	19.0	9.11	**28.6**	6.7	28.8
1000	7.6	13.8	12.8	21.4	11.2	**21.1**
500	6.4	13.1	12.1	**18.1**	15.7	18.7
100	4.8	12.0	12.0	**15.6**	20.7	16.0

Table 1. Results for the Colon dataset, over 500 resamplings. First column indicates the number of selected genes. For each selection method (Oracle, Swapping and Cross-Valid.) the mean number of neighbors (N) and the mean test error (R) are computed.

5 Discussion

The methods proposed in this paper to estimate the conditional error rate are connected to some recent papers. A review of the field of prediction error estimation in a quite general context has been made by [Efron, 2004] who divides the methods into two classes: covariance penalties, assuming a parametric model, and nonparametric methods such as cross validation and bootstrap. The swapping method is clearly a covariance penalty method, but it may be applied to non parametric statistical methods. Its only requirement is that a conditional probability $P(Y = 1/X = x)$ may be estimated for each observed value x. This is true because the field is reduced to the error rate in classification, where the p.d.f. of the response variable Y reduces to only one parameter.

The swapping expression in Theorem 1 was present in an earlier paper of [Efron, 1986], but the idea of estimating $E_Y\left(B(\Phi_n^*)\right)$ by its sample estimate (Corollary 1), and the application to model selection in classification are new. Moreover we propose a robust estimate of p_x, which attempts to correct the over-learning bias. In this study n_0 was fixed to 10, but simulations performed with values ranging from 5 to 20 give similar results. However the choice of n_0 is an open problem.

References

[Alon *et al.*, 1999]U. Alon, N. Barkai, D. A. Notterman, K. Gish, S. Ybarra, D. Mack, and A. J. Levine. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. 96(12):6745–6750, 1999.

[Efron, 1986]B. Efron. How biased is the apparent error rate of a prediction rule? pages 461–470, 1986.

[Efron, 2004]B. Efron. The estimation of prediction error: covariance penalties and cross-validation. 99, 2004.

[Fort and Lambert-Lacroix, 2004]G. Fort and S. Lambert-Lacroix. Classification using partial least squares with penalized logistic regression. *Bioinformatics*, 2004.

[Kearns *et al.*, 1997]M. Kearns, Y. Mansour, A.Y. Ng, and D. Ron. An experimental and theoretical comparison of model selection methods. *Machine Learning*, 27(1):7–50, 1997.

High Dimensional Discriminant Analysis

Charles Bouveyron[1,2], Stéphane Girard[1], and Cordelia Schmid[2]

[1] LMC – IMAG, BP 53, Université Grenoble 1, 38041 Grenoble cedex 9 – France (e-mail: charles.bouveyron@imag.fr, stephane.girard@imag.fr)

[2] LEAR – INRIA Rhône-Alpes, 655 avenue de l'Europe, Montbonnot, 38334 Saint-Ismier Cedex – France (e-mail: Cordelia.Schmid@inrialpes.fr)

Abstract. We propose a new method of discriminant analysis, called High Dimensional Discriminant Analysis (HHDA). Our approach is based on the assumption that high dimensional data live in different subspaces with low dimensionality. Thus, HDDA reduces the dimension for each class independently and regularizes class conditional covariance matrices in order to adapt the Gaussian framework to high dimensional data. This regularization is achieved by assuming that classes are spherical in their eigenspace. HDDA is applied to recognize object in real images and its performances are compared to classical classification methods.

Keywords: Discriminant analysis, Dimension reduction, Regularization.

1 Introduction

In this paper, we introduce a new method of discriminant analysis, called High Dimensional Discriminant analysis (HDDA) to classify high dimensional data, as occur for example in visual object recognition. We assume that high dimensional data live in different subspaces with low dimensionality. Thus, HDDA reduces the dimension for each class independently and regularizes class conditional covariance matrices in order to adapt the Gaussian framework to high dimensional data. This regularization is based on the assumption that classes are spherical in their eigenspace. It is also possible to make additional assumptions to reduce the number of parameters to estimate. This paper is organized as follows. We first remind in section 2 the discrimination problem and classical discriminant analysis methods. Section 3 presents the theoretical framework of HDDA. Section 4 is devoted to the inference aspects. Our method is then compared to reference methods on a real images dataset in section 5.

2 Discriminant analysis framework

In this section, we remind the general framework of the discrimination problem and present the main methods of discriminant analysis.

2.1 Discrimination problem

The goal of discriminant analysis is to assign an observation $x \in \mathbb{R}^p$ with unknown class membership to one of k classes $C_1, ..., C_k$ known *a priori*. To this end, we have a learning dataset $A = \{(x_1, c_1), ..., (x_n, c_n)/x_j \in \mathbb{R}^p$ and $c_j \in \{1, ..., k\}\}$, where the vector x_j contains p explanatory variables and c_j indicates the index of the class of x_i. It is a statistical decision problem and the learning dataset allows to construct a decision rule which associates a new vector $x \in \mathbb{R}^p$ to one of the k classes. The optimal decision rule, called *Bayes decision rule*, affects the observation x to the class C_{i^*} which has the *maximum a posteriori* probability which is equivalent, in view of the Bayes formula, to minimize a cost function $K_i(x)$ *i.e.* $i^* = \operatorname{argmin}_{i=1,...,k} K_i(x)$, with

$$K_i(x) = -2\log(\pi_i\, f_i(x)),$$

where π_i is the *a priori* probability of class C_i and $f_i(x)$ denotes the class conditional density of x, $\forall i = 1, ..., k$.

2.2 Classical discriminant analysis methods

Some classical discriminant analysis methods can be obtained by combining additional assumptions with the Bayes decision rule. We refer to [Celeux, 2003] and [Saporta, 1990, chap. 18] for further informations on this topic. For instance, Quadratic discriminant analysis (QDA) assumes that, $\forall i = 1, ..., k$, the class conditional density f_i for the class C_i is Gaussian $\mathcal{N}(\mu_i, \Sigma_i)$ which leads to the cost function

$$K_i(x) = (x - \mu_i)^t \Sigma_i^{-1}(x - \mu_i) + \log(\det \Sigma_i) - 2\log(\pi_i).$$

This decision rule makes quadratic separations between the classes. In practice, this method is penalized in high-dimensional spaces since it requires the estimation of many parameters. For this reason, particular rules of QDA exist in order to regularize the estimation of Σ_i. As an example, it can be assumed that covariance matrices are proportional to the identity matrix, *i.e.* $\Sigma_i = \sigma_i^2 Id$. In this case, classes are spherical and this method is referred to as QDAs. One can also assume that covariance matrices are equal, *i.e.* $\Sigma_i = \Sigma$, which yields the framework of the linear discriminant analysis (LDA). This method makes linear separations between the classes. If, in addition, covariance matrices are assumed equal and proportional to the identity matrix, we obtain the so-called LDAs method.

2.3 Dimension reduction and regularization

Classical discriminant analysis methods have disappointing behavior when the size n of the training dataset is small compared to the number p of variables. In such cases, a dimension reduction step and/or a regularization of the discriminant analysis are introduced.

Fisher discriminant analysis (FDA) This approach combines a dimension reduction step and a discriminant analysis procedure and is in general efficient on high dimensional data. FDA provides the $(k-1)$ discriminant axes maximizing the ratio between the inter class variance and the intra class variance. It is then possible to perform one of the previous methods on the projected data (usually LDA).

Regularized discriminant analysis (RDA) In [Friedman, 1989] a regularization technique of discriminant analysis is proposed. RDA uses two regularization parameters to design an intermediate classifier between LDA and QDA. The estimation of the covariance matrices depends on a complexity parameter and on a shrinkage parameter. The complexity parameter controls the ratio between Σ_i and the common covariance matrix Σ. The other parameter controls shrinkage of the class conditional covariance matrix toward a specified multiple of the identity matrix.

Eigenvalue decomposition discriminant analysis (EDDA) This other regularization method [Bensmail and Celeux, 1996] is based on the re-parametrization of the covariance matrices: $\Sigma_i = \lambda_i D_i A_i D_i^t$, where D_i is the matrix of eigenvectors of Σ_i, A_i is a diagonal matrix containing standardized and ordered eigenvalues of Σ_i and $\lambda_i = |\Sigma_i|^{1/p}$. Parameters λ_i, D_i and A_i respectively control the volume, the orientation and the shape of the density contours of class C_i. By allowing some but not all of these quantities to vary, the authors obtain geometrical interpreted discriminant models including QDA, QDAs, LDA and LDAs.

3 High Dimensional Discriminant Analysis

The *empty space phenomena* [Scott and Thompson, 1983] enables us to assume that high-dimensional data live in subspaces with dimensionality lower than p. In order to adapt discriminant analysis to high dimensional data and to limit the number of parameters to estimate, we propose to work in class subspaces with lower dimensionality. In addition, we assume that classes are spherical in these subspaces, in other words class conditional covariance matrices have only two different eigenvalues.

3.1 Definitions and assumptions

Similarly to classical discriminant analysis, we assume that class conditional densities are Gaussian $\mathcal{N}(\mu_i, \Sigma_i)$ $\forall i = 1, ..., k$. Let Q_i be the orthogonal matrix of eigenvectors of the covariance matrix Σ_i and $\mathcal{B}_i$ be the eigenspace of Σ_i, *i.e.* the basis made of eigenvectors of Σ_i. The class conditional covariance matrix Δ_i is defined in the basis $\mathcal{B}_i$ by $\Delta_i = Q_i^t \, \Sigma_i \, Q_i$. Thus, Δ_i is diagonal and made of eigenvalues of Σ_i. We assume in addition that Δ_i has only two different eigenvalues $a_i > b_i$. Let $\mathbb{E}_i$ be the affine space generated by the eigenvectors associated to the eigenvalue a_i with $\mu_i \in \mathbb{E}_i$, and let $\mathbb{E}_i^\perp$ be $\mathbb{E}_i \oplus \mathbb{E}_i^\perp = \mathbb{R}^p$ with $\mu_i \in \mathbb{E}_i^\perp$. Thus, the class C_i is both spherical in $\mathbb{E}_i$ and in $\mathbb{E}_i^\perp$. Let $P_i(x) = \tilde{Q}_i\tilde{Q}_i^t(x - \mu_i) + \mu_i$ be the projection of x on $\mathbb{E}_i$, where $\tilde{Q}_i$ is made of the d_i first raws of Q_i and supplemented by zeros. Similarly, let $P_i^\perp(x) = (Q_i - \tilde{Q}_i)(Q_i - \tilde{Q}_i)^t(x - \mu_i) + \mu_i$ be the projection of x on $\mathbb{E}_i^\perp$.

3.2 Decision rule

The preceding assumptions lead to the cost function:

$$K_i(x) = \frac{\|\mu_i - P_i(x)\|^2}{a_i} + \frac{\|x - P_i(x)\|^2}{b_i} + d_i \log(a_i) + (p - d_i)\log(b_i) - 2\log(\pi_i),$$

(*cf.* [Bouveyron *et al.*, 2005] for the proof). In order to interpret the decision rule the following notations are needed: $\forall i = 1, ..., k$, $a_i = \frac{\sigma_i^2}{\alpha_i}$ and $b_i = \frac{\sigma_i^2}{(1-\alpha_i)}$ with $\alpha_i \in]0,1[$ and $\sigma_i > 0$. The cost function can be rewritten:

$$K_i(x) = \frac{1}{\sigma_i^2}\left(\alpha_i\|\mu_i - P_i(x)\|^2 + (1-\alpha_i)\|x - P_i(x)\|^2\right)$$
$$+ 2p\log(\sigma_i) + d_i \log\left(\frac{1-\alpha_i}{\alpha_i}\right) - p\log(1-\alpha_i) - 2\log(\pi_i).$$

The Bayes formula allows to compute the classification error risk based on the *a posteriori* probability

$$p(C_i|x) = \exp\left(-\frac{1}{2}K_i(x)\right) \Big/ \sum_{j=1}^{k} \exp\left(-\frac{1}{2}K_j(x)\right).$$

Note that some particular cases of HDDA reduce to classical discriminant analysis. If $\forall i = 1, ..., k$, $\alpha_i = 1/2$: HDDA reduces to QDAs. If moreover $\forall i = 1, ..., k$, $\sigma_i = \sigma$: HDDA reduces to LDAs.

3.3 Particular rules

By allowing some but not all of HDDA parameters to vary between classes, we obtain 24 particular models which some ones have easily geometrically interpretable rules and correspond to different types of regularization (see [Bouveyron *et al.*, 2005]). Due to space restrictions, we present only two methods: HDDAi and HDDAh.

Isometric decision rule (HDDAi) The following additional assumptions are made: $\forall i = 1, ..., k$, $\alpha_i = \alpha$, $\sigma_i = \sigma$, $d_i = d$ and $\pi_i = \pi_*$, leading to the cost function

$$K_i(x) = \alpha\|\mu_i - P_i(x)\|^2 + (1-\alpha)\|x - P_i(x)\|^2.$$

Case $\alpha = 0$: HDDAi affects x to the class C_{i^*} if $\forall i = 1, ..., k$, $d(x, \mathbb{E}_{i^*}) < d(x, \mathbb{E}_i)$. From a geometrical point of view, the decision rule affects x to the class associated to the closest subspace $\mathbb{E}_i$.
Case $\alpha = 1$: HDDAi affects x to the class C_{i^*} if $\forall i = 1, ..., k$, $d(\mu_{i^*}, P_{i^*}(x)) < d(\mu_i, P_i(x))$. It means that the decision rule affects x to the class for which the mean is closest to the projection of x on the subspace.
Case $0 < \alpha < 1$: the decision rule affects x to the class realizing a compromise between the two previous cases. The estimation of α is discussed in the following section.

Homothetic decision rule (HDDAh) This method differs from the previous one by removing the constraint $\sigma_i = \sigma$. The corresponding cost function is:

$$K_i(x) = \frac{1}{\sigma_i^2}(\alpha\|\mu_i - P_i(x)\|^2 + (1-\alpha)\|x - P_i(x)\|^2) + 2p\log(\sigma_i).$$

It favours classes with large variance. Indeed, if the point x is equidistant to two classes, it is natural to affect x to the class with the larger variance.

Removing constraints on d_i and π_i The two previous methods assume that d_i and π_i are fixed. However, these assumptions can be too restrictive. If these constraints are removed, it is necessary to add the corresponding terms in $K_i(x)$: if d_i are free, then add $d_i \log(\frac{1-\alpha}{\alpha})$ and if π_i are free, then add $-2\log(\pi_i)$.

4 Estimators

The methods HDDA, HDDAi and HDDAh require the estimation of some parameters. These estimators are computed through maximum likelihood (ML) estimation based on the learning dataset A. In the following, the *a priori* probability π_i of the class C_i is estimated by $\hat{\pi}_i = n_i/n$, where $n_i = card(C_i)$ and the class covariance matrix Σ_i is estimated by $\hat{\Sigma}_i = \frac{1}{n_i}\sum_{x_j \in C_i}(x_j - \hat{\mu}_i)^t(x_j - \hat{\mu}_i)$ where $\hat{\mu}_i = \frac{1}{n_i}\sum_{x_j \in C_i} x_j$.

4.1 HDDA estimators

Starting from the log-likelihood expression found in [Flury, 1984, eq. (2.5)], and assuming for the moment that the d_i are known, we obtain the following ML estimates:

$$\hat{a}_i = \frac{1}{d_i}\sum_{j=1}^{d_i}\lambda_{ij} \text{ and } \hat{b}_i = \frac{1}{(p-d_i)}\sum_{j=d_i+1}^{p}\lambda_{ij},$$

where $\lambda_{i1} \geq \cdots \geq \lambda_{ip}$ are the eigenvalues of $\hat{\Sigma}_i$. Moreover, the jth column of Q_i is estimated by the unit eigenvector of $\hat{\Sigma}_i$ associated to the eigenvalue

λ_{ij}. Note that parameters a_i and b_i are estimated by the empirical variances of C_i respectively in $\hat{\mathbb{E}}_i$ and in $\hat{\mathbb{E}_i^\perp}$. The previous result allows to deduce the maximum likelihood estimators of α_i and σ_i^2:

$$\hat{\alpha}_i = \hat{b}_i/(\hat{a}_i + \hat{b}_i) \text{ and } \hat{\sigma}_i^2 = \hat{a}_i\hat{b}_i/(\hat{a}_i + \hat{b}_i).$$

4.2 Estimation of the intrinsic dimension

Estimation of the dataset intrinsic dimension is a difficult problem which we can find for example in the choice of the factor number in PCA. Our approach is based on the eigenvalues of the class conditional covariance matrix Σ_i. The jth eigenvalue of Σ_i corresponds to the fraction of the full variance carried by the jth eigenvector of Σ_i. Consequently, we propose to estimate dimensions d_i, $i = 1, ..., k$, by the empirical method of the scree-test of Cattell [Cattell, 1966] which analyses the differences between eigenvalues in order to find a break in the scree. The selected dimension is the dimension for which the following differences are very small compared to the maximum of differences.

4.3 Particular model estimators

Among the 24 particular models, 9 benefit from explicit ML estimators (see [Bouveyron *et al.*, 2005]). The computation of the ML estimates associated to the 15 other particular rules requires iterative algorithms. We do not reproduce them here by lack of space.

5 Application to object recognition

Object recognition is one of the most challenging problems in computer vision. In the last few years, many successful object recognition approaches use local images descriptors. However, local descriptors are high-dimensional and this penalizes classification methods and consequently recognition. For this reason, HDDA seems well adapted to this application. In the following, we show that HDDA outperform existing techniques in this context.

5.1 Framework of the object recognition

In our framework, small scale-invariant regions are detected on a learning image set and they are then characterized by the local descriptor Sift [Lowe, 2004]. The object is recognized in a test image if a sufficient number of matches with the learning set is found. The recognition step is done using supervised classification methods. Frequently used methods are LDA and, more recently, kernel methods (SVM) [Hastie *et al.*, 2001, chap. 12]. In our approach, the object is represented as a set of object parts. For the motorbike, we consider three parts: wheels, seat and handlebars.

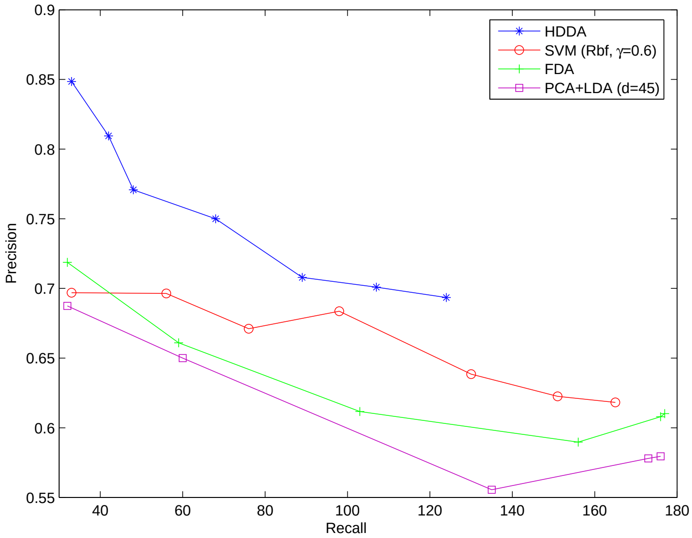


Fig. 1. Comparison of classification results between HDDA method and reference methods.

5.2 Data and protocol

Sift descriptors are computed on 200 motorbike images and 1000 descriptors of motorbike features and of the background were preserved. Consequently, the dataset is made of descriptors in 128 dimensions divided into 4 classes: wheels, seat, handlebars and background. The learning and test dataset are respectively made of 500 and 500 descriptors. Class proportions are respectively: $\forall i = 1, ..., 3$, $\pi_i = 1/6$ and $\pi_4 = 1/2$.

5.3 Results

Figure 1 presents classification results obtained on test data. In order to synthesize the results, only two classes were considered to plot recall-precision curve: motorbike (positive) and background (negative). We remind that the *precision* is the ratio between the number of true positives and the number of detected positives, and the *recall* is the number of detected positives. The different values for each method corresponds to different classifiers. For SVM, the parameter γ is fixed to the best value (0.6) while the parameter C varies. For the other methods, the decision rule varies according to the *a posteriori* probability. In addition, for LDA, we reduced the dimension of data to 45 using PCA in order to obtain the best results for this method. It appears that HDDA outperforms the other methods. In addition, HDDA method

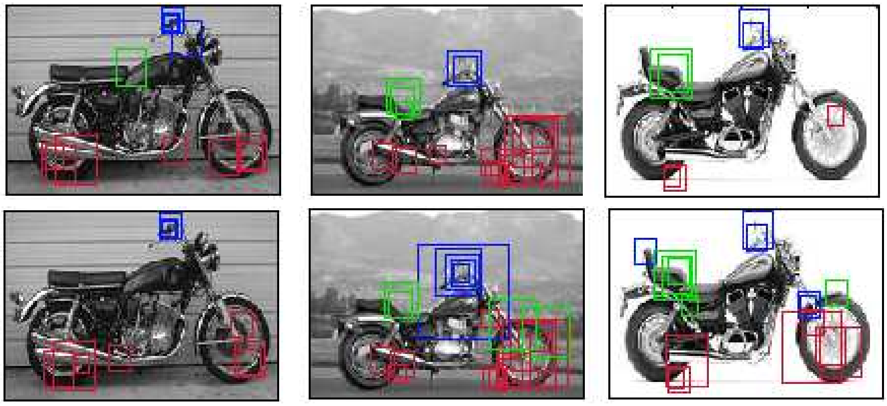


Fig. 2. Recognition of the class "motorbike" using HDDA (top) and SVM (bottom) classifiers. Only descriptors classified as motorbike are displayed. The colors blue, red and green are respectively associated to handlebars, wheels and seat.

is as fast as classical discriminant analysis (computation time $\simeq$ 1 sec. for 1000 descriptors) and much faster than SVM ($\simeq$ 7 sec.). Figure 2 presents recognition results obtained on 5 motorbike images. These results show that HDDA gives better recognition results than SVM. Indeed, the classification errors are significantly lower for HDDA compared to SVM. For example, on the 3th image, HDDA recognizes the motorbike parts without error whereas SVM makes five errors.

6 Conclusion and further work

We presented in this paper a new generative model to classify high-dimensional data in the Gaussian framework. This new model estimates the intrinsic dimension of each class and uses this information to reduce the number of parameters to estimate. In addition, classes are assumed spherical in both subspaces in order to reduce again the number of parameters to estimate and to obtain easily geometrically interpretable rules. In the supervised framework, this model gives very good results without dimension reduction of the data and with a small learning set. Another advantage of this generative model is that it can be used either in supervised or in unsupervised classification. In unsupervised classification, the model presented here arises to a new clustering method based on the EM algorithm. In addition, it is possible to combine unsupervised and supervised classification to recognize an object in a natural image without human interaction. Indeed, the clustering method associated to our model can be used to learn automatically the discriminant part of the object, and then HDDA can be used to recognize the object on a new natural image. First results obtained using this approach are very promising.

Acknowledgments

This work was supported by the French department of Research through the *ACI Masse de données* (MoViStaR project).

References

[Bensmail and Celeux, 1996]H. Bensmail and G. Celeux. Regularized gaussian discriminant analysis through eigenvalue decomposition. *Journal of the American Statistical Association*, 91:1743–1748, 1996.

[Bouveyron *et al.*, 2005]C. Bouveyron, S. Girard, and C. Schmid. High dimensional discriminant analysis. Technical Report 5470, INRIA, January 2005.

[Cattell, 1966]R. B. Cattell. The scree test for the number of factors. *Multivariate Behavioral Research*, 1(2):140–161, 1966.

[Celeux, 2003]G. Celeux. Analyse discriminante. In G. Govaert, editor, *Analyse de Données*, pages 201–233. Hermes Science, Paris, France, 2003.

[Flury, 1984]B. W. Flury. Common principal components in k groups. *Journal of the American Statistical Association*, 79:892–897, 1984.

[Friedman, 1989]J.H. Friedman. Regularized discriminant analysis. *Journal of the American Statistical Association*, 84:165–175, 1989.

[Hastie *et al.*, 2001]T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning.* Springer, New York, 2001.

[Lowe, 2004]D. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.

[Saporta, 1990]G. Saporta. *Probabilités, analyse des données et statistique.* Editions Technip, Paris, France, 1990.

[Scott and Thompson, 1983]D. Scott and J. Thompson. Probability density estimation in higher dimensions. In *Proceedings of the Fifteenth Symposium on the Interface, North Holland-Elsevier Science Publishers*, pages 173–179, 1983.

Learning fitness function in a combinatorial optimization process

Frédéric Clerc[1 2], Ricco Rakotomalala[1], and David Farrusseng[2]

[1] Laboratoire ERIC – Université Lyon2
5 avenue Pierre Mendès France
69676 Bron CEDEX, France
(e-mail: ricco.rakotomalala@univ-lyon2.fr)
[2] Institut de Recherches sur la Catalyse – CRNS
2 avenue Albert Einstein
69626 Villeurbanne CEDEX, France
(e-mail: fclerc@catalyse.cnrs.fr, farrusseng@catalyse.cnrs.fr)

Abstract. Combinatorial optimization is a well known technique to solve problems in various fields such as jet engine design, factory and project scheduling or image recognition. Evolutionary computation and particularly genetic algorithms are commonly used to solve problems defined by complex and high dimensional mathematical expressions. Nevertheless, in some cases, domain experts cannot define this function exactly because of its complexity. In this paper we show that it is possible to solve such optimization problems, where the so called fitness function is unknown. To do this, we hybridize a classic genetic algorithm with a knowledge discovery system which extracts information from a database containing known observations allowing to build a model replacing the fitness function. We use the k nearest neighbours algorithm to solve such a problem sat in heterogeneous catalysis, a division of chemical science where a compound shall be optimized to favour a reaction.
Keywords: datamining, combinatorial optimization, genetic algorithm, fitness function.

1 Introduction

In drug design for medical applications as well as in catalyst development for oil refinery, the discovery and optimization of new formulations is based on the trial and error process. The state of knowledge in both biochemistry and solid state chemistry does not enable to build a model which would give guidelines for the design of formulations with targeted performance. In the vocabulary of optimization it means that the fitness function is not *a priori* known : each formulation must be first synthesized and then its performance measured with specific equipments. At the light of the results, chemists can draw new hypothesis and can design new formulations. A cycle is usually a day and years are required to end up with a final formulation. The new research methodology named high-throughput experimentation now enables to synthesize and test several dozen to hundreds of samples in parallel fashion in order to speed up the research process [B.Jandeleit *et al.*, 1998]. But now the question is : what are the experiments to be performed among an infinite possible number, which maximizes the chance of discovery and/or speeds up the optimization process? [Isar *and al.*, 2002], [Isar and Moga, 2004].

Computer assisted issues were recently reported to develop new catalysts. In [Wolf *et al.*, 2000], libraries of samples corresponding to populations are synthesized and tested in an iterative manner, using an evolutionary strategy. After typically 10 generations, the targeted compound presenting the best performance, the *optimum*, was found. Nevertheless, the total number of catalysts synthesized is still too high and shall be reduced. We present a system which enable to save experiments by hybridizing an optimization process with a knowledge discovery (KD) system. The concept was already reported in [Farrusseng *et al.*, 2003] and [Hanagandi and Kargupta, 1996]. The starting point consists in a real catalyst library which is synthesized and then tested. The corresponding information is stored in a database (DB) which is used by a KD algorithm to *estimate* new virtual individuals. The best estimated are *evaluated* (synthesized and tested) and the resulting information is added to DB so the prediction will be finer. This process is repeated until the checking of a given criterion. The creation of statistical models after each generation shall enable to direct the design of the libraries (i.e. population) by a virtual pre-screening.

In a first section we describe the hybrid optimization process, in a second section, the constraints and issues of the learning process are detailed. The experimental methodology and the results are presented in the third section, before concluding.

2 Hybridizing an optimization process with a knowledge discovery algorithm

Among several optimization processes such as tabu search [Laguna and Glover, 1998] and simulated annealing [S.Kirkpatrick *et al.*, 1983], we decided to use genetic algorithms (GA) [Holland, 1975],[Goldberg, 1989] as this technique was already known and used in the field of heterogeneous catalysis. The mechanics of a genetic algorithm are conceptually simple: (1) maintain a population of individuals (library or generation), (2) select the better for crossover operator, (3) perform mutation operator, and (4) use the offspring to replace poorer individuals. The hybridization consists at inserting a learning process in the genetic algorithm as described in Fig.1.

1. Initialization : Generating randomly a first population of n individuals which are potential solutions to the catalysis problem.
2. Evaluation : Giving a real value to each individual by synthesizing and testing the catalyst. The information produced is stored in DB. Each catalyst is defined by (1) a set of parameters and (2) its performance.

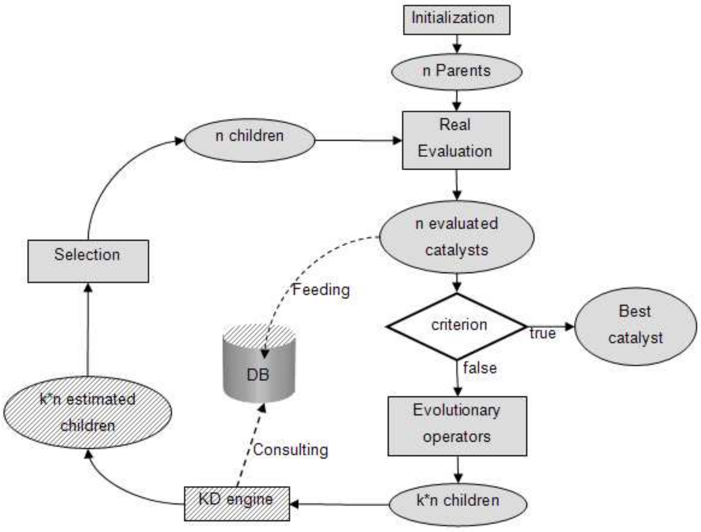


Fig. 1. Hybrid GA with KD system. The hatched bricks are the elements of the KD engine, the remainder are traditional elements of GA.

(number of loops*size of population) individuals are really evaluated and stored in DB.

3. Criterion : Stopping the evolution if verified. Usually, a number of loops is used.
4. Evolutionary operators : Applying traditional GA operators (crossover and mutation) on the parameters of the catalysts which has just been evaluated. During this operation, M*n virtual catalysts are generated in order to maximize the chances of obtaining the optimum quickly. This operation is costless as no individual is really synthesized.
5. KD engine : Mining the database DB so as to estimate the virtual library proposed by evolutionary operators. This quantitative prediction of the fitness involves the use of a supervised learning technique which is described in the next section. This *virtual screening* which is used as a first pass filter is the added value to classical GA.
6. Selection : Applying a selection which extracts n individuals among M*n from the virtual estimated ones. The two best are always picked up and the remaining (n-2) ones are selected using their rank. The selection and its role in genetic algorithms is complex and of importance. It is developed in [Miller and Goldberg, 1996].
7. Loop : Returning at step (2), individuals resulting from selection will be evaluated.

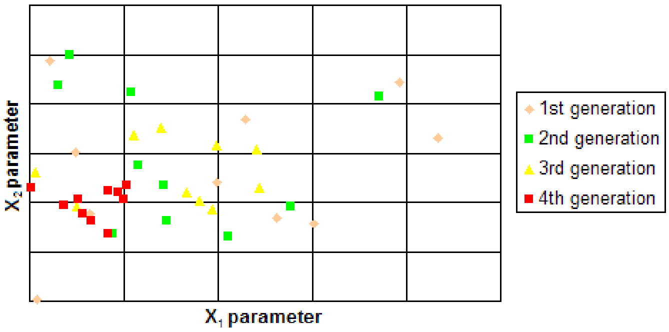


Fig. 2. Schematic evolution of database during a 4 generation process. From a generation to the other, the individuals get closer to the optimum. Thus, some zones of the space are well known, others are almost unknown. This training sample is not homogeneous and the algorithm must take this into account

3 The knowledge discovery algorithm

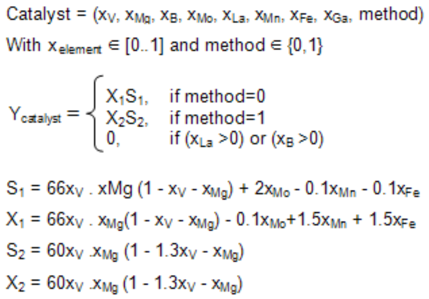
The knowledge discovery algorithm is integrated in an optimization process, so it needs to be adapted to this particular use. In the field of application the constraints are the following: (1) the search space is usually defined by 10 to 20 predictive continuous or categorical variables, (2) the representation space is non linear, (3) a maximum of 400 individuals can be screened and the less the better (4) the predicted variable is continuous. In addition, the learning algorithm has to face the issue of non homogeneous sampling of the search space. Indeed, because the optimization process focuses on specific zones of the search space (see Fig 2) the data set is usually biased.

Among various datamining algorithms [D.Hand *et al.*, 2001], we use the k nearest neighbours algorithm (k-nn)[D.W.Aha *et al.*, 1991]. To estimate a new individual, the algorithm searches among the known individuals (DB) its k nearest neighbours and attributes it their average performance. This algorithm fulfils the main requirements e.g. the learning is (1) adapted to overcome the problem of evolution and convergence as k-nn algorithm itself doesn't require complex update like neural networks or decision trees, (2) nonlinear.

4 Methodology and experiments

4.1 Benchmark

Because there is no open database in the field of catalysis, and because of the cost, the validation of optimization algorithms is performed through simulation using virtual benchmarks. We consider in this study the one presented in



Catalyst = (x_V, x_{Mg}, x_B, x_{Mo}, x_{La}, x_{Mn}, x_{Fe}, x_{Ga}, method)

With $x_{element} \in [0..1]$ and method $\in \{0,1\}$

$$Y_{catalyst} = \begin{cases} X_1S_1, & \text{if method=0} \\ X_2S_2, & \text{if method=1} \\ 0, & \text{if } (x_{La} > 0) \text{ or } (x_B > 0) \end{cases}$$

$$S_1 = 66x_V \,.\, xMg\,(1 - x_V - x_{Mg}) + 2x_{Mo} - 0.1x_{Mn} - 0.1x_{Fe}$$

$$X_1 = 66x_V \,.\, x_{Mg}(1 - x_V - x_{Mg}) - 0.1x_{Mo} + 1.5x_{Mn} + 1.5x_{Fe}$$

$$S_2 = 60x_V \,.x_{Mg}\,(1 - 1.3x_V - x_{Mg})$$

$$X_2 = 60x_V \,.x_{Mg}\,(1 - 1.3x_V - x_{Mg})$$

Fig. 3. The virtual benchmark

[Wolf *et al.*, 2000]. It is composed of 9 predictive variables : 8 percentages of elements for the composition of the catalyst (V, Mg, B, Mo, La, Mn, Fe and Ga) represented by continuous variables from zero to one and a preparation method (coprecipitation or impregnation) represented by a discrete variable (0 or 1). The performance, named Y for Yield, is the continuous variable to predict and is defined in Fig.3. The optimum is a compound containing 32% of Vanadium, 32% of Magnesium and 36% of Molybdenum, the method being coprecipitation. According to the benchmark, we calculate that Y(0.32 , 0.32 , 0 , 0.36 , 0 , 0 , 0, 0) = 7.55

4.2 Conditions

We hybridize a very simple and classical GA for two major reasons. First, the application of computer based optimization methods in the field of heterogeneous catalysis is something quite new and before examining complex issues, we have to experiment the simple ones. Second, in this paper, we're aiming at measuring the performance of an hybrid GA and specially the KD algorithm. This GA, used to generate relevant new virtual individuals, uses a rank selection associated with an elitist selection (2 best individuals kept), a 3 point crossover (probability = 0.8) and a bit-flip mutation (probability = 0.01). Furthermore, the value of the multiplier M is arbitrarily fixed at 15.

In real conditions, the evaluation of a single catalyst is very costly so we have a strong constraint to respect. We consider the optimization finished at the end of 10 generations of 40 individuals, meaning 400 individuals evaluated during the whole experiment. This constitutes the stopping criterion we used. We call one optimization experiment a *run*. GA being stochastic processes only an average value is significative. Thus all the following results are based on 30 runs.

The behaviour of the k-nn algorithm is compared to the behaviour of trivial learning algorithms, the "learning limits" : no and perfect estimation. The upper limit is the perfect learning : the real value of the individual. The

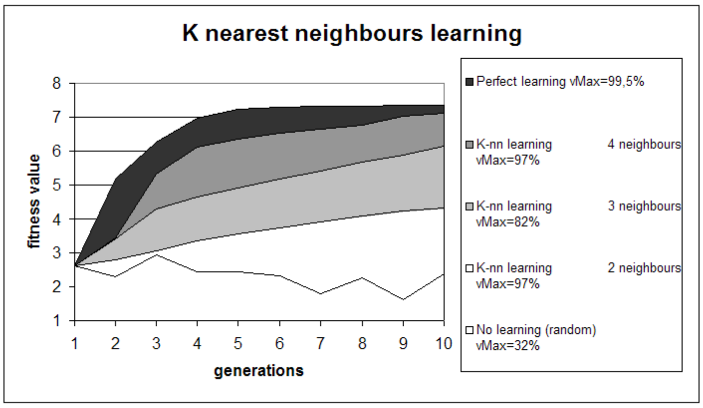


Fig. 4. Evolutionary behaviour of a hybridized GA with a k-nn algorithm, $k \in \{2, 3, 4\}$. Comparison with the learning limits. Note: the starting value of each curve depends on the benchmark and on the size of the population. Here, its value is 2.6 ± 0.5

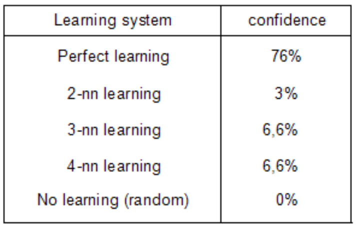
lower limit is the absence of learning, a random value with respect to the range of the benchmark (from 0 to 7.5).

4.3 Results

The quality of the algorithm is assessed by 2 criteria. First, the *performance* (vMax) is the average maximum value reached at the tenth generation. It is a percentage of the real optimum, for instance vMax = 50% means $7.5/2 = 3.75$. The Fig.4 presents the results of a hybrid for various k values. Whatever it is, the performance is manifestly better than using no learning. We expect that the upper bound is unreachable.

Second, the reliability of each algorithm is computed. Indeed a stochastic algorithm presents a different behaviour from one run to another. The *confidence* (conf) illustrates the percentage of runs where at least 98% of the optimum is really obtained on the whole the 30 runs. For instance, if 3 runs out of 30 reached at least 7.4 (98% of the optimum) then conf = 10%. The Fig.5 summarizes the values of this indicator according to the learning algorithm. The hybridization of a GA with no learning never reaches the optimum. In the opposite, the perfect learning fully benefits from the multiplication mechanism, the optimum is reached 23 times out of 30 at the tenth. Our proposal using k nearest neighbours occupies an intermediate position, whatever the value of k.

The results are in average better than no learning, either in terms of performance or in terms of confidence and the use of 4 neighbours gives the best results considering both criteria. In a real experiment, we would favour the confidence because the cost of a catalyst would not allow failure. The use



Learning system	confidence
Perfect learning	76%
2-nn learning	3%
3-nn learning	6,6%
4-nn learning	6,6%
No learning (random)	0%

Fig. 5. Percentage of runs which reaches at least 98% of the global optimum. Confidence of the hybrid GA/KD algorithm.

of a KD system makes it possible to improve the behaviour of a simple optimization algorithm, without increasing the global cost of experimentation. It makes it possible to choose in a relevant way among multiplied virtual individuals those which present truly good real performances.

5 Conclusion

We empirically studied in this article the hybridization of a genetic algorithm with a knowledge discovery system and its application to a heterogeneous catalysis problem whose fitness function is unknown. Its objective is to estimate the value of a potential solution to a problem which is not defined by a mathematical expression but by a set of observations, each of high monetary cost. We compare the results obtained by hybridizing a genetic algorithm with (1) a learning process using k nearest neighbours algorithm, (2) a perfect learning and (3) no learning. We show that the use of k-nn increases the optimization speed and improves the robustness compared to random learning.

There remains opened interrogations concerning the role of the number of neighbours. Increasing k value means that the k-nn algorithm is more linear and so the hybrid GA/KD would become less efficient for this application, but this remains to be demonstrated. Another question concerns the population multiplier, one expects that the higher, the more the chances to gain the optimum quickly are large. But however, this postulate is limited by the learning process. A ceiling value probably exists giving the best results possible for each KD algorithm.

Combinatorial catalysis is a vast field of investigation for applying new types of computer based optimizations and knowledge discovery systems. The actors of the domain are currently acquiring and storing data in vast databases. Combinatorial optimization methods are in total adequacy with experts needs and the expansion of such techniques is ensured. For this kind of hybrids, we are particularly interested in knowledge discovery methods which extract association rules. Indeed, the interactions between the predictive variables are often badly known and their description would be of a great interest. This constitutes the Knowledge Discovery in Genetic Algorithms (KDGA) project, materialized by a self made free software : OptiCat [IRC and ERIC, 2005] which has been used to perform all experiments presented here.

References

[B.Jandeleit *et al.*, 1998]B.Jandeleit, D.J.Schaefer, T.S.Powers, H.W.Turner, and W.H.Weinberg. Combinatorial materials science and catalysis. *Angew. Chem*, pages 2494,2532, 1998.

[D.Hand *et al.*, 2001]D.Hand, H.Mannila, and P.Smyth. *Principles of Data Mining.* Massachusetts Institute of Technology, 2001.

[Duff *et al.*, 2002]DG Duff, A Ohrenberg, S Voelkening, and M Boll. A screening workflow for synthesis and testing of 10,000 heterogeneous catalysts per day – lessons learned. *Macromolecular Rapid Communications*, pages 169–177, 2002.

[D.W.Aha *et al.*, 1991]D.W.Aha, D.Kibler, and M.K.Albert. *Instance-based learning algorithms.* Machine Learning, 1991.

[Farrusseng *et al.*, 2003]D Farrusseng, L Baumes, and C Mirodatos. In high-throughput analysis: A tool for combinatorial materials science. *Potyrailo*, pages 551–579, 2003.

[Goldberg, 1989]DE Goldberg. *Genetic Algorithms in Search, Optimization and machine learning.* Addison- Wesley, 1989.

[Hanagandi and Kargupta, 1996]V Hanagandi and H Kargupta. Unconstrained blackbox optimization: The search perspective. *Institute for Operations Research and the Management Sciences (INFORMS)*, 1996.

[Holland, 1975]J Holland. Adaptation in natural and artificial systems. 1975.

[IRC and ERIC, 2005]IRC and ERIC. `http://eric.univ-lyon2.fr/~fclerc/` OptiCat, 2005.

[Klanner *et al.*, 2003]C Klanner, D Farrusseng, L Baumes, C Mirodatos, and F Schuth. How to design diverse libraries of solid catalysts? *QSAR & Combinatorial Science*, 2003.

[Laguna and Glover, 1998]M Laguna and F Glover. *Handbook of Combinatorial Optimization.* Kluwer, Colorado Business Review, 1998.

[Miller and Goldberg, 1996]BL Miller and DE Goldberg. Genetic algorithms, selection schemes, and the varying effects of noise. *Evolutionary Computation*, 1996.

[S.Kirkpatrick *et al.*, 1983]S.Kirkpatrick, C.D.Gelatt, and M.P.Vecchi. Optimization by simulated annealing. *Science*, 1983.

[Wolf *et al.*, 2000]D Wolf, OV Buyevskaya, and M Baerns. An evolutionary approach in the combinatorial selection and optimization of catalytic materials. *Applied Catalysis A: General*, pages 63–77, 2000.

Invariances in Classification: an efficient SVM implementation

Gaëlle Loosli[1], Stéphane Canu[1], S.V.N. Vishwanathan[2], and Alex J. Smola[2]

[1] Laboratoire Perception, Systèmes, Information - FRE CNRS 2645
B.P. 08 - Place Emile Blondel
76131 - Mont Saint Aignan Cedex - France
(e-mail: gloosli@insa-rouen.fr, scanu@insa-rouen.fr)

[2] National ICT for Australia.
Canberra, ACT 0200 - Australia
(e-mail: vishy@axiom.anu.edu.au, alex.smola@anu.edu.au)

Abstract. Often, in pattern recognition, complementary knowledge is available. This could be useful to improve the performance of the recognition system. Part of this knowledge regards invariances, in particular when treating images or voice data. Many approaches have been proposed to incorporate invariances in pattern recognition systems. Some of these approaches require a pre-processing phase, others integrate the invariances in the algorithms. We present a unifying formulation of the problem of incorporating invariances into a pattern recognition classifier and we extend the SimpleSVM algorithm [Vishwanathan *et al.*, 2003] to handle invariances efficiently.

Keywords: SVM, Invariances, Classification, Active Constraints.

1 Introduction

The problem of invariances has been widely studied from a signal processing point of view for pattern recognition (see [Wood, 1996] for a review). Proposed methods in this area mainly consists in invariant features extraction before feature classification. To do so, Fourier transforms and similar transforms are used, as well as moment methods like Zernike moments [Wood, 1996]. In 1993 the invariances where taken into account in neural networks [Simard *et al.*, 1993] with the idea to modify the metric distance and use one that allows the variations of a pattern to be *close* the one from the others. In 1996 the invariances appeared for SVMs. In [Schölkopf *et al.*, 1996] the authors propose to generate some virtual examples to enlarge the dataset and thus make the algorithm learn invariances. Our approach is to provide a unifying framework for invariances in Support Vector Machines. First we define a general view of invariances in pattern recognition and show how to incorporate it in SVMs. The next part shows the connexion between our method and the existing ones. Finally we give details on *Invariant SimpleSVM* algorithm and some results obtained on the USPS database.

2 Invariant SimpleSVM

In this section we will propose a general formalisation of invariances in pattern recognition.

Definition. We need to define what a transformation is and how to apply it to any kind of patterns. A pattern x belongs to a level space $\mathcal{L}$ (for instance the contrast) and relies on its support space $\mathcal{S}$ (for instance the background). A pattern transformation is an application that maps a pattern and some parameters to a *transformed* pattern:

$$\begin{aligned} T : \mathcal{L} \times \Theta &\to \mathcal{L} \\ x, \theta \quad &\mapsto T(x, \theta) \end{aligned}$$

Moreover we require $T(x, 0) = x$.

If we now consider the binary classification purpose, we define the decision function $D(x)$ as a mapping from $\mathcal{X}^d$ to $\{0, 1\}$ that maps x to $D(x)$. If we want the decision function to be invariant with respect to the rotation, we will require that it gives the same decision for an image and its rotations:

$$\begin{aligned} &D(\{I(\ell_i, L_i), i = [1, d]\}) = D(\{I(R_\theta(\ell_i, L_i)), i = [1, d], \forall\theta\}) \\ &D(x_0) = D(x_\theta) \end{aligned}$$

2.1 Formulation

Integrating invariances into SVMs requires us to distinguish between separable (with no error) and non separable (with errors) cases. The first case is quite easily solvable while the second requires some non trivial constraints to be satisfied.

A kernel $k(x, y)$ is a positive and symmetric function of two variables (for more details see [Atteia and Gaches, 1999]) lying in a Reproducing Kernel Hilbert Space with the scalar product:

$$\langle f, g \rangle_{\mathcal{H}} = \sum_{i=1}^{k} \sum_{j=1}^{l} f_i g_j k(\mathbf{x}_i, \mathbf{x}'_j).$$

Hard-margins. In the separable case we can formulate the SVM problem with invariances as follows:

$$\begin{cases} \quad \min_{f,b} \frac{1}{2} \|f\|^2_{\mathcal{H}} \\ \text{s.t.} \quad\quad y_i(f(T(\mathbf{x}_i, \theta)) + b) \geq 1 \; i \in [1, m], \theta \in \Theta \end{cases} \tag{1}$$

where b is a scalar called bias. From this we can deduce the dual formulation (Wolfe's dual):

$$\begin{cases} \max\limits_{\boldsymbol{\alpha}} -\dfrac{1}{2}\displaystyle\sum_{i,j=1}^{m}\int\int_{\Theta}\alpha_i(\theta_1)\alpha_j(\theta_2)y_iy_jk(T(\mathbf{x}_i,\theta_1),T(\mathbf{x}_j,\theta_2))d\theta+\sum_{i=1}^{m}\int_{\Theta}\alpha_i(\theta)d\theta \\ \text{s.t. } \displaystyle\sum_{i=1}^{m}\int_{\Theta}\alpha_i(\theta)y_id\theta=0 \\ \text{and } \alpha_i(\theta)\geq 0 \qquad\qquad i\in[1,m],\theta\in\Theta \end{cases} \tag{2}$$

The Lagrange multipliers will be 0 for all points except the support vectors. Because of the nature of the hypothesis space, it is reasonable to assume that $\alpha_i(\theta)$ will have non-zero values only for a few finite number of parameters θ, thus we can simplify the writing:

$$\begin{cases} \max\limits_{\boldsymbol{\alpha}} -\dfrac{1}{2}\displaystyle\sum_{i,j,\theta_1,\theta_2}\alpha_i(\theta_1)\alpha_j(\theta_2)y_iy_jk(T(\mathbf{x}_i,\theta_1),T(\mathbf{x}_j,\theta_2))+\sum_{i,\theta}\alpha_i(\theta) \\ \text{s.t. } \displaystyle\sum_{i,\theta}\alpha_i(\theta)y_i=0 \\ \text{and } \alpha_i(\theta)\geq 0 \qquad i\in[1,m],\theta\in\Theta \end{cases} \tag{3}$$

$$\begin{cases} \max\limits_{\boldsymbol{\gamma}\in\mathbb{R}^{m\times p}} & -\frac{1}{2}\boldsymbol{\gamma}^\top G\boldsymbol{\gamma}+\mathbf{e}^\top\boldsymbol{\gamma} & \\ \text{s.t.} & \boldsymbol{\gamma}^\top\mathbf{y}=0 & \\ \text{and} & \gamma_i\geq 0 & i\in[1,mp] \end{cases} \tag{4}$$

where $\boldsymbol{\gamma}=[\boldsymbol{\alpha}_1(\theta);\boldsymbol{\alpha}_2(\theta);\ldots;\boldsymbol{\alpha}_m(\theta)]$, G is the block matrix defined as $G_{IJ}=y_iy_jK^{ij}$ with $K^{ij}_{kl}=k(T(x_i,\theta_k),T(x_j,\theta_l))$ and p is the size of Θ.

Soft-margin. Considering the case of soft-margin, i.e. the non separable case, we face another problem. Adding a slack variable to allow errors in the solution makes the last condition of system 2 $\alpha_i(\theta)\geq 0$ become $0\leq\int_{\Theta}\alpha_i(\theta)d\theta\leq C$ where C is a trade-off acting on the regularity of the decision function. This is quite similar to the hard-margin case. However we cannot make the assumption that we will end up with a finite number of non-zero valued Lagrange multipliers. Indeed if the trajectory of a transformation goes through the margin, then we have an infinite number of Lagrange multipliers bounded so that $\int_{\Theta}\alpha_i(\theta)d\theta=C$.

3 A Unifying Formulation

We can roughly identify three main streams in the algorithms dealing with invariances and support vector methods. Some are based on an artificial enlargement of the dataset, others rely on the modification of the cost function to incorporate the invariances (and thus using a different metric) and there are also methods using polynomial approximations to represent trajectories.

3.1 Enlarging artificially the dataset

A very intuitive way to learn invariances is to incorporate them in the training set. This operation can be processed before the use of the learning algorithm by artificially generating transformations of the sample data. The enlarged dataset (actual samples and virtual samples) contains thus the prior knowledge. Despite its simplicity, one major drawback of this method is the size of the resulting problem.

Virtual-SV. This idea was applied in SVM in [Schölkopf *et al.*, 1996] with the V-SV. The authors propose a way to reduce the size problem. Knowing that all the information needed for the classification task in SVM is contained in the support vectors, the authors make the assumption that one do not need to take into account the non-support vector's variations, since they are supposed to be far from the frontier between the classes. So basically the idea is to run a first time a classical SVM to retrieve the support vectors. Then the virtual vectors are generated from these support vectors and another SVM is run on the enlarged database. Experiments show that applying transformations on support vectors only gives at least as good results as enlarging the complete dataset.

Invariant SimpleSVM can achieve the same task if the transformation $T(x, \theta)$ is approximated by a finite number of point by fixing a finite number of values for θ.

3.2 Adapting the distance metric to invariances

Introduced in 1993 in [Simard *et al.*, 1993] and referred as the tangent distance, the motivation was to find a better distance measure than the Euclidean distance for the purpose of invariances treatment. In the field of support vector algorithm this idea has been used in various methods and we briefly describe some of them in the following parts.

Invariant SVM Let's now introduce the tangent vectors. The idea is to associate each training vector with one or several tangent vectors and to incorporate the invariances in the cost function [Chapelle and Schölkopf, 2002]. Optimising this cost function turns out to be equivalent to run a classic SVM with pre-processed data with a particular linear mapping. In the non-linear case, the results are similar except one would train a linear SVM on pre-processed data with a non-linear mapping.

Tangent Distance and Tangent Vector Kernels. Following the idea of taking a tangent measure (TD-measure), kernels embedding the invariances has been proposed. In [Pozdnoukhov and Bengio, 2003] the authors propose kernel functions between trajectories rather than between points. They define a function that measures the proximity between a point and the transformation trajectory of another point.

Invariant SimpleSVM is similar to these methods if the transformation is approximated by a first order polynomial ($T(x, \theta) \simeq x + \nabla_\theta T(x, 0)\theta$).

3.3 Polynomial Approximation

Semi-Definite Programming Machines.

Presenting the SDPM [Graepel and Herbrich, 2004], the authors are trying to learn data that are trajectories instead of the usual points. The aim is then to separate trajectories that represent the (differentiable) transformations of the original training points.

They show that this problem is solvable for transformations that can be represented or approximated by polynomials. Basing their approach on Nesterov's theorem they formulate the problem of learning a maximum margin classifier with an SDP formulation under polynomial constraints. Using the SD-representability of non negative polynomials they replace the usual non-negativity constraints in SVM by positive semi-definite constraints. Doing so they show that it is possible to learn to classify trajectories. However this approach is rather intractable since it requires to solve large SDP.

Invariant SimpleSVM also contains this approach if the transformation is approximated by a second order polynomial ($T(x, \theta) \simeq x + \nabla_\theta T(x, \theta)\theta + \frac{1}{2}\theta^\top H_\theta \theta$).

In the separable case, we can solve directly the problem and our solution is thus more tractable than the SDPM. Nevertheless we are penalised in the non separable case since we need to discretise the parameter space. Despite the complexity of SDPM, it always works in the original space Θ.

4 Algorithm and applications

We present in this section the SimpleSVM algorithm. We extend this method for invariances because of its structure. Briefly, SimpleSVM adds points to the solution one by one and this let us incorporate some treatment to each point separately. This strong property breaks down the computing time that would occur if one would apply the equivalent treatment to the whole database. The main idea is to transform points only when adding them to the solution, which excludes from this treatment all the points that are far from the frontier between classes.

4.1 SimpleSVM

The SimpleSVM algorithm [Vishwanathan *et al.*, 2003] is based on the decomposition of the database into three groups (the *working* set, the *inactive* set and the *bounded* set). Assuming the groups are known, it solves the SVM optimisation problem on the working set only. Having a solution, it checks whether the group repartition is relevant. If not the groups are updated (by adding a violator point in the working set) and it iterates over these two steps (see algorithm 1). A detailed explanation can be found in [Loosli *et al.*, 2004].

Algorithm 3 : *simpleSVM*

```
1. $(I_s, I_0, I_c) \leftarrow$ initialise
while minimumReached=FALSE
    2. $(\alpha,\lambda) \leftarrow$ solve the system without constraints($I_s$)
    if $\exists\alpha_i \leq 0$ or $\exists\alpha_i \geq C$
        3.1 project $\alpha$ inside the admissible set
        3.2 transfers the associated point from $I_s$ to $I_0$ or $I_c$
    else
        4. look for the best candidate $x_{cand}$ in $I_c$ and $I_0$
        if $x_{cand}$ is found
            5. transfer $x_{cand}$ to $I_s$
        else
            6. minimumReached $\leftarrow$ TRUE
        end if
    end if
end while
```

The Matlab implementation of this algorithm as well as the *invariant SimpleSVM* are available at:
`http://asi.insa-rouen.fr/~gloosli/simpleSVM.html` [Loosli, 2004].

4.2 Invariant SimpleSVM

Invariant SimpleSVM integrates invariances like virtual vectors, first order polynomials and so on. In the implementation we present here we chose to deal with virtual vectors. The idea is to add virtual vectors that are derived from potential support vectors only. Doing so we can achieve the same task as V-SV in only one run of the algorithm. Compared to SimpleSVM, only the step 4 in algorithm 1 is modified. While in SimpleSVM the best candidate is the point that violates the most the constraints in the dual space (or is the worst classified in the primal space), for *Invariant SimpleSVM* the best candidate is chosen among the transformations of one vector. Here we can come up with several heuristics, depending on how the transformations are represented. Let's take the case we choose to discretize the space of parameters Θ:

- complete search: each step considers only one point and all its transformations and searches whether one violates the constraints,
- incomplete search: each step considers a group of points and searches whether one violates the constraints. If so, it also considers all the transformations and looks if one is worse than the original point,
- random search: can be applied to both of the previous heuristics. Instead of taking all the transformations, pick randomly one or several transformations. This way is faster but does not necessarily reach the best solution.

4.3 Application to character recognition

It is known that incorporating invariances improves the results of a recognition task. In our experiments we first to get an idea of the efficiency of the method, in other words to monitor the actual improvements. We present the results for the complete USPS database in order to compare our method to the published results.

Experiments settings. All the results here are obtained on the USPS database. This database contains 7291 training pictures 16 × 16 pixels, valued in $[-1, 1]$ and 2007 test pictures. Pictures represent digits from 0 to 9 collected from handwritten postcodes. This dataset is widely used to benchmark recognition methods and is known as a difficult set. Indeed the human performance is 2.5% of error.

The nature of the data induces the choice of the transformations to apply. A digit means the same regardless of translations, small rotations, line thickness for instance. Hence the transformation used for experiments are vertical and horizontal translation, rotation with angle 10° clockwise and anti-clockwise, line thinning and thickening. These transformations are computed on-the-fly for any point point that is about to reach the working set. The main advantage of this choice is that we do not need to store all the transformations of all the points. However it increases the training time.

The experiments on the USPS database were done with several objectives. The first one was to show our algorithm was efficient and fast. The second one was to explore different combinations of transformations (for instance published results with SVM methods are applied with only the translation of one pixel). The results are shown in table 1. The parameters are obtained from a cross-validation. In table 2 we give the main published results on USPS.

Kernel	bandwidth	C	Transformation	Error	Time (train and test)
Poly 5	0.1	10^-5	none	4.09	235 sec
Poly 5	0.1	10^-4	tr	3.44	1800 sec
Poly 5	0.1	10^-4	tr+rot	3.19	3200 sec
Poly 5	0.1	10^-4	tr+er	3.14	-
Poly 5	0.1	10^-4	tr+er+dil	3.24	2400 sec
Poly 5	*0.1*	10^-4	*tr+rot+er*	*2.99*	*2300 sec*
Poly 5	0.1	10^-4	tr+rot+er+dil	3.24	4800 sec

Table 1. results on USPS: Here we present results obtained with *Invariant SimpleSVM*. The applied transformations are translation of 1 pixel (*tr*), rotation (*rot*), erosion and dilatation (respectively *er* et *dil*). The best result is obtained in less than 40 minutes. Note that the computation time depends on the number of support vectors, thus adding a transformation may improve computing time if it generates good support vectors that are eliminating many candidates (for instance this happens between *tr + rot* and *tr + rot + er*, where erosion clearly gives good support vectors and the algorithm converges faster).

Method	Error
Tangant Vector and Local Rep. [Keysers *et al.*, 2002]	2.0 %
Virtual SVM [Schölkopf *et al.*, 1996]	3.2 %
Invariance Hyperplane + V-SV	3.0 %
Invariant SimpleSVM (this paper)	3.0 %
Human performance	2.5 %

Table 2. Some published results on USPS

4.4 Discussion

We show here that *Invariant SimpleSVM* is efficient. Let's note that we have implemented the transformations with the discrete point of view, which is equivalent to the V-SV method. However we achieve a better performance on USPS. This can be explained by the fact we method is more flexible concerning the points which generates virtual vectors. Indeed we consider the transformations of each point that could be support vector, but not necessarily is support vector in the end. That way we consider more transformations. Taking into account the invariances considerably increases the training time (from less than 6 min without transformations up to 1 hour if we consider all the listed transformations) but it is still very fast compared to other methods. As for the effect of the different transformations, it is hard to conclude. We noticed that the translation is the most influential one. The others have small effects and the differences between different combinations are not significant enough.

5 Conclusion

Based on the unifying approach for invariances with SVM proposed, an efficient implementation for the virtual vector case has been developed. This implementation is an interesting evolution of the SimpleSVM algorithm and is available on our website [Loosli, 2004]. The efficiency of our method has been illustrated on the USPS database. Our results outperform the equivalent algorithm Virtual-SVM in a significantly shorter computational time. We are now carrying on a deeper study of the comparison with the SDPM.

This work was supported in part by the IST Program of the European Community, under the PASCAL Network of Excellence, IST-2002-506778. This publication only reflects the authors' views.

References

[Atteia and Gaches, 1999]Marc Atteia and Jean Gaches. *Approxiation Hilbertienne.* Presses Universitaires de Grenoble, 1999.

[Chapelle and Schölkopf, 2002]O. Chapelle and B. Schölkopf. Incorporating invariances in nonlinear SVMs. In Dietterich T. G.and Becker S. and Ghahramani Z., editors, *Advances in Neural Information Processing Systems*, volume 14, pages 609–616, Cambridge, MA, USA, 2002. MIT Press.

[DeCoste and Schölkopf, 2002]Dennis DeCoste and Bernhard Schölkopf. Training invariant support vector machines. *Machine Learning*, 46:161–190, 2002.

[Graepel and Herbrich, 2004]Thore Graepel and Ralf Herbrich. Invariant pattern recognition by semi-definite programming machines. In Sebastian Thrun, Lawrence Saul, and Bernhard Schölkopf, editors, *Advances in Neural Information Processing Systems 16*. MIT Press, Cambridge, MA, 2004.

[Keysers *et al.*, 2002]D. Keysers, R. Paredes, H. Ney, and E. Vidal. Combination of tangent vectors and local representation for handwritten digit recognition. In *Lecture Notes in Computer Science*, volume LNCS 2396, pages 538–547. SPR2002, International Workshop on Statistical Pattern Recognition, Windsor, Ontario, Canada, springer-vertag edition, Aug 2002.

[Loosli *et al.*, 2004]G. Loosli, S. Canu, S.V.N. Vishwanathan, Alexander J. Smola, and Monojit Chattopadhyay. Une boîte à outils rapide et simple pour les SVM. In Michel Liquière and Marc Sebban, editors, *CAp 2004 - Conférence d'Apprentissage*, pages 113–128. Presses Universitaires de Grenoble, 2004.

[Loosli, 2004]G. Loosli. Fast SVM toolbox in MATLAB based on SimpleSVM algorithm, 2004. `http://asi.insa-rouen.fr/~gloosli/simpleSVM.html`.

[Pozdnoukhov and Bengio, 2003]A. Pozdnoukhov and S. Bengio. Tangent vector kernels for invariant image classification with SVMs. IDIAP-RR 75, IDIAP, Martigny, Switzerland, 2003. Submitted to International Conference on Pattern Recognition 2004.

[Schölkopf *et al.*, 1996]B. Schölkopf, C. Burges, and V. Vapnik. Incorporating invariances in support vector learning machines. In J.C. Vorbrüggen C. von der Malsburg, W. von Seelen and B. Sendhoff, editors, *Artificial Neural Networks — ICANN'96*, volume 1112, pages 47–52, Berlin, 1996. Springer Lecture Notes in Computer Science.

[Simard *et al.*, 1993]P. Simard, Y. LeCun, and Denker J. Efficient pattern recognition using a new transformation distance. In S. Hanson, J. Cowan, and L. Giles, editors, *Advances in Neural Information Processing Systems*, volume 5. Morgan Kaufmann, 1993.

[Vishwanathan *et al.*, 2003]S. V. N Vishwanathan, A. J. Smola, and M. Narasimha Murty. SimpleSVM. In *Proceedings of the Twentieth International Conference on Machine Learning*, 2003.

[Wood, 1996]Jeffrey Wood. Invariant pattern recognition: A review. *Pattern Recognition*, 29 Issue 1:1–19, 1996.

Learning numbers from Graphs

Aurélie Goulon[1], Arthur Duprat[1,2], Gérard Dreyfus[1]

Ecole Supérieure de Physique et de Chimie Industrielles de la Ville de Paris (ESPCI-ParisTech)
[1]Laboratoire d'Électronique, [2]Laboratoire de Chimie Organique (CNRS UMR 7084)
10, rue Vauquelin
75005 PARIS, France
(e-mail agoulon@libertysurf.fr, Arthur.Duprat@espci.fr, Gerard.Dreyfus@espci.fr)

Abstract. The recent developments of statistical learning focused mainly on vector machines, i.e. on machines that learn from examples described by a vector of features. There are many fields where structured data must be handled; therefore, it would be desirable to learn from examples described by graphs. The presentation describes graph machines, which learn real numbers from graphs. Applications in the field of Quantitative Structure-Activity Relations (QSAR), which aim at predicting properties of molecules from their (graph) structures, are described.
Keywords: Graph machines, Vector machines.

1 Introduction

The present paper describes graph machines, i.e. machines that learn numbers from structured data, which can be described by graphs, in contrast to conventional approaches such as neural networks, kernel machines, support vector machines, which handle vectors. Unlike recursive neural networks, graph machines can handle any type of graph, whether cyclic or not. The first part of the paper is devoted to definitions. The second part is devoted to examples of applications; first, academic validations are described, showing that graph machines are indeed able to learn numbers related to the graph structure itself, such as graph diameters or Wiener indices. We proceed to show that graph machines are very efficient for QSAR and QSPR applications; comparisons with results obtained by other authors on the same data show that graph machines outperform standard machine learning techniques and recursive neural networks, with the computational advantage of exempting the model designer from performing the steps of computing and selecting descriptors, which are generally at least as costly as the training of the machine.

2 Graph machines

Before describing graph machines, some facts and definitions pertaining to vector machines are described cursorily.

2.1 Vector machines

Conventional numerical machine learning methods aim at learning applications from $\mathcal{R}^n$ to $\mathcal{R}^m$: data is in the form of pairs of vectors, the input vector being of dimension n, and the output vector of dimension m. In all the following we consider that $m = 1$ without loss of generality. When the task to be learnt is a classification task, the output is often binary; for process modeling, whether static or dynamic, the output is real. The techniques of machine learning for static modeling are very similar to statistical regression techniques: the main difference is the fact that statistical regression is essentially interested in the values of the *parameters* of the models, whereas modeling by machine learning is essentially interested in the *predictions* of the models. Support vector machines and neural networks are typical vector machines; support vector machines were designed mainly for classification tasks, with excellent performances; neural networks are more suitable for modeling, whether static (feedforward neural networks, also termed Multilayer Perceptrons), or dynamic (recurrent neural networks).

In all the following, we focus on static modeling, i.e. learning from examples an application of $\mathcal{R}^n$ to $\mathcal{R}$. The model is sought within a family of parameterized functions $g_\theta(x)$, where x is the vector of variables (of dimension n) and θ is the vector of parameters (of dimension p). Training is performed by minimizing a cost function, which is usually the least squares cost function, with respect to the parameters:

$$J(\theta) = \sum_{i=1}^{N} (y_p^i - g_\theta(x_i))^2 \tag{1}$$

where the summation runs on all N examples of the training set, y_p^i is the value of the quantity to be modeled for example i, and x_i is the vector of variables for example i.

2.2 Graph machines

2.2.1 Definition We turn now to the problem of learning an application between a set of graphs and a corresponding set of real (or possibly binary) numbers. To start with, we consider directed acyclic graphs only. A natural idea is to build a model whose mathematical structure is the same as the structure of the input graph: each node of the graph is a parameterized function, and the model is a composition of that function, which reflects the structure of the graph. In the field of neural networks, such models are known as *folding networks* (the function present at each node is a feedforward neural network), but the idea can be extended to other types of machines (for a review see [Hammer, 2003]). Consider, as an illustration, the graphs shown on Figure 1:

- the graph machine associated to graph 1 is:
$f^1_{\theta,\Theta} = G_\Theta\{g_\theta(x), g_\theta(x), g_\theta(x)\}$;
- the graph machine associated to graph 2 is:
$f^2_{\theta,\Theta} = G_\Theta\{g_\theta(g_\theta(g_\theta(x), g_\theta(x)), 0), g_\theta(g_\theta(x), g_\theta(g_\theta(x), g_\theta(x)))\}$;
- the graph machine associated to graph 3 is:
$f^2_{\theta,\Theta} = G_\Theta\{g_\theta(g_\theta(g_\theta(x), g_\theta(g_\theta(g_\theta(x), 0), g_\theta(g_\theta(x), g_\theta(x)))), 0), 0\}$.

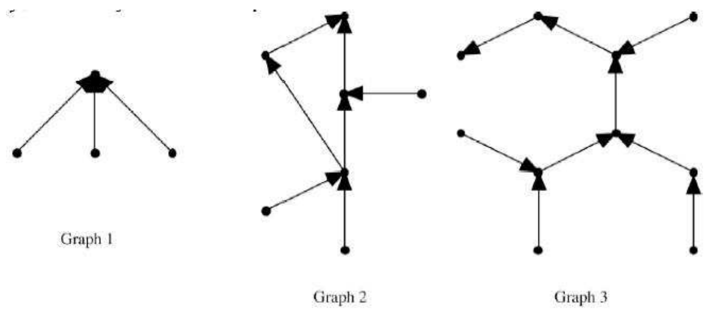


Fig. 1.

In the above examples, the size of x must be at least equal to the maximal in-degree d_m of the nodes of the graph. For a node of in-degree $d < d_m$, $d_m - d$ components of x are arbitrary, and may be taken equal to 1 for instance.

For generality, in the above examples, the function G_Θ associated to the final root is different from the function g_θ of the other nodes. That is by no means necessary; in all examples described in the present paper, all nodes including the root were assigned the same function.

The size of vector θ (and the size of Θ) depends on the complexity of the mapping, just as for vector machines.

Definition: a graph machine is a set G of parameterized functions, constructed as described above from the same functions $g_\theta(x)$ (and $G_\Theta(x)$), which are representations of the graphs of the training set. The size of x is lower bounded by the maximal in-degree of the nodes of the graphs.

2.2.2 The training of a graph machine

The training of a graph machine is performed by minimizing a cost function with respect to the parameters θ (and Θ); in all the examples described below, the least squares cost function was used:

$$J(\theta, \Theta) = \sum_{i=1}^{N} (y^i_p - f^i_{\theta,\Theta})^2 \qquad (2)$$

where y_p^i is the quantity to be learnt, associated to graph i. Note the difference with the cost function (1) that is minimized during the training of a vector machine: .

$$J(\theta) = \sum_{i=1}^{N} (y_p^i - g_\theta(x_i))^2$$

Instead of training a single parameterized function with different input output-pairs, different parameterized functions, *sharing the same set of parameters*, are trained with a single example each.

As mentioned above, the fact that two sets of parameters, θ and Θ, are used in graph machines is unimportant. A single parameter vector θ is often sufficient.

In practice, training is performed much in the same way as vector machines. One has

$$\frac{\delta J}{\delta \Theta_k} = \sum_{i=1}^{N} \frac{\delta J^i}{\delta \Theta_k} \text{ where } J^i = (y_p^i - f_{\theta,\Theta}^i)^2 \tag{3}$$

and Θ_k denotes the k-th component of vector Θ. For neural networks, the gradient of J^i with respect to each parameter $\frac{\delta J^i}{\delta \Theta_{k_j}}$ is computed by backpropagation on the network that represents graph i; $\delta\Theta_{k_j}$ denotes the j-th occurrence of parameter Θ_k in graph i. Denoting by $n_{\Theta_k}^i$ the number of occurrences of parameter Θ_k in graph i, the shared weight trick consists in setting

$$\frac{\delta J^i}{\delta \Theta_k} = \sum_{j=1}^{n_{\Theta_k}^i} \frac{\delta J^i}{\delta \Theta_{k_j}} \tag{4}$$

(if the root node has the same parameters as the other nodes, then $n_{\Theta_k}^i$ is equal to the number of nodes in graph i). Therefore, one obtains:

$$\frac{\delta J}{\delta \Theta_k} = \sum_{i=1}^{N} \sum_{j=1}^{n_{\Theta_k}^i} \frac{\delta J^i}{\delta \Theta_{k_j}} \tag{5}$$

Finally, the cost function (2) is minimized by any appropriate gradient optimization method (Levenberg-Marquardt, BFGS, conjugate gradient, etc.), using gradient (5).

2.2.3 Model selection All the tricks-of-the-trade that are usually applied to vector machines can be applied to graph machines as well: validation, cross-validation, leave-one-out, bootstrap estimates of the generalization error, etc. In the following examples, cross-validation is used for model selection; the root mean square error on a set (training or validation) is defined as:

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_p^i - f_{\theta,\Theta}(x_i))^2} \tag{6}$$

where N is the size of the set.

3 Examples

3.1 Neural-network-based graph machines

In all examples described below, the function g_θ is a neural network. Therefore, a machine is made of identical neural networks, connected with the same structure as the nodes in the graph. For example, consider a node A with n parent vertices B_i, $i = 1, \dots, n$. An elementary neural network (A) is assigned to that node: its inputs are (i) the outputs of the networks (B_i), and (ii) additional inputs that provide information on the node (e.g. its degree). If the graph is cyclic, the degree of the node is provided by one such input, so that a cyclic graph is first turned into a directed acyclic graph by deleting as many edges as necessary, while retaining the information about the original graph structure.

3.2 Learning graph properties

In order to validate the approach in an academic way, graph machines were trained to learn graph properties. For all examples described below, a data base of 150 randomly generated graphs, featuring 2 to 15 nodes and 0 to 9 cycles was created. Various splits between training and validation sets were performed on that data base.

3.2.1 Learning the number of nodes and cycles of a graph The easiest problem consists in learning the number of nodes of a graph, since it is a linear problem. The number of vertices N is equal to the number of elementary functions g_θ in the graph machine, and the number of cycles of a connected graph is given by:

$$C = E - N + 1$$

where E is the number of edges. Therefore, graph machines with linear elementary functions

$$g_\theta(x) = \sum_i \Theta_i x_i$$

should learn those tasks. As expected, for all splits between training and validation sets, the task was perfectly learnt and the error was equal to zero.

3.2.2 Learning the diameter of a graph The diameter of a graph is the length of the shortest path between its most distant nodes:

$$D = \max_{u,v} d(u,v)$$

where $d(u,v)$ is the distance (the shortest path) between nodes u and v. In the database under investigation, that index ranges from 1 to 9. That is clearly a non-linear property; therefore, the elementary function was a neural network with four hidden neurons. The RMS error (relation (6)) on the training set is 0.36, and the RMS validation error (10-fold cross-validation) is 0.53. Since the index is an integer ranging from 1 to 9, the prediction is excellent given the complexity of the graphs.

3.2.3 Learning the Wiener Index of a graph The Wiener Index of a graph G is the sum of the distances between the vertices of G. That index was first defined by the chemist H. Wiener, in order to investigate the relationships between the structure of chemicals and their properties:

$$W(G) = \frac{1}{2} \sum_{u,v} d(u,v)$$

In our database, that index ranges from 1 to 426. 10-fold cross-validation was performed with a 4-hidden neuron elementary neural network, leading to a RMS validation error of 7.9.

The above examples (together with other examples not reported here) show the ability of graph machines to learn from the sole data structure, without any need for extraneous descriptors.

In addition, they prove that indices such as the Wiener index need not be used as descriptors (e.g. in QSAR as described in the next section) since the information is present in the structure of the machine.

3.3 Application to the prediction of chemical properties of molecules

3.3.1 Graph machines for QSPR/QSAR Graph machines are particularly appropriate for the prediction of molecular properties in QSPR (Quantitative Structure-Property Relations) and QSAR (Quantitative Structure-Activity Relations). A molecule can be described as a directed graph by associating each non-hydrogen atom to a node and each bond to an edge. The original graphs are preprocessed in order to turn them to acyclic graphs as described in section 3.1. Provision is made for extraneous inputs that code (in a one-out-of-*n* code) for the nature of the atoms, their degree or their stereochemistry for example. Therefore, any type of molecule can be handled, be it acyclic, cyclic or even aromatic. Figure 2 shows how an aromatic compound can be described as a graph; the digits are the degrees of the nodes.

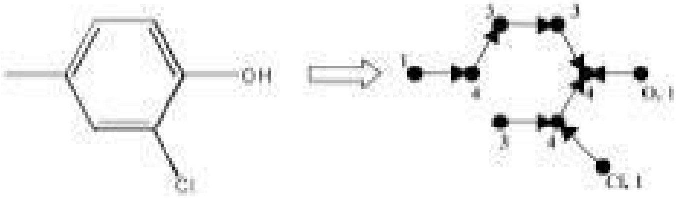


Fig. 2.

3.3.2 Learning and predicting boiling points of alkanes

Graph machines were first tested on the prediction of the boiling points of a set of linear or branched acyclic alkanes. Table 1 compares the results obtained by graph machines to those found in the literature.

	RMSE (K)	RMSE (K)
Graph machines	1.0	1.5
Recursive neural networks [Bianucci et al., 2000]	2.0	3.0
Conventional neural networks [Cherqaoui and Villemin, 1994]	2.2	2.7

Table 1.

3.3.3 Predicting the toxicity of phenols

Phenols are a family of chemicals that are of current industrial use as biocides or disinfectants. Most synthetic phenols are toxic and considered as dangerous pollutants. We studied a set of 153 of these phenols, whose toxicity to a particular kind of cells, Tetrahymena pyriformis, was available ([Schultz, 1997]). This database is especially interesting since it contains complex molecules with 6 kinds of heteroatoms, and it deals with a property that is close to pharmacological properties.

In order to compare the performances of graph machines to those obtained with other methods - Multiple Linear Regression (MLR), Support Vector Machines (SVM) and Radial Basis Function Neural Networks (RBFNN) - the same protocol as used in [Yao *et al.*, 2004] was implemented: the database was split into training and validation sets of 131 and 22 examples respectively. The results obtained with graph machines built with 4 and 5 hidden neuron (GM-4N and GM-5N) elementary neural networks are summarized in Table 2, where they are compared to those obtained with the previously cited methods.

	Method	Learning	Validation
RMSE	GM-4N	0.17	0.29
	GM-5N	0.09	0.32
	MLR	0.30	0.46
	RBFNN	0.19	0.29
	SVM	0.22	0.36

Table 2.

For a more rigorous assessment, 7-fold cross-validation was also performed on the same set with a 4 hidden neuron network. The RMSE obtained were then respectively 0.16 and 0.29 in learning and validation. No test set was provided in the referenced articles.

The above results show that graph machines compare favorably with other QSAR methods for the prediction of that biological activity, with an accuracy that is at least as good as the accuracy of the methods investigated by other authors, be it on the learning or on the validation sets. However, whereas the other methods require the prior selection and measurement or computation of descriptors such as hydrophobicity (log Kow), acidity constant (pKa), and frontier orbital energies (HOMO and LUMO), the structure of the molecules is the only information required for graph machines to perform accurate predictions. This is a twofold advantage. First, graph machines are much less computationally expensive than methods that require the design, selection and computation of descriptors. Furthermore, since no specific descriptors are selected, a graph machine implemented for a given molecule can be used for the prediction of *any* property of that molecule: the only requirement is a re-training of the machine, whereas conventional vector machines require the selection and computation of an appropriate set of descriptors for each property to be predicted.

3.3.4 A classification task: classification of the molecules as aromatics/ non aromatics Graph machines can perform classification tasks, just as neural networks or SVM's do. As a test example, 240 molecules were classified into two classes: molecules that feature no aromatic cycle and molecules that feature at least one aromatic cycle. 10-fold cross-validation was performed on that database, leading to a training classification error rate of 0% and a validation error rate of 2% with a 4 hidden neuron elementary neural network. A test set of 40 examples lead to an error rate of 0%. Again, no descriptor whatsoever was computed prior to performing classification.

4 Conclusion

In the present paper, graph machines have been described, and some of their applications have been outlined. The results presented here show that graph machines outperform vector machines and recursive neural networks. The prediction of the properties of molecules from their structure is obviously an important field of application of our approach, but it can be conjectured that graph machines may be beneficial in all fields where learning must be performed from structured data.

References

[Bianucci *et al.*, 2000]A.M. Bianucci, A. Micheli, A. Sperduti, and A. Starita. Application of cascade correlation networks for structures to chemistry. Applied Intelligence, pages 115-145, 2000.

[Cherqaoui and Villemin, 1994]. Cherqaoui and D. Villemin. Use of a neural network to determine the boiling point of alkanes. Journal of the Chemical Society, Faraday Transactions, pages 97-102, 1994.

[Hammer, 2003]. Hammer, Perspectives on Learning Symbolic Data with Connectionistic Systems. In R.Kühn, R.Menzel, W.Menzel, U.Ratsch, M.M.Richter, I.-O.Stamatescu, eds., , Adaptivity and Learning, pages 141-160, Springer, 2003.

[Schultz, 1997].W. Schultz. TETRATOX: The Tetrahymena pyriformis population growth impairment endpoint - A surrogate for fish lethality. Toxicological Methods 7, pages 289-309, 1997.

[Yao *et al.*, 2004]. J. Yao, A. Panaye, J.P. Doucet, R.S. Zhang, H.F. Chen, M.C. Liu, Z.D. Hu, and B.T. Fan. Comparative Study of QSAR/QSPR Correlations Using Support Vector Machines, Radial Basis Function Neural Networks, and Multiple Linear Regression. Journal of Chemical Information and Computer Sciences, pages 1257-1266, 2004.

Testing the number of parameters of multidimensional MLP

Joseph Rynkiewicz

SAMOS - MATISSE
Université de Paris I
72 rue Regnault, 75013 Paris, France
(e-mail: joseph.rynkiewicz@univ-paris1.fr)

Abstract. This work concerns testing the number of parameters in one hidden layer multilayer perceptron (MLP). For this purpose we assume that we have identifiable models, up to a finite group of transformations on the weights, this is for example the case when the number of hidden units is know. In this framework, we show that we get a simple asymptotic distribution, if we use the logarithm of the determinant of the empirical error covariance matrix as cost function.

Keywords: Multilayer Perceptron, Statistical test, Asymptotic distribution.

1 Introduction

Consider a sequence $(Y_t, Z_t)_{t\in\mathbb{N}}$ of i.i.d.[1] random vectors (i.e. identically distributed and independents). So, each couple (Y_t, Z_t) has the same law that a generic variable $(Y, Z) \in \mathbb{R}^d \times \mathbb{R}^{d'}$.

1.1 The model

Assume that the model can be written

$$Y_t = F_{W^0}(Z_t) + \varepsilon_t$$

where

- F_{W^0} is a function represented by a one hidden layer MLP with parameters or weights W^0 and sigmoidal functions in the hidden unit.
- The noise, $(\varepsilon_t)_{t\in\mathbb{N}}$, is sequence of i.i.d. centered variables with unknown invertible covariance matrix $\Gamma(W^0)$. Write ε the generic variable with the same law that each ε_t.

Notes that a finite number of transformations of the weights leave the MLP functions invariant, these permutations form a finite group (see [Sussman, 1992]). To overcome this problem, we will consider equivalence classes of

[1] It is not hard to extend all what we show in this paper for stationary mixing variables and so for time series

MLP : two MLP are in the same class if the first one is the image by such transformation of the second one, the considered set of parameter is then the quotient space of parameters by the finite group of transformations.

In this space, we assume that the model is identifiable, this can be done if we consider only MLP with the true number of hidden units (see [Sussman, 1992]). Note that, if the number of hidden units is over-estimated, then such test can have very bad behavior (see [Fukumizu, 2003]). We agree that the assumption of identifiability is very restrictive, but we want emphasize the fact that, even in this framework, classical test of the number of parameters in the case of multidimensional output MLP is not satisfactory and we propose to improve it.

1.2 testing the number of parameters

Let q be an integer lesser than s, we want to test "$H_0 : W \in \Theta_q \subset \mathbb{R}^q$" against "$H_1 : W \in \Theta_s \subset \mathbb{R}^s$", where the sets Θ_q and Θ_s are compact. H_0 express the fact that W belongs to a subset of Θ_s with a parametric dimension lesser than s or, equivalently, that $s-q$ weights of the MLP in Θ_s are null. If we consider the classic cost function : $V_n(W) = \sum_{t=1}^{n} \|Y_t - F_W(Z_t)\|^2$ where $\|x\|$ denotes the Euclidean norm of x, we get the following statistic of test :

$$S_n = n \times \left(\min_{W \in \Theta_q} V_n(W) - \min_{W \in \Theta_s} V_n(W) \right)$$

It is shown in [Yao, 2000], that S_n converges in law to a ponderated sum of χ_1^2

$$S_n \xrightarrow{\mathcal{D}} \sum_{i=1}^{s-q} \lambda_i \chi_{i,1}^2$$

where the $\chi_{i,1}^2$ are $s-q$ i.i.d. χ_1^2 variables and λ_i are strictly positives values, different of 1 if the true covariance matrix of the noise is not the identity. So, in the general case, where the true covariance matrix of the noise is not the identity, the asymptotic distribution is not known, because the λ_i are not known and it is difficult to compute the asymptotic level of the test.

To overcome this difficulty we propose to use instead the cost function

$$U_n\left(W\right) := \ln \det \left(\frac{1}{n} \sum_{t=1}^{n} (Y_t - F_W(Z_t))(Y_t - F_W(Z_t))^T \right). \tag{1}$$

we will show that, under suitable assumptions, the statistic of test :

$$T_n = n \times \left(\min_{W \in \Theta_q} U_n(W) - \min_{W \in \Theta_s} U_n(W) \right) \tag{2}$$

will converge to a classical χ_{s-q}^2 so the asymptotic level of the test will be very easy to compute. The sequel of this paper is devoted to the proof of this property.

2 Asymptotic properties of T_n

In order to investigate the asymptotic properties of the test we have to prove the consistency and the asymptotic normality of $\hat{W}_n = \arg\min_{W\in\Theta_s} U_n(W)$. Assume, in the sequel, that ε has a moment of order at least 2 and note

$$\Gamma_n(W) = \frac{1}{n}\sum_{t=1}^{n}(Y_t - F_W(Z_t))(Y_t - F_W(Z_t))^T$$

remark that these matrix $\Gamma_n(W)$ and it inverse are symmetric. in the same way, we note $\Gamma(W) = \lim_{n\to\infty}\Gamma_n(W)$, which is well defined because of the moment condition on ε

2.1 Consistency of $\hat{W}_n$

First we have to identify contrast function associated to $U_n(W)$

<u>Lemma</u> 1

$$U_n(W) - U_n(W^0) \overset{a.s.}{\to} K(W, W^0)$$

with $K(W, W^0) \geq 0$ *and* $K(W, W^0) = 0$ *if and only if* $W = W^0$.

Proof : By the strong law of large number we have

$$U_n(W) - U_n(W^0) \overset{a.s.}{\to} \ln\det(\Gamma(W)) - \ln\det(\Gamma(W^0)) = \ln\frac{\det(\Gamma(W))}{\det(\Gamma(W^0))} = \ln\det\left(\Gamma^{-1}(W^0)\left(\Gamma(W) - \Gamma(W^0)\right) + I_d\right)$$

where I_d denotes the identity matrix of $\mathbb{R}^d$. So, the lemme is true if $\Gamma(W) - \Gamma(W^0)$ is a positive matrix, null only if $W = W^0$. But this property is true since

$$\Gamma(W) = E\left((Y - F_W(Z))(Y - F_W(Z))^T\right) =$$
$$E\left((Y - F_{W^0}(Z) + F_{W^0}(Z) - F_W(Z))(Y - F_{W^0}(Z) + F_{W^0}(Z) - F_W(Z))^T\right) =$$
$$E\left((Y - F_{W^0}(Z))(Y - F_{W^0}(Z))^T\right) +$$
$$E\left((F_{W^0}(Z) - F_W(Z))(F_{W^0}(Z) - F_W(Z))^T\right) =$$
$$\Gamma(W^0) + E\left((F_{W^0}(Z) - F_W(Z))(F_{W^0}(Z) - F_W(Z))^T\right) \blacksquare$$

We deduce then the theorem of consistency :

<u>Theorem</u> 1 *If* $E\left(\|\varepsilon\|^2\right) < \infty$,

$$\hat{W}_n \overset{P}{\to} W^0$$

Proof Remark that it exist a constant B such that

$$sup_{W\in\Theta_s}\|Y - F_W(Z)|^2 < \|Y\|^2 + B$$

because Θ_s is compact, so $F_W(Z)$ is bounded. For a matrix $A \in \mathbb{R}^{d\times d}$, let $\|A\|$ be a norm, for example $\|A\|^2 = tr\left(AA^T\right)$. We have

$$\begin{array}{l} \liminf_{W\in\Theta_s} \|\Gamma_n(W)\| = \|\Gamma(W^0)\| > 0 \\ \limsup_{W\in\Theta_s} \|\Gamma_n(W)\| := C < \infty \end{array}$$

and since the function :

$$\Gamma \mapsto \ln\det\Gamma, \text{ for } C \geq \|\Gamma\| \geq \|\Gamma(W^0)\|$$

is uniformly continuous, by the same argument that example 19.8 of [Van der Vaart, 1998] the set of functions $U_n(W),\ W \in \Theta_s$ is Glivenko-Cantelli.

Finally, the theorem 5.7 of [Van der Vaart, 1998], show that $\hat{W}_n$ converge in probability to W^0 ■.

2.2 Asymptotic normality

For this purpose we have to compute the first and the second derivative with respect to the parameters of $U_n(W)$. First, we introduce a notation : if $F_W(X)$ is a d-dimensional parametric function depending of a parameter W, write $\frac{\partial F_W(X)}{\partial W_k}$ (resp. $\frac{\partial^2 F_W(X)}{\partial W_k \partial W_l}$) for the d-dimensional vector of partial derivative (resp. second order partial derivatives) of each component of $F_W(X)$.

First derivatives : if $\Gamma_n(W)$ is a matrix depending of the parameter vector W, we get from [Magnus and Neudecker, 1988]

$$\frac{\partial}{\partial W_k}\ln\det\left(\Gamma_n(W)\right) = tr\left(\Gamma_n^{-1}(W)\frac{\partial}{\partial W_k}\Gamma_n(W)\right)$$

Hence, if we note

$$A_n(W_k) = \frac{1}{n}\sum_{t=1}^{n}\left(-\frac{\partial F_W(z_t)}{\partial W_k}(y_t - F_W(z_t))^T\right)$$

using the fact

$$tr\left(\Gamma_n^{-1}(W)A_n(W_k)\right) = tr\left(A_n^T(W_k)\Gamma_n^{-1}(W)\right) = tr\left(\Gamma_n^{-1}(W)A_n^T(W_k)\right)$$

we get

$$\frac{\partial}{\partial W_k}\ln\det\left(\Gamma_n(W)\right) = 2tr\left(\Gamma_n^{-1}(W)A_n(W_k)\right) \tag{3}$$

Second derivatives : We write now

$$B_n(W_k, W_l) := \frac{1}{n}\sum_{t=1}^{n}\left(\frac{\partial F_W(z_t)}{\partial W_k}\frac{\partial F_W(z_t)}{\partial W_l}^T\right)$$

and

$$C_n(W_k, W_l) := \frac{1}{n}\sum_{t=1}^{n}\left(-(y_t - F_W(z_t))\frac{\partial^2 F_W(z_t)}{\partial W_k \partial W_l}^T\right)$$

We get

$$\begin{array}{l}\frac{\partial^2 U_n(W)}{\partial W_k \partial W_l} = \frac{\partial}{\partial W_l} 2tr\left(\Gamma_n^{-1}(W)A_n(W_k)\right) = \\ 2tr\left(\frac{\partial \Gamma_n^{-1}(W)}{\partial W_l}A_n(W_k)\right) + 2tr\left(\Gamma_n^{-1}(W)B_n(W_k, W_l)\right) + 2tr\left(\Gamma_n(W)^{-1}C_n(W_k, W_l)\right)\end{array}$$

Now, [Magnus and Neudecker, 1988], give an analytic form of the derivative of an inverse matrix, so we get

$$\begin{array}{l}\frac{\partial^2 U_n(W)}{\partial W_k \partial W_l} = 2tr\left(\Gamma_n^{-1}(W)\left(A_n(W_k) + A_n^T(W_k)\right)\Gamma_n^{-1}(W)A_n(W_k)\right) + \\ 2tr\left(\Gamma_n^{-1}(W)B_n(W_k, W_l)\right) + 2tr\left(\Gamma_n^{-1}(W)C_n(W_k, W_l)\right)\end{array}$$

so

$$\begin{array}{l}\frac{\partial^2 U_n(W)}{\partial W_k \partial W_l} = 4tr\left(\Gamma_n^{-1}(W)A_n(W_k)\Gamma_n^{-1}(W)A_n(W_k)\right) \\ +2tr\left(\Gamma_n^{-1}(W)B_n(W_k, W_l)\right) + 2tr\left(\Gamma_n^{-1}(W)C_n(W_k, W_l)\right)\end{array} \quad (4)$$

Asymptotic distribution of $\hat{W}_n$ *:* The previous equations allow us to give the asymptotic properties of the estimator minimizing the cost function $U_n(W)$, namely from equation (3) and (4) we can compute the asymptotic properties of the first and the second derivatives of $U_n(W)$. If the variable Z has a moment of order at least 3 then we get the following lemma :

Theorem 2 *Assume that* $E\left(\|\varepsilon\|^2\right) < \infty$ *and* $E\left(\|Z\|^3\right) < \infty$, *let* $\Delta U_n(W^0)$ *be the gradient vector of* $U_n(W)$ *at* W^0 *and* $HU_n(W^0)$ *be the Hessian matrix of* $U_n(W)$ *at* W^0.

Write finally

$$B(W_k, W_l) := \frac{\partial F_W(Z)}{\partial W_k}\frac{\partial F_W(Z)}{\partial W_l}^T$$

We get then

1. $HU_n(W^0) \overset{a.s.}{\to} 2I_0$
2. $\sqrt{n}\Delta U_n(W^0) \overset{Law}{\to} \mathcal{N}(0, 4I_0)$
3. $\sqrt{n}\left(\hat{W}_n - W^0\right) \overset{Law}{\to} \mathcal{N}(0, I_0^{-1})$

where, the component (k, l) *of the matrix* I_0 *is :*

$$tr\left(\Gamma_0^{-1}E\left(B(W_k^0, W_l^0)\right)\right)$$

proof : We can show easily that, for all $x \in \mathbb{R}^d$, we have :

$$\begin{array}{l}\|\frac{\partial F_W(Z)}{\partial W_k}\| \le Cte(1 + \|Z\|) \\ \|\frac{\partial^2 F_W(Z)}{\partial W_k \partial W_l}\| \le Cte(1 + \|Z\|^2) \\ \|\frac{\partial^2 F_W(Z)}{\partial W_k \partial W_l} - \frac{\partial^2 F_W^0(Z)}{\partial W_k \partial W_l}\| \le Cte\|W - W^0\|(1 + \|Z\|^3)\end{array}$$

Write

$$A(W_k) = \left(-\frac{\partial F_W(Z)}{\partial W_k} (Y - F_W(Z))^T \right)$$

and $U(W) := \log \det(Y - F_W(Z))$.

Note that the component (k, l) of the matrix $4I_0$ is:

$$E\left(\frac{\partial U(W^0)}{\partial W_k} \frac{\partial U(W^0)}{\partial W_l^0} \right) = E\left(2tr\left(\varGamma_0^{-1} A^T(W_k^0)\right) \times 2tr\left(\varGamma_0^{-1} A(W_l^0)\right)\right)$$

and, since the trace of the product is invariant by circular permutation,

$$\begin{aligned}
&E\left(\tfrac{\partial U(W^0)}{\partial W_k}\tfrac{\partial U(W^0)}{\partial W_l^0}\right) = \\
&4E\left(-\tfrac{\partial F_{W^0}(Z)^T}{\partial W_k}\varGamma_0^{-1}(Y - F_{W^0}(Z))(Y - F_{W^0}(Z))^T\varGamma_0^{-1}\left(-\tfrac{\partial F_{W^0}(Z))}{\partial W_l}\right)\right) \\
&= 4E\left(\tfrac{\partial F_{W^0}(Z)^T}{\partial W_k}\varGamma_0^{-1}\tfrac{\partial F_{W^0}(Z)}{\partial W_l}\right) \\
&= 4tr\left(\varGamma_0^{-1}E\left(\tfrac{\partial F_{W^0}(Z)}{\partial W_k}\tfrac{\partial F_{W^0}(Z)^T}{\partial W_l}\right)\right) \\
&= 4tr\left(\varGamma_0^{-1}E\left(B(W_k^0, W_l^0)\right)\right)
\end{aligned}$$

Now, the derivative $\frac{\partial F_W(Z)}{\partial W_k}$ is square integrable, so $\varDelta U_n(W^0)$ fulfills Lindeberg's condition (see [Hall *et al.*, 2001]) and

$$\sqrt{n}\varDelta U_n(W^0) \overset{Law}{\to} \mathcal{N}(0, 4I_0)$$

For the component (k, l) of the expectation of the Hessian matrix, remark first that

$$\lim_{n\to\infty} tr\left(\varGamma_n^{-1}(W^0)A_n(W_k^0)\varGamma_n^{-1}(W^0)A_n(W_k^0)\right) = 0$$

and

$$\lim_{n\to\infty} tr\varGamma_n^{-1}C_n(W_k^0, W_l^0) = 0$$

so

$$\begin{aligned}
&\lim_{n\to\infty} H_n(W^0) = \lim_{n\to\infty} 4tr\left(\varGamma_n^{-1}(W^0)A_n(W_k^0)\varGamma_n^{-1}(W^0)A_n(W_k^0)\right) + \\
&2tr\varGamma_n^{-1}(W^0)B_n(W_k^0, W_l^0) + 2tr\varGamma_n^{-1}C_n(W_k^0, W_l^0) = \\
&= 2tr\left(\varGamma_0^{-1}E\left(B(W_k^0, W_l^0)\right)\right)
\end{aligned}$$

Now, since $\|\frac{\partial^2 F_W(Z)}{\partial W_k \partial W_l}\| \leq Cte(1 + \|Z\|^2)$ and $\|\frac{\partial^2 F_W(Z)}{\partial W_k \partial W_l} - \frac{\partial^2 F_W^0(Z)}{\partial W_k \partial W_l}\| \leq Cte\|W - W^0\|(1 + \|Z\|^3)$, by standard arguments found, for example, in [Yao, 2000] we get

$$\sqrt{n}\left(\hat{W}_n - W^0\right) \overset{Law}{\to} \mathcal{N}(0, I_0^{-1})$$

■

2.3 Asymptotic distribution of T_n

In this section, we write $\hat{W}_n = \arg\min_{W \in \Theta_s} U_n(W)$ and $\hat{W}_n^0 = \arg\min_{W \in \Theta_q} U_n(W)$, where Θ_q is view as a subset of $\mathbb{R}^s$. The asymptotic distribution of T_n is then a consequence of the previous section, namely, if we have to replace $nU_n(W)$ by its Taylor expansion around $\hat{W}_n$ and $\hat{W}_n^0$, following [Van der Vaart, 1998] chapter 16 we have :

$$T_n = \sqrt{n}\left(\hat{W}_n - \hat{W}_n^0\right)^T I_0 \sqrt{n}\left(\hat{W}_n - \hat{W}_n^0\right) + o_P(1) \overset{\mathcal{D}}{\rightarrow} \chi^2_{s-q}$$

3 Conclusion

It has been show that, in the case of multidimensional output, the cost function $U_n(W)$ leads to a test for the number of parameters in MLP simpler than with the traditional mean square cost function. In fact the estimator $\hat{W}_n$ is also more efficient than the least square estimator (see [Rynkiewicz, 2003]). We can also remark that $U_n(W)$ matches with twice the "concentrated Gaussian log-likelihood" but we have to emphasize, that its nice asymptotic properties need only moment condition on ε and Z, so it works even if the distribution of the noise is not Gaussian. An other solution could be to use an approximation of the covariance error matrix to compute generalized least square estimator :

$$\frac{1}{n}\sum_{t=1}^{n}\left(Y_t - F_W\left(Z_t\right)\right)^T \Gamma^{-1}\left(Y_t - F_W\left(Z_t\right)\right),$$

assuming that Γ is a good approximation of the true covariance matrix of the noise $\Gamma(W^0)$. However it take time to compute a good the matrix Γ and if we try to compute the best matrix Γ with the data, it leads to the cost function $U_n(W)$ (see for example [Gallant, 1987]).

Finally, as we see in this paper, the computation of the derivatives of $U_n(W)$ is easy, so we can use the effective differential optimization techniques to estimate $\hat{W}_n$ and numerical examples can be found in [Rynkiewicz, 2003].

References

[Fukumizu, 2003]K. Fukumizu. Likelihood ratio of unidentifiable models and multilayer neural networks. *Annals of Statistics*, 31:3:533–851, 2003.

[Gallant, 1987]R.A. Gallant. *Non linear statistical models.* J. Wiley and Sons, New-York, 1987.

[Hall and Heyde, 1980]P. Hall and C. Heyde. *Martingale limit theory and its applications.* Academic Press, New-York, 1980.

[Magnus and Neudecker, 1988]Jan R. Magnus and Heinz Neudecker. *Matrix differential calculus with applications in statistics and econometrics.* J. Wiley and Sons, New-York, 1988.

[Rynkiewicz, 2003]J. Rynkiewicz. Estimation of multidimensional regression model with multilayer perceptrons. In J. Mira and J.R. Alvarez, editors, *Computational methods in neural modeling*, volume 2686 of *Lectures notes in computer science*, pages 310–317, 2003.

[Sussman, 1992]H.J. Sussman. Uniqueness of the weights for minimal feedforward nets with a given input-output map. *Neural Networks*, pages 589–593, 1992.

[Van der Vaart, 1998]A.W. Van der Vaart. *Asymptotic statistics.* Cambridge University Press, Cambridge, UK, 1998.

[Yao, 2000]J. Yao. On least square estimation for stable nonlinear ar processes. *The Annals of Institut of Mathematical Statistics*, 52:316–331, 2000.

Independent Factor Discriminant Analysis

Angela Montanari, Daniela Giovanna Caló, and Cinzia Viroli

Statistics Department
University of Bologna,
via Belle Arti 41, 40126, Bologna, Italy
(e-mail: montanari@stat.unibo.it, calo@stat.unibo.it,
viroli@stat.unibo.it)

Abstract. In the general classification context the recourse to the so-called Bayes decision rule requires to estimate the class conditional probability density functions. In this paper we propose a mixture model for the observed variables which is derived by assuming that the data have been generated by an independent factor model. Independent factor analysis is in fact a generative latent variable model whose structure closely resembles the one of ordinary factor model but it assumes that the latent variables are mutually independent and not necessarily Gaussian. The method therefore provides a dimension reduction together with a semiparametric estimate of the class conditional probability density functions. This density approximation is plugged into the classic Bayes rule and its performance is evaluated both on real and simulated data.

Keywords: Classification, Independent Factor Analysis, Mixture Models.

1 Introduction

In the general classification context the goal is to define a rule for the assignment of one new unit, on which a p-variate vector of variables $\mathbf{X}$ has been observed, to the class, out of G unordered ones, from which it comes. The training sample on which the rule is built consists of an indication of the class membership and of the p predictors for a set of n units. Denoted by f_g, with $g = 1, \ldots, G$, the class conditional densities and by π_g the *a priori* probability of observing an individual from population g, the so-called Bayes decision rule suggests to allocate $\mathbf{x}$ to the population $\hat{g}$ such that

$$\hat{g} = \arg\max_{g=1,\ldots,G} \{f_g(\mathbf{x})\pi_g\} \tag{1}$$

If the class conditional densities are Gaussian, the expression (1) simply yields the well known linear or quadratic discriminant functions according to whether the condition of homoscedasticity is fulfilled or not. But in most applications neither $f_g(x)$ nor π_g $(g = 1, \ldots, G)$ are known and the recourse to the Gaussian based approach may be strongly misleading.

When the training sample data may be considered as a random sample from the pooled population, the prior probabilities may be easily estimated by the relative frequencies of the g classes in the sample $\hat{\pi}_g = n_g/n$ where n_g is the number of units from class g observed in the training sample. The estimation of the unknown densities f_g is on the contrary a more complex task.

The solution most often used in the statistical literature is based on kernel density estimation ([Hand, 1982] and [Silverman, 1986]) and on the use of the estimated densities in the classification rule (1), which therefore becomes a nonparametric one. It is well known however that kernel methods deeply suffer from the curse of dimensionality when applied in the multidimensional context (as the one we are dealing with is). They also tend to produce poor estimates of the density tails, whose role may on the contrary be crucial for classification purposes. Amato *et al.* [Amato *et al.*, 2002] suggest to overcome the problem by transforming the data into independent components [Comon, 1994]. Exploiting the independence condition they rephrase the multivariate density estimation task as a sequence of univariate ones, *i.e.* the estimation of the marginal densities, whose product yields the multivariate density in the transformed space. This density is then back-transformed in order to obtain an estimate of the probability density function of the observed variables. The method seems to outperform linear, quadratic and flexible discriminant analysis in the training set, but its performance is quite poor in the test one.

Other approaches to nonparametric density estimation for classification, such as the one due to Polzehl [Polzehl, 1995], who suggests a discrimination oriented version of projection pursuit density estimation, seem to produce quite good results but at a high computational cost and many aspects, at least from an algorithmic point of view, still need improvement (for instance, the selection of the bandwidth parameters in univariate kernel density estimation, which should be optimal from a classification perspective).

A more recent approach is based on mixture models. In particular, in [McLachlan and Peel, 2000] each class conditional density, f_g, is modeled as a mixture of m_g normally distributed components (Gaussian Mixture Model, GMM):

$$\hat{f}_g(\mathbf{x}) = \sum_{l=1}^{m_g} w_{gl}\phi(\mathbf{x}; \boldsymbol{\mu}_{gl}, \boldsymbol{\Sigma}_{gl}) \qquad (2)$$

where $\phi(\mathbf{x}, \boldsymbol{\mu}_{gl}, \boldsymbol{\Sigma}_{gl})$ denotes the p-variate normal density function with vector mean $\boldsymbol{\mu}_{gl}$ and covariance matrix $\boldsymbol{\Sigma}_{gl}$ $(l = 1, \ldots, m_g)$, and w_{gl} are the mixing proportions. The density estimation involves therefore the estimation of $\boldsymbol{\mu}_{gl}$, $\boldsymbol{\Sigma}_{gl}$ and w_{gl} for $l = 1, \ldots, m_g$ and $g = 1, \ldots, G$; this is a quite large number of parameters as it is

$$h_g^{GMM} = m_g p + m_g \frac{p(p+1)}{2} + (m_g - 1), \qquad \text{for } g = 1, \ldots, G$$

which may be difficult to estimate for relatively small sample sizes.

Hastie and Tibshirani [Hastie and Tibshirani, 1996] introduce what they call mixture discriminant analysis (MDA) which exploits Gaussian mixtures for classification purposes by imposing some constraints which make estimation and interpretation easier. A completely different solution, aimed at reducing the number of free parameters in a mixture model, is due to McLachlan *et al.* [McLachlan *et al.*, 2002] who propose to assume a factor model for each mixture component, thus modeling the density as a mixture of factor analyzers.

In this paper we derive an approach for modeling class conditional densities which combines the potentialities of the independence condition in a low dimensional latent space (in the spirit of Amato *et al.*) with the semi-parametric structure of mixture models. The method which simultaneously allows to address both aspects is Independent Factor Analysis [Attias, 1999].

2 Independent Factor Analysis

Independent Factor Analysis has been recently introduced by Attias (1999) in the context of signal processing and only recently it has been given a solid statistical foundation [Montanari and Viroli, 2004]. Its aim is to describe p observed variables x_j, which are generally correlated, in terms of a smaller set of k unobserved independent latent variables y_i and an additive specific term u_j:

$$x_j = \sum_{i=1}^{k} \lambda_{ji} y_i + u_j,$$

where $j = 1, ..., p$, $i = 1, ..., k$. In a more compact form the model is

$$\mathbf{x} = \Lambda \mathbf{y} + \mathbf{u} \tag{3}$$

where the factor loading matrix $\Lambda = \{\lambda_{ji}\}$ is also termed as *mixing matrix.* Its structure closely resembles the classical factor model but it differs from it as far as the properties of the latent variables it involves is concerned. The random vector $\mathbf{u}$ representing the noise is assumed to be normally distributed, $\mathbf{u} \sim \mathcal{N}(\mathbf{0}, \Psi)$ with Ψ allowing for correlations between the error terms. The latent variables $\mathbf{y}$ are assumed to be mutually independent and not necessarily normally distributed; their densities are indeed modeled as mixtures of Gaussians. The independence assumption allows to model the density of each y_i in the latent space separately. In more formal terms each factor is thus described as a mixture of m_i gaussians with mean $\mu_{i,q}$, variance $\nu_{i,q}$ and mixing proportions $w_{i,q}$ $(q = 1, ..., m_i)$:

$$f(y_i) = \sum_{q=1}^{m_i} w_{i,q} \phi\left(y_i; \mu_{i,q}, \nu_{i,q}\right) \tag{4}$$

The mixing proportions $w_{i,q}$ are constrained to be non-negative and sum to unity.

A particular characterization of the IFA model is that it involves two layers of latent variables: besides the factors, $\mathbf{y}$, an *allocation variable*, $\mathbf{z}$, must be introduced, as always when dealing with mixture models. With reference to a particular factor i, the mixture can be thought of as the density of an heterogeneous population consisting of m_i subgroups. For each observation the allocation variable denotes the identity of the subgroup from which it is drawn. In the k-dimensional space, the multivariate allocation variable, $\mathbf{z}$, follows a multivariate multinomial distribution. The density of the observed data can be constructed by conditioning to these two latent layers:

$$\begin{aligned} f\left(\mathbf{x}|\Theta\right) &= \sum_{\mathbf{z}} \int f(\mathbf{x}\text{,}\mathbf{y}\text{,}\mathbf{z}|\Theta) d\mathbf{y} \\ &= \sum_{\mathbf{z}} \int f(\mathbf{z}|\Theta) f(\mathbf{y}|\mathbf{z},\Theta) f(\mathbf{x}|\mathbf{y},\mathbf{z},\Theta) d\mathbf{y} \\ &= \sum_{\mathbf{z}} f(\mathbf{z}|\Theta) f(\mathbf{x}|\mathbf{z},\Theta) \end{aligned} \tag{5}$$

where Θ denotes the whole set of the IFA model parameters.

It is not difficult to derive that the conditional density $f(\mathbf{x}|\mathbf{z},\Theta)$ follows a Gaussian distribution since it is the convolution of two Gaussian densities: .

$$\mathbf{x}|\mathbf{y},\mathbf{z} \sim \mathcal{N}(\Lambda\mathbf{y},\Psi) \tag{6}$$

and

$$\mathbf{y}|\mathbf{z} \sim \mathcal{N}(\boldsymbol{\mu}_{\mathbf{z}},\mathbf{V}_{\mathbf{z}}) \tag{7}$$

where $\boldsymbol{\mu}_{\mathbf{z}}$ and $\mathbf{V}_{\mathbf{z}}$ are respectively defined as:

$$\boldsymbol{\mu}_{\mathbf{z}} = \left[\prod_{q=1}^{m_1} \mu_{1,q}^{z_{1,q}}, ..., \prod_{q=1}^{m_k} \mu_{k,q}^{z_{k,q}}\right] \qquad \mathbf{V}_{\mathbf{z}} = \operatorname{diag}\left[\prod_{q=1}^{m_1} \nu_{1,q}^{z_{1,q}}, ..., \prod_{q=1}^{m_k} \nu_{k,q}^{z_{k,q}}\right].$$

For more details see [Montanari and Viroli, 2004].

Therefore the expression (5) indicates that the density of the observed data given the IFA model, *i.e.* the likelihood function $f(\mathbf{x}|\Theta)$, is a finite mixture of p-variate normals. Its generic component is given by

$$f(\mathbf{x}|\mathbf{z},\Theta) = \phi\left(\mathbf{x}|\mathbf{z};\Lambda\boldsymbol{\mu}_{\mathbf{z}},\Lambda\mathbf{V}_{\mathbf{z}}\Lambda^T + \Psi\right) \tag{8}$$

Implicit in the IFA estimation problem (which is solved by the EM-algorithm) are the two assumptions regarding the correct number of factors and the number of mixture components for modeling each factor. Assessing the correct specification of the model is an important but as jet unsolved problem; this issue has been addressed in [Montanari and Viroli, 2004]. Once the number of factors, k, and the number of components for each of them, m_i, have been correctly specified, the total number of the IFA model parameters for f_g is

$$h_g^{IFA} = pk + \frac{p(p+1)}{2} + 3\sum_{i=1}^{k} m_i - k \qquad g = 1, \ldots, G.$$

As a consequence of this *a priori* double choice, the number of the mixture components in (5) is univocally determined as $m_g = \prod_{i=1}^{k} m_i$.

An advantage of this approach is that by applying mixture models not directly to the observed variables but onto the reduced latent space the density of the observed variables is still a Gaussian mixture model that generally involves a smaller set of parameters. For instance, a mixture model for the class conditional density f_g with $m_g = 4$ components can be obtained by estimating $k = 2$ factors with $m_i = 2$ components each. The resulting number of parameters

$$h_g^{IFA} = \frac{p^2}{2} + \frac{5}{2}p + 10$$

is smaller than the one to be estimated in (2)

$$h_g^{GMM} = 2p^2 + 6p + 3$$

for $p \geq 2$.

A further appealing feature of the proposed solution is that the formulation given by (4) and (5) does not rely on any constraints on the parameters, allowing for a very flexible density approximation.

3 Empirical results

3.1 Simulated data

The discrimination performance of Independent Factor Discriminant Analysis (IFDA) has been tested on the popular waveform data. This example has been taken from [Breiman *et al.*, 1984] and subsequently used in many works on classification, since it is considered a difficult pattern recognition problem. It is a three class problem with 21 variables, which are defined by

$$\begin{aligned} x_i &= uh_1(i) + (1-u)h_2(i) + \varepsilon_i \quad \text{Class 1} \\ x_i &= uh_1(i) + (1-u)h_3(i) + \varepsilon_i \quad \text{Class 2} \\ x_i &= uh_2(i) + (1-u)h_3(i) + \varepsilon_i \quad \text{Class 3} \end{aligned}$$

where $i = 1, \ldots, 21$, u is uniform on [0,1], ε_i are standard normal random variables and h_1, h_1 and h_3 are the following shifted triangular forms:

$h_1(i) = \max(6 - |i - 11|, 0)$, $h_2(i) = h_1(i - 4)$ and $h_3(i) = h_1(i + 4)$. The method discussed here is compared with the following classification procedures: linear discriminant analysis (LDA), quadratic discriminant analysis (QDA), mixture discriminant analysis (MDA), flexible discriminant analysis (FDA), penalized discriminant analysis (PDA) and the CART procedure. The training sample consists of 300 observations and the test sample has size 500. Both of them have been generated with equal priors.

Technique	Error rates	
	Training	Test
LDA	0.121(.006)	0.191(.006)
QDA	0.039(.004)	0.205(.006)
CART	0.072(.003)	0.289(.004)
FDA/MARS (degree=1)	0.100(.006)	0.191(.006)
FDA/MARS (degree=2)	0.068(.004)	0.215(.002)
MDA (3 subclasses)	0.087(.005)	0.169(.006)
MDA (3 subclasses, penalized 4df)	0.137(.006)	0.157(.005)
PDA (penalized 4df)	0.150(.005)	0.171(.005)
IFDA (2 factors)	0.054(.010)	0.133(.004)

Table 1. Results for waveform data. The values are averages over 10 simulations, with the standard error of the average in parentheses. The eight entries above the line are taken from Hastie and Tibshirani (1996). The last line indicates the error rates in the IFDA with 2 components for each factor.

Table 1 indicates the classification results taken from Hastie and Tibshirani [Hastie and Tibshirani, 1996] and includes the performances of IFDA over 10 simulations. Independent Factor Discriminant Analysis shows the lowest classification error rate in the test samples.

3.2 Real data

We applied the proposed method on the thyroid data [Coomans *et al.*, 1983]. The example consists of 5 measurements (T3-resin uptake test, Total Serum thyroxin, Total serum triiodothyronine, Basal thyroid-stimulating hormone and maximal absolute difference of TSH value after injection of 200 micro grams of thyrotropin-releasing hormone as compared to the basal value) on 215 patients, that are distinguished in three groups on the basis of their thyroid status (normal, hyper and hypo). The data have been randomly divided into a training sample of size 143 and a test sample that consists of the remaining patients. Table 2 shows a summary of the performance of several classification procedures. In order to compare our results with those published in a technical report which represents an extended version of [Hastie *et al.*, 1994], only one split into training and test set has been considered.

Independent Factor Discriminant Analysis performs very well and it is competitive with respect to non linear methods such as neural networks and the MDA/FDA procedure.

Technique	Error rates	
	Training	Test
LDA	0.091	0.083
MDA	0.028	0.042
MDA/FDA	0.049	0.014
FDA	0.049	0.042
Neural network (10 hidden units)	0.000	0.027
IFDA (2 factors)	0.056	0.027

Table 2. Results for Thyroid data. The first five lines are taken from an extended version (technical report) of the paper by Hastie and Tibshirani (1996). The last entry indicates the error rates in the IFDA with 2 components for each factor.

4 Conclusion

In this paper we have proposed a new approach to classification by Gaussian mixtures. Its main assumption is that the observed data have been generated by an independent factor model. In this way we obtain a very flexible density approximation, which, for a given number of mixture components, is often based on a lesser number of parameters than the classic mixture model solution and allows for heteroscedastic components. Its performance seems to be very competitive with respect to the main classification procedures proposed in the statistical literature.

References

[Amato *et al.*, 2002]U. Amato, A. Antoniadis, and Gréfoire G. Independent component discriminant analysis. *International Mathematical Journal*, pages 735–753, 2002.

[Attias, 1999]H. Attias. Independent factor analysis. *Neural Computation*, pages 803–851, 1999.

[Breiman *et al.*, 1984]L. Breiman, J. Friedman, R. Olshen, and C. Stone. *Classification and Regression Trees.* Wadsworth, Belmont, California, 1984.

[Comon, 1994]P. Comon. Independent component analysis, a new concept?. *Signal Processing*, pages 287–314, 1994.

[Coomans *et al.*, 1983]D. Coomans, M. Broeckaert, and D.L. Broeckaert. Comparison of multivariate discriminant techniques for clinical data - application to the tyroid functional state. *Meth. Inform. Med.*, pages 93–101, 1983.

[Hand, 1982]D.J. Hand. *Kernel Discriminant Analysis.* Research Studies Press, Letchworth, 1982.

[Hastie and Tibshirani, 1996]T. Hastie and R. Tibshirani. Discriminant analysis by gaussian mixtures. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 155–176, 1996.

[Hastie *et al.*, 1994]T. Hastie, R. Tibshirani, and A. Buja. Flexible discriminant analysis by optimal scoring. *Journal of the American Statistical Association*, pages 1255–1270, 1994.

[McLachlan and Peel, 2000]G.J. McLachlan and D. Peel. *Finite Mixture Models.* Wiley, New York, 2000.

[McLachlan *et al.*, 2002]G.J. McLachlan, R.W. Bean, and D. Peel. A mixture model-based approach to the clustering of microarray expression data. *Bioinformatics*, pages 413–422, 2002.

[Montanari and Viroli, 2004]A. Montanari and C. Viroli. The independent factor analysis approach to latent variable modeling. *Submitted*, 2004.

[Polzehl, 1995]J. Polzehl. Projection pursuit discriminant analysis. *Computational Statistics and Data Analysis*, pages 141–157, 1995.

[Silverman, 1986]B.W. Silverman. *Density Estimation for Statistics and Data Analysis.* Chapman and Hall, London, 1986.

Measuring Distance from a Training Data Set

Ilmari Juutilainen and Juha Röning

Computer Engineering Laboratory
PO BOX 4500
90014 University of Oulu, Finland
(e-mail: ilmari.juutilainen@ee.oulu.fi, juha.roning@ee.oulu.fi)

Abstract. In this paper, a new method is proposed for measuring the distance between a training data set and a single, new observation. The novel distance measure reflects the expected squared prediction error, when the prediction is based on the k nearest neighbours of the training data set. The simulation shows that the distance measure correlates well with the true expected squared prediction error in practice. The distance measure can be applied, for example, to assessing the uncertainty of prediction.

Keywords: Distance measure, Model uncertainty,
Distance weighted k-nearest-neighbour, Novelty detection.

1 Introduction

In some applications, such as in evaluation of the reliability of prediction at a query point, it is interesting to measure the information given by the training data set about a new observation via the current prediction model. In this work, we propose a novel measure for the distance between a single observation and a data set. The distance measure reflects the expected uncertainty of the new observation being predicted based on the data set. The distance measure is a linear function of the approximated expected squared prediction error, when the new observation is predicted with the distance weighted k-nearest-neighbour method.

There has been much discussion about measuring the distance between two observations. We refer to a review paper [Wettschereck *et al.*, 1997] that discusses the different methods. Often, Euclidean distance or Manhattan distance is used, and the problem lies in the weighting or scaling of the variables. The input variables that have a large effect on the response should have large weights in the distance measure. Global distance measures use constant weights, unlike local distance measures. Some distance measures take the correlations between the explanatory variables into account.

The measurement of the distance between a set of observations and a single observation has also been widely discussed. Different distance measures have been applied in clustering and in prototype methods. In these applications, the aim in defining the distance has been to assign the observation to the nearest cluster or prototype. Examples of the different methods include the average pairwise distance, the Mahalanobis distance and the Euclidian distance to the cluster centroid. We refer to [Kaufman and Rousseeuw, 1990] for these methods. However, these methods have been planned to measure the distance between a cluster and a single observation and not the distance between a data set and a single observation.

Novelty detection aims to find abnormal observations from a data set. Abnormal observations can indicate that the modelled system is in an abnormal state, which needs to be reported. In classification, detection of novel observations is needed to identify new classes and observations that cannot be classified reliably. Novelty detection can be used to differentiate novel information from existing information when only the novel information needs to be shown to the learners. For novelty detection methods, we refer to the review [Markou and Singh, 2003].

The usual approach in novelty detection is to measure somehow the similarity with the training data and to use some threshold to interpret the observations as novel. The most common method is to model the joint density function of input variables to judge the observations with low density as novel [Markou and Singh, 2003]. Our approach differs in that we do not construct any distribution model for the inputs. Our distance measure tries to measure the uncertainty about the expected response value at a new query point, which is quite a novel approach to the problem. The standard errors of predictions measure the uncertainty with variance, but we take also bias into account.

[Angiulli and Pittuzi, 2005] suggested a method for detecting outliers in a data set. They calculated the sum of the Euclidean distances to the k nearest neighbours for measuring the distance, which approach is quite similar to our proposal. [Mahamud and Hebert, 2003] discussed the optimal distance measures in k-nearest-neighbour prediction, and we constructed our distance measure using a similar optimality principle.

2 Distance between two single observations

Let $x_{(j)}$ refer to the jth explanatory variable and x_{ij} denote the ith observation of $x_{(j)}$, y_i denote the ith observation of the response and T denote the training data set consisting of N observations $(x_1, y_1), (x_2, y_2), \ldots, (x_N, y_N)$. Let (x_0, y_0) be a new test data observation and $d_i = d(x_0, x_i)$ measure the distance between x_0 and $x_i \in T$. We assume that the response depends on the inputs via a regression function $f(\cdot)$, and that the additive error term has a constant variance

$$y_i = f(x_i) + \varepsilon_i,\ E(\varepsilon_i) = 0,\ \mathrm{Var}\ (\varepsilon_i) = \sigma^2. \tag{1}$$

[Mahamud and Hebert, 2003] discussed the optimal distance measures in nearest-neighbour classification. The optimal distance measure in 1-nearest-neighbour prediction minimises the expected loss function $E_{y_0,x_0,T}L(y_0, y')$, where y' is the measured response at x', which is the nearest neighbour of x_0

using the distance measure d. The distance measure $d(x_0, x_i) = EL(y_0, y_i)$ is optimal, because the nearest neighbour $x' = \arg\min_{x_i} EL(y_0, y_i)$ minimises the expected loss $L(y_0, y') \; \forall x_0 \; \forall T$ [Mahamud and Hebert, 2003]. The same reasoning holds for k-nearest-neighbour prediction. All order-preserving transformations of the expected loss function are optimal, because the nearest neighbours remain the same. We use the expected squared error loss $EL(\mu_0, y_i) = E(\mu_0 - y_i)^2 = E(y_0 - y_i)^2 - \sigma^2$ related to the true expectation $\mu_0 = E(y|x_0) = f(x_0)$ without losing optimality.

The optimal distance measure cannot be used directly because the conditional expectation of the response is not known, and the true expected loss cannot be solved. The optimal distance measure is not monotonic, which implies an interpretational disadvantage: The nearest neighbours may lie far away from the query point on the scale of explanatory variables. To eliminate this problems, we must be content with a coarse approximation of the expected loss: We use the sum of the expectations of squared differences in the true regression function, when one input variable at a time is set to the measured values x_0 and x_i, and other input variables are drawn randomly,

$$E(\mu_0 - y_i)^2 = \sigma^2 + [f(x_0) - f(x_i)]^2 \approx \sigma^2 + \sum_{j=1}^{p} E_x\big\{f[w_i^{(j)}(x)] - f[w_0^{(j)}(x)]\big\}^2. \tag{2}$$

In the formula, x is a randomly drawn input observation, $w_0^{(j)}(x)$ is otherwise identical with x but the jth element is altered $w_{0j}^{(j)} = x_{0j}$, and $w_i^{(j)}(x)$ is otherwise identical with x but the jth element $w_{ij}^{(j)} = x_{ij}$.

In the case of continuous input variables, we further approximate the squared differences in y with squared differences in the input variable values $E_x\big\{f[w_i^{(j)}(x)] - f[w_0^{(j)}(x)]\big\}^2 \approx \alpha_j (x_{0j} - x_{ij})^2$. [Mahamud and Hebert, 2003] proposed to estimate the α-coefficients by fitting a regression model to a data set of pairs of training data instances using the response $L(y_i, y_j)$. The advantage of their direct method is that the regression function need not be estimated. We propose a different method. Let our prediction model be $\widehat{y} = \widehat{f}(x) = \widehat{f}(x_{(1)}, x_{(2)}, \ldots, x_{(p)})$, and let $\widehat{\sigma}^2$ be the corresponding error variance estimate. Let now $x_c \in T$ denote a training data observation lying near x_0, and let $\widehat{f}'(x_c) = \Big(\frac{\partial \widehat{f}(x)}{\partial x_{(1)}}, \frac{\partial \widehat{f}(x)}{\partial x_{(2)}}, \ldots, \frac{\partial \widehat{f}(x)}{\partial x_{(p)}}\Big)_{(x=x_c)}$ denote the gradient of the fitted response surface at point x_c. Motivated by the first-order Taylor approximation around x_c, we suggest that $\alpha_1, \ldots, \alpha_p$ are defined as the average squared partial derivative over the training data set

$$\alpha_j = \frac{1}{N} \sum_{i=1}^{N} \Big(\frac{\partial \widehat{f}(x)}{\partial x_{(j)}}_{\,(x=x_i)} \Big)^2. \tag{3}$$

For large N, it is enough to calculate the average over a sample. The regression function can be fitted using any learning method, for example, neural networks or additive models. The partial derivatives of the fitted response surface can be approximated numerically with $\frac{\partial \widehat{f}(x)}{\partial x_{(j)}}{}_{(x=x_i)} = \frac{\widehat{f}(x_i)-\widehat{f}(x_i+o_j)}{|o_j|}$, where o_j is a vector of zeros elsewhere but a small constant at the jth element.

When $x_{(j)}$ is a categorical variable with class levels $\gamma_{j1}, \gamma_{j2}, \ldots, \gamma_{jq_j}$, we can estimate the expected squared difference $E_x\big\{f[w_i^{(j)}(x)]-f[w_0^{(j)}(x)]\big\}^2$ between each two class levels γ_{jl} and γ_{jm} using the fitted prediction model with $\frac{1}{|J|}\sum_{j\in J}\left(\widehat{f}(x_i)-\widehat{f}(w_i^{(j)})\right)^2$. The input vectors $w_i^{(j)}$ are otherwise identical to x_i but the jth element is altered: $w_{ij}^{(j)} = \gamma_{jm}$, if $x_{ij} = \gamma_{jl}$, and $w_{ij}^{(j)} = \gamma_{jl}$, if $x_{ij} = \gamma_{jm}$. The squared differences in the prediction are averaged over the index set $J = \left\{i|\widehat{f}(x_i), \widehat{f}(w_i^{(j)}) \text{ are reliable and } (x_{ij} = \gamma_{jl} \text{ or } x_{ij} = \gamma_{jm})\right\}$. For binary variables we can notate

$$\alpha_j = \frac{1}{|J|}\sum_{j\in J}\left(\widehat{f}(x_i)-\widehat{f}(w_i^{(j)})\right)^2. \tag{4}$$

We propose to use an approximate optimal distance measure that is the approximated expected squared error loss

$$d(x_0, x_i) = \alpha_0 + \sum_{j=1}^{p}\alpha_j(x_{0j}-x_{ij})^2. \tag{5}$$

The coefficient α_0 is the error variance estimate $\widehat{\sigma}^2$. The notation (Eq. 5) is applicable for continuous and binary variables, but categorical variables can be taken into account as explained previously.

3 Distance between a single observation and a data set

We suggest that the distance of a single observation from a set of k observations, S_k, is measured on the basis of the expected squared error when the single observation is predicted based on S_k. This can be seen as the generalisation of the pairwise optimal distance measure. The true expected loss at x_0 is not known and has to be approximated. We predict $\mu_0 = E(y_0)$ with a distance-weighted linear combination of the y values measured in S_k, which results in measurement of the distance with the harmonic sum of pairwise distances.

Let $S_k = (x_1, x_2, \ldots, x_k)$ with the distances $d_1, d_2, \ldots, d_k$ from x_0, and let each distance be known $d_i = E(\mu_0 - y_i)^2$. Let us now estimate μ_0 with a weighted linear combination $\widehat{y}_0 = \omega_1 y_1 + \omega_2 y_2 + \cdots + \omega_k y_k$. Under the symmetry assumption $E(\mu_0 - y_i) = 0$, the minimum variance unbiased estimator gives weights proportional to the inverses of the variances and sums the weights to unity $\omega_j = \frac{1}{d_j}\Big/\sum_{i=1}^k \frac{1}{d_i}$. We use this distance-weighted estimator

$$\widehat{y}_0 = \Big(\sum_{i=1}^k \frac{1}{d_i}\, y_j\Big)\Big/\sum_{i=1}^k \frac{1}{d_i} \tag{6}$$

to predict y_0 based on S_k. We keep the estimator (Eq. 6) as a natural basis for the interpretation of our distance measure because the approach does not make any assumption about the form of the regression function. The expected squared loss of our estimator is the harmonic sum of pairwise distances d_i plus a bias term

$$\begin{aligned}
E(\widehat{y}_0 - \mu_0)^2 &= E\Big(\sum_{i=1}^k(\omega_i y_i) - \mu_0\Big)^2 = E\Big(\sum_{i=1}^k \omega_i(y_i - \mu_0)\Big)^2 \\
&= \sum_{i=1}^k \omega_i^2 E(y_i - \mu_0)^2 + 2\sum_{j=1}^k\sum_{i\neq j} E(y_i-\mu_0)E(y_j-\mu_0)\omega_i\omega_j \\
&= \Big(\frac{1}{\sum_{i=1}^k \frac{1}{d_i}}\Big)^2\Big[\sum_{i=1}^k d_i/d_i^2 + 2\sum\sum_{i\neq j}\frac{E(y_i-\mu_0)E(y_j-\mu_0)}{d_i d_j}\Big] \\
&= \frac{1}{\sum_{i=1}^k \frac{1}{d_i}} + \Big(\frac{1}{\sum_{i=1}^k \frac{1}{d_i}}\Big)^2 2\sum\sum_{i\neq j}\frac{E(y_i-\mu_0)E(y_j-\mu_0)}{d_i d_j}. \qquad (7)
\end{aligned}$$

We take the expectations (Eq. 7, 8, 9 and 10) over x_i, also, which means that x_i are assumed to be random points satisfying the condition $d(x_0, x_i) = d_i$. If the assumption $E_{Y,x_i|d_i}(\mu_0 - y_i) = 0\,\forall i$ holds, the bias term would be zero, and the expected squared error would be the harmonic sum of the pairwise distances. However, that is not a realistic assumption. Some query points x_0 may lie in a 'symmetric' position where the assumption holds. But some query points may lie at the bottom of a valley or on the top of a hill, where the expectation $E(y_i - \mu_0)$ is negative for all possible neighbours x_i.

Let us now us examine the bottom of a valley scenario in more detail. Because $E\ (y_i - \mu_0)^2 = d_i$ and Var $y_i = \sigma^2$, it holds that $E\ (y_i - \mu_0) = \sqrt{d_i - \sigma^2}$. Let $\bar{d}$ denote the average inverse distance $\frac{1}{k}\sum_{i=1}^k 1/d_i$. We can derive an upper bound for the bias term

$$\begin{aligned}
2\sum\sum_{i\neq j}\frac{E\ (y_i-\mu_0)E\ (y_j-\mu_0)}{d_i d_j} &= 2\sum\sum_{i\neq j}\frac{\sqrt{d_i-\sigma^2}\sqrt{d_j-\sigma^2}}{d_i d_j} \\
= \sum_{i=1}^k\Big[\frac{\sqrt{d_i-\sigma^2}}{d_i}\sum_{j\neq i}\frac{\sqrt{d_j-\sigma^2}}{d_j}\Big] &<= \sum_{i=1}^k \bar{d}\sqrt{\frac{1}{\bar{d}}-\sigma^2}\sum_{j\neq i}\bar{d}\sqrt{\frac{1}{\bar{d}}-\sigma^2} \\
= k(k-1)\bar{d}^2(\frac{1}{\bar{d}}-\sigma^2) &= (k-1)\sum_{i=1}^k\frac{1}{d_i} - \sigma^2\frac{k-1}{k}\Big[\sum_{i=1}^k\frac{1}{d_i}\Big]^2 (8)
\end{aligned}$$

The result follows from the Jensen inequality and concavity of the function $q(x) = x\sqrt{1/x - \sigma^2}$. Equality is achieved if the distances to all the neighbours are constant $1/d_i = \bar{d}\,\forall i$. When all the k neighbours are roughly equally distant, and x_0 lies at the bottom of a valley or on the top of a hill, the bias term can be approximated as a linear function of the harmonic sum $1/\sum_{i=1}^{k}\frac{1}{d_i}$ and k

$$\left(\frac{1}{\sum_{i=1}^{k}\frac{1}{d_i}}\right)^2 2\sum_{j=1}^{k}\sum_{i\neq j}\frac{E(y_i-\mu_0)E(y_j-\mu_0)}{d_i d_j}\approx\frac{k-1}{\sum_{i=1}^{k}\frac{1}{d_i}}-\sigma^2\frac{k-1}{k}. \qquad (9)$$

Simulation studies showed that this approximation holds well in practice: In all of the simulated data sets, the correlation between the harmonic sum and the bias term was over 0.99 when pairwise distances depending only on x were used (Eq. 5) and over 0.94 when the true distances $d_i = E(\mu_0 - y_i)^2$ were used and $k \leq 50$.

At all query points x_0, the true bias can be expressed in relation to the maximum bias with $E_{Y,x_i|x_0,d_i}(y_i - \mu_0) = c(x_0)\sqrt{d_i - \alpha_0}$. When x_0 lies in a symmetric position, $c(x_0) = 0$, at the bottom of the valley $c(x_0) = 1$, and on the top of the hill $c(x_0) = -1$. When we assume that $c(x_0)$ does not depend on the distance d_i and denote $E_{x_0}c(x_0)^2 = \delta^2$, the expected squared prediction error can be approximated with

$$\begin{aligned}E_{Y,x|d_1\ldots d_k}(\mu_0-\widehat{y}_0)^2 &= E_{x_0}\left(\frac{1}{\sum_{i=1}^{k}\frac{1}{d_i}}\right)^2 2\sum\sum_{i\neq j}\frac{c(x_0)^2\sqrt{d_i-\sigma^2}\sqrt{d_j-\sigma^2}}{d_i d_j}\\ &+\frac{1}{\sum_{i=1}^{k}\frac{1}{d_i}}\approx\frac{1}{\sum_{i=1}^{k}\frac{1}{d_i}}+\delta^2(k-1)\frac{1}{\sum_{i=1}^{k}\frac{1}{d_i}}-\sigma^2\delta^2\frac{k-1}{k}. \qquad (10)\end{aligned}$$

This is a linear transformation of the harmonic sum when k is kept fixed. Thus, the harmonic sum $1/\sum_{i=1}^{k}\frac{1}{d_i}$ can be used as a measure of the uncertainty about μ_0 when $y_1,\ldots,y_k$ and $d_1,\ldots,d_k$ are given. On the basis of simulated data, the approximation seems to work well in practice: The correlations between the approximation and the true expected loss were about 0.9.

We propose that the distance between a single observation x_0 and a set of observations $S_k = (x_1, x_2, \ldots, x_k)$ is measured with the harmonic sum of pairwise distances $d_i = d(x_i, x_0)$. When the pairwise distances correspond to the expected squared error $d_i \approx E(\mu_0 - y_i)^2$, our distance measure $d(x_0, S_k)$ approximates an increasing linear function of the expected squared prediction error $E(\mu_0 - \widehat{y}_0(S_k))^2$. We suggest that the distance between x_0 and S_k is measured with

$$\begin{aligned}d(x_0,S_k) &= \frac{1}{\sum_{i=1}^{k}\frac{1}{d_i}}\\ d_i &= \sum_{j=1}^{p}\alpha_0+\alpha_j(x_{0j}-x_{ij})^2. \qquad (11)\end{aligned}$$

4 Measuring the distance to a training data set

Our method could be used directly to measure the distance between a single observation and the training data set by letting $S_k = T$. However, when the training data set is large, it makes more sense to use only the k nearest observations. In the k-nearest-neighbour method, typically 5 to 100 neighbours are used to obtain the most accurate prediction. Thus, the observations lying far away from x_0 should not have an effect on the distance measure, because they do not affect the prediction. Let $d^{(k)}$ be the kth smallest distance $d(x_0, x_i)$. Our suggestion for the distance between the training data set and a single observation is

$$d(x_0, T) = d(x_0, S_k),\ S_k = \left\{x_i \in T \mid d(x_0, x_i) \leq d^{(k)}\right\}, \tag{12}$$

Our distance measure is problem-dependent. If we have the same inputs and several responses, the distance measure has to be defined separately for each response. The distance measure adapts itself to the regression function. The variables that do not affect the response do not affect the distance, either. The distance measure is invariant for linear transformations and approximately invariant for order-preserving transformations of the inputs. The distance measure also has a reasonable interpretation as the approximate measure of the expected loss function, which is an informative and novel way to measure the uncertainty about a new observation. The distance measure uses the squared error loss function, but can also be used for non-gaussian responses. If μ_0 were estimated with the unweighted k-nearest-neighbour method, the result would be the sum of single distances, just as proposed in [Angiulli and Pittuzi, 2005].

After the distance measure has been initialised by defining the α-coefficients, the major computational task is to find the k nearest training data observations. The computation of a single distance to the training data set requires about $N(p+2)+k^2$ operations. Initialision of the distance measure consists of fitting a prediction model and defining the α-coefficient for each explanatory variable, which is not a computational problem even in large data sets.

When prediction using some novel input values is needed, there rises the question of whether the model gives a reliable prediction or not. If the query point has enough training data instances nearby, the prediction can be kept reliable. If the query point is far away from the training data instances, the model will give a poor prediction with a high probability. The distance between the query point and the training data set gives information about the uncertainty of the prediction, see the example in Figure 1. The prediction accuracy of the model for validation data observations distant from the training data gives some information about the interpolation ability of model. In the example shown in Figure 1, the smoothed prediction accuracies of a linear regression model, a quadratic regression model and an additive spline model are plotted as functions of distance from the training data set.

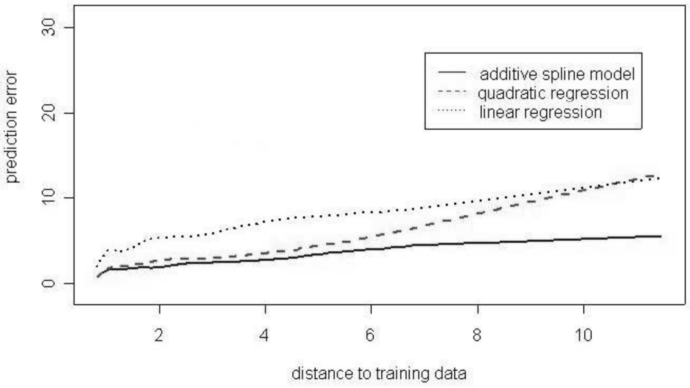


Fig. 1. Average prediction error (rMSE) in a simulated data set.

5 Performance in simulated data sets

The proposed distance measure reflects the expected squared error loss function $d(x_0, S_k) \approx c_1 + c_2 E(\mu_0 - \widehat{y}_0(S_k))^2$. We evaluated the correlation between the distance measure and the true expected loss using simulated data sets. The simulated data sets tried to represent a range of data sets which could arise from an industrial process of production. The observations occurred in clusters of different sizes, and the input variables were correlated. The true expected response was defined as a sum of 24 random effects of the form $\nu|b_0+\beta_1 x_{(1)}+\beta_2 x_{(2)}+\cdots+\beta_{16} x_{(16)}|^b s$, where $b = e^{0.5 z_b}$, $z_b \sim N(0, 1)$, making the typical effect rather linear, and the signum s turns the effect monotone with a probability of 0.7. Only 1, 2, 3 or 4 of β_i differs from 0, which means that interactions are restricted to the 4th order. The observed response was normally distributed around the true expected response. One simulated data set consisted of 10 000 observations and 16 input variables.

We simulated 20 data sets. We split all simulated data sets randomly into a learning data set and a validation data set. Out of the 2000 observations in the validation data, we calculated the distances to the learning data set using the proposed method. For each data set, we fitted an additive model with univariate thin plate regression splines as basis functions to define the α- coefficients of our distance measure. We defined the true pairwise distance as the true expectation $E(\mu_0 - y_i)^2$ and the true distance to the training data as the true expected squared prediction error

$$E\left(\mu_0 - \frac{\sum_{i=1}^{k} y_i/d(x_i, x_0)}{\sum_{i=1}^{k} 1/d(x_i, x_0)}\right)^2. \tag{13}$$

We examined the accuracy of our pairwise distance measure in approximating the expected loss $E(\mu_0 - y_i)^2 \approx \alpha_0 + \sum_{j=1}^{p} \alpha_j (x_{0j} - x_{ij})^2 = d(x_0, x_i)$ based on the correlations between the pairwise distance measure $d(x_i, x_j)$

and its theoretical reciprocal $(\mu_i - \mu_j)^2 + \sigma^2$. In the simulated data sets, the correlation varied between 0.19 and 0.81, the average correlation being 0.47. When neighbourhood size $k = 30$ was used, the correlation between the distance measure (Eq. 11) and the true expected squared error $E_Y(\mu_0 - \widehat{y}_0)^2$ ($\widehat{y}_0$ is defined in Eq. 6) varied between 0.41 and 0.66, the average correlation being 0.52. Thus, our distance measure $d(x_0, T)$ reflects relatively well its theoretical reciprocal, the expected squared error loss when x_0 is predicted based on T using distance-weighted k-nearest-neighbour. The deviation between the true expected squared error and our distance measure is mainly the consequence of the difficulty in approximating pairwise expected loss based only on x. If the true pairwise expected losses were known, the approximation would work much better: The correlation between the true expected loss $E(y_0 - \widehat{y}_0)^2$, $\widehat{y}_0 = \sum_{i=1}^{k} y_i / E(y_i - \mu_0)^2$ and the harmonic sum of the true pairwise distances $1/\sum_{i=1}^{k} 1/E(y_i - \mu_0)^2$ was typically about 0.93 and over 0.83 in all simulated data sets for $k \leq 200$. The size of the neighbourhood had a relatively small effect on the results, and all alternatives between $k = 5$ and $k = 500$ gave satisfactory correlations, and the best size of the neighbourhood varied greatly between the simulation runs. We suggest the use of $k = 30$, because that seemed to work best, and no larger neighbourhood was needed for k-nearest-neighbour prediction. Also, it seems intuitively reasonable that the distance to the training data can be defined based on the distances to the 30 nearest neighbours.

We compared our distance measure to the sum of pairwise distances of [Angiulli and Pittuzi, 2005]. Using $k = 30$, our distance measure was slightly better in 90 % of the simulation runs, and the average difference in correlation was 0.035. We also examined the effect of the method on defining α-coefficients for the distance measure. The average correlation between the pairwise distances based on a fitted additive model and on the true response surface was 0.92. The correlations between distances calculated based on two different learning methods were around 0.95, which means that the model fitting had only a small effect on the results. The method of [Mahamud and Hebert, 2003] for specifying α-coefficients gave poor results: The average pairwise correlation was only 0.30.

In the simulated data sets, the distance measure reflected the uncertainty about a new observation pretty well. We applied the distance measure to real industrial process data. We used a training data set having 90 000 observations, 26 continuous input variables and 6 binary input variables without any computational problems. In the test data set containing 60 000 observations, the average prediction error increased along with the distance from the training data (Figure 2). The correlations between the measured loss and the distance measure were between 0.25 and 0.5, depending on the response variable and the prediction model.

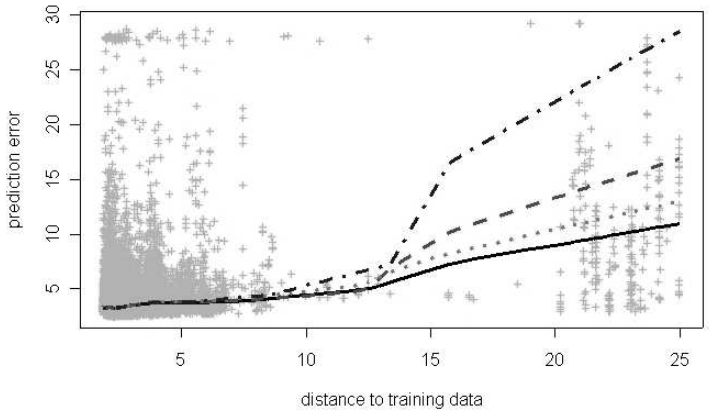


Fig. 2. Prediction error plotted against distance from the training data set in the real data set. The lines are the smoothed medians for four different prediction models.

6 Conclusion

We proposed a novel distance measure for the distance between a data set and a single observation. The distance measure can be interpreted to reflect the expected squared error loss when the single observation is predicted based on the data set using distance-weighted k-nearest-neighbour. Measurement of the distance from a data set has many potential applications, such as evaluation of the uncertainty of prediction and discovery of outliers.

References

[Angiulli and Pittuzi, 2005]F. Angiulli and C. Pittuzi. Outlier mining in large high-dimensional data sets. *IEEE Transactions on Knowledge and Data Engineering*, pages 203–215, 2005.

[Kaufman and Rousseeuw, 1990]L. Kaufman and P.J. Rousseeuw. *Finding Groups in Data: An Introduction to Cluster Analysis.* Wiley, New York, 1990.

[Mahamud and Hebert, 2003]S. Mahamud and M. Hebert. Minimum risk distance measure for object recognition. In *Proceedings of the ninth IEEE International Conference on Computer Vision (ICCV)*, pages 242–248, 2003.

[Markou and Singh, 2003]M. Markou and S. Singh. Novelty detection: a review. *Signal Processing*, pages 2481–2521, 2003.

[Wettschereck *et al.*, 1997]D. Wettschereck, D.W. Aha, and T. Mohri. Review and empirical evaluation of feature weighting methods for a class of lazy learning algorithms. *Artificial Intelligence Review*, pages 273–314, 1997.

What is at stake in the construction and use of credit score?

Mireille Bardos

Banque de France
Companies Observatory
(email: `mireille.bardos@banque-France.fr`)

Abstract. ”Statistical inference techniques, if not applied to the real world, will lose their import and appear to be deductive exercises. Furthermore, it is my belief that a statistical course emphasis should be given to both mathematical theory of statistics and to application of the theory to practical problems. A detailed discussion on the application of a statistical technique facilitates better understanding of the theory behind the technique.” C. Radhakrishna RAO in Linear Statistical Inference and Its Applications

The following summary presents important topics currently being debated for companies risk assessment and the main problems to be solved in the construction and use of credit scoring. Many examples and statistical issues will be presented during my presentation.

Keywords: discriminant analysis, credit risk forecasting, accuracy of probability of failure, stability of risk classes, transition matrices, credit risk models.

1 Introduction

The need for better control of credit risk by banks has led to a stepping-up of research concerning credit scoring. Several types of technique make possible the early detection of payment defaults by companies. These techniques fall within the field of discriminant analysis.

One of banks’ major objectives is to estimate expected loss and, using an extreme quantile, unexpected loss, for a population of companies, for example the customers of a bank. In order to do so, it is necessary to know the probability of default for each company for a given horizon. It is then possible to determine homogenous classes of risk.

Such an objective gives rise to several questions about the properties of the score made available. These relate to:

- the accuracy of estimates of probability and homogeneity of risk classes
- the stability over time of risk classes and their properties
- the dependence of the risk measurement on the business cycle
- the stability of transition matrices
- the correlation of risks

In order to get to grips with these questions, it is necessary to investigate the process of constructing the score and to examine the sensitive stages of the process. The practice of constructing and using scores leads to a second set of questions regarding:

- the type of variables
- the historical period of the files used to construct the scores
- the process of selecting variables
- the choice of discriminant analysis technique
- the forecasting horizon
- the stability of companies within risk classes
- the frequency with which the tool is revised
- the interaction between the business cycle, forecasting and revision

These issues are important for the quality of risk forecasting. They have increasingly been the subject of research and it appears that they are highly interdependent. The successive stages of a score's construction have an impact on the robustness and effectiveness of the tool eventually developed. In examining them, we shall set out the choices made at the Banque de France. Various different uses of the tool will be looked at.

2 Construction

2.1 How appropriate is the model for the data

2.1.1 The data Defining the event to be detected constitutes the first difficulty: Should this be legal proceedings or payment default? How serious does the payment default need to be? For the statistician, the criterion chosen is dependent on the information available. The question then arises of the correlation between these events for the same company. The population of target companies. For the statistical work to be of good quality, it requires: homogeneity of the population, representativeness of samples and their possible adjustment.

The forecasting horizon is determined by the needs of the banking system, but is dependent on how recent the data are and the impact of the economic cycle. The way in which the data files are organised is the result of a compromise. The choice of explanatory variables is also determined by availability and reliability. Qualitative variables are especially fragile and often better suited to expert assessment. Among quantitative variables, monitoring companies' bank accounts is probably very revealing over the short term, but this option is not available to the Banque de France. Economic and financial ratios constructed using accounting variables are widely available and relatively homogenous thanks to the existence of a chart of accounts. They are based on an underlying theory: financial analysis.

They are tricky and time-consuming to prepare. Abnormal, extreme or "bizarre" values are examined, as well as their law of probability, discriminant capacity, correlations and linearity or non-linearity with respect to the problem in question. This last characteristic may only be known via an examination of ratio distributions. It also determines the choice of technique (linear or non-linear). Once the correct ratios have been identified, the discriminant capacity is assessed by tests on quantiles [Vessereau, 1987]. Some statisticians use the stochastic dominance test [Davidson and Duclos, 1999].

2.1.2 The models The aim in constructing a score may be confined to the desire to identify risk signals. However, if one wishes to obtain an operational tool, its construction needs to be based on a decision rule, but its practical use requires knowing the probability of failure at a given horizon. The methods that result in linear combination of ratios are by far the most robust and are easy to interpret.

Indeed, corporate failure is a complex phenomenon for which the actual causal variables are difficult to access and to identify. The score functions therefore make use of symptoms such as descriptors of the company's situation before its failure. In other words, it is impossible to accurately define the phenomenon of company failure, contrary to what occurs in other fields of application of discriminant analysis that are closer to physical science, such as shape recognition, where overlearning is easier to master and techniques such as neural networks are successfully applied. It is, therefore, the very traditional linear discriminant analysis (LDA) of Fisher that is used at the Banque de France.

Estimating probability may be associated with the theoretical model used or may be done on the basis of empirical distributions and Bayes' theorem. The choice between the two will depend on how representative the files are and the extent to which the data correspond to the assumptions in the model.

2.2 Some thoughts on models

With detailed theoretical comparisons having been made in several studies [Baesens *et al.*, 2003], here we suggest some thoughts about suitability for companies' economic data and robustness over time. The main models will be looked at: Fisher's linear or quadratic discriminant analysis; logistical regression; Disqual [Lelogeais, 2003]; decision tree; neural networks; and other non-parametric methods.

The much-debated comparison between Fisher LDA and LOGIT warrants some further investigation – in terms of the theoretical properties [Amemiya and Powell, 1983], interpretability (the great advantage of Fisher's LDA: contributions of variables to the value of the score), sensitivity to the sampling plan of the logistical regression [Celeux and Nakache, 1994], estimation of the probability of failure (either via a theoretical formula or by use of Bayes' theorem on empirical distributions).

2.2.1 Some arguments regarding the choice between models The model's appropriateness for the data derives from the following properties and the corresponding choices:

- linearity or non-linearity,
- robustness to loss of parametric assumptions
- sensitivity to extreme values
- robustness over time (problem of thresholds for economic variables)
- interpretability of results for the user.

2.3 Probability of failure

The probability of failure provides a measure of the intensity of risk. It is much more informative than a decision threshold. Several crucial issues determine the quality of the tool:

1. The forecasting horizon must be consistent with the nature of the data. There is by definition a lag of a few months between balance sheet variables and the time at which the company is assessed, and these variables describe what has occurred over the course of the past year; they are consequently better suited to a medium-term forecast than to a short-term one. Balance sheets undoubtedly provide useful and robust information, provided that the assessment and the forecasting horizon are well matched. With a one-year horizon, it might be thought that it would be possible to create a short-term indicator, which, if it were re-estimated sufficiently often, would allow us to track the conditions under which companies are operating. But such an indicator would then follow the business cycle closely. However, this kind of perspective is very difficult to work with as frequent re-estimation in a changing environment is liable to lead to functions that always lag the current situation. It was therefore decided to work on a medium-term horizon with quantitative variables based on balance sheets and which are submitted to a method of financial analysis whose quality is long established. Given that balance sheet structures are related to the sector to which a company belongs, scores are created according to the major sectors (industry, wholesale trade, retail trade, transport, construction, business services).
2. An estimate of posterior failure probabilities well suited to the empirical data using Bayes' theorem is closely associated with the determination of risk classes. Robustness over time must be ensured for the average probability per risk class. The confidence interval of this average indicates the accuracy and provides a measure of what can happen in the worst case scenario.
3. The stability of companies in risk classes is studied using transition matrices. This paper is participating in the currently heated debate about "through the cycle" vs. "point in time" estimates.

3 Use

3.1 Individual diagnosis

Credit scoring is the first stage in the analysis of individual cases. A whole range of tools is made available. The scores are accompanied by aids to interpretation for the user, who is not a statistician but rather a financial analyst: failure probabilities, contributions of ratios to the score, and the company situation relative to the sector as a whole.

It is a great asset for the statistician to be able to identify ways in which the tool is unsuitable thanks to the analyst users who point out concrete cases where there are measurement difficulties. Their observations make it possible to improve the statistical measure of concepts of financial analysis and understanding of corporate failure processes.

3.1.1 Risk assessment for a given population Progress reports for a given group of customers are recommended by the Basel Committee. The Banque de France has produced some examples aimed at monitoring a particular population: IRISK method; economic impact of corporate failure.

3.2 Research under way

The scores constructed at the Banque de France cover a wide range of sectors. Applied to a representative sample of firms whose turnover exceeds EUR 0.75 million, they allow us to study many questions related to credit risk.

Risk contagion [Stili, 2003] [Stili, 2005] can be studied using the Banque de France's database of payment incidents involving trade bills.

Risk correlation [Foulcher *et al.*, 2003] between companies has a substantial impact on the assessment of potential losses.

If the **link between risk and the business cycle** [Bataille *et al.*, 2005] can be clarified, it would make it possible to better anticipate the risk of future failures in the light of macroeconomic variables or specific factors.

Looking at the **paths followed by companies** [Bardos, 1998b] makes possible the dynamic study of risk.

The **transition matrices** between classes of risk allow us to study the Markovian character or otherwise of failure processes.

Concentration of debt [Bardos and Plihon, 1999] is a source of major risk for banks.

Furthermore, investigation of credit risk models requires comparisons between company rating systems. Statistical research on the simulation of **distributions of default rates by rating** [Tiomo, 2002] make it possible to establish scales of reference for these comparisons.

References

[Altman and Saunders, 1998]E. I. Altman and A. Saunders. Credit risk measurement: developments over the last 20 years. *Journal of Banking and Finance*, 21:1721–1742, 1998.

[Amemiya and Powell, 1983]T. Amemiya and J. Powell. *Karlin, Amemiya, Goodman, Studies in econometrics, time series and multivariate statistics.* Academic Press New York, 1983.

[Anderson, 1984]T. W. Anderson. *An introduction to multivariate statistical analysis*, chapter 6: Classification of observations. Wiley, 1984.

[Baesens *et al.*, 2003]Baesens, T. Van Gestel, S. Viaene, M. Stepanova, J. Suykens, and J. Vanthienen. Benchmarking state-of-art classification algorithms for credit scoring. *Journal of the Operational Research Society*, 2003.

[Bardos and Plihon, 1999]M. Bardos and D. Plihon. Détection des secteurs risqués - la méthode IRISK. *Bulletin de la Banque de France*, 69, 1999.

[Bardos and Zhu, 1998]M. Bardos and W. H. Zhu. *Bio-mimetic approaches in management science*, chapter Comparison of linear discriminant analysis and neural networks, application for the detection of company failures, pages 77–100. Kluwer Academic Publishes, 1998.

[Bardos *et al.*, 2004]M. Bardos, S. Foulcher, and E. Bataille. Banque de France scores: method, results, practice. Technical report, Banque de France, 2004.

[Bardos, 1998a]M. Bardos. Detecting the risk of company failure. *The Journal of Banking and Finance*, 22, 1998.

[Bardos, 1998b]M. Bardos. Le score BDFI: du diagnostic individuel à l'analyse de portefeuille, les études de l'observatoire des entreprises. Technical report, 1998.

[Bardos, 2001]M. Bardos. *Analyse discriminante: application au risque et scoring financier.* Dunod, 2001.

[Barret and Donald, 2002]G. Barret and S. G. Donald. Consistent tests for stochastic dominance. 2002.

[Bataille *et al.*, 2005]E. Bataille, C. Bruneau, and F. Michaud. Use of principal components method to follow the link between business cycle and risk of companies' failure. Technical report, 2005.

[Blockwitz and Hohl, 2001]S. Blockwitz and S. Hohl. Reconciling ratings. *Risk Magazine*, June 2001.

[Breiman *et al.*, 1984]L. Breiman, J. H. Freidman, R. A. Ohlson, and C. J. Stone. *Classification and regression trees.* Edition Wadswoth International Group, California, 1984.

[Celeux and Nakache, 1994]G. Celeux and J. P. Nakache. *Analyse discriminante sur variables qualitatives.* Polytechnica, 1994.

[Davidson and Duclos, 1999]R. Davidson and J.Y. Duclos. Statistical inference for stochastic dominance and for the measurement of poverty and inequality. 1999.

[Efron and Tibshirani, 1993]B. Efron and R. J. Tibshirani. *An introduction to the Bootstrap.* Chapman & Hall, 1993.

[Efron, 1975]B. Efron. The efficiency of logistic regression compared to normal discriminant analysis. *Journal American Statistical Society*, 70:892–898, 1975.

[Foulcher *et al.*, 2003]S. Foulcher, C. Gouriéroux, and A. Tiomo. La structure par terme des taux de défauts et ratings, études et recherches de l'observatoire des entreprises. Technical report, 2003.

[Foulcher *et al.*, 2004]S. Foulcher, C. Gouriéroux, and A. Tiomo. études et recherches de l'observatoire des entreprises :la corrélation de migration: méthode d'estimation et application aux historiques de notation des entreprises françaises. Technical report, 2004.

[Gnanadesikan and *al*, 1989]R. Gnanadesikan and *al.* Discriminant analysis and clustering. *Statistical Science*, 4 (1):34–69, 1989.

[Gouriéroux, 1989]C. Gouriéroux. économétrie des variables qualitatives. *Économica*, 1989.

[Hand, 1981]D. J. Hand. Discrimination and classification. *Wiley series in probability and mathematical statistics*, 1981.

[Hristache *et al.*, 2004]Hristache, Delcroix, and Patiléa. On semi parametric m-estimation. *Journal of Statistic companies' economic data and robustess over time*, 2004.

[Kendall and Stuart, 1961]M. G. Kendall and A. Stuart. *The advanced Theory of Statistics.* Griffin, London, 1961.

[Lebart *et al.*, 1998]L. Lebart, A. Morineau, and M. Piron. *Statistique exploratoire multidimensionnelle.* Dunod, 1998.

[Lelogeais, 2003]L. Lelogeais. Un score sur variables qualitatives pour la détection précoce des défaillances d'entreprises. *Bulletin de la Banque de France*, 114, 2003.

[Lo, 1986]A. W. Lo. Logit versus discriminant analysis. *Journal of Econometrics*, 31:151–178, 1986.

[Maddala, 1999]G. S. Maddala. *Limited-dependent and qualitative variables in econometrics.* Cambridge University Press, 1999.

[Mc Lachlan, 1992]J. Mc Lachlan. *Discriminant Analysis and Statistical Pattern Recognition.* Wiley, New York, 1992.

[Michaud, 2004]F. Michaud. Mémoire de stage. Technical report, 2004.

[Rao, 1973]C. R. Rao. *Linear statistical inference and its applications.* Wiley, 1973.

[Saporta, 1990]G. Saporta. *Probabilités, analyse des données et statistique.* Technip, 1990.

[Stili, 2003]D. Stili. Etudes et recherches de l'observatoire des entreprises : Les incidents de paiement sur effets de commerce. Technical report, 2003.

[Stili, 2005]D. Stili. *Détection précoce des défaillances d'entreprises et contagion du risque.* PhD thesis, Université Paris I, 2005.

[Thiria *et al.*, 1997]S. Thiria, Y. Lechevallier, O. Gascuel, and S. Canu. *Statistique et méthodes neuronales.* Dunod, 1997.

[Tiomo, 2002]A. Tiomo. Risque de crédit et variabilité des taux de défaut: une analyse empirique par simulation, études et recherches de l'observatoire des entreprises. Technical report, 2002.

[Vessereau, 1987]Vessereau. Une propriété peu connue: l'intervalle de confiance de la médiane. *Revue de Statistique Appliquée*, XXXV (1):5–8, 1987.

A review of some semiparametric regression models with application to scoring

Jean-Loïc Berthet[1] and Valentin Patilea[2]

[1] ENSAI
Campus de Ker-Lann
Rue Blaise Pascal - BP 37203
35172 Bruz cedex, France
(e-mail: jean-loic.berthet@ensai.fr)
[2] CREST-ENSAI
Campus de Ker-Lann
Rue Blaise Pascal - BP 37203
35172 Bruz cedex, France
(e-mail: patilea@ensai.fr)

Abstract. Some semiparametric models for binary response data are reviewed: single-index models, generalized partially linear models, generalized partially linear models single-index models and multiple-index models. All these models can be seen as extensions of the classical logistic regression. We test and compare these models using data on bankruptcy of French companies and data from credit business.

Keywords: Scoring, semiparametric regression, iterative methods, single and multiple-index, bandwidth choice.

1 Introduction

Classification techniques are used in many statistical applications. The objective of any classification model is to classify individuals in two or more groups based on a predicted outcome associated with each individual. Here, we are interested in statistical models classifying individuals in two groups: 'good' (or 'not default') and 'bad' (or 'default') individuals. Such models can be applied in banking and credit control, marketing, medicine, *etc.* The classification rule for an individual must be based on the information about the individual at the time of the decision. This information is contained in a vector of explanatory variables (factors, indicators, characteristics, ...) $\mathbf{X} = (X_1, ..., X_p)^\top$. Usually, the available information for an individual is synthesized into a single value usually called the *score*. The score aims to reflect the probability that the individual will 'not default'.

Various parametric and nonparametric methods can be used to solve classification problems (see, e.g., [Hand and Henley, 1997] for a review). Discriminant analysis, linear regression and logistic regression are the standard parametric techniques, while $k-$nearest neighbors, classification trees, neural networks and, more recently, support vector machine are some common nonparametric (distribution-free) procedures.

The simple, user friendly and easily interpretable character of the parametric regression models make them the most popular classification techniques in many application fields. The nonparametric methods, unlike the parametric methods, make no (or mild) assumption about the distribution of the observations and are therefore attractive when data on hand does not meet strict statistical assumptions. The price of this flexibility can be high, however. First, estimation precision decreases rapidly as the dimension of $\mathbf{X}$, the vector of explanatory variables, increases. This is the so-called curse of dimensionality. A second problem with nonparametric methods is that the results can be difficult to display, communicate, and interpret when $\mathbf{X}$ is multidimensional. A further problem with nonparametric methods is the difficulty to extrapolate the prediction to individuals with characteristics that are very different from the characteristics of the individuals that served for estimation.

The semiparametric methods represent an appealing compromise for constructing statistical models. By making assumptions that are of intermediate strength between the parametric and nonparametric approaches, the semiparametric models reduce the risk of misspecification relative to a parametric model and avoid at least in part the aforementioned drawbacks of the nonparametric methods.

In this paper we review some semiparametric regression methods that apply to *scoring*, that is to determine how likely an individual will 'not default'. The starting point of the review is the logistic regression. The power of the semiparametric methods is investigated using data on bankruptcy of French companies and publicly available data on credit-scoring from a German bank.

2 Semiparametric models for binary response variables

Let Y be a random variable taking the values 0 ('bad' or 'default' individual) or 1 ('good' or 'not default' individual). The problem on hand to estimate the probability of the event $\{Y = 1\}$ given a vector of explanatory variables $\mathbf{X}$. The logistic regression is a particular case of the so-called *generalized linear model* (see [McCullagh and Nelder, 1989]) where the conditional mean of Y given $\mathbf{X}$ has the form

$$E(Y \mid \mathbf{X}) = G(c + \mathbf{X}^\top \beta) \tag{1}$$

with a known monotone function G ($G(x) = \{1 + \exp(-x)\}^{-1}$ for the logistic regression) and an unknown parameters $(c, \beta^\top)^\top$. This model can be interpreted as follows: there exists a latent variable Y^* that can be related to $\mathbf{X}$ through a linear model $Y^* = c + \mathbf{X}^\top \beta + u$ with u an error term with cumulative distribution function (cdf) G. The observation Y is nothing but $\mathbf{1}_{\{Y^* \geq 0\}}$ where $\mathbf{1}_{\{\cdot\}}$ equals one if the condition inside the curly brackets holds, and zero otherwise. The model (1) is purely parametric in the sense that one only has to estimate the vector of coefficients $(c, \beta^\top)^\top$.

Several semiparametric extensions of model (1) have been proposed. A natural idea is to relax the hypothesis of a linear regression model for the latent variable Y^*. [Härdle *et al.*, 1998] proposed to separate the explanatory variables into two groups, that is $\mathbf{X} = (\mathbf{Z}^\top, \mathbf{T}^\top)^\top$ with $\mathbf{Z} \in \mathbf{R}^{p_1}$, $\mathbf{T} \in \mathbf{R}^{p_2}$, and to suppose that $Y^* = \mathbf{Z}^\top\beta + m(\mathbf{T}) + u$, where the error term has a logistic law (the constant c appearing in model (1) is absorbed by the function $m(\cdot)$). The function m is unknown and it must be estimated nonparametrically. In this settings one has a *generalized partially linear model*

$$E(Y \mid \mathbf{Z}, T) = G(\mathbf{Z}^\top\beta + m(\mathbf{T})) \tag{2}$$

with $G(x) = \{1 + \exp(-x)\}^{-1}$. This model is semiparametric in the sense that in addition to the finite dimensional vector β, one has to estimate also the function m. If one wants to assume that several explanatory variables have a nonlinear effect on the conditional mean of Y^*, one has to estimate nonparametrically a multivariate function m. In order to avoid the course of dimensionality, [Härdle *et al.*, 2004] considered that $m(\mathbf{T}) = m(T_1, ..., T_{p_2}) = m_1(T_1) + ... + m(T_{p_2})$. Another approach that avoids nonparametric estimation of a multivariate function is to suppose that there exists a vector $(\alpha_1, ..., \alpha_{p_2})^\top$ (identifiable up to a scaling factor) such that

$$m(T_1, ..., T_{p_2}) = m(\alpha_1 T_1 + ... + \alpha_{p_2} T_{p_2}).$$

See [Carroll *et al.*, 1997]. In all these models the coefficients β (β and α for the model of [Carroll *et al.*, 1997]) can be estimated with a precision of order $n^{-1/2}$ where n is the sample size, that is the usual precision of a parametric model.

Another natural extension of the parametric model goes as follows. Assume that $Y^* = c + \mathbf{X}^\top\beta + u$ with u an error term with *unknown* law independent of $\mathbf{X}$ given $\mathbf{X}^\top\beta$. Then,

$$E(Y \mid \mathbf{X}) = r(\mathbf{X}^\top\beta) \tag{3}$$

with $r(\cdot)$ an unknown function that has to be estimated nonparametrically. The constant c is absorbed by $r(\cdot)$. Moreover, the vector β can only be determined up to a scaling factor. The model (3) belongs to a general class of semiparametric models called *single-index models (SIM)*. In such models one only assumes that when computing the conditional expectation of Y given $\mathbf{X}$, all the relevant information carried by $\mathbf{X}$ is contained in a linear combination of the components of $\mathbf{X}$. In the following we shall concentrate on inference methods for model (3). Note that model (3) can be obtained as a particular case of the model of [Carroll *et al.*, 1997] by taking $\beta = 0$ and setting $r = G \circ m$ (and relabelling the explanatory variables).

Several semiparametric approaches for consistent estimation of β in SIM have been proposed including $M-$estimation, average derivative methods and iterative methods. See [Delecroix *et al.*, 2004] for a review. Here, we focus on $M-$estimation. Typically, if $(Y_1, \mathbf{X}_1^\top)^\top, ..., (Y_n, \mathbf{X}_n^\top)^\top$ denote the observations, a semiparametric M-estimator of β is defined as

$$\widehat{\beta} = \arg\min_\beta \frac{1}{n} \sum_{i=1}^n \psi\left(Y_i, \widehat{r}\left(\mathbf{X}_i^\top \beta; \beta\right)\right) \tau_n(\mathbf{X}_i), \tag{4}$$

where $\widehat{r}\,(t;\beta)$ is a nonparametric estimator, for instance the Nadaraya-Watson estimator, of the regression function $r\,(t;\beta) = E\left(Y \mid \mathbf{X}^\top\beta = t\right)$, ψ is a contrast function and $\tau_n(\cdot)$ is a so-called trimming function introduced to guard against small values of the denominators appearing in the nonparametric estimator. Finally, the conditional mean of Y is estimated by $\widehat{r}(x^\top\widehat{\beta};\widehat{\beta})$. [Klein and Spady, 1993] considered the case $\psi(y,r) = -\{y\log(r) + (1-y)\log(1-r)\}$ which yields the semiparametric maximum likelihood estimate of β. [Dominitz and Sherman, 2003] considered the case $\psi(y,r) = (y-r)^2$ and proposed a nice iterative method that avoids optimization with respect to both occurrences of β in equation (4). [Delecroix *et al.*, 2004] suggested other choices for $\psi(y,r)$ that improve the performances of the estimator $\widehat{\beta}$ in the presence of outliers.

The large sample properties of the estimates $\widehat{\beta}$ and $\widehat{r}(x^\top\widehat{\beta};\widehat{\beta})$ obtained from optimization procedures as (4) are now well known in the case of independent, identically distributed observations of $(Y, \mathbf{X}^\top)^\top$. In particular, this allows to obtain significance tests for the coefficients β and confidence intervals for the conditional probability of 'not default' given $\mathbf{X}$. Extensions to the case of dependent data have been also studied. See [Xia *et al.*, 2002] for a description of the techniques of proof that apply for dependent data and for a list of references.

A crucial problem associated with the estimator $\widehat{\beta}$ is the choice of the smoothing parameter for the nonparametric estimator of the regression function $r\,(t;\beta)$. One may consider the smoothing parameter as another parameter of interest which can be estimated at the same time as β, that is one can optimize the objective function in (4) simultaneously with respect to β and the smoothing parameter. In general, to avoid degenerate problems when optimizing simultaneously with respect to β and the smoothing parameter, a leave-one-out version of the nonparametric estimator should replace $\widehat{r}$ in equation (4). See, e.g., [Delecroix *et al.*, 2004] for the theoretical properties of the simultaneous optimization approach.

Despite the fact that the regression function is supposed unknown, a SIM still imposes that all the relevant information carried by $\mathbf{X}$ is contained in one factor that is obtained as a linear combination of the components of $\mathbf{X}$. A natural idea is to investigate whether more than one factor is necessary to capture the information contained in $\mathbf{X}$. For instance, one may consider the model

$$E(Y \mid \mathbf{X}) = r(\mathbf{X}^\top\beta^1, \mathbf{X}^\top\beta^2) \tag{5}$$

with $r(\cdot,\cdot)$ an unknown bivariate function that has to be estimated nonparametrically and β^1, β^2 two vectors of unknown coefficients. (Suitable normalization conditions are necessary to make the vectors β^1, β^2 identifiable.) The unknown parameters can be estimated by an extension of (4), that is

$$(\widehat{\beta^1},\widehat{\beta^2}) = \arg\min_{(\beta^1,\beta^2)} \frac{1}{n}\sum_{i=1}^{n} \psi\left(Y_i, \widehat{r}\left(\mathbf{X}_i^\top \beta^1, \mathbf{X}_i^\top \beta^2; (\beta^1,\beta^2)\right)\right)\tau_n(\mathbf{X}_i), \quad (6)$$

where $\widehat{r}\left(t,s;(\beta^1,\beta^2)\right)$ is a nonparametric estimator of the regression function $r(t,s;(\beta^1,\beta^2)) = E\left(Y \mid (\mathbf{X}^\top\beta^1, \mathbf{X}^\top\beta^2) = (t,s)\right)$. The smoothing parameters of the bivariate estimator of r can be selected by simultaneous optimization in (6) with respect to (β^1,β^2) and the smoothing parameters. See [Xia *et al.*, 2002] and [Delecroix *et al.*, 2004].

An alternative procedure for finding β^1, β^2 is to search these directions one by one: first, search $\widehat{\beta^1}$ like in (4); second, search $\widehat{\beta^2}$ orthogonal to $\widehat{\beta^1}$ and solution of the problem

$$\widehat{\beta^2} = \arg\min_{\beta^2} \frac{1}{n}\sum_{i=1}^{n} \psi\left(Y_i, \widehat{r}\left(\mathbf{X}_i^\top \widehat{\beta^1}, \mathbf{X}_i^\top \beta^2; (\widehat{\beta^1},\beta^2)\right)\right)\tau_n(\mathbf{X}_i).$$

This procedure simplifies the optimization problem. It can be shown that, under mild conditions, the linear subspace generated by directions obtained by sequential search is the same as the linear subspace generated by the directions obtained from (6). One may search for more than two directions β, either by joint maximization as in (6) or by sequential search after finding the first two directions $\widehat{\beta^1}$, $\widehat{\beta^2}$, but the results will become much more difficult to interpret.

The last theoretical issue we shall discuss here is the problem of testing in the semiparametric models mentioned above. There are at least two types of test problems one may consider. First, it is important to be able to test the traditional parametric binary response regression models, typically the logistic regression, using semiparametric models. [Härdle *et al.*, 1998] started from model (2) and tested the logistic regression model by setting the null hypothesis $m(\mathbf{T}) = c + \mathbf{T}^\top\gamma$ for some constant c and some vector γ. [Härdle and Spokoiny, 1997] considered the SIM framework described by equation (3) and proposed a test procedure for checking whether the function r has a given form (typically, whether r is the logistic function or not).

If the parametric model is rejected in favor of a more flexible semiparametric specification, the next step is to test the semiparametric model itself against more general semiparametric or nonparametric alternatives. It is only recently that promising testing procedures for SIM have been proposed. See [Xia *et al.*, 2004] and [Stute and Wang, 1994].

3 The data

The stakes of a reliable, interpretable, easy to implement and easy to update scoring method are important. The semiparametric methods represent an alternative stream of dealing with these aspects. Their power was relatively little investigated in scoring applications. Our aim is to provide additional empirical evidence on the utility of the semiparametric methods in scoring problems. The semiparametric techniques mentioned above are tested and compared with the benchmark parametric models. For this purpose we use two types of data. First, we work with a sample from a database of *Banque de France*. Our dataset contains the accounting balances of French companies from one economic sector during several years. The task is to asses the risk of bankruptcy for a company given the information provided by the company.

For our second application we use data on private loans from a German bank. The data are presented in [Fahrmeir and Tutz, 1994] and are publicly available. In credit business, banks are interested in information whether prospective consumers will pay back their credit or not. The aim of credit scoring is to predict the probability that a consumer with certain characteristics is to be considered as a potential risk. The dataset we consider consists of 1000 consumer credits. For each consumer the binary response variable "creditability" is available, together with a set of covariates that are assumed to influence creditability.

References

[Carroll *et al.*, 1997]R.J. Carroll, J. Fan, I. Gijbels, and M.P. Wand. Generalized partially linear single-index models. *Journal of the American Statistical Association*, 92(438):477–489, 1997.

[Delecroix *et al.*, 2004]M. Delecroix, M. Hristache, and V. Patilea. On semiparametric m-estimation. *Journal of Statistical Planning and Inference*, in press, 2004.

[Dominitz and Sherman, 2003]J. Dominitz and R.P. Sherman. Some convergence theory for iterative estimation procedures. *Working Paper, California Institute of Technology*, 2003.

[Fahrmeir and Tutz, 1994]L. Fahrmeir and G. Tutz. *Multivariate Statistical Modelling Based on Generalized Linear Models.* Springer, New-York, 2nd edition, 1994.

[Hand and Henley, 1997]D.J. Hand and W.E. Henley. Statistical classification methods in consumer credit scoring: a review. *Journal of the Royal Statistical Society, Series A*, 160(4):523–541, 1997.

[Härdle and Spokoiny, 1997]W. Härdle and S. Spokoiny, V.and Sperlich. Semiparametric single index versus fixed function modelling. *The Annals of Statistics*, 25:212–243, 1997.

[Härdle *et al.*, 1998]W. Härdle, E. Mammen, and M. Müller. Testing parametric versus semiparametric modeling in generalized linear models. *Journal of the American Statistical Association*, 93(444):1461–1474, 1998.

[Härdle *et al.*, 2004]W. Härdle, S. Huet, E. Mammen, and M. Sperlich. Bootstrap inference in semiparametric generalized additive models. *Econometric Theory*, 20:265–300, 2004.

[Klein and Spady, 1993]R.W. Klein and R.H. Spady. An efficient semiparametric estimator for binary response models. *Econometrica*, 61:387–421, 1993.

[McCullagh and Nelder, 1989]P. McCullagh and J.A. Nelder. *Generalized Linear Models*. Monographs on Statistics and Applied Probabilites. Chapman and Hall, London, 2nd edition, 1989.

[Stute and Zhu, 2004]W. Stute and L.X. Zhu. Nonparametric checks for single-index models. *The Annals of Statistics*, in press, 2004.

[Xia *et al.*, 2002]Y. Xia, H. Tong, W.K. Li, and L.X. Zhu. An adaptive estimation of dimension reduction space (with discussions). *Journal of the Royal Statistical Society, Series B*, 64(3):363–410, 2002.

[Xia *et al.*, 2004]Y. Xia, W.K. Li, H. Tong, and D. Zhang. A goodness-of-fit test for single index models. *Statistica Sinica*, 14:1–39, 2004.

Business cycle and Corporate Failure in France: is there a link?

Eric Bataille[1], Catherine Bruneau[1,2], Alexis Flageollet[1,2], and Frédéric Michaud[1,3]

[1] Banque de France
(e-mail: eric.bataille@banqe-france.fr,
alexis.flageollet@banque-france.fr)
[2] THEMA
Université Paris X
(e-mail: cbruneau@u-paris10.fr)
[3] Crédit Agricole du Morbihan
(e-mail: Frederic.michaud@ca-morbihan.fr)

Abstract. The aim of this paper is to extract cyclical factors, first from companies' data used to build the score functions estimated by the Bank de France and, second, from these functions themselves. The constraints are those of a database including a large number of variables and companies and a small number of time periods. The method chosen is the "principal components analysis" adapted by [Bai and Ng, 2000] and [Bai and Ng, 2004a] in the context of large N and limited T. We show that the factorial structure could be useful to immunize the score functions and the related decisions against the cyclical variations in the state of economy.
Keywords: Panel data, common factors, principal components analysis, scoring.

1 Introduction

Over the last 30 years, many research papers have focused on the early detection of corporate failures. The Banque de France has developed several scores for use by financial analysts in the branches and at head office. Indeed, for reasons of robustness and for the seak of simplicity, the Bank has chosen to implement the linear Fisher classification method to build its scores. The analysis is static: the classification is conducted with a cross-section estimation over one year and adjusted so as to be robust to changes over time, even if the score functions need to be regularly adapted. In certain cases, a complete reestimation is needed: this is the case, when the nature of corporate failure has significantly changed and the related structural change cannot be modelled ex ante. In the other cases, the score function remains valid but the discriminiant threshold has to be adjusted. Thus, this one seems to be dependent on the position of the economy in the business cycle.

In that purpose, we choose to extract an endogeneous cyclical component directely from the corporate database, because it allows to answer directly the question about immunization of the scores against cyclical variations and because the whole analysis remains at a microeconomic level, which avoids considering difficult questions about data agregation.

The business cycle can be represented as an unobservable component which can be identified with the principal components method presented in [Bai and Ng, 2000] and [Bai and Ng, 2004a].

After extracting cyclical component from corporate database and from the scores themselves, we look for interpreting these factors as cyclical comovements, by comparing them to macroeconomic series which usually enter in the characterization of the French business cycle ([Bruneau *et al.*, 2002a]).

2 The dynamic factor analysis (DFA) in the lines of [Stock and Watson, 1998]

In this section, we recall the main contributions in Dynamic Factor Analysis starting with the work by [Stock and Watson, 1998] continuing with the papers by [Bai and Ng, 2000], [Bai and Ng, 2004a] and [Bai and Ng, 2004b].

2.1 The main assumptions

From now on, X_t will denote a N-dimensional multiple time series. The factor structure is as follows:

$$X_t = \Lambda_t F_t + u_t \tag{1}$$

where the dimensions are respectively : Nx1, Nxr, rx1 and Nx1. The common part of X_t is $\Lambda_t F_t$ and u_t denotes its idiosyncratic part. Note that, in the previous model, the dynamics is introduced in three ways:

1) the factors are assumed to evolve according to a time series (multivariate) process which is not observable;

2) the idiosyncratic error terms are serially correlated;

3) the factors can enter with lags (or even with leads).

Note also that the dynamic factor model can be rewritten such that Λ_t is constant by suitable redefinition of the factors and the idiosyncratic disturbances.

The factors as well as the loadings (Λ_t) are considered as parameters that are estimated by solving a non-linear least squares problem which is decomposed into two successive ordinary least squares minimizations, which finally lead to solve an eigenvalue problem.

It is important to recall the assumptions:

i) $\Lambda_t = \Lambda_0$

ii) the disturbances u_t are i.i.d. independent across series, normally distributed so that the covariance matrix Σ of the vector of residuals $u = (u_1, ..., u_T)$ is diagonal. (Its seems to be possible to allow a weak correlation sructure between the u_{jt} for any date t ([Chamberlain and Rothschild, 1983]).

Thus the estimator of $(\varLambda_0, F)$ solves the non-linear least squares problem with the objective function:

$$V_{NT}(\varLambda_0, F) = \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} I_{it}(X_{it} - \lambda_{i0}F_t)^2 \tag{2}$$

where $I_{it} = 1$ if the variable is observed at time t and equal to 0, otherwise.

The previous analysis is a standard principal component analysis with the only difference that dynamic features are taken into account.

Recently, [Bai and Ng, 2004a] have proposed a statistical procedure to extract factors without considering the degree of persistance in the series. It is the so-called PANIC approach (Panel Analysis of Non-stationary in Idiosyncratic and Common components).

2.2 PANIC analysis ([Bai and Ng, 2004a])

The model is the following:

$$\begin{aligned} X_{it} &= c_i + \beta_i t + \lambda_i' F_t + e_{it} \\ (1-L)F_t &= C(L)u_t \\ (1-\rho_i L)e_{it} &= D_i(L)\varepsilon_{it} \end{aligned}$$

with $C(L) = \sum_{j=0}^{\infty} C_j L^j$ and $D_i(L) = \sum_{j=0}^{\infty} D_{i,j} L^j$. The idiosyncratic e_{it} is $I(1)$ if $\rho_i = 1$ and is stationary if $|\rho_i| < 1$.

When the residuals e_{it} are $I(0)$ it is possible to get consistent estimates of the factors. When it is not the case, e_{it} are $I(1)$, one has to work with the first differences of the series. The model allows r_0 stationary factors and r_1 common trends with $r = r_0 + r_1$ [1]. Equivalently, the rank of $C(1)$ is equal to r_1.

Instead of testing for the presence of a unit root in X_{it}, the approach proposed here is to test the common factor and the idiosyncratic separately. PANIC has two objectives: first, to determine if non-stationarity comes from the common or from the idiosyncratic component. Second, to construct valid pooled tests for panel data when units are correlated; that is under the cross-sectional dependence (CSD) assumption.

More precisely, the objective of PANIC is to determine r_1 and test if $\rho_i = 1$ when neither F_t nor e_{it} is observed and is estimated by the method of principal components.

The large N permits consistent estimation of the factor and idiosyncratic components, whether or not they are $I(1)$ or $I(0)$. A large T enables application of relevant central limit theorems so that limiting distributions of the tests can be obtained.

[1] The number of factors r is supposed to be given. Recently, [Bai and Ng, 2000] have proposed to use relevant information criteria to determine the number of factors in the S&P framework.

A important aspect of PANIC is that the idiosyncratic errors can be analysed (more specifically their stationarity) without knowing if the factors are stationary and vice-versa. More precisely, the tests on the factors are asymptotically (large N and T) independent of the tests on the idiosyncratic terms.

When the idiosyncratic part is non stationary, [Bai and Ng, 2004a] recommend to deal with the firts differenced series.The rank parameter r is identified by using an information criterium like the previous one, applied for the model in first differences.

Simulations show that the proposed tests have good finite sample properties even for panels with only 40 units.

It is worth noting that the factors are estimated more efficiently from the series in levels, if the idiosyncratic components are $I(0)$. The procedure we use can be find in [Bai and Ng, 2004a] and because computing individual p-values requires simulation, for that purpose we use the table computed by S. Ng for the DF distribution.

To give an economic intrepretation of the factors extacted in the lines of [Bai and Ng, 2004a], we use the methodology presented in the paper by [Bai, 2003]. The point is to estimate a confidence interval around each of the (true) factors and check if an observed series lies or not in this interval.

3 Data and results

We first comment the main contents in the database before presenting the results.

3.1 Corporate Database

The Banque de France built 8 scores to detect corporate failures. Estimates of score functions are based on data from company balance sheets in the Banque de France's Fiben banking database[2]. This database is used to construct a pool of ratios. Some of them enter the score functions. We work on the basis of ratio and on ones of the scores themselves. In all cases, we work on the averages of ratios or scores[3].

The database of ratios covers 10 industries defined by the NES classification. It reports 91 ratios, that are usually employed in financial analysis and scoring decisions. They are estimated from the data characterizing firms over the 1989–2002 period. We formed their average on the 10 NES groups.

[2] For a complet description of the Banque de France scoring methodology, see [Bardos *et al.*, 2004].

[3] The restrictive choice of average measures can be *ex post* justified by the Fisher classification analysis employed at the Banque de France for the construction of the scores.

The statistical results can be summarized as follows, by focusing first on the stationarity properties of the series and, next, on the co-movements of these series.

3.2 The results

As some of the series exhibit a trend, we regresse them on a linear function of time and replace them by the corresponding residuals when the trend is significant. In what follows, the residuals are designed as the detrended series[4].

First, we focus on the idiosyncratic components, estimated from the first differences of the series. The test statistic $P^c_{\hat{e}}$ takes the value 19.6, which leads to reject the stationarity of the idiosyncratic components e_i. However, as the time dimension is very low ($T = 12$), we work with the level of the series to extract the factors. *A contrario*, the three extracted factors appear stationary.

According to the [Bai and Ng, 2000] criterion, 13 factors appear to be necessary to summarize the panel. Since 14 is the time dimension in the present analysis, this criterion does not appear to be relevant here. As the first three factors account for 74% of the variance (43.5% for the first one, 23.5% for the second and 7% for the third), we retain them to summarize the co-movements of the series at hand.

In addition, the contributions of the ten sectors appeared to be very homogenous[5]. It justifies the choice of implementing a global analysis and, more precisely, a business cycle analysis.

To make easier the interpretation of the contributions of the 91 ratios, we group them together in 10 financial-type ratios. The contribution of each synthetic ratio is just the sum of the contributions of the underlying ratios it summarizes. To take into account the number of contributing ratios to each synthetic ratio, we compare their effective contribution with the average contribution of all ratios. So, the first factor appears strongly associated with mark-up variables, the second one with solvability features and the third one with indebtedness characteristics.

Since the Principal Component Analysis is essentially static, the second and third factors may be lags or leads of the first factor. Correlation estimations accredit this hypothesis. The maximum correlations are achieved for a lag of 2 years for the second factor (0.89) and 4 years for the third one (0.84).

The use of the correlation procedure developed by [Bai, 2003] seems to confirme it. We should conclude that there is a unique factor, rather than

[4] The analysis was also conducted on the raw series to control the robustness of the method. It appeared that the first factor extracted in the later context looked like the trend of the series and the following factors like the factors extracted from the detrended series. Indeed, one finds high correlations between the factors of both analysis, around 0.8 (results available on request).

[5] The detailed results will be presented in the complet paper.

three, which really summarizes the co-movements of our series. But this result has to be considered with caution, because of the low time dimension of the series. Focusing on the contributions of each synthetic ratio to the factors supports this interpretation. Indeed, we can suppose that the cyclical co-movement, which we interpret as a business cycle effect (see the following paragraph), affects beforehand the profitability of the firms and, consequently, the balance sheet structure, which is weakened via the increase in debts or the degradation of stockholders' equity.

In what follows, we give a precise interpretation of the factors by comparing them to observable series that are usually considered as representative of the business cycle in France. Three macroeconomic series were chosen: the annual variation of French GDP in value, the output gap of the French GDP in volume obtained by the Hodrick-Prescott filter and the industrial production capacity utilization (TUC) as calculated by the Banque de France. We use confidence intervals in the lines of [Bai, 2003].

The results confirm the relationship between our three factors and the business cycle[6].But at the last, we have to decide if using the factors to represent business cycle is more relevant, or efficient, than using particular macroeconomic series, as the output gap, for example. It would be also interesting to investigate the possibility of using variables to partially forescast the cyclical co-movements.

Indeed, such a forecasting power would allow setting up scenarii characterizing different states of economy.

The scoring method as it is implemented up to now at the Banque de France does not cover all activity sectors and we are not sure that this limitation does not influence the results of the factorial analysis. This question is taken into account in this study.

Moreover, as we aime at measuring the influence of the business cycle on the corporate ratios and especially on the failure risk, we have to take into account information on failure. However, the failure rate in our sample is very weak -around 2 and 5%. The weight of the failing firms is consequently very weak within the sample. Their influence on the detection of business cycle could be out of measure. To overcome this difficulty, we performe weighted averages for each sector/ratio.

To refine the diagnosis, we also use separately the samples of failing and non-failing firms. The PCA gives the results summarized in the following table. In spite of the rebalancing of sample, the contributions of the first three factors for the failing firms significantly differ from those obtained for the other firms.

In order to investigate if the restriction of the database or/and the distinction between failing and non-failing firms modified the results obtained over the whole sample, we projecte the first factors of every partial ACP on the space spanned by the first three factors stemming from the complete

[6] The detailed results and graphs will be presented in the complet paper.

sample. The results indicate a very strong similarity of the common factors[7], indicating that the state of economy influences failing and non-failing firms in a similar way, whaterver the sector considered.

After extracting common cyclical facors from corporate, we focuse on the score functions themselves. Indeed, nothing says that the cyclical components are detectable from the scores, which have been precisely adapted to be immunized against cyclical variations.

The same plan of study as previously was thus applied to the averages of the score functions. To increase the cross-section dimension of the sample, the desagregation is made according to the NAF classification which is finer than the NES classification used before.

We had thus a sample of 8 averaged score functions over the same time period (1989 to 2003) for 49 sectors. So it includes 392 variables over 14 years. We then apply the principal components estimation on the 392x392 dimensional matrix. The first three factors account for 58.8% of the variance (33.6% for the first one, 17.5% for the second and 7.7% for the third). To compare these factors with the first three factors obtained before, we use again confidence intervals estimated in the lines of Bay [2003].

The results[8] show that the cyclical component is common to the ratios and the function scores. The projection of each height average score on the same space has come to the same conclusion.

The functions scores do not modify substantially the cyclical common component which is present in the original ratios.

So, we have to conclude that we should take into account the information about the cycle of activity to improve modeling of the corporate failure risk.

To finish, a question remains un-answered: how to implement scoring so as to account for cyclical variations? Have we just to adjust the decision thresholds or should we modify the score functions according to the state of economy? To give a first answer to these questions, we compute, for every year and for every score, the optimal decision threshold. Optimality equalizes both first and second type errors. Thus, we regresse the different optimal thresholds onto the first three factors extracted from the scores database.

What do we observe? The decision threesholds are significantly correlated with the factors most of the time, as they generally belong to the confidence interval around each factor. But it is not always the case, so that we have to conclude that the analysis should be deepened in order to decide how to modify the scoring decision and more precisely the threeshold, so as to include the cyclical variations observed in the score functions.

[7] The detailed results and graphs will be presented in the complet paper.
[8] The detailed results and graphs will be presented in the complet paper.

4 Conclusion

To summarize, we have to claim that the corporate ratios of the firms in our data base display significant cyclical comovements, which are quite similar when they are extracted from the whole corporate database and from the database just including the firms for which a score function has been estimated up to now in Banque de France. Moerover, the score functions themselves display the same cyclical comovements and, as a consequence, we have to conclude that these score functions are not immunized against cyclical variations in the state of economy.

Indeed, macroeconomic series, that are usually recognized as proxy variables to characterize the cyclical behavior of the French economy, appear to be significantly correlated with the estimated factors extracted from the database of the financial ratios as well as from the database of the scores.

Finally, we examine what kind of consequence this dependency of the score functions on the cyclical movements in economy may have on the scoring decision process itself.

We compared the decision thresholds to the estimated factors for each sector. We observed that the thresholds are correlated with the factors most of the time. Accordingly, the threshold should be vary over time like the underlying cyclical components in order to improve the scoring procedure. However, there are subperiods where the correlation disappears, indicating that changing the threshold is not always sufficient to account for changes in the economic environment. So, the analysis has to be deepened and this is left for further research.

References

[Bai and Ng, 2000]J. Bai and S. Ng. Determining the number of factors in approximate factor models. *Working Paper*, 2000.

[Bai and Ng, 2004a]J. Bai and S. Ng. A PANIC attack on unit roots and cointegration. *Econometrica*, pages 1127–1177, 2004a.

[Bai and Ng, 2004b]J. Bai and S. Ng. Evaluating latent and observed factors in macroeconomics and finance. *Working Paper*, 2004b.

[Bai, 2003]J. Bai. Inferential theory for factor models of large dimensions. *Econometrica*, pages 135–172, 2003.

[Bardos *et al.*, 2004]M. Bardos, Foulcher S., and E. Bataille. Banque de France scores: methode, results and applications. *Banque de France*, 2004.

[Bruneau *et al.*, 2002a]C. Bruneau, O. de Bandt, A. Flageollet, and E. Michaux. Forecasting inflation using economic indicators: the case of france. *NER - Banque de France*, 2002a.

[Chamberlain and Rothschild, 1983]G. Chamberlain and M. Rothschild. Arbitrage, factor structure, and mean-variance analysis on large asset markets. *Econometrica*, pages 1281–1304, 1983.

[Stock and Watson, 1998]J.H. Stock and M.W. Watson. Testing for common trends. *Journal of the American Statistical Association*, pages 1097–1107, 1998.

Part VII

Data Analysis

A Solution to the Discrete-Time Linear Prediction Problem Using PCA*

Rosa Fernández-Alcalá, Jesús Navarro-Moreno, and Juan C. Ruiz-Molina

Department of Statistics and Operations Research
University of Jaén
23071 Jaén, Spain
(e-mail: rmfernan@ujaen.es, jnavarro@ujaen.es, jcruiz@ujaen.es)

Abstract. A recursive solution is given to the linear one-stage prediction problem in discrete-time systems involving correlated signal and noise. Using Principal Component Analysis of stochastic processes, a suboptimum filter is designed. The main advantage of this solution is that it can be computed through a Kalman-like filter in those situations in which the signal does not verify a state-space model. The efficiency of the proposed methodology lies in the possibility of representing adequately the processes involved by a sample of points not excessively large.

Keywords: Linear Prediction Problem, PCA.

1 Introduction

In this paper we treat the discrete linear one-stage prediction problem involving correlated signal and noise. This estimation problem is useful in applications to feedback control and feedback communications. Thus, let $\{x(t_i), t_1 \leq t_i \leq t_n\}$ be a signal process which is a real second-order stochastic process, with zero-mean and correlation function R_x. Let $\{z(t_i), t_1 \leq t_i \leq t_n\}$ be a second-order stochastic process with zero-mean and correlation function R_z.

We assume that the signal process is observed corrupted by an additive white noise through the equation

$$y(t_i) = x(t_i) + v(t_i), \qquad t_1 \leq t_i \leq t_n$$

where $v(t_i)$ is a zero-mean white noise process with $E[v(t_i)v(t_j)] = r_i\delta_{ij}$ and correlated with both the signal $x(t_i)$ and the process $z(t_i)$. Let $R_{x_1x_2}(t_i, t_j)$ denote the correlation function between any two processes $x_1(t_i)$ and $x_2(t_j)$.

Under the above hypotheses, we consider the problem of finding the linear minimum variance estimator $\hat{z}(t_{k+1}/t_k)$ of the process $z(t_{k+1})$, based on the set of observations $\{y(t_1), \ldots, y(t_k)\}$, with $k < n$.

[*] This work was supported in part by Project MTM2004-04230 of the Plan Nacional de I+D+I, Ministerio de Educación y Ciencia, Spain. This project is financed jointly by the FEDER.

According to the projection theorem, this element $\hat{z}(t_{k+1}/t_k)$ exists, is unique and can be expressed as a linear transform of the observations set $\{y(t_1), \ldots, y(t_k)\}$ of the form [Poor, 1994]

$$\hat{z}(t_{k+1}/t_k) = \mathbf{h}'_k(t_{k+1})\mathbf{y}_k \tag{1}$$

where $\mathbf{y}_k = [y(t_1), \ldots, y(t_k)]'$ and the vector of optimum coefficients $\mathbf{h}_k(t_{k+1}) = [h_1(t_{k+1}), \ldots, h_k(t_{k+1})]'$ satisfies the Wiener-Hopf equation

$$\boldsymbol{\sigma}_k(t_{k+1}) = \boldsymbol{\Sigma}_{k\times k}(t_k)\mathbf{h}_k(t_{k+1}) \tag{2}$$

where $\boldsymbol{\sigma}_k(t_{k+1}) = [R_{zy}(t_{k+1}, t_1), \ldots, R_{zy}(t_{k+1}, t_k)]'$, with $R_{zy}(t_{k+1}, t_i) = R_{zx}(t_{k+1}, t_i) + R_{zv}(t_{k+1}, t_i)$, and $\boldsymbol{\Sigma}_{k\times k}(t_k)$ is the correlation matrix of the vector $\mathbf{y}_k$ whose elements are $R_y(t_i, t_j) = R_x(t_i, t_j) + R_{xv}(t_i, t_j) + R_{vx}(t_i, t_j) + r_i\delta_{ij}$.

Then, the estimation problem is basically that of solving the equation (2) involving the correlation functions of the signal process and the process to be estimated. In principle, this equation is easy to solve and its solution is given by

$$\mathbf{h}_k(t_{k+1}) = \boldsymbol{\Sigma}^{-1}_{k\times k}(t_k)\boldsymbol{\sigma}_k(t_{k+1}) \tag{3}$$

Unfortunately, from the practical point of view, the determination of these optimum coefficients through the equation (3) can lead to a computational difficulty since the inversion of the matrix $\boldsymbol{\Sigma}_{k\times k}(t_k)$ makes that the number of basic computational operations grows linearly with the number of observations considered.

Recently, an extensive literature concerning the design of a more efficient computational procedure has been developed. One of the most used techniques consists in imposing additional structural conditions on the correlations involved such as, stationarity [Poor, 1994], state-space models which lead to the Kalman filter [Kalman and Bucy, 1961], semi-degenerate kernel forms [Sugisaka, 1983], among others. Although this approach is widely applied, there is a great number of physical phenomena that do not satisfy these assumptions. In these situations, an alternative methodology is possible by using Principal Component Analysis (PCA) of stochastic processes [Aguilera *et al.*, 1995, Aguilera *et al.*, 1996].

In this paper, we propose a new recursive one-stage prediction procedure following this second perspective. In this framework, by considering any truncated series representation for the involved processes in terms of their principal components, the vector of optimum coefficients (3), and then the optimum one-stage predictor (1), can be approximated. Although a suboptimum one-stage predictor is provided, the main advantage of this via of solution is that it can be efficiently computed through a recursive algorithm without imposing any structural assumption on the processes involved. In fact, they can be applied under the only hypothesis that the involved correlation functions are known. This occurs frequently in applications to system identification problems or in statistical communication theory, where the relevant statistics of the problem are initially known in terms of correlation functions derived from measurements or mathematical models [Gardner and Franks, 1971]. In particular, these results can be applied in detection problems [Kailath, 1970] and in feedback communication systems [Gardner, 1975].

Then, the rest of the paper is structured as follows. In the next subsection, a brief description about the orthogonal representation of a stochastic process in terms of its principal components is included. The main characteristic of these series expansions is that they allow us to represent adequately a process through a short number of terms. Next, in Section 2, a new methodology based on these series representations is developed with the aim of designing a suboptimum one-stage predictor which can be efficiently computed through a Kalman-like recursive algorithm.

1.1 Approximate Series Expansions Using PCA

Let us consider the random vector $\mathbf{z}_{2n} = [z(t_1), \ldots, z(t_n), x(t_1), \ldots, x(t_n)]'$.

Let $\mathbf{a}_{2n}(i) = [a_1(i), \ldots, a_{2n}(i)]'$ and λ_i denote the principal values and the principal factors, respectively. Let also b_i be the principal components obtained from the principal factors as $b_i = \mathbf{a}'_{2n}(i)\mathbf{z}_{2n}$[1].

Then, $\mathbf{z}_{2n}$ admits the following orthogonal representation in terms of its principal components:

$$\mathbf{z}_{2n} = \sum_{i=1}^{2n} \mathbf{a}_{2n}(i) b_i \tag{4}$$

Moreover, this representation is optimal in the sense of being the best $2n$-dimensional linear model for $\mathbf{z}_{2n}$ in the least squares sense [Fukunaga and Koontz, 1970].

From (4), we have that the processes $z(t_j)$ and $x(t_j)$ can be expressed through finite series expansions in terms of their principal components as follows:

$$z(t_j) = \sum_{i=1}^{2n} a_j(i) b_i, \quad x(t_j) = \sum_{i=1}^{2n} a_{j+n}(i) b_i, \qquad j = 1, \ldots, n$$

On the other hand, the correlation functions involved in (2) can be expressed by the following product of matrices:

$$\begin{aligned} R_x(t_k, t_j) &= \mathbf{d}'_{2n}(t_k) \boldsymbol{\Lambda}_{2n\times 2n} \mathbf{d}_{2n}(t_j) \\ R_{xv}(t_k, t_j) &= \mathbf{d}'_{2n}(t_k) \mathbf{f}_{2n}(t_j) \\ R_{zx}(t_k, t_j) &= \mathbf{c}'_{2n}(t_k) \boldsymbol{\Lambda}_{2n\times 2n} \mathbf{d}_{2n}(t_j) \\ R_{zv}(t_k, t_j) &= \mathbf{c}'_{2n}(t_k) \mathbf{f}_{2n}(t_j) \end{aligned} \tag{5}$$

[1] Note that, $E[b_i] = 0$ and $E[b_i b_j] = \lambda_i \delta_{ij}$.

where $\mathbf{d}_{2n}(t_j) = [a_{j+n}(1), \ldots, a_{j+n}(2n)]'$, $\mathbf{c}_{2n}(t_j) = [a_j(1), \ldots, a_j(2n)]'$, $\boldsymbol{\Lambda}_{2n\times 2n}$ is the $2n$-dimensional diagonal matrix whose elements are the principal values λ_i, and $\mathbf{f}_{2n}(t_j)$ is the $2n$-dimensional vector with elements $f_i(t_j) = E[v(t_j)b_i]$, for $i = 1, \ldots, 2n$.

Finally, note that a suitable representation of any stochastic process is possible without taking all the samples but that it is sufficient to select an adequate subset of them [Fukunaga and Koontz, 1970]. Then, we can select $m < n$ instants of times, $t_1 \leq t_{i_1} < t_{i_2} <, \ldots, < t_{i_m} < t_n$, and consider the vector $[z(t_{i_1}), \ldots, z(t_{i_m}), x(t_{i_1}), \ldots, x(t_{i_m})]'$. Next, using the principal values $\tilde{\lambda}_i$, the principal factors $\tilde{\mathbf{a}}_{2m}(i) = [\tilde{a}_1(i), \ldots, \tilde{a}_{2m}(i)]'$ and the principal components $\tilde{b}_i$ associated with this vector, $\mathbf{z}_{2n}$ can be approximated by the series expansion

$$\mathbf{z}_{2n} \approx \tilde{\mathbf{z}}_{2n} = \sum_{i=1}^{2m} \tilde{\mathbf{g}}_{2n}(i)\tilde{b}_i \tag{6}$$

where $\tilde{\mathbf{g}}_{2n}(i)$ is a $2n$-dimensional vector whose elements are of the form

$$\tilde{g}_j(i) = \frac{1}{\tilde{\lambda}_i} E\left[z(t_j)\tilde{b}_i\right] = \frac{1}{\tilde{\lambda}_i} \sum_{k=1}^{m} \left(\tilde{a}_k(i)R_z(t_j, t_{i_k}) + \tilde{a}_{m+k}(i)R_{zx}(t_j, t_{i_k})\right)$$

$$\tilde{g}_{j+n}(i) = \frac{1}{\tilde{\lambda}_i} E\left[x(t_j)\tilde{b}_i\right] = \frac{1}{\tilde{\lambda}_i} \sum_{k=1}^{m} \left(\tilde{a}_k(i)R_{xz}(t_j, t_{i_k}) + \tilde{a}_{m+k}(i)R_x(t_j, t_{i_k})\right)$$

for $j = 1, \ldots, n$.

The main advantage of the series expansion (6) with respect to (4) is the reduction of the computational burden. In fact, the amount of computation required depends on the number of points selected, m, and a criterion for determining a suitable m can be found in [Fukunaga and Koontz, 1970].

Now, the processes $z(t_j)$ and $x(t_j)$ can be approximated by finite series expansions with less number of terms as follows:

$$z(t_j) \approx z_m(t_j) = \sum_{i=1}^{2m} \tilde{g}_j(i)\tilde{b}_i, \quad x(t_j) \approx x_m(t_j) = \sum_{i=1}^{2m} \tilde{g}_{j+n}(i)\tilde{b}_i, \quad j = 1, \ldots, n$$

Moreover, the correlation functions given in (5) can be approximated by the product of matrices of reduced dimension. Specifically,

$$\begin{aligned} R_x(t_k, t_j) \approx R_{x_m}(t_k, t_j) =& \tilde{\mathbf{d}}'_{2m}(t_k)\tilde{\boldsymbol{\Lambda}}_{2m\times 2m}\tilde{\mathbf{d}}_{2m}(t_j) \\ R_{xv}(t_k, t_j) \approx R_{x_m v}(t_k, t_j) =& \tilde{\mathbf{d}}'_{2m}(t_k)\tilde{\mathbf{f}}_{2m}(t_j) \\ R_{zx}(t_k, t_j) \approx R_{z_m x_m}(t_k, t_j) =& \tilde{\mathbf{c}}'_{2m}(t_k)\tilde{\boldsymbol{\Lambda}}_{2m\times 2m}\tilde{\mathbf{d}}_{2m}(t_j) \\ R_{zv}(t_k, t_j) \approx R_{z_m v}(t_k, t_j) =& \tilde{\mathbf{c}}'_{2m}(t_k)\tilde{\mathbf{f}}_{2m}(t_j) \end{aligned} \tag{7}$$

where $\tilde{\mathbf{d}}_{2m}(t_j) = [\tilde{g}_{j+n}(1), \ldots, \tilde{g}_{j+n}(2m)]'$, $\tilde{\mathbf{c}}_{2m}(t_j) = [\tilde{g}_j(1), \ldots, \tilde{g}_j(2m)]'$, $\tilde{\boldsymbol{\Lambda}}_{2m\times 2m}$ is the $2m$-dimensional diagonal matrix with i-th entry $\tilde{\lambda}_i$, and

$\tilde{\mathbf{f}}_{2m}(t_j)$ is the $2m$-dimensional vector with elements $\tilde{f}_i(t_j)$, for $i = 1, \ldots, 2m$, of the form

$$\tilde{f}_i(t_j) = E[v(t_j)\tilde{b}_i] = \sum_{k=1}^{m} \left(\tilde{a}_k(i) R_{vz}(t_j, t_{i_k}) + \tilde{a}_{m+k}(i) R_{vx}(t_j, t_{i_k}) \right)$$

2 Suboptimum Predictor

In this section, a recursive suboptimum solution to the linear least mean-square one-stage prediction problem in discrete-time systems involving correlated signal and noise is devised. For that, the following approximate version of (2) is considered by taking the approximate representations (7) for the correlation functions involved:

$$\tilde{\boldsymbol{\sigma}}_k(t_{k+1}) = \tilde{\boldsymbol{\Sigma}}_{k \times k}(t_k)\tilde{\mathbf{h}}_k(t_{k+1}) \tag{8}$$

where $\tilde{\boldsymbol{\sigma}}_k(t_{k+1}) = [R_{z_m y_m}(t_{k+1}, t_1), \ldots, R_{z_m y_m}(t_{k+1}, t_k)]'$, and $\tilde{\boldsymbol{\Sigma}}_{k \times k}(t_k)$ is the correlation matrix of $\tilde{\mathbf{y}}_k = [y_m(t_1), \ldots, y_m(t_k)]'$, with $y_m(t_i) = x_m(t_i) + v(t_i)$.

From (7), we obtain that

$$\tilde{\boldsymbol{\sigma}}_k(t_{k+1}) = \mathbf{L}_{k \times 4m}(t_k)\mathbf{A}_{4m \times 4m}\mathbf{q}_{4m}(t_{k+1})$$

and

$$\tilde{\boldsymbol{\Sigma}}_{k \times k}(t_k) = \mathbf{L}_{k \times 4m}(t_k)\mathbf{A}_{4m \times 4m}\mathbf{L}'_{k \times 4m}(t_k) + \mathbf{R}_{k \times k}(t_k)$$

where $\mathbf{q}_{4m}(t_k) = [\tilde{\mathbf{c}}'_{2m}(t_k), \mathbf{0}'_{2m}]'$, with $\mathbf{0}_{2m}$ the $2m$-dimensional vector whose elements are zero, $\mathbf{L}_{k \times 4m}(t_k) = \left[\mathbf{D}'_{2m \times k}(t_k), \mathbf{F}'_{2m \times k}(t_k)\right]$ with $\mathbf{D}_{2m \times k}(t_k) = \left[\tilde{\mathbf{d}}_{2m}(t_1), \ldots, \tilde{\mathbf{d}}_{2m}(t_k)\right]$ and $\mathbf{F}_{2m \times k}(t_k) = \left[\tilde{\mathbf{f}}_{2m}(t_1), \ldots, \tilde{\mathbf{f}}_{2m}(t_k)\right]$, $\mathbf{R}_{k \times k}(t_k)$ is a diagonal matrix with i-th entry r_i, and

$$\mathbf{A}_{4m \times 4m} = \left[\begin{array}{c|c} \tilde{\boldsymbol{\Lambda}}_{2m \times 2m} & \mathbf{I}_{2m \times 2m} \\ \hline \mathbf{I}_{2m \times 2m} & \mathbf{0}_{2m \times 2m} \end{array} \right]$$

being $\mathbf{I}_{2m \times 2m}$ the $2m \times 2m$-dimensional identity matrix and $\mathbf{0}_{2m \times 2m}$ the $2m \times 2m$-dimensional matrix with zero elements.

Then, the solution of (8) is of the form

$$\begin{aligned} \tilde{\mathbf{h}}_k(t_{k+1}) = \left[\mathbf{L}_{k \times 4m}(t_k)\mathbf{A}_{4m \times 4m}\mathbf{L}'_{k \times 4m}(t_k) + \mathbf{R}_{k \times k}(t_k)\right]^{-1} \\ \times \mathbf{L}_{k \times 4m}(t_k)\mathbf{A}_{4m \times 4m}\mathbf{q}_{4m}(t_{k+1}) \end{aligned} \tag{9}$$

From (9) we can define the suboptimum one-stage predictor

$$\hat{z}_m(t_{k+1}/t_k) = \tilde{\mathbf{h}}'_k(t_{k+1})\mathbf{y}_k \tag{10}$$

At a first sight, in comparison with the optimum one-stage predictor (1), the proposed solution (10) does not show an improvement from the computational standpoint since both estimates require the computation of the product of two k-dimentional vectors. However, the suboptimum coefficients (9) lead to a reduction in the computational burden with respect to solving directly the Wiener-Hopf equation.

In the following result, a recursive algorithm similar to the Kalman filter is designed for the computation of the proposed suboptimum one-stage predictor (10).

Theorem 1

$$\hat{z}_m(t_{k+1}/t_k) = \mathbf{q}'_{4m}(t_{k+1})\mathbf{e}_{4m}(t_k) \tag{11}$$

where $\mathbf{e}_{4m}(t_k)$ *is recursively computed through the equation*

$$\mathbf{e}_{4m}(t_k) = \mathbf{e}_{4m}(t_{k-1}) + \mathbf{k}_{4m}(t_k)\left[y(t_k) - \mathbf{l}'_{4m}(t_k)\mathbf{e}_{4m}(t_{k-1})\right]$$

with the initialization $\mathbf{e}_{4m}(t_0) = \mathbf{0}_{4m}$*, and where* $\mathbf{l}'_{4m}(t_k) = \left[\tilde{\mathbf{d}}'_{2m}(t_k), \tilde{\mathbf{f}}'_{2m}(t_k)\right]$ *and the vector* $\mathbf{k}_{4m}(t_k)$ *is given by*

$$\mathbf{k}_{4m}(t_k) = \mathbf{P}_{4m\times 4m}(t_{k-1})\mathbf{l}_{4m}(t_k)\left[\mathbf{l}'_{4m}(t_k)\mathbf{P}_{4m\times 4m}(t_{k-1})\mathbf{l}_{4m}(t_k) + r_j\right]^{-1} \tag{12}$$

with

$$\mathbf{P}_{4m\times 4m}(t_k) = \mathbf{P}_{4m\times 4m}(t_{k-1}) - \mathbf{k}_{4m}(t_k)\mathbf{l}'_{4m}(t_k)\mathbf{P}_{4m\times 4m}(t_{k-1})$$

where $\mathbf{P}_{4m\times 4m}(t_0) = \mathbf{A}_{4m\times 4m}$.

Proof. From (9), we have that the suboptimum one-stage predictor (10) is given by

$$\hat{z}_m(t_{k+1}/t_k) = \mathbf{q}'_{4m}(t_{k+1})\mathbf{A}_{4m\times 4m}\mathbf{L}'_{k\times 4m}(t_k) \times \left[\mathbf{L}_{k\times 4m}(t_k)\mathbf{A}_{4m\times 4m}\mathbf{L}'_{k\times 4m}(t_k) + \mathbf{R}_{k\times k}(t_k)\right]^{-1}\mathbf{y}_k$$

Then, introducing the vector

$$\mathbf{e}_{4m}(t_k) = \mathbf{A}_{4m\times 4m}\mathbf{L}'_{k\times 4m}(t_k) \times \left[\mathbf{L}_{k\times 4m}(t_k)\mathbf{A}_{4m\times 4m}\mathbf{L}'_{k\times 4m}(t_k) + \mathbf{R}_{k\times k}(t_k)\right]^{-1}\mathbf{y}_k \tag{13}$$

the equation (11) for $\hat{z}_m(t_k)$ is obtained.

Next, applying the matrix inversion lemma [Anderson and Moore, 1979, p. 138] in (13), we have that

$$\mathbf{e}_{4m}(t_k) = \mathbf{P}_{4m\times 4m}(t_k)\mathbf{L}'_{k\times 4m}(t_k)\mathbf{R}^{-1}_{k\times k}(t_k)\mathbf{y}_k$$

where

$$\mathbf{P}_{4m\times 4m}(t_k) = \left[\mathbf{A}^{-1}_{4m\times 4m} + \mathbf{L}'_{k\times 4m}(t_k)\mathbf{R}^{-1}_{k\times k}(t_k)\mathbf{L}_{k\times 4m}(t_k)\right]^{-1} \tag{14}$$

Finally, taking into account the matrix inversion lemma in (14) and defining the vector $\mathbf{k}_{4m}(t_k)$ as in (12), the theorem holds.

Remark 1 *Note that, from* (5)*, a similar recursive algorithm can be designed for the optimum one-stage predictor. However, the amount of computation required with the resulting recursive formulas makes that this algorithm loss interest in practical applications.*

Remark 2 *From the PCA, the convergence of the proposed suboptimum predictor toward the optimum one is guaranteed. Then, the suboptimum one-stage predictor becomes a better approximation of the optimum one as the number m increases. On the other hand, a suitable m must be selected in order to reduce the computational burden. In fact, the efficiency of the proposed suboptimum estimate will be more relevant when the signal can be represented by a short series expansion. Some examples of such signals can be found in [Ghanem and Spanos, 1991].*

References

[Aguilera *et al.*, 1995]A.M. Aguilera, R. Gutiérrez, F.A. Ocaña, and M.J. Valderrama. Computational Approaches to Estimation in the Principal Component Analysis of a Stochastic Process. *Applied Stochastic Models and Data Analysis*, 11(4):279–299, 1995.

[Aguilera *et al.*, 1996]A.M. Aguilera, R. Gutiérrez, and M.J. Valderrama. Approximation of Estimators of the PCA of a Stochastic Process Using B-Splines. *Communications in Statistics (Simulation and Computation)*, 25(3):671–690, 1996.

[Anderson and Moore, 1979]B.D.O. Anderson and J.B. Moore. *Optimal Filtering.* Prentice-Hall, New Jersey, 1979.

[Fukunaga and Koontz, 1970]K. Fukunaga and L.G. Koontz. Representation of Random Processes Using the Finite Karhunen-Loève Expansion. *Informat. and Control*, 16:85–101, 1970.

[Gardner and Franks, 1971]W.A. Gardner and L.E. Franks. An Alternative Approach to Linear Least Squares Estimation of Continuous Random Processes. In *5th Ann. Princeton Conf. Information Sciences and Systems*, 1971.

[Gardner, 1975]W.A. Gardner. A Series Solution to Smoothing, Filtering, and Prediction Problems Involving Correlated Signal and Noise. *IEEE, Trans. Information Theory*, IT-21:698–699, 1975.

[Ghanem and Spanos, 1991]R.G. Ghanem and P.D. Spanos. *Stochastic Finite Elements: A Spectral Approach.* Springer-Verlag, New York, 1991.

[Kailath, 1970]T. Kailath. Likelihood Ratios for Gaussian Processes. *IEEE, Trans. Information Theory*, 16(3):276–288, 1970.

[Kalman and Bucy, 1961]R. E. Kalman and R. S. Bucy. New Results in Linear Filtering and Prediction Theory. *Trans. ASME, J. Basic Engineering, Ser. D*, 83:95–108, 1961. In: Epheremides, A. and Thomas, J.B. (Ed.) 1973. Random Processes. Multiplicity Theory and Canonical Decompositions.

[Poor, 1998]H.V. Poor. *An Introduction to Signal Detection and Estimation.* Springer-Verlag, New York, 2nd edition, 1998.

[Sugisaka, 1983]M. Sugisaka. The Design of On-line Least-Squares Estimators Given Covariance Specifications Via an Imbedding Method. *Applied Mathematics and Computation*, (13):55–85, 1983.

Total Least Squares for Functional Data

Christophe Crambes

Université Paul Sabatier
Laboratoire de Statistique et Probabilités
UMR C5583
118 route de Narbonne
31062 Toulouse Cedex, France
(e-mail:crambes@cict.fr)

Abstract. We are interested in the functional linear regression when the covariates are subject to errors, for instance measurement errors. The aim of this paper is to propose a procedure giving a spline estimator of the functional coefficient of the model with noisy covariates. The functional coefficient is the solution of an ill-conditioned minimization problem, so a penalization approach is used. Indeed, we present an extension of the penalized total least squares algorithm to the case where the covariates are curves. Then, this estimation procedure is evaluated by the way of simulations.

Keywords: functional linear regression, errors-in-variables, total least squares, penalization, spline functions.

1 Introduction

In many fields of applications, it is frequent to deal with the problem of the explanation of a random variable Y (response), usually scalar, using information from a random variable X (covariate), belonging to some Hilbert space E. Then, a way to formulate this problem is to consider the linear regression of Y on X that, in case of existence and unicity, allows us to write

$$Y = \mu + \langle \alpha, X \rangle + \epsilon, \tag{1}$$

where $\langle .,. \rangle$ stands for the inner product of the Hilbert space E and ϵ is a real random variable satisfying $\mathbb{E}(\epsilon) = 0$ and $\mathbb{E}(\epsilon X) = 0$. Implicitly, in (1), the variable X is supposed to be observed without error, and all errors are into the variable Y by the way of ϵ. However, in practice, this assumption seems to be quite unrealistic, for example because of instrument measurement errors. That is why it should be natural to consider that the variable X is not directly observed, but we observe instead a variable W such that

$$W = X + \delta. \tag{2}$$

In the case where E is $\mathbb{R}$ or $\mathbb{R}^p$, that is to say when X is an univariate or a multivariate random variable, this problem of *errors-in-variables* model has already been studied. Some theoretical approaches have been proposed, using the maximum likelihood method (see [Fuller, 1987]) or deconvolution techniques (see [Carroll *et al.*, 1995]). A practical point of view is given under the name of *Total Least Squares* (TLS) in [Van Huffel and Vandewalle, 1991]. However, in many fields of applications (chemistry, climatology, teledetection, linguistics, ...), the data do not belong to the frame of univariate or multivariate variables. Indeed, the data can come from the observation of continuous phenomenons (that is to say continuous functions of time, space, ...), then they are comparable to curves. These data, called *functional data* in the literature, are the object of many studies (see [Ramsay and Silverman, 1997] and [Ramsay and Silverman, 2002] for a functional data analy! sis overview). Our goal is to study the problem of *errors-in-variables* model in the framework where X is a functional random variable, in other words when E is an infinite dimension space.

In the following, we consider n couples of random variables $(X_i, Y_i)_{i=1,\dots,n}$ independant and identically distributed, with the same distribution as (X, Y), where X is a random variable taking values in some functional space E and Y belongs to $\mathbb{R}$. For sake of simplicity, we consider that E is the space $L^2(I)$ of the functions of square integrable defined on an interval I of $\mathbb{R}$. We still denote by $\langle .,. \rangle$ the usual inner product of $L^2(I)$ and by $\|.\|$ the associated norm. We rewrite (1) taking the point of view of the *functional linear regression* introduced in [Ramsay and Dalzell, 1993], hence we assume that

$$Y = \mu + \int_I \alpha(t)X(t)\, dt + \epsilon, \tag{3}$$

where $\mu \in \mathbb{R}$ and $\alpha \in L^2(I)$ are the unknown parameters of the model and ϵ is a real random variable such that $\mathbb{E}(\epsilon) = 0$ and $\mathbb{E}(\epsilon X) = 0$. We assume conditions for existence and unicity of α (see [Cardot *et al.*, 2003]). Let us remark that, if we denote by Γ_X the covariance operator of X (defined by $\Gamma_X u = \mathbb{E}\left(\langle X - \mathbb{E}(X), u\rangle (X - \mathbb{E}(X))\right)$ for all $u \in L^2(I)$) and by Δ_{XY} the cross covariance operator of X and Y (defined by $\Delta_{XY} u = \mathbb{E}\left(\langle X - \mathbb{E}(X), u\rangle Y\right)$ for all $u \in L^2(I)$), then we easily see that $\langle \Gamma_X \alpha, u\rangle = \Delta_{XY} u$ for all function $u \in L^2(I)$. One of the properties of Γ_X is that it is a nuclear operator (see [Loève, 1963] for details). So Γ_X^{-1} is not bounded and estimation of ! α is an *ill-conditioned* problem. A possibility to deal with this problem is to introduce a penalization approach (this is done in [Cardot *et al.*, 2003]), and to find μ and α as solutions of the minimization problem

$$\min_{\mu \in \mathbb{R}, \alpha \in L^2(I)} \left\{ \frac{1}{n} \sum_{i=1}^{n} \left(Y_i - \mu - \langle \alpha, X_i \rangle\right)^2 + \rho \left\| \alpha^{(m)} \right\|^2 \right\}, \tag{4}$$

where $\alpha^{(m)}$ stands for the derivative of order m of the function α and ρ is a smoothing parameter allowing to control the regularity of the estimator of the function α.

Now coming back to our *errors-in-variables* setting, we suppose that the curve X is not directly available. In practice, the curves $X_1, \ldots, X_n$ are observed in p discretization points $t_1, \ldots, t_p \in I$ such that $t_1 \leq \ldots \leq t_p$. So, the data are

$$W(t_j) = X(t_j) + \delta(t_j), \quad j = 1, \ldots, p, \tag{5}$$

where $(\delta(t_j))_{j=1,\ldots,p}$ is a sequence of real random variables independent and identically distributed, centered and with variance σ_δ^2. We also assume that $\delta(t_j)$ and ϵ are independant for all $j = 1, \ldots, p$. These variables represent the error made on X at each measure point. The random variables W and δ give us the corresponding samples $(W_i)_{i=1,\ldots,n}$ and $(\delta_i)_{i=1,\ldots,n}$. The aim of this paper is to build an estimator of μ and α. In section 2, we generalize the TLS algorithm to our functional framework. In section 3, this estimator is evaluated by the way of simulations. Finally in section 4, we make some concluding remarks.

2 Functional Total Least Squares

The aim of this section is to adapt the *Total Least Squares* algorithm introduced in [Van Huffel, 2004] when the covariate X is of functional nature.

2.1 Total Least Squares in the multivariate case

When X is a multivariate random variable, the linear regression is written

$$Y = \mu + {}^t\mathbf{X}\boldsymbol{\alpha} + \epsilon, \tag{6}$$

where $\mathbf{X} = {}^t(X_1, \ldots, X_p)$ belongs to $\mathbb{R}^p$. We have to estimate $\mu \in \mathbb{R}$ and $\alpha \in \mathbb{R}^p$, assuming we observe Y_i and $\mathbf{W}_i = \mathbf{X}_i + \boldsymbol{\delta}_i$ for $i = 1, \ldots, n$. We denote by $\mathbf{Y}$ the vector ${}^t(Y_1, \ldots, Y_n)$, $\boldsymbol{\epsilon}$ the vector ${}^t(\epsilon_1, \ldots, \epsilon_n)$, $\mathbf{X}$ and $\mathbf{W}$ the matrices of respective elements X_{ij} and W_{ij}. Under an hypothesis of normality for the errors (that is to say if $\epsilon \sim \mathcal{N}(0, \sigma_\epsilon^2)$ and $\delta(t_j) \sim \mathcal{N}(0, \sigma_\delta^2)$ for all $j = 1, \ldots, p$), the likelihood function is proportional to

$$\exp\left\{ -\sum_{i=1}^n \left[\frac{1}{\sigma_\epsilon^2} \left(Y_i - \mu - {}^t\mathbf{X}_i\boldsymbol{\alpha}\right)^2 + \frac{1}{\sigma_\delta^2}\, {}^t(\mathbf{X}_i - \mathbf{W}_i)(\mathbf{X}_i - \mathbf{W}_i) \right] \right\}. \tag{7}$$

Without any more condition, the model (6) with $\mathbf{W}_i = \mathbf{X}_i + \boldsymbol{\delta}_i$ is not identifiable and another condition needs to be imposed (see [Van Huffel, 2004]). In the following, we choose to assume that the ratio of the variances $\sigma_\epsilon^2/\sigma_\delta^2$ is known. Indeed, we can suppose that this ratio is equal to 1 (if the ratio is $\eta = \sigma_\epsilon^2/\sigma_\delta^2$, we consider the scaled variable $\widetilde{\mathbf{X}} = \sqrt{\eta}\mathbf{X}$ and then $\boldsymbol{\alpha} = \sqrt{\eta}\widetilde{\boldsymbol{\alpha}}$).

Then, the maximization of (7) comes back to the resolution of the minimization problem

$$\min_{\mu\in\mathbb{R},\boldsymbol{\alpha}\in\mathbb{R}^p,\mathbf{X}_i\in\mathbb{R}^p}\left\{\frac{1}{n}\sum_{i=1}^{n}\Big[(Y_i-\mu-\mathbf{X}_i\boldsymbol{\alpha})^2+\ {}^t(\mathbf{X}_i-\mathbf{W}_i)(\mathbf{X}_i-\mathbf{W}_i)\Big]\right\}. \quad (8)$$

The TLS algorithm given in [Van Huffel, 2004] follows the two steps below:

- step 1: we make the singular value decomposition (SVD) of the matrix $[\,\mathbf{1}\mid\mathbf{W}\mid\mathbf{Y}\,]$, that is to say $[\,\mathbf{1}\mid\mathbf{W}\mid\mathbf{Y}\,]=\mathbf{U}\boldsymbol{\Sigma}\,{}^t\mathbf{V}$ with ${}^t\mathbf{U}\mathbf{U}=\mathbf{I}_n$ and ${}^t\mathbf{V}\mathbf{V}=\mathbf{I}_{p+2}$, where $\mathbf{I}_n$ and $\mathbf{I}_{p+2}$ are respectively the $n\times n$ and $(p+2)\times(p+2)$ identity matrices,
- step 2: if the elements of the matrix $\mathbf{V}$ are denoted by v_{jl}, then the TLS estimator of μ and $\boldsymbol{\alpha}$ is given by

$$\begin{pmatrix}\widehat{\mu}_{TLS}\\ \widehat{\boldsymbol{\alpha}}_{TLS}\end{pmatrix}=-\frac{1}{v_{p+2,p+2}}\,{}^t(v_{1,p+2},\dots v_{p+1,p+2}). \quad (9)$$

However, the problem of this algorithm is that it can not be used directly when the minimization problem (8) is *ill-conditioned* and needs a regularization. The minimization problem we consider is then (see [Golub *et al.*, 1999])

$$\min_{\mu\in\mathbb{R},\boldsymbol{\alpha}\in\mathbb{R}^p,\mathbf{X}_i\in\mathbb{R}^p}\left\{\frac{1}{n}\sum_{i=1}^{n}\Big[(Y_i-\mu-\mathbf{X}_i\boldsymbol{\alpha})^2+\ {}^t(\mathbf{X}_i-\mathbf{W}_i)(\mathbf{X}_i-\mathbf{W}_i)\Big]\right.$$
$$\left.+\rho\ {}^t\boldsymbol{\alpha}{}^t\mathbf{L}\mathbf{L}\boldsymbol{\alpha}\right\}, \quad (10)$$

where $\mathbf{L}$ is a $p\times p$ matrix. Using the properties of the SVD, it can be shown (see [Golub and Van Loan, 1996]) that

$$\begin{pmatrix}\widehat{\mu}_{TLS}\\ \widehat{\boldsymbol{\alpha}}_{TLS}\end{pmatrix}=\left({}^t[\,\mathbf{1}\mid\mathbf{W}\,][\,\mathbf{1}\mid\mathbf{W}\,]-\sigma_{p+2}^2\mathbf{I}_{p+1}\right)^{-1}{}^t[\,\mathbf{1}\mid\mathbf{W}\,]\mathbf{Y}, \quad (11)$$

where σ_{p+2} is the smallest singular value of the matrix $[\,\mathbf{1}\mid\mathbf{W}\mid\mathbf{Y}\,]$ and $\mathbf{I}_{p+1}$ is the $(p+1)\times(p+1)$ identity matrix. From this expression, the TLS solution to the minimization problem (10) is given by

$$\begin{pmatrix}\widehat{\mu}_{TLS}\\ \widehat{\boldsymbol{\alpha}}_{TLS}\end{pmatrix}=\left({}^t[\,\mathbf{1}\mid\mathbf{W}\,][\,\mathbf{1}\mid\mathbf{W}\,]-\lambda\mathbf{I}_{p+1}+\rho\,{}^t\mathbf{M}\mathbf{M}\right)^{-1}{}^t[\,\mathbf{1}\mid\mathbf{W}\,]\mathbf{Y}, \quad (12)$$

where $\mathbf{M}$ is the $(p+1)\times(p+1)$ matrix defined by $\mathbf{M}=\begin{pmatrix}0\;0\dots0\\ 0\\ \vdots\quad\mathbf{L}\\ 0\end{pmatrix}$.

2.2 Total Least Squares in the functional case

All that has been done in the previous paragraph can be adapted to the case where X is of functional type. The minimisation problem considered is a combination of (4) and (10), that we write

$$\min_{\mu\in\mathbb{R},\alpha\in L^2(I),X_i\in L^2(I)} \left\{ \frac{1}{n}\sum_{i=1}^{n} \left[(Y_i-\mu-\langle\alpha,X_i\rangle)^2 + \|X_i-W_i\|^2 \right] + \rho \left\| \alpha^{(m)} \right\|^2 \right\}. \quad (13)$$

We choose to build a spline estimator of α. We have to fix a degree $q \in \mathbb{N}$ and a number $k \in \mathbb{N}^\star$ of knots (taken equispaced) giving a subdivision of the interval I (see [de Boor, 1978] for details on spline functions). These spline functions have well-known properties, in particular, this space of spline functions is a vectorial space of dimension $k+q$. A usual basis is the set of the so-called B-spline functions, that we denote by $\mathbf{B}_{k,q} = {}^t(B_1 \dots B_{k+q})$. Then, we estimate α as a linear combination of the B-spline functions, that is to say we have to find a vector $\widehat{\boldsymbol{\theta}} = {}^t(\widehat{\theta}_1 \dots \widehat{\theta}_{k+q}) \in \mathbb{R}^{k+q}$ such that $\widehat{\alpha} = {}^t\mathbf{B}_{k,q}\widehat{\boldsymbol{\theta}}$ with $\widehat{\mu}$ and $\widehat{\boldsymbol{\theta}}$ solutions of the minimization problem

$$\min_{\mu\in\mathbb{R},\boldsymbol{\theta}\in\mathbb{R}^{k+q},X_i\in L^2(I)} \left\{ \frac{1}{n}\sum_{i=1}^{n} \left[\left(Y_i-\mu-\langle\, {}^t\mathbf{B}_{k,q}\widehat{\boldsymbol{\theta}},X_i\rangle\right)^2 + \|X_i-W_i\|^2 \right] + \rho \left\| \left({}^t\mathbf{B}_{k,q}\widehat{\boldsymbol{\theta}}\right)^{(m)} \right\|^2 \right\}. (14)$$

Using the work in [Cardot *et al.*, 2003] for the spline estimator of the functional coefficient and what has been done in the multivariate case (see equation (12)), it is possible to find an explicit solution to the minimization problem (14), given by

$$\begin{pmatrix} \widehat{\mu}_{FTLS} \\ \widehat{\boldsymbol{\theta}}_{FTLS} \end{pmatrix} = \frac{1}{n}(\frac{1}{n}\,{}^t\mathbf{D}\mathbf{D} - \lambda\mathbf{I}_{k+q+1} + \rho\mathbf{K})^{-1}\,{}^t\mathbf{D}\mathbf{Y}, \quad (15)$$

with

$$\mathbf{D} = \begin{pmatrix} 1 & \langle B_1,W_1\rangle & \dots & \langle B_{k+q},W_1\rangle \\ \vdots & \vdots & & \vdots \\ 1 & \langle B_1,W_n\rangle & \dots & \langle B_{k+q},W_n\rangle \end{pmatrix},$$

and

$$\mathbf{K} = \begin{pmatrix} 0 & 0 & \dots & 0 \\ 0 & \langle B_1^{(m)}, B_1^{(m)} \rangle & \dots & \langle B_1^{(m)}, B_{k+q}^{(m)} \rangle \\ \vdots & \vdots & & \vdots \\ 0 & \langle B_{k+q}^{(m)}, B_1^{(m)} \rangle & \dots & \langle B_{k+q}^{(m)}, B_{k+q}^{(m)} \rangle \end{pmatrix}.$$

3 A simulation study

The aim of this simulation is to see the behaviour of this TLS estimator and to compare it with the spline estimator given in [Cardot *et al.*, 2003] by

$$\begin{pmatrix} \widehat{\mu}_{FLS} \\ \widehat{\boldsymbol{\theta}}_{FLS} \end{pmatrix} = \frac{1}{n}(\frac{1}{n}\,{}^t\mathbf{DD} + \rho\mathbf{K})^{-1}\,{}^t\mathbf{DY}. \tag{16}$$

We choose to take

- $n = 200$: the initial sample will be splitted into a learning sample of length $n_l = 100$ (to estimate μ and α) and a test sample of length $n_t = 100$ (to see the quality of prediction),
- $p = 50$ discretization points on $I = [0, 1]$,
- X is either a standard brownian motion or an Ornstein-Uhlenbeck process on I,
- $\mu = 2$,
- $\alpha(t) = 10\sin(2\pi t)$,
- $\epsilon \sim \mathcal{N}(0, \sigma_\epsilon^2)$ with $\sigma_\epsilon = 0.1$,
- $\delta(t_j) \sim \mathcal{N}(0, \sigma_\delta^2)$ with either $\sigma_\delta = 0.05$, $\sigma_\delta = 0.1$, or $\sigma_\delta = 0.2$.

Concerning the choice of the different parameters of the model, we have taken $k = 8$, $q = 3$ and $m = 2$. Moreover, in the functional least squares estimation, ρ is fixed by generalized cross validation (see [Wahba, 1990]). For the total least squares estimation, we have made the estimation for different values of λ and ρ among the values $10^{-2}, 10^{-3}, \dots, 10^{-10}$, and we have kept the best values for these two parameters in terms of prediction.

We have given in table 1 the mean relative errors on 50 simulations for the different models tested when X is a standard brownian motion on I and in table 2 the same errors when X is an Ornstein-Uhlenbeck process on I. The estimation of the curve X_i, noted $\widehat{X}_i$, is given by

$$\widehat{X}_i = W_i + \frac{Y_i - \widehat{\mu} - \langle \widehat{\alpha}, W_i \rangle}{1 + \|\widehat{\alpha}\|^2}\,\widehat{\alpha}, \tag{17}$$

as the generalization of $\widehat{\mathbf{X}}_i$ in the multivariate case (see [Fuller, 1997]), obtained by differentiation of equation (13) with respect to X_i. An example of the estimation of α is plotted on figure 1 in the case where X is a standard brownian motion on I with the variance noise $\sigma_\delta = 0.1$. These results show that the corrected estimator constructed with the TLS approach improves the estimation of α compared to the uncorrected estimator defined by (16).

	$\frac{(\widehat{\mu}-\mu)^2}{\mu^2}$			$\frac{\|\widehat{\alpha}-\alpha\|^2}{\|\alpha\|^2}$			$\frac{1}{n}\sum_{i=1}^{n}\left(\langle\widehat{\alpha},\widehat{X}_i\rangle-\langle\alpha,X_i\rangle\right)^2$		
	$\sigma_\delta = 0.05$	$\sigma_\delta = 0.1$	$\sigma_\delta = 0.2$	$\sigma_\delta = 0.05$	$\sigma_\delta = 0.1$	$\sigma_\delta = 0.2$	$\sigma_\delta = 0.05$	$\sigma_\delta = 0.1$	$\sigma_\delta = 0.2$
FLS	0.0002	0.0009	0.0017	0.21	0.41	0.78	0.007	0.010	0.018
FTLS	0.0002	0.0009	0.0016	0.12	0.27	0.56	0.006	0.008	0.015

Table 1. Errors on μ, α and prediction - case where X is a standard brownian motion on I.

	$\frac{(\widehat{\mu}-\mu)^2}{\mu^2}$			$\frac{\|\widehat{\alpha}-\alpha\|^2}{\|\alpha\|^2}$			$\frac{1}{n}\sum_{i=1}^{n}\left(\langle\widehat{\alpha},\widehat{X}_i\rangle-\langle\alpha,X_i\rangle\right)^2$		
	$\sigma_\delta = 0.05$	$\sigma_\delta = 0.1$	$\sigma_\delta = 0.2$	$\sigma_\delta = 0.05$	$\sigma_\delta = 0.1$	$\sigma_\delta = 0.2$	$\sigma_\delta = 0.05$	$\sigma_\delta = 0.1$	$\sigma_\delta = 0.2$
FLS	0.0004	0.0007	0.0017	0.07	0.19	0.39	0.006	0.011	0.021
FTLS	0.0004	0.0007	0.0015	0.02	0.11	0.26	0.005	0.010	0.019

Table 2. Errors on μ, α and prediction - case where X is an Ornstein-Uhlenbeck process on I.

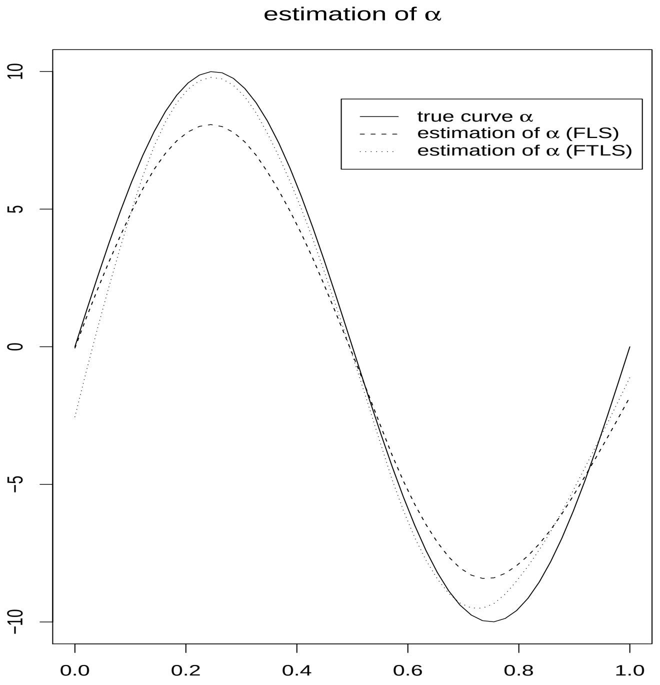


Fig. 1. Example of estimation of α (solid line) with functional least squares (dashed line) and functional total least squares (dotted line).

4 Conclusion and openings

This adaptation of the *Total Least Squares* method to the functional framework seems to give encouraging results on simulations. A theoretical work is needed to get the statistical properties of the estimator we have built. Moreover, it could also be interesting to compare this method to other ones. In particular, another idea to deal with noisy functional covariates (which is a work in progress) is to smooth the noisy curves (for instance by the way of a kernel method) and to estimate α by a procedure equivalent to a functional principal component regression used in the work of [Kneip and Utikal, 2001].

References

[Cardot *et al.*, 2003]H. Cardot, F. Ferraty, and P. Sarda. Spline estimators for the functional linear model. *Statistica Sinica*, pages 571–591, 2003.

[Carroll *et al.*, 1995]R.J. Carroll, D. Ruppert, and L.A. Stefanski. *Measurement Error in Nonlinear Models.* Chapman & Hall, London, 1995.

[de Boor, 1978]C. de Boor. *A Practical Guide to Splines.* Springer, New-York, 1978.

[Fuller, 1987]W.A. Fuller. *Measurement Error Models.* Wiley, New-York, 1987.

[Fuller, 1997]W.A. Fuller. Estimated true values for errors-in-variables models. In S. Van Huffel, editor, *Recent Advances in Total Least Squares Techniques and Errors-in-Variables Modeling*, pages 51–57, 1997.

[Golub and Van Loan, 1996]G.H. Golub and C.F. Van Loan. *Matrix Computations.* Johns Hopkins University Press, Baltimore, 1996.

[Golub *et al.*, 1999]G.H. Golub, P.C Hansen, and D.P. O'Leary. Tikhonov regularization and total least squares. *SIAM, J. Matrix Anal. Appl.*, pages 185–194, 1999.

[Kneip and Utikal, 2001]A. Kneip and K.J. Utikal. Inference for density families using functional principal component analysis. *J. Amer. Statist. Assoc.*, pages 519–542, 2001.

[Loève, 1963]M. Loève. *Probability Theory, 3^{rd} edition.* D. Van Nostrand, New Jersey, 1963.

[Ramsay and Dalzell, 1993]J.O. Ramsay and C.J. Dalzell. Some tools for functional data analysis. *J. Roy. Statist. Soc. Ser. B*, pages 539–572, 1993.

[Ramsay and Silverman, 1997]J.O. Ramsay and B.W. Silverman. *Functional Data Analysis.* Springer-Verlag, New-York, 1997.

[Ramsay and Silverman, 2002]J.O. Ramsay and B.W. Silverman. *Applied Functional Data Analysis.* Springer-Verlag, New-York, 2002.

[Van Huffel and Vandewalle, 1991]S. Van Huffel and J. Vandewalle. *The Total Least Squares Problem: Computational Aspects and Analysis.* SIAM, Philadelphia, 1991.

[Van Huffel, 2004]S. Van Huffel. Total least squares and errors-in-variables modeling: Bridging the gap between statistics, computational mathematics and engineering. In J. Antoch, editor, *Compstat Proceedings in Computational Statistics*, pages 539–555, 2004.

[Wahba, 1990]G. Wahba. *Spline Models for Observational Data.* Society for Industrial and Applied Mathematics, Philadelphia, 1990.

Forecasting binary longitudinal data by a functional PC-ARIMA model

A. M. Aguilera[1], M. Escabias[2], and M. J. Valderrama[2]

[1] Universidad de Granada
Dpto. de Estadística e I.O
Facultad de Ciencias
Campus de Fuentenueva
18071-Granada, Spain
(e-mail: aaguiler@ugr.es)

[2] Universidad de Granada
Dpto. de Estadística e I.O
Facultad de Ciencias
Campus de Fuentenueva
18071-Granada, Spain
(e-mail: escabias@ugr.es, valderra@ugr.es)

Abstract. The purpose of this paper is to forecast the time evolution of a binary response variable from an associated continuous time series observed only at discrete time points that usually are unequally spaced. In order to solve this problem we are going to use a functional logit model based on functional principal component analysis of the predictor time series that takes into account its continuous nature, close to classical ARIMA modelling of the associated discrete time series of principal components.

Keywords: Logistic Regression, Funcional Principal Components, ARIMA models.

1 Problem formulation

In this paper we propose a functional logit model based in mixed ARIMA-FPCA modelling of the functional predictor that allows to forecast the time evolution of a binary response from discrete time observations of a continuous time series. FPCA [Ramsay and Silverman, 1997] is a generalization of the classic principal component analysis (PCA) of a sample of data vectors for the reduction of dimension of a set of sample curves obtained in our case by cutting the predictor series in periods of the same amplitude. Mixed ARIMA-FPCA models [Valderrama *et al.*, 2002] allows not only to forecast a continuous time series in a whole future interval but also to reconstruct it between the discretization time points in the past.

Let us suppose that we have observations of a continuous time series $\{x(t)\}$ at discrete time points in the interval $(0, NT]$ and one observation Y_w of a related binary response Y at each period $((w-1)T, wT], w = 1, \ldots, N$. Then the purpose of this paper is to estimate a functional logit model to forecast the binary response in future periods $((w^* - 1)T, w^*T](w^* > N)$ from the forecasting of the series $x(t)$ in such periods provided by a mixed ARIMA-FPCA model.

In order to formulate and to estimate a functional logit model based on functional principal component analysis, we propose to cut the observed series $x(t)$ in N periods of amplitude T, so that we have N sample paths of the following functional predictor (continuous time process):

$$\{X_w(s) = x((w-1)T + s) : s \in (0, T]; w = 1, \ldots, N\}, \tag{1}$$

and a sample of size N of the binary response given by $\{Y_w : w = 1, \ldots, N\}$ (see Figure 1).

Let us observe that the choice of the amplitude T is simple enough in practice when there is a well defined seasonal period as in the case of many real time series.

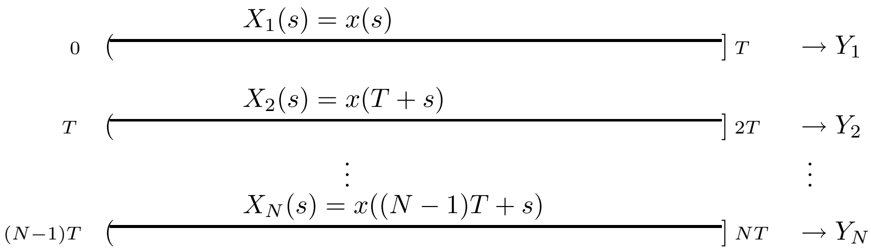


Fig. 1. Sample information obtained after cutting the original continuous time series

2 Functional logistic regression

The objective of the functional logistic regression (FLR) model is to explain a binary response variable Y in terms of a functional variable $X(s)$ whose sample information is given by a set of curves measured without error.

Let $X_1(s), \ldots, X_N(s)$ be a sample of curves of a functional variable $\{X(s) : s \in (0, T]\}$, obtained by cutting in periods of amplitude T the original predictor series $x(t)$, and let $Y_w(w = 1, \ldots, N)$ be the random observations of the binary response variable Y associated with the sample curves. Then, the FLR model is given by $Y_w = \pi_w + \varepsilon_w$, where ε_w are zero mean independent random errors with variance $\pi_w(1 - \pi_w)$, and π_w is the probability of response $Y = 1$ for a specific curve $X_w(s)$ modelled as

$\pi_w = \exp(l_w)/(1+\exp(l_w))$, with l_w being the logit transformation given by

$$l_w = \alpha + \int_0^T X_w(s)\,\beta(s)\,ds,\ w = 1,\ldots,N, \tag{2}$$

where α is a real parameter and $\beta(s)$ is a parameter function that has to be estimated. In terms of the logit transformations, the model can be equivalently seen as a functional generalized linear model [James, 2002].

As in the functional linear model [Ramsay and Silverman, 1997], it is impossible to obtain a direct estimation of the FLR model by using the usual likelihood or least squares methods. In addition functional data are usually observed only in a finite set of time points so that its true functional form has to be reconstructed from its discrete time observations by using an approximating procedure. Then, the most used solution for solving this estimation problem is based on assuming that the parameter function and the sample curves belong to a finite dimension space generated by a basis of functions $\{\phi_1(t),\ldots,\phi_p(t)\}$, so that they can be expressed in terms of the basis as

$$\beta(s) = \sum_{k=1}^{p} \beta_k \phi_k(s) \quad \text{and} \quad X_w(s) = \sum_{j=1}^{p} a_{wj}\phi_j(s). \tag{3}$$

Then, the functional model given by equation (2) is equivalent to a multiple logit model given in matrix form by $L = \mathbf{1}\alpha + A\Psi\beta$, with $L = (l_1,\ldots,l_N)'$, A the matrix that has the basis coefficients of the sample curves as rows, $\Psi = (\psi_{jk})_{p\times p}$ the one that has the L^2-usual inner products between the basic functions as entries, $\left(\psi_{jk} = \int_0^T \phi_j(s)\,\phi_k(s)\,dt\right)$, and $\beta = (\beta_1,\ldots\beta_p)'$ the vector of the parameter function basis coefficients.

Before estimating by likelihood the vector β, we have to compute the matrix A of sample curves basis coefficients. Let $x_w = (x_{w1},\ldots,x_{wm_w})'$ be the vector of observations of the *wth* sample curve $X_w(s)$ at m_w time points of the interval $((w-1)T, wT]$, $\forall w = 1,\ldots,N$. When discrete-time observations are considered to be measured without error, $x_{wk} = x_w(t_{wk})$, an interpolation method to estimate the basis coefficients can be used. On the other hand, if some error is considered in the observations, $x_{wk} = x_w(t_{wk}) + \varepsilon_{wk}$, least squares approximation is usually used for estimating the basis coefficients for a specific curve as $a_w = (a_{w1},\ldots,a_{wp})' = (\Phi'\Phi)^{-1}\Phi' x_w$, with $\Phi_{m_w\times p} = (\phi_j(t_{wk}))$. Let us observe that least squares approximation can be also applied when the functional variable is recorded at different time points for each individual (missing longitudinal data). On the other hand, taking into account the underlying nature of curves, different basis have been used in literature as for example, Fourier, Wavelets or Spline functions.

The problem is that likelihood estimation of the parameters of the logit model with design matrix $A\Psi$ is very unaccurate due to multicollinearity so that the estimated parameter function can not be used to stablish the true relationship between the response and predictor variables [Escabias *et al.*, 2004].

3 Functional principal component logit model

In order to reduce dimension and to obtain better estimations of the parameter function, two different approaches based on FPCA of sample paths have been proposed in literature [Escabias *et al.*, 2004], so that the FLR model is reduced to a multiple one with a reduced number of functional principal components as covariates. In this paper we are going to perform FPCA of the sample paths $X_w(s)$ with respect to the usual inner product in $L^2\left((0,T]\right)$.

Functional principal components of $X_w(s)$ are defined as N-dimensional vectors $\xi_j(j = 1, \ldots, N-1)$ with components

$$\xi_{wj} = \int_0^T (X_w\left(s\right) - \bar{x}(s))f_j\left(s\right)ds,\ w = 1, \ldots, N,$$

where $\bar{x}(s)$ is the sample mean of the sample curves and the weight functions $f_j\left(s\right)(j = 1, \ldots, N-1)$ that define the functional pc's are the eigenfunctions of the sample covariance function of the sample curves whose associated positive eigenvalues $\lambda_1 > \lambda_2 > \cdots > \lambda_{n-1} \geq 0$ are the variances of the corresponding principal components (pc's).

Then, the sample curves admit the following orthogonal representation in terms of the sample pc's:

$$X_w\left(s\right) = \sum_{j=1}^{N-1} \xi_{wj} f_j\left(s\right),\ w = 1, \ldots, N.$$

By truncating this expression we obtain a reconstruction of the sample paths in terms of a reduced number of pc's that accumulate a certain percentage of the total variance $TV = \sum_{j=1}^{N-1} \lambda_j$.

It can be shown that if the sample paths belong to a finite space of $L^2(0,T]$ generated by a basis, their functional pc's are given by the standard principal components of the matrix $A\Psi^{1/2}$. If we denote by $\Gamma = \left(\xi_{ij}\right)_{N\times p}$ the matrix whose columns are the pc's of the $A\Psi^{1/2}$ matrix, and G the one whose columns are its associated eigenvectors, then $\Gamma = \left(A\Psi^{1/2}\right)G$ and the weight functions that define the functional pc's are given by

$$f_j\left(s\right) = \sum_{k=1}^{p} f_{jk}\phi_k\left(s\right),\ j = 1, \ldots, p \tag{4}$$

with $F = \left(f_{jk}\right)_{p\times p} = \Psi^{-1/2}G$.

Then, FLR model (2) can be equivalently expressed in terms of the pc's as

$$l_w = \alpha + \sum_{j=1}^{p} \xi_{wj}\gamma_j,\ w = 1, \ldots, N. \tag{5}$$

[Escabias *et al.*, 2004]

The functional principal component logistic regression (FPCLR) model is obtained by truncating model (5) in terms of a subset of pc's. If we consider the matrices defined before partitioned as follows

$$\Gamma = \left(\Gamma_{(q)} \middle| \Gamma_{(r)} \right), \; F = \left(F_{(q)} \middle| F_{(r)} \right), \; r + q = p,$$

then, the FPCLR model is defined by taking as covariates the first q principal components

$$L_{(q)} = \alpha_{(q)} \mathbf{1} + \Gamma_{(q)} \gamma_{(q)},$$

where $\alpha_{(q)}$ is a real parameter and $L_{(q)} = \left(l_{1(q)}, \ldots, l_{N(q)} \right)'$ with

$$l_{w(q)} = ln \left[\frac{\pi_{w(q)}}{1 - \pi_{w(q)}} \right] = \alpha_{(q)} + \sum_{j=1}^{q} \xi_{wj} \gamma_{j(q)}, \; i = 1, \ldots, N. \tag{6}$$

Finally, the likelihood estimation of the parameter function given by

$$\widehat{\beta}_{(q)} \left(s \right) = \sum_{j=1}^{p} \widehat{\beta}_{j(q)} \phi_j \left(s \right), \tag{7}$$

with the coefficient vector $\widehat{\beta}_{(q)} = F_{(q)} \widehat{\gamma}_{(q)}$ is more accurate than the one obtained with the original $A\Psi$ design matrix [Escabias *et al.*, 2004].

4 Mixed ARIMA-FPCA logit model

Let us observe that functional PCA provides an orthogonal expansion of the functional predictor $\{X(s)\}$ in terms of a set of deterministic functions (the principal factors) and random variables (the principal components).

In our case, the values ξ_{wj} $(w = 1, \ldots, N)$ of each sample principal component ξ_j can be seen as observations of a discrete time series at each period $((w-1)T, wT]$ of amplitude T where the original series $x(t)$ is observed. Then, in order to forecast the binary response in future periods $((w^* - 1)T, w^*T](w^* > N)$, we propose the modelization of each principal component by an ARIMA model [Box and Jenkins, 1970]. The general expression of an ARIMA(p,d,q) model for the *jth* principal component ξ_j is given by $\Phi(B)(1-B)^d \xi_{wj} = \theta(B)\epsilon_{wj}$, where B is the backward shift operator, $\Phi(B)$ is the autoregressive operator defined as $\Phi(B) = 1 - \Phi_1 B - \Phi_2 B^2 - \cdots - \Phi_p B^p$, $\theta(B)$ is the moving average operator given by $\theta(B) = 1 - \theta_1 B - \theta_2 B^2 - \cdots - \theta_p B^p$, and ϵ_{wj} is a white noise process for each j $(j = 1, \ldots, q)$.

The prediction model proposed in this paper is based on ARIMA forecasting of each of the q principal components selected for estimating the functional logit model. After estimating in the usual form these q ARIMA models, we will be able to obtain forecasts for each principal component in the future periods $((w^* - 1)T, w^*T]$, denoted by $\widetilde{\xi}_{w^*j}$.

Finally, the original series $x(t)$ is predicted in all the interval of time $((w^* - 1)T, w^*T]$, by the principal component reconstruction of the process $\{X(s)\}$ in terms of the predicted principal component values

$$\widetilde{x}((w^* - 1)T + s) = \widetilde{X}^q_{w*}(s) = \bar{x}(s) + \sum_{j=1}^{q} \widetilde{\xi}_{w^*j} f_j(s) \quad s \in [0, T],$$

and the estimated probabilities of success in the future periods $((w^* - 1)T, w^*T]$ are predicted from the logit transformations

$$\tilde{l}_{w*(q)} = \hat{\alpha}_{(q)} + \sum_{j=1}^{q} \tilde{\xi}_{w^*j} \hat{\gamma}_{j(q)},$$

in terms of the ARIMA forecasts of the principal components $\tilde{\xi}_{w^*j}$.

5 Predicting the risk of drought

In order to illustrate the proposed Mixed ARIMA-FPCA logit model, we are going to predict the risk of drought in the future in terms of its evolution in the past by using as predictor the past evolution of temperatures, as in [Escabias *et al.*, 2005]. With this objective, let us consider a specific zone where drought has been tested monthly for several years by classifying a month as dry or not dry according to the definition of drought based on the amount of precipitations observed in this zone. That is, if it rains less that a certain percentile during a specific month, it is considered as a dry month meanwhile in the opposite case the month is considered as not dry. Then, if we define the binary variable $Y = \{0, 1\}$ as the one that takes value one in a specific month if it is not a dry month and zero in the opposite case, we have a monthly time series of binary values.

We have daily temperatures and precipitations observed in the *Estación Meteorológica del Departamento de Botánica de la Universidad de Granada* from 01/01/1992 to 12/31/2001. In this period the precipitation have been monthly accumulated (30 days period) and each month has been classified as dry if the accumulated precipitations in this month have been lower than a specific percentile of the precipitations observed in same month over all the years. In order to test the forecasting performance of mixed ARIMA-FPCA logit models we have considered different examples by using different percentiles (0.25 and 0.50) for defining the binary time series of drought.

As predictor time series $x\,(t)$ we have considered the daily temperatures cut at 30 days periods (T=30). In order to obtain the functional form of temperatures in each month we have considered the expansion of such functions as in (3) in terms of the basis of B-splines defined from the knots

$\{1, 3, 5, 7, 9, 11, 13, 15, 17, 19, 21, 23, 25, 27, 30\}$, and we have obtained the basis coefficients of each curve by least squares approximation from the discrete time observations of daily temperatures.

Once the predictor curves have been approximated from their discrete-time observations and the response observed in each one of the considered examples (percentiles 0.25 and 0.50), we have considered the first $N = 108$ observations to fit the Mixed ARIMA-FPCA logit model and the last 12 to validate the results. ¿From the fist $N = 108$ observations of monthly temperatures (predictor) and drought (binary response) we have fitted the FPCLR model with different number of functional pc's in the model. The percentages of variance explained by the first four functional pc's can be seen in Table 1. Once the functional pc's have been computed, we have modeled them as ARIMA's obtaining that only the two first pc's have such structure. We have considered the rest as white noises. ARIMA modelling of pc's can be seen in Table 1. After modelling the pc's we have obtained 12 steps ahead forecasts (12 months) of such time series.

pc	Exp. Var.	Cum. Var.	ARIMA Model	Estimated Parameters
ξ_1	87.17%	87.17%	$SARIMA(0,0,1) \times (0,1,1)_{12}$	$\theta_1 = -0,351239$ $\Theta_1 = 0,567944$
ξ_2	4.13%	91.30%	$SARIMA(0,0,0) \times (0,1,1)_{12}$	$\Theta_1 = 0,894991$
ξ_3	2.26%	93.56%	White Noise ($\sigma = 5.44$)	
ξ_4	1.46%	95.02%	White Noise ($\sigma = 4.37$)	

Table 1. Percentages of explained variances (Exp. Var.), cumulated variances (Cum. Var.), and ARIMA modelling for the first four pc's.

In order to test the performance of mixed ARIMA-FPCA logit models we have obtained the estimated probabilities for the response (risk of drought) from the 12 predictions provided by ARIMA modelling of the first pc's (see Table 2). These probabilities have been obtained by using the estimated parameters of the logistic models with the first 1, 2, 3 and 4 pc's in the models with each one of the responses. All adjusted logit models have high deviance statistics with low p-values what shows that the models fit well and that the logit model is a good election for estimating this response. In each case the Mean Squared Error (MSE) between predictions and observed values have been obtained. The results can be seen in Table 2. It can be observed that the MSE of the models with the components that are modelled as ARIMA are always lower than the ones that include not modelled principal components.

2002	Y defined by 0.25 precentile					Y defined by 0.50 precentile				
Months	Dry	1 cp	2 cp's	3 cp's	4 cp's	Dry	1 cp	2 cp's	3 cp's	4 cp's
Jan	1	0.684	0.684	0.779	0.670	1	0.425	0.425	0.446	0.282
Feb	1	0.678	0.679	0.522	0.537	1	0.408	0.408	0.377	0.395
Mar	1	0.690	0.674	0.627	0.581	1	0.441	0.438	0.427	0.369
Apr	1	0.713	0.698	0.616	0.719	1	0.507	0.504	0.486	0.628
May	0	0.717	0.707	0.798	0.827	0	0.520	0.518	0.539	0.591
Jun	1	0.741	0.716	0.597	0.623	0	0.593	0.588	0.562	0.598
Jul	1	0.774	0.765	0.787	0.821	1	0.691	0.689	0.692	0.743
Oct	1	0.794	0.793	0.824	0.836	1	0.748	0.748	0.753	0.768
Sep	1	0.791	0.803	0.937	0.950	1	0.742	0.744	0.785	0.823
Oct	1	0.779	0.797	0.818	0.753	0	0.705	0.709	0.711	0.596
Nov	1	0.743	0.758	0.768	0.606	1	0.600	0.603	0.603	0.363
Dec	1	0.717	0.748	0.822	0.847	1	0.519	0.525	0.543	0.586
MSE		0.108	0.107	0.130	0.142		0.248	0.247	0.247	0.267

Table 2. Observed values of the response (no drought) and estimated probabilities of no drought for the mixed ARIMA-FPCA logit model in each one of the selected responses (percentiles 0.25 and 0.50) for the models with the first 1, 2, 3 and 4 pc's.

6 Acknowledgments

This research has been supported in part by Project MTM2004-5992 of the Ministerio de Ciencia y Tecnología, Spain.

References

[Box and Jenkins, 1970]G.E.P. Box and G.M. Jenkins. *Time Series Analysis Forecasting and Control.* Holden Day, San Francisco, 1970.

[Escabias *et al.*, 2004]M. Escabias, A.M. Aguilera, and M.J. Valderrama. Principal component estimation of functional logistic regression: discussion of two different approaches. *Journal of Nonparametric Statistics*, pages 365–384, 2004.

[Escabias *et al.*, 2005]M. Escabias, A.M. Aguilera, and M.J. Valderrama. Modeling environmental data by functional principal component logistic regression. *Environmetrics*, pages 95–107, 2005.

[James, 2002]G.M. James. Generalized linear models with functional predictors. *Journal of the Royal Statistical Society, Series B*, pages 411–432, 2002.

[Ramsay and Silverman, 1997]J.O. Ramsay and B.W. Silverman. *Functional Data Analysis.* Springer-Verlag, New York, 1997.

[Valderrama *et al.*, 2002]M. J. Valderrama, F. A. Ocaña, and A. M. Aguilera. Forecasting pc-arima models for functional data. In W. Härdle and B. Rönz, editors, *Proceedings in Computational statistics*, pages 25–36, 2002.

Classification in Hilbert Spaces with Support Vector Machines

Fabrice Rossi[1] and Nathalie Villa[2]

[1] Projet AxIS, INRIA,
Domaine de Voluceau, Rocquencourt, BP 105,
78153 Le Chesnay cedex - FRANCE
(e-mail: Fabrice.Rossi@inria.fr)

[2] Equipe GRIMM - Université Toulouse Le Mirail,
5 allées A. Machado,
31058 Toulouse cedex 1 - FRANCE
(e-mail: villa@univ-tlse2.fr)

Abstract. In many applications, input data are in fact sampled functions rather than standard high dimensional vectors. Most of the traditional data analysis tools for regression, classification and clustering have been adapted to handle functional inputs under the general name of Functional Data Analysis (FDA). In general, the major problem is to overcome the issue of infinite dimensional input. This is done by introducing regularity constraints on the studied functions, thanks to penalization or to projection on finite dimensional functional spaces.

Support Vector Machine (SVM) are large margin classifier tools that have the interesting property of being less sensitive to the curse of dimensionality than other tools. On the contrary, they are based on implicit non linear mappings of the considered data into high dimensional spaces (sometimes with infinite dimension) thanks to kernel functions.

In this paper, we investigate the use of Support Vector Machine for functional data analysis. We define simple kernels that take into account the functional nature of the data and lead to consistent classification. Experiments conducted on real world data emphasize the benefit of taking into account some functional aspects of the problems.

Keywords: Functional Data Analysis, Support Vector Machine, Classification.

1 Introduction

This paper deals with functional classification: let (X, Y) be a pair of random variables in which Y takes values in $\{-1; 1\}$ and X in a functional space. Y is the label (the class) associated to X. The goal of classification is to predict the value of Y given an observed value for X. The difficulty in functional data analysis [Ramsay and Silverman, 1997], compared to the traditionnal setting, is that X does not take values in $\mathbb{R}^d$ but in a functional space.

In this paper, we investigate how Support Vector Machine (SVM) can be used for functional data classification. The paper is organized as follows: Section 2 explains why functional SVM leads to particular problems and proposes solutions to overcome them. Section 3 develops several functional kernels and explains how some of them lead to consistent classifier. Finally, Section 5 illustrates the various approaches on real data sets.

2 Support Vector Machine For FDA

2.1 Hard margin functional SVM

We assume given a learning set, i.e. N examples $(x_1, y_1), \ldots, (x_N, y_N)$ which are i.i.d. realizations of (X, Y). As explained before, X is a function valued random variable. More formally, X takes its values in a separable Hilbert space $\mathcal{X}$, for example a subspace of $L^2(\mu)$ where μ denotes a finite Borel measure on $\mathbb{R}$. We denote $\langle ., . \rangle$ the inner product of $\mathcal{X}$.

The principle of SVM is to perform an affine discrimination of the observations with the largest margin as possible, that is to find a function $w \in \mathcal{X}$ with a minimum norm and a real value b, such that $y_i(\langle w, x_i \rangle + b) \geq 1$ for all i. The classification rule associated to (w, b) is simply $\phi(x) = \mathrm{sign}(\langle w, x \rangle + b)$. We therefore request the rule to have zero error on the learning set.

In functional spaces, it is always possible to find such a discrimination, provided the $(x_i)_{1 \leq i \leq N}$ are in general position, i.e. provided they span a vector space of dimension N. However it is well known that the obtained classification rule do not behave in a satisfactory way unless a regularization method is used (see [Hastie and Mallows, 1993], [Marx and Eilers, 1996], [Ramsay and Silverman, 1997] and [Cardot *et al.*, 1999]).

2.2 Soft margin functional SVM

While SVM introduces a form of regularization by looking for large margin (i.e., minimal norm for w), additional regularization can be obtained by solving the following optimization problem:

$$(P_C) \qquad \begin{array}{l} \min_{w,b,\xi} \langle w, w \rangle + C \sum_{i=1}^N \xi_i, \\ \text{subject to } \ y_i(\langle w, x_i \rangle + b) \geq 1 - \xi_i, \\ \qquad\qquad\quad \xi_i \geq 0, \ \text{for all } i = 1, \ldots, N, \end{array}$$

for an appropriate $C \geq 0$. Using the slack variables ξ_i allows to relax the very strong condition that the classification rule should make no error on the learning set. It is well known (see e.g., [Hastie *et al.*, 2004]) that this form of regularization is needed to achieve good performances for classification in high dimensional spaces.

In order to solve this problem, we use results from [Chih-Jen, 2001] that apply to any Hilbert space. Problem (P_C) is indeed equivalent to the dual optimization problem:

$$(D_C) \qquad \begin{array}{l} \min_\alpha \sum_{i=1}^N \alpha_i - \sum_{i,j=1}^N \alpha_i \alpha_j y_i y_j \langle x_i, x_j \rangle, \\ \text{subject to } \sum_{i=1}^N \alpha_i y_i = 0 \text{ and } 0 \leq \alpha_i \leq C \text{ for all } i = 1, \ldots, N. \end{array}$$

The advantage of (D_C) versus (P_C) in the infinite dimensional context is that the optimization problem (D_C) has to be solved in $\mathbb{R}^N$ whereas (P_C) needs an optimization procedure in $\mathcal{X}$. Moreover inner products in functional spaces such as $L^2(\mu)$ are easy to approximate using classical quadrature or Monte Carlo methods. Finaly, the classification rule is obtained as $\phi(x) = \text{sign}(\sum_{i=1}^N y_i\alpha_i\langle x_i, x\rangle + b)$ which is only based on inner products. In practice, this means that any SVM software can be used to provide functional classification as long as inner products can be calculated and used in the software.

It should be noted that C is a free parameter. It has therefore to be chosen so has to provide good performances. We will provide a possible solution in section 4.1.

3 Functional kernels

3.1 Kernels for SVM

A major difference between standard multivariate data and functional data is that the former are seldom linearly separable whereas the latter often are. In finite dimensional settings, this motivates the use of kernels to replace the inner product that is used in problem (D_c). A kernel corresponds to an implicit mapping from the input space to another feature space. In general this feature space has a high dimension so that the data become linearly separable in it. Thanks to the dual formulation of the SVM optimization problem, the implicit mapping is not calculated: everything is based on the kernel.

For functional data, the use of kernels might seem worthless. However, despite the regularization provided by using slack variables, it happens in practice for linear functional SVM to have very bad performances. A possible solution consists in using functional transformation and functional kernels, as proposed in this section.

3.2 Using an orthogonal basis

A natural functional kernel can be constructed thanks to the general functional classification framework proposed in [Biau *et al.*, 2005]. The methods proceeds as follows:

1. choose a complete orthonormal system of $\mathcal{X}$, $\{\Psi_j\}_{j\geq 1}$, and express each observation x_i as a series expansion $x_i = \sum_{j\geq 1} x_{ij}\,\Psi_j$;
2. approximate each observation x_i by the sum $\sum_{j=1}^d x_{ij}\,\Psi_j$;
3. perform a classical $\mathbb{R}^d$ SVM on the coefficients $\mathbf{x}_i^{(d)} = (x_{i1}, \ldots, x_{id}) \in \mathbb{R}^d$ for all $i = 1, \ldots, N$.

This procedure is equivalent to working with a functional kernel which can be written as

$$\mathcal{K}_d(x, x') = K(\mathcal{P}_d(x), \mathcal{P}_d(x'))$$

where $\mathcal{P}_d$ denotes the projection onto the the space spanned by $\{\Psi_j\}_{j=1,\ldots,d}$ and K is any standard SVM kernel. Of course, d has to be chosen appropriately. As recalled in section 4.1, [Biau *et al.*, 2005] proposes to use a split sample approach.

3.3 Using a B-Spline basis

Another way of choosing a projection space consists in using spline spaces and their B-spline bases. Results from [Biau *et al.*, 2005] are still applicable, but with major restriction. Indeed, a B-spline basis is not a basis of L^2: it only spans a subspace of L^2. Nevertheless, they perform efficiently in practice.

An interesting property of B-spline bases if they can be use to provide additional transformation on the input data: using a B-Spline expansion, an estimation of $x^{(q)}$, the q-th derivative of x, can be easily obtained. Then any kernel can be used on the derivatives. This method allows to focus on some particular aspects of the underlying functions, such as the curvature for the second derivative. It is well known that in some application domain such as spectrometry, such kind of features might be more interesting than the original curves. We give in Section 5.3 an application of this approach.

4 Consistency of functional SVM

4.1 Choice of the parameters

Performing a functional SVM leads to choose three types of parameters:

1. parameters due to the functional pre-processing: d, the dimension of the projection if we use a orthogonal basis as in section 3.2 or the order of the B-Splines basis, the number of knots and the order q of the derivative(s) chosen in the case of the pre-processing described in section 3.3;
2. C, the regularization parameter of the SVM (see section 2.2);
3. K, which is indeed the kernel: we can both choose the type of kernel (linear, gaussian, ...) but also the parameter of this kernel such as σ for the gaussian kernel $K(x, x') = e^{-\|x-x'\|^2/\sigma}$.

In order to select these parameters, we follow [Biau *et al.*, 2005] and use a data splitting device. To do that, let us introduce some notations: a denotes the parameters that we have to chosen in a set $\mathcal{A}$ of relevant parameters and $\mathcal{P}$ the preprocessing performed on the original data set. The data are then split into two sets. First, for a fixed value of the parameters, a, a training set $\{(x_i, y_i), i = 1, \ldots, l\}$ is used to calculate the SVM classification rule

$\phi_a^l = \text{sign}(\sum_{i=1}^l \alpha_i^* y_i K(\mathcal{P}(.), \mathcal{P}(x_i)) + b^*)$ where $(\{\alpha_i^*\}_i, b^*)$ are the solution of (D_C) in which we replace the classical dot product by $K \circ \mathcal{P}$. Then a validation set $\{(x_i, y_i), i = l+1, \dots, N\}$ is used to select a optimally in $\mathcal{A}$:

$$a^* = \arg\min_{a \in \mathcal{A}} \left\{ \hat{L}(\phi_a^l) + \frac{\lambda_a}{\sqrt{N-l}} \right\}.$$

where $\hat{L}(\phi_a^l) = \frac{1}{m} \sum_{i=l+1}^N \mathbf{1}_{\{\phi_a^l(x_i) \neq y_i\}}$ and $\frac{\lambda_a}{\sqrt{N-l}}$ is a penalty term.

4.2 Consistency

We now restrict ourselves to the case of the functional kernels of section 3.2. Then, as pointed out by [Biau *et al.*, 2005], a necessary and sufficient condition of consistency for the procedure described in sections 3.2 and 4.1 is that classical SVM are consistent in $\mathbb{R}^d$. [Steinwart, 2002] shows the universal consistency of some SVMs when two conditions are fulfilled: the input data must belong to a compact subset of $\mathbb{R}^d$ and the regularization parameter for N observations must be equal to $C_N = N^{\beta-1}$ (see Corollary 1 of [Steinwart, 2002]). This consistency result holds as long as the kernel used to perform it is *universal*; that is : if Φ is the feature map of the kernel, then the set of all the functions of the form $\langle w, \Phi(.)\rangle$ has to be dense in the set of all continuous functions defined on the considered compact subset. In particular, the gaussian kernel with any $\sigma > 0$ is universal for all compact subsets of $\mathbb{R}^d$.

Therefore, for this procedure, the choice of $a = (d, C, K)$ leads to a consistent classifier provinding some simple facts: for any fixed dimension d, K has to be chosen in a finite set $\mathcal{K}_d$ which contains, at least, one universal kernel. C can be chosen in a finite grid search (as this is the case in our applications) but recent progresses (see [Hastie *et al.*, 2004]) allows to choose C in an interval of the form $\mathcal{I}_d = [0; \mathcal{C}_d]$ by an automatic recurrent procedure.

The consistency result of [Biau *et al.*, 2005] is obtained for a k-nn classifier but, as stated in the paper, the result can be extended to any classifier. When choosing C in a infinite set, an adaptation of the proof is needed. As the original proof is constructed thanks to an oracle inequality that gives an upper bound for $EL(\phi_{d^*,C^*,K^*}) - L^*$ in finite dimension (L^* denotes the Bayes error), we have to obtain a similar oracle inequality: this can be done by the use of the shatter coefficient of a particular class of linear classifiers which provides the behavior of the classification rule on a set of $N - l$ points (see [Devroye *et al.*, 1996]). A limitation of SVM that does not appear in [Biau *et al.*, 2005] for k-nn, is that the input functions must belong to a compact subset of the functional space.

5 Applications

5.1 Speech recognition in very high dimensional space

We compare SVM to k-nn by applying exactly the procedure described in [Biau *et al.*, 2005] to the data used in the paper. The only difference is the replacement of the k-nn classifier by a regular SVM. The problem considered in [Biau *et al.*, 2005] consists in classifying speech samples. There are three two classes problems: classifying "yes" against "no", "boat" against "goat" and "sh" against "ao". For each problem, we have 100 functions. Each function is described by a vector in $\mathbb{R}^{8192}$. Performances of the algorithms are obtained thanks to a leave-one-out procedure: 99 functions are used as the learning set (to which the split sample procedure is applied to choose the parameters) and the remaining function provide a test example. We use the Fourier functional basis. We report the percentage of error for each problem in the following table:

Problem	k-nn	QDA	Gaussian SVM	linear SVM
yes/no	10%	7%	10%	58%
boat/goat	21%	35%	8%	46%
sh/ao	16%	32%	12%	47%

The first two columns have been reproduced from [Biau *et al.*, 2005] (QDA corresponds to Quadratic Discriminant Analysis). The "Gaussian SVM" column corresponds to the functional kernel obtained thanks to the projection of the Fourier basis combined to a Gaussian kernel in $\mathbb{R}^d$. The "linear SVM" corresponds to the direct application of the procedure described in 2.2, without any prior projection. In general the functional kernel give very satisfactory results, whereas the direct linear approach obtain extremely bad results (they corresponds to a random classification). This shows that the regularization provided by the slack variables is not adapted to functional data, a fact that was already known in the context of linear discriminant analysis [Hastie *et al.*, 1995].

The functional SVM performs in general better than k-nn and QDA, but the training time of the methods are not comparable. Indeed, solving problem (D_C) can cost up to $O(N^3)$ operations, whereas there is no training time for k-nn.

5.2 Using wavelet basis

In order to investigate the limitation of the direct use of the linear SVM, we have applied them to another speech recognition problem. We studied a part of TIMIT database which was investigated in [Hastie *et al.*, 1995]. The data are log-periodograms corresponding to recording phonemes of 32 ms duration. We have chosen to restrict ourselves to classifying "aa" against "ao", because this is the most difficult sub-problem in the database. The database is a multi-speaker database. Each speaker (325 in the training set and 112 in the test set) is recorded at a 16-kHz sampling rate; and we retain only the first 256 frequencies. We have 519 examples for "aa" in the training set (759 for "ao") and 176 in the test set (263 for "ao"). We use the split sample approach to choose the parameters on the training set (50% of the training examples are used for validation) and we report the classification error on the test set. The projection basis is here a hierarchical wavelet basis (see e.g., [Mallat, 1989]). We obtain the following results:

Functional Gaussian SVM	Functional linear SVM	Linear SVM
22%	19.4%	20%

It appears that functional kernels are not as useful here as in the previous example, as linear SVM applied directly to the discretized functions (in $\mathbb{R}^{256}$) performs as well as linear SVM on the wavelet coefficients. A natural explanation is that the actual dimension of the input space (256) is smaller than the number of learning examples (1278) which means that evaluating the optimal coefficients of the SVM is less difficult than in the previous example. Therefore, the additional regularization provided by the projection is not really useful in this context.

5.3 Spectrometric data set

The data presented in this section are 215 near infrared spectra of a meat sample recorded on a Tecator Infrared Food and Feed Analyser[1]. The classification problem consists in separating meat samples with a high fat content (more than 20%) from sample with a low fat content (less than 20%). It is well known that in some spectrometric problem, the curvature of the spectrum is more relevant for the prediction of the sample content than the spectrum itself. This drives us to construct a classifier based on the curvature of the spectra i.e. on the second derivative as explained in section 3.3.

We then decide to compare: a linear and a gaussian kernel performed on the original data and a linear and a gaussian kernel on the second derivatives. The training set contains 120 spectra (randomly chosen) and the testing set 95 spectra. The parameters (C and σ for the gaussian kernel) are chosen using a 10-fold cross validation procedure rather than a simple cross validation procedure. The following table gives the performances of the various methodologies:

Kernel	Learning set error rate	Test set error rate
Linear	0.83%	2.11%
Gaussian	0%	4.21%
Linear on second derivatives	0%	0%
Gaussian on second derivatives	0.83%	1.05%

It appear that the functional pre-processing slightly improves the results: in both linear and gaussian kernels, the use of the second derivatives introduces a kind of expert knowledge and overcomes the limitation of the original kernel. This is specially the case for the gaussian kernel which is norm dependant and is then dominated by the mean value of the spectra (which is not a good feature for spectrometric problems as we already said).

[1] available on statlib: `http://lib.stat.cmu.edu/datasets/tecator`

6 Conclusion

We have proposed in this paper functional kernels that provide consistent classification in Hilbert spaces with Support Vector Machines. When the considered functions are represented by very high dimensional vectors, projection based kernels provide regularization that enhance SVM classification rates. In other contexts, transformation based kernels allow to integrate expert knowledge in the SVM.

References

[Biau *et al.*, 2005]Gérard Biau, Florentina Bunea, and Marten Wegkamp. Functional classification in hilbert spaces. *IEEE Transactions on Information Theory*, 2005. To be published.

[Cardot *et al.*, 1999]Hervé Cardot, Frédéric Ferraty, and Pascal Sarda. Functional linear model. *Statist. & Prob. Letters*, 45:11–22, 1999.

[Chih-Jen, 2001]L. Chih-Jen. Formulation of support vector machines: a note from an optimization point of view. *Neural Computation*, 2(13):307–317, 2001.

[Devroye *et al.*, 1996]L. Devroye, L. Györfi, and G. Lugosi. *A Probabilistic Theory of Pattern Recognition.* Springer Verlag, New York, 1996.

[Hastie and Mallows, 1993]T. Hastie and C. Mallows. A discussion of "a statistical view of some chemometrics regression tools" by i.e. frank and j.h. friedman. *Technometrics*, 35:140–143, 1993.

[Hastie *et al.*, 1995]T. Hastie, A. Buja, and R. Tibshirani. Penalized discriminant analysis. *Annals of Statistics*, 23:73–102, 1995.

[Hastie *et al.*, 2004]Trevor Hastie, Saharon Rosset, Robert Tibshirani, and Ji Zhu. The entire regularization path for the support vector machine. *Journal of Machine Learning Research*, 5:1391–1415, October 2004.

[Mallat, 1989]Stéphane Mallat. Multiresolution approximation and wavelet orthonormal bases of l2. *Transaction of the American Mathematical Society*, 315:69–87, September 1989.

[Marx and Eilers, 1996]B. D. Marx and P. H. Eilers. Generalized linear regression on sampled signals with penalized likelihood. In R. Hatzinger A. Forcina, G. M. Marchetti and G. Galmacci, editors, *Statistical Modelling. Proceedings of the 11th International workshop on Statistical Modelling*, Orvietto, 1996.

[Ramsay and Silverman, 1997]Jim Ramsay and Bernard Silverman. *Functional Data Analysis.* Springer Series in Statistics. Springer Verlag, June 1997.

[Steinwart, 2002]I. Steinwart. Support vector machines are universally consistent. *J. Complexity*, 18:768–791, 2002.

Functional Learning with Wavelets

Laurent Rouvière

Institut de Mathématiques et de Modélisation de Montpellier,
UMR CNRS 5149, Equipe de Probabilités et Statistique,
Université Montpellier II, Cc 051,
Place Eugène Bataillon, 34095 Montpellier Cedex 5, France
(e-mail: `rouviere@ensam.inra.fr`)

Abstract. Let X be a random variable taking values in $L_2\big([0,1]\big)$ and let Y be a random label with values in $\{0,1\}$. Given a class of classifiers and n independent copies (X_i, Y_i) of the pair (X, Y), we show how to select optimally a particular classifier in the class and derive its consistency properties. To build our classifier, we first reduce the dimension of the functional observations using a particular thresholding on the coefficients of the curves X_i expressed in a wavelet basis. Then a classification rule working in finite dimension is performed on the selected coefficients. The dimension is automatically selected by data-splitting and empirical risk minimization. An application of this technique to a signal discrimination problem involving speech recognition is presented.

Keywords: Functional Data Analysis, Classification, Wavelets.

1 Introduction

The problem of pattern recognition (or classification or discrimination) is about guessing or predicting the unknown class of an observation. An observation is usually a collection of numerical measurements represented by a d-dimensional vector. In many real-life problems however, input data are in fact sampled functions rather than standard high dimensional vectors, and this casts the classification problem into the class of Functional Data Analysis.

Although standard pattern recognition techniques appear to be feasible, the intrinsic infinite dimensional structure of the observations makes learning suffer from the curse of dimensionality (see [Abraham *et al.*, 2003] for a detailed discussion, examples and counterexamples). In practice, before applying any learning technique to model real data, a preliminary dimension reduction or model selection step reveals crucial for appropriate smoothing and circumscription of the dimensionality effect. As a matter of fact, filtering is a popular dimension reduction method in signal processing, and this is the central approach we take in this paper.

Roughly, filtering reduces the infinite dimension of the observations by considering only the first d coefficients of the data on an appropriate basis. This approach was followed by [Kirby and Sirovich, 1990], [Comon, 1994], [Belhumeur *et al.*, 1997], [Hall *et al.*, 2001], or [Amato *et al.*, 2005]. Given a collection of functions we wish to classify, [Biau *et al.*, 2005] propose to use first Fourier filtering on each function and then perform k-nearest neighbor classification in $\mathbb{R}^d$. These authors study finite sample and asymptotic properties of a data-driven procedure that selects simultaneously both the dimension d and the optimal number of neighbors k.

The aim of the present paper is to extend the data-based filtering approach of [Biau *et al.*, 2005] to wavelet bases and general discrimination rules. Our motivation is twofold.

- First, as pointed out for example in [Amato *et al.*, 2005], wavelet bases offer some significant advantages over other bases. Indeed, wavelets can be used successfully for compression of a stochastic process, in the sense that the sample paths can be accurately reconstructed from a fraction of the full set of wavelet coefficients. Further, the wavelet decomposition of the sample paths is a local one, so that if the information relevant to the classification problem is contained in a particular part of the sample functions, as typically it is, this information will be carried by a very small number of wavelet coefficients. Moreover, the ability of wavelets to model the signal at different levels of resolution means that we have the option of selecting from the paths at a range of bandwidths.
- Second, we seek for general performance bounds and consistency results when using (finite dimensional approximations of) the sample data in the selection of a discrimination rule and/or its parameters. This article offers both a practical methodology and general performance results for all those who are willing to use wavelet filtering as a dimension reduction step before effective classification.

Throughout the manuscript, we will adopt the point of view of automatic pattern recognition described, to a large extend, in [Devroye, 1988]. In this setup, one uses a test sequence to select the best rule from a rich class of discrimination rules defined in terms of a training sequence. For the clarity of the paper, all important concepts regarding this classification paradigm are summarized in the next section. In Section 3, we outline the method and state consistency of our classification rule. Section 4 offers some experimental results on real-life data.

2 Automatic pattern recognition

This section gives a brief exposition and set up terminology of automatic pattern recognition. For a detailed introduction, the reader is referee to [Devroye, 1988].

To model the automatic learning problem, we introduce a probabilistic setting. Denote by $\mathcal{F} = L_2([0,1])$ the space of all square integrable functions on $[0,1]$. The data consist of a sequence of $n+m$ i.i.d. $\mathcal{F} \times \{0,1\}$-valued random variables $(X_1, Y_1), \ldots, (X_{n+m}, Y_{n+m})$. The X_i's are the *observations*, and the $Y_i's$ are the *labels*[1]. Note that the data are artificially split by us into two independent sequences, one of length n, and one of length m: we call the n sequence the *training sequence*, and the m sequence the *testing sequence*. A discrimination rule is a function $g : \mathcal{F} \times (\mathcal{F} \times \{0,1\})^{n+m} \rightarrow \{0,1\}$. It classifies a new observation $x \in \mathcal{F}$ as coming from class $g\left(x, (X_1, Y_1), \ldots, (X_{n+m}, Y_{n+m})\right)$. We will write $g(x)$ for the sake of convenience.

The probability of error of a given rule g is

$$L_{n+m}(g) = \mathbf{P}\left\{g(X) \neq Y | (X_1, Y_1), \ldots, (X_{n+m}, Y_{n+m})\right\},$$

where (X, Y) is independent of the data sequence and is distributed as (X_1, Y_1). Although we would like $L_{n+m}(g)$ to be small, we know that it cannot be smaller than the Bayes probability of error

$$L^* = \inf_{s:\mathcal{F} \rightarrow \{0,1\}} \mathbf{P}\{s(X) \neq Y\},$$

(see [Devroye *et al.*, 1996], Theorem 2.1, page 10). In the learning process, we aim at constructing rules with small probability of error. To do this, we employ the learning sequence to design a class of data-dependent discrimination rules, and we use the testing sequence as an impartial judge in the selection process. More precisely, we denote by $\mathbf{D}_n$ a (possibly infinite) collection of functions $g : \mathcal{F} \times (\mathcal{F} \times \{0,1\})^n \rightarrow \{0,1\}$, from which a particular function $\hat{g}$ is selected by minimizing the *empirical risk* based upon the testing sequence:

$$\hat{L}_{n,m}(\hat{g}) = \frac{1}{m} \sum_{i=n+1}^{n+m} \mathbf{1}_{[\hat{g}(X_i) \neq Y_i]} = \min_{g \in \mathbf{D}_n} \frac{1}{m} \sum_{i=n+1}^{n+m} \mathbf{1}_{[g(X_i) \neq Y_i]}.$$

At this point, observe that

$$g(X_i) = g\left(X_i, (X_1, Y_1), \ldots, (X_n, Y_n)\right)$$

and

$$\hat{g}(X_i) = \hat{g}\left(X_i, (X_1, Y_1), \ldots, (X_n, Y_n)\right),$$

i.e., the discriminators themselves are based upon the training sequence only. Observe however that $\hat{g}$ depends on the *entire data set*, as the rest of the data is used for selecting the classifiers.

[1] In this study we restrict our attention to binary classification. The reason is simplicity and that the binary problem already captures many of the main features of more general problems. Even though there is much to say about multiclass classification, we will not approach this increasing field of research.

3 Dimension reduction for classification

3.1 Wavelet-based expansion of the observations

The theory of wavelets has recently undergone a period of rapid development with exciting implications for nonparametric function estimation. Wavelets are orthonormal basis functions that cut up signals into different frequency components, and then study each component with a resolution matched to its scale. The books of [Daubechies, 1992], [Meyer, 1992] and [Mallat, 1999] give detailed expositions of the mathematical aspects of wavelets.

To summarize in our context, we recall that $L_2([0,1])$ is approximated by a multiresolution analysis, *i.e.*, a ladder of closed subspaces

$$V_0 \subset V_1 \subset \ldots \subset L_2([0,1])$$

whose union is dense in $L_2([0,1])$, and where each V_j is spanned by 2^j orthonormal scaling functions $\phi_{j,k}, k = 0, \ldots, 2^j - 1$, such that $\mathrm{supp}(\phi_{j,k}) \subset [k2^{-j}, (k+1)2^{-j}]$. At each resolution level $j \geq 0$, the orthonormal complement W_j between V_j and V_{j+1} is generated by 2^j orthonormal wavelets $\psi_{j,k}, k = 0, \ldots, 2^j - 1$. Thus, the family

$$\bigcup_{j \geq 0} \{\psi_{j,k}\}_{k=0,\ldots,2^j-1}$$

completed by $\{\phi_{0,0}\}$ forms an orthonormal basis of $L_2([0,1])$. As a consequence, any observation X in $L_2([0,1])$ reads

$$X(t) = \sum_{j=0}^{\infty} \sum_{k=0}^{2^j-1} \zeta_{j,k} \psi_{j,k}(t) + \eta \phi_{0,0}(t), \quad t \in [0,1],$$

where

$$\zeta_{j,k} = \int_0^1 X(t) \psi_{j,k}(t) \mathrm{d}t \quad \text{and} \quad \eta = \int_0^1 X(t) \phi_{0,0}(t) \mathrm{d}t.$$

3.2 Consistent functional classification

In this paragraph, we present the construction of our classifier and discuss its consistency properties. Using the notation of Section 2, the data consist of a sequence of $n + m$ i.i.d. $L_2([0,1]) \times \{0,1\}$-valued random observations $(X_1, Y_1), \ldots, (X_{n+m}, Y_{n+m})$. Given a multiresolution analysis of $L_2([0,1])$ as explicited above, each observation X_i is expressed as a series expansion

$$X_i(t) = \sum_{j=0}^{\infty} \sum_{k=0}^{2^j-1} \zeta_{j,k}^i \psi_{j,k}(t) + \eta^i \phi_{0,0}(t), \quad t \in [0,1]. \tag{1}$$

For the sake of coherence, it will be convenient to reindex the sequence $\{\phi_{0,0}, \psi_{0,0}, \psi_{1,0}, \psi_{1,1}, \psi_{2,0}, \psi_{2,1}, \psi_{2,2}, \psi_{3,0}, ...\}$ into $\{\psi_1, \psi_2, \psi_3, ...\}$. With this scheme, expression (1) may be rewritten as

$$X_i(t) = \sum_{j=1}^{\infty} X_{ij}\psi_j(t), \quad t \in [0,1], \tag{2}$$

hence the random coefficients

$$X_{ij} = \int_0^1 X_i(t)\psi_j(t)\mathrm{d}t.$$

Let $\mathbf{X}_i = (X_{i1}, X_{i2}, \ldots)$ be the coefficients associated with X_i. Recall that the Hilbert space $L_2([0,1])$ is isomorphic with $\ell_2 = \big\{\mathbf{x} = (x_1, x_2, \ldots) : \sum_{j=1}^{\infty} x_j^2 < \infty\big\}$. Consequently, knowing X_i is the same as knowing $\mathbf{X}_i = (X_{i1}, X_{i2}, \ldots)$. In our quest of dimension reduction, we first fix in (1) a maximum resolution level J ($J \geq 0$, possibly function of n) so that

$$X_i(t) \approx \sum_{j=0}^{J-1}\sum_{k=0}^{2^j-1} \zeta_{j,k}^i \psi_{j,k}(t) + \eta^i\phi_{0,0}(t), \quad t \in [0,1]$$

or equivalently, using (2),

$$X_i(t) \approx \sum_{j=1}^{2^J} X_{ij}\psi_j(t), \quad t \in [0,1].$$

At this point, we could try to use these finite-dimensional approximations of the observations, and let the data select optimally one of the $2^{2^J} - 1$ subbases of $\{\psi_1, \ldots, \psi_{2^J}\}$. By doing so, we would face with an unreasonable overall algorithmic complexity, and therefore catastrophic subsequent performance bounds. Thus, in order to reduce the overall complexity of the problem, we suggest the following procedure.

First, for each $d = 1, \ldots, 2^J$, we assume to be given beforehand a (possibly infinite) collection $\mathbf{D}_n^{(d)}$ of rules $g^{(d)} : \mathbb{R}^d \times (\mathbb{R}^d \times \{0,1\})^n \to \{0,1\}$ working in $\mathbb{R}^d$ and using n d-dimensional learning data as input. For fixed training sequence $(x_1, y_1), \ldots, (x_n, y_n)$, denote by $\mathbf{C}_n^{(d)}$ the collection of all sets

$$\Big\{\{x \in \mathbb{R}^d : \phi(x) = 1\} : \quad \phi \in \mathbf{D}_n^{(d)}\Big\},$$

and define the shatter coefficient as

$$\mathbb{S}_{\mathbf{C}_n^{(d)}}(m) = \max_{z_1,\ldots,z_m \in \mathbb{R}^d} \text{Card}\,\big\{\{z_1, \ldots, z_m\} \cap C : C \in \mathbf{C}_n^{(d)}\big\}.$$

With a slight abuse of notation, we will denote by $\mathbb{S}_{\mathbf{C}_n}^{(J)}(m)$ the shatter coefficient corresponding to the collection of all rules $\{g^{(d)} : d = 1, \ldots, 2^J\}$ embedded in $\mathbb{R}^{2^J}$. Observe that

$$\mathbb{S}_{\mathbf{C}_n}^{(J)}(m) \leq \sum_{d=1}^{2^J} \mathbb{S}_{\mathbf{C}_n^{(d)}}(m). \tag{3}$$

Second, we let the n training data reorder the first 2^J basis functions $\{\psi_1, \ldots, \psi_{2^J}\}$ into $\{\psi_{j_1}, \ldots, \psi_{j_{2^J}}\}$ via the scheme

$$\sum_{i=1}^{n} X_{ij_1}^2 \geq \sum_{i=1}^{n} X_{ij_2}^2 \geq \ldots \geq \sum_{i=1}^{n} X_{ij_{2^J}}^2. \tag{4}$$

In other words, we just let the training sample decide by itself which basis functions carry the most significant information.

We finish the procedure by a **third** selection step: pick the *effective* dimension $d \leq 2^J$ and a classification rule $g^{(d)}$ in $\mathbf{D}_n^{(d)}$ by approximating each X_i by $\mathbf{X}_i^{(d)} = (X_{ij_1}, \ldots, X_{ij_d})$ (without loose of generality, we assume implicitly that the sequence (j_k) is ordered – if not, just reorder it).

We select the dimension d and the rule simultaneously, using the data-splitting device described in Section 2. Precisely, we select both d and $g^{(d)}$ optimally by minimizing the empirical probability of error based on the independent validation set, that is

$$\left(\hat{d}, \hat{g}^{(\hat{d})}\right) = \operatorname*{argmin}_{d=1,\ldots,2^J, g^{(d)} \in \mathbf{D}_n^{(d)}} \left[\frac{1}{m} \sum_{i=n+1}^{n+m} \mathbf{1}_{[g^{(d)}(X_i^{(d)}) \neq Y_i]}\right]. \tag{5}$$

Apart from being conceptually simple, this method leads to the classifier $\hat{g}(\mathbf{x}) = \hat{g}^{(\hat{d})}(\mathbf{x}^{(\hat{d})})$ with a probability of misclassification

$$L_{n+m}(\hat{g}) = \mathbf{P}\big\{\hat{g}(\mathbf{X}) \neq Y \,|\, (\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_{n+m}, Y_{n+m})\big\}.$$

The selected rule $\hat{g}$ satisfies the following optimal inequality.

Theorem 1

$$\mathbf{E}\big\{L_{n+m}(\hat{g})\big\} - L^* \leq L_{2^J}^* - L^* + \mathbf{E}\big\{\inf_{\substack{d=1,\ldots,2^J \\ g^{(d)} \in \mathbf{D}_n^{(d)}}} L_n(g^{(d)})\big\} - L_{2^J}^*$$
$$+ 2\,\mathbf{E}\left\{\sqrt{\frac{8 \log\left(4\mathbb{S}_{\mathbf{C}_n}^{(J)}(2m)\right)}{m}} + \frac{1}{\sqrt{(m/2)\log\left(4\mathbb{S}_{\mathbf{C}_n}^{(J)}(2m)\right)}}\right\}.$$

Here

$$L^*_{2^J} = \inf_{s:\mathbb{R}^{2^J}\to\{0,1\}} \mathbf{P}\{s(\mathbf{X}^{(2^J)}) \neq Y\}$$

stands for the Bayes probability of error when the feature space is $\mathbb{R}^{2^J}$*.*

We may view the first term, $L^*_{2^J} - L^*$, on the right of the inequality as an approximation term – the price to be paid for using a finite dimensional approximation – and it converges to zero. The second term,

$$\mathbf{E}\big\{ \inf_{\substack{d=1,\dots,2^J \\ g^{(d)}\in\mathbf{D}_n^{(d)}}} L_n(g^{(d)})\big\} - L^*_{2^J}$$

can be handled by standard results on classifications. Let us first recall the definition of a *consistent* rule: a rule g is consistent if $\mathbf{E}\big\{L_n(g)\big\} \to L^*$ as $n \to \infty$.

Corollary 1 *Let $J \geq 0$ be a fixed integer. Assume that from each $\mathbf{D}_n^{(2^J)}$, $n \geq 1$, we can pick one $g_n^{(2^J)}$ such that the sequence $(g_n^{(2^J)})_{n\geq 1}$ is consistent for a certain class of distributions. Then the automatic rule $\hat{g}$ defined in (5) is consistent for the same class of distributions,* i.e.,

$$\mathbf{E}\big\{L_{n+m}(\hat{g})\big\} \to L^* \quad \textit{as } n \to \infty$$

if

$$\lim_{n\to\infty} J = \infty, \quad \lim_{n\to\infty} m = \infty, \quad \textit{and} \quad \lim_{n\to\infty} \mathbf{E}\left\{\frac{\log \mathbb{S}^{(J)}_{\mathbf{C}_n}(2m)}{m}\right\} = 0.$$

This consistency result is new and is especially valuable since few theoretical results have been established for functional classification. Corollary 1 shows that a consistent rule is selected if, for each fixed $J \geq 0$, the sequence of $\mathbf{D}_n^{(2^J)}$'s contains a consistent rule, even if we do not know which functions from $\mathbf{D}_n^{(2^J)}$ lead to consistency. If we are just worried about consistency, Corollary 1 reassures us that nothing is lost as long as we take m much larger than $\log \mathbf{E}\left\{\mathbb{S}_{\mathbf{C}_n^{(J)}}(2m)\right\}$. Often, this reduces to a very weak condition on the size m of the testing set and the maximum resolution J. Note also that it is usually possible to find upper bounds on the random variable $\mathbb{S}_{\mathbf{C}_n^{(J)}}(2m)$ that depend on n,m and J, but not on the actual values of the random variables $(X_1, Y_1), \dots, (X_n, Y_n)$. In this case, the bound is distribution-free, and the problem is purely combinatorial: count $\mathbb{S}^{(J)}_{\mathbf{C}_n}(2m)$. For example, if $\mathbf{D}_n^{(d)}$ contains all nearest-neighbor rules, a trivial bound is

$$\mathbb{S}_{\mathbf{C}_n^{(d)}}(2m) \leq n$$

because there are only n members in $\mathbf{D}_n^{(d)}$. Consequently

$$\mathbb{S}_{\mathbf{C}_n}^{(J)}(2m) \leq 2^J n\,.$$

[Stone, 1977] proved the striking result that k-nearest neighbor classifiers are consistent if $X \in \mathbb{R}^d$, provided $k \to \infty$ and $k/n \to \infty$. Thus we see that our strategy leads to a consistent rule whenever $J/m \to 0$ and $\log n/m \to 0$ as $n \to \infty$. For other examples, we refer to [Devroye, 1988].

4 Application to a speech recognition problem

In this section, we illustrate performance of our method. To this aim, we study a part of TIMIT database which was investigated in [Hastie *et al.*, 1995]. The data are log-periodograms corresponding to recording phonemes of 32 ms duration. We are concerned with the discrimination of five speech frames corresponding to five phonemes transcribed as follows : “aa” as the vowel in “dark” (695 items), “a0” as the first vowel in “water” (1022 items), “dcl” as in “dark” (757 items), “iy” as the vowel in “she” (1163 items) and “sh” as in “she” (872 items). The database is a multispeaker database. Each speaker is recorded at a 16k-Hz sampling rate and we retain only the first 256 frequencies (see Figure 1). Thus, the data consist of 4509 series of length 256 with known class membership.

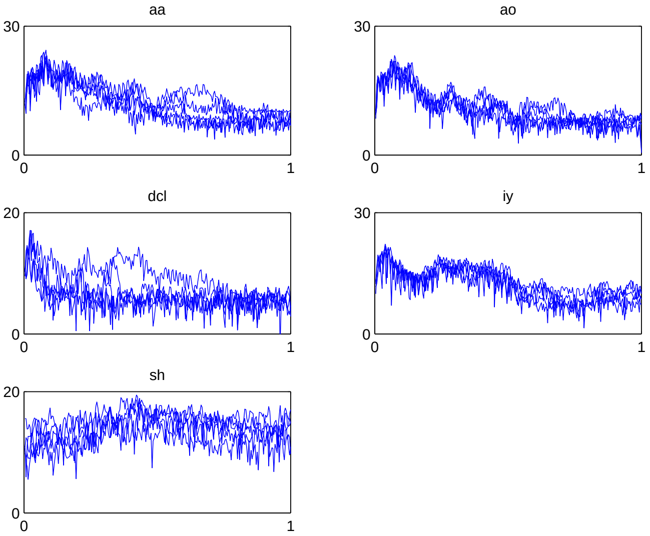


Fig. 1. A sample of 5 log-periodograms per class.

We first compute the wavelet filtering approach described in Section 3 using three collections of rules $\mathbf{D}_n^{(d)}$ working in $\mathbb{R}^d$. Precisely:

- W-LDA denotes the wavelet filtering followed by the class $\mathbf{D}_n^{(d)}$ of all linear discrimination rules.
- W-NN denotes the wavelet filtering followed by the class $\mathbf{D}_n^{(d)}$ of all nearest-neighbor rules.
- W-T denotes the wavelet filtering followed by the class $\mathbf{D}_n^{(d)}$ of all binary trees in which each internal node corresponds to a split perpendicular to one of the axes [Devroye *et al.*, 1996].

In addition, we propose to compare our algorithm with two existing alternative approaches:

- F-NN refers to the Fourier filtering approach combined with the k nearest-neighbor rule described in [Biau *et al.*, 2005].
- MPLSR refers to the multivariate partial least square regression. This approach is studied in detail in [Preda and Saporta, 2002] and is used as a benchmark in our context. The number of PLS components is selected by minimizing the empirical probability of error based on the testing sequence.

We use the split sample approach presented in Section 2 to select the free parameters. The *training sequence* and the *testing sequence* both contain 250 observations. The error rate ($e.r.$) for classifying new observations is unknown, but it can be estimated using the rest of the data:

$$e.r. = \frac{1}{3509} \sum_{i=501}^{4509} \mathbf{1}_{[\hat{g}(X_i) \neq Y_i]} \,,$$

where $\hat{g}$ denotes the selected rule. Table 1 displays the estimated error rates for the different methods together with the dimensions selected (number of PLS components for MPLSR). Results are averaged over 50 random partitions of the data.

Method	$e.r.$	$\hat{d}$
W-LDA	0.0854	18.70
W-NN	0.1096	19.52
W-T	0.1253	9.10
F-NN	0.1277	48.76
MPLSR	0.0904	5.96

Table 1. Estimated error rates.

We see that method W-LDA achieves the best estimated error rates, and that its results are slightly inferior to method MPLSR. The results of the

Fourier-based algorithm are still acceptable, because of a good localisation of the signal.

References

[Abraham *et al.*, 2003]C. Abraham, G. Biau, and B. Cadre. On the kernel rule for function classification. Technical report, University Montpellier II, 2003. http://www.math.univ-montp2.fr/∼biau/publications.html.

[Amato *et al.*, 2005]U. Amato, A. Antoniadis, and I. De Feis. Dimension reduction in functional regression with applications. *Computational Statistics and Data Analysis*, 2005. In press.

[Belhumeur *et al.*, 1997]P.N. Belhumeur, J.P. Hepana, and D.J. Kriegman. Eigenfaces vs. fisherfaces: Recognition using class specific linear projection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19:711–720, 1997.

[Biau *et al.*, 2005]G. Biau, F. Bunea, and M. Wegkamp. Functional classification in hilbert spaces. *IEEE Transactions on Information Theory*, 2005. In press.

[Comon, 1994]P. Comon. Independent component analysis, a new concept? *Signal Processing*, 36:287–314, 1994.

[Daubechies, 1992]I. Daubechies. *Ten Lectures on Wavelets.* Society for Industrial and Applied Mathematics, 1992.

[Devroye *et al.*, 1996]L. Devroye, L. Györfi, and G. Lugosi. *A Probabilistic Theory of Pattern Recognition.* Springer–Verlag, New-York, 1996.

[Devroye, 1988]L. Devroye. Automatic pattern recognition: a study of the probability of error. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 10:530–543, 1988.

[Hall *et al.*, 2001]P. Hall, D.S. Poskitt, and B. Presnell. A functional data-analytic approach to signal discrimination. *Technometrics*, 43:1–9, 2001.

[Hastie *et al.*, 1995]T. Hastie, A. Buja, and R. Tibshirani. Penalized discriminant analysis. *Annals of Statistics*, 23:73–102, 1995.

[Kirby and Sirovich, 1990]M. Kirby and L. Sirovich. Application of the karhunen-loeve procedure for the characterization of human faces. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12:103–108, 1990.

[Mallat, 1999]S. Mallat. *A Wavelet Tour of Signal Processing.* Academic Press, 1999. 2nd edition.

[Meyer, 1992]Y. Meyer. *Wavelet and Operators.* Cambridge University Press, Cambridge, 1992.

[Preda and Saporta, 2002]C. Preda and G. Saporta. Régression PLS sur un processus stochastique. *Revue de Statistique Appliquée*, 50(2), 2002.

[Stone, 1977]C.J. Stone. Consistent nonparametric regression. *Annals of Statistics*, 5:595–645, 1977.

PLS discriminant analysis for functional data

Cristian Preda[1] and Gilbert Saporta[2]

[1] Dept. de Statistique
CERIM - Faculté de Médecine
Université de Lille 2,
59045 Lille Cedex, France
(e-mail: cpreda@univ-lille2.fr)
[2] Chaire de Statistique Appliquée
CEDRIC, CNAM - Paris
292, Rue Saint Martin,
75141 Paris Cedex 03, France
(e-mail: saporta@cnam.fr)

Abstract. Partial least squares regression on functional data is applied in the context of linear discriminant analysis with binary response. The discriminant coefficient function is then used to compute scores which allow to assign a new curve to one of the two classes. The method is applied to gait data and the results are compared with those given by linear discriminant analysis and logistic regression on the principal components of predictors.

Keywords: PLS, Second order stochastic process, Functional data, Linear discriminant analysis.

1 Introduction

Functional data analysis extends the classical multivariate methods when data are functions or curves. Examples of functional data can be found in different fields of application such as medicine, economics, chemometrics and many others (see [Ramsay and Silverman, 2002] for an overview). Figure 1 gives an example of such data. A well accepted model for this kind of data is to consider it as paths of a stochastic process $X = \{X_t\}_{t \in T}$ taking values in a Hilbert space of functions on some set T.

In this paper we consider X to be a second order stochastic process $X = \{X_t\}_{t \in [0,1]}$, L_2–continuous and with sample paths in $L_2([0,1])$. Let also Y be a binary random variable, for instance, $Y \in \{0,1\}$, defined on the same probability space as X.

As formulated by Fisher in the classical setting (finite dimensional predictor), the aim of the linear discriminant analysis (LDA) of (X, Y) is to find the linear combination $\Phi(X) = \int_0^1 X_t \beta(t) dt$, $\beta \in L_2[0,1]$, such that the between-class variance is maximized relative to the total variance, i.e.

$$\max_{\beta \in L_2[0,1]} \frac{\mathbb{V}(\mathbb{E}(\Phi(X)|Y))}{\mathbb{V}(\Phi(X))}. \tag{1}$$

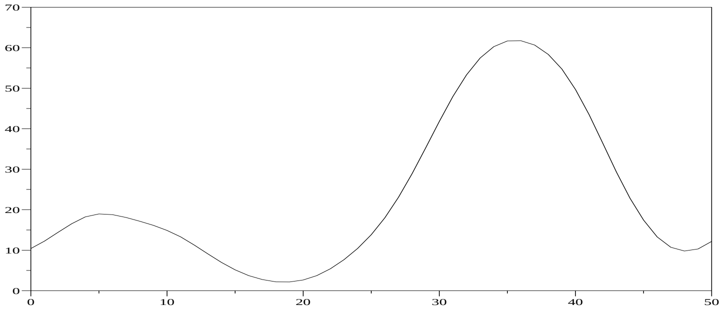


Fig. 1. Knee angular rotation over a complete gait cycle for one subject.

The random variable $\Phi(X)$ is referred as discriminant variable and the function β as discriminant coefficient function ([Hastie *et al.*, 2001]).

In the context of functional data, the estimation problem for the discriminant coefficients function, β, is generally an ill–posed one. Indeed, is well known that the optimization problem (1) is equivalent to find the regression coefficients of the regression of Y (after a convenient encoding) on the stochastic process X under the least-squares criterion. [Cardot *et al.*, 1999], [Preda and Saporta, 2002] point out the inconsistency of such a criterion for this kind of predictors and propose solutions to overcome this difficulty. From practical point of view, a large number of predictors (relatively to the size of the learning sample) as well as the multicollinearity of predictors, lead to inconsistent estimators. Nonparametric approaches for functional discriminant analysis are proposed in [Ferraty and Vieu, 2003] and [Biau *et al.*, 2004]. Logistic regression for functional data using the projection method [Aguilera *et al.*, 1998] is given in [Escabias *et al.*, 2004] and [Araki and Sadanori, 2004].

The aim of this paper is to perform LDA using the Partial Least Squares (PLS) approach developed in [Preda and Saporta, 2002]. The paper is organized as follows. In section 2 we introduce some basic results on the linear regression on functional data and the PLS approach. The relationship between LDA and linear regression is given in section 3. The section 4 presents an application of the PLS approach for LDA using gait data provided by the Center of Neurophysiology of the Regional Hospital of Lille (France). The goal is to separate young and senior patients from the curve given by the knee angular rotation over a complete gait cycle. The results are compared with those given by the LDA and the logistic regression using as predictors the principal components of data. The comparison of methods is made using the criterion based on the area under the ROC curve.

2 Some tools for linear regression on a stochastic process

As stated above, let $X = \{X_t\}_{t\in[0,1]}$ be a second order stochastic process L_2-continuous and with sample paths in $L_2[0,1]$ and Y a real random variable. Without loss of generality we assume also that $E(X_t) = 0$, $\forall t \in [0,1]$ and $E(Y) = 0$.

It is well known that the approximation of Y obtained by the classical linear regression on X, $\hat{Y} = \int_0^1 \beta(t)X_t dt$ is such that β is in general a distribution rather than a function of $L_2([0,1])$ ([Saporta, 1981]). This difficulty appears also in practice when one tries to estimate the regression coefficients, $\beta(t)$, using a sample of size N. Indeed, if $\{(Y_1, X_1, (Y_2, X_2), \ldots (Y_N, X_N)\}$ is a finite sample of (Y, X), the system

$$Y_i = \int_0^1 X_i(t)\beta(t)dt, \quad \forall i = 1, ..., N,$$

has an infinite number of solutions ([Ramsay and Silverman, 1997]). Regression on principal components (PCR) of $(X_t)_{t\in[0,1]}$ ([Aguilera *et al.*, 1998]) and PLS approach ([Preda and Saporta, 2002]) give satisfactory solutions to this problem.

2.1 Linear regression on principal components

Also known as Karhunen-Loève expansion, the principal component analysis (PCA) of the stochastic process $(X_t)_{t\in[0,1]}$ consists in representing X_t as :

$$X_t = \sum_{i\geq 1} f_i(t)\xi_i, \quad \forall t \in [0,1], \tag{2}$$

where the set $\{f_i\}_{i\geq 1}$ (the principal factors) forms an orthonormal system of deterministic functions of $L_2([0,1])$ and $\{\xi_i\}_{i\geq 1}$ (principal components) are uncorrelated zero-mean random variables. The principal factors $\{f_i\}_{i\geq 1}$ are solution of the eigenvalue equation :

$$\int_0^1 C(t,s)f_i(s)ds = \lambda_i f_i(t), \tag{3}$$

where $C(t,s) = \mathrm{cov}(X_t, X_s)$, $\forall t, s \in [0,1]$. Therefore, the principal components $\{\xi_i\}_{i\geq 1}$ defined as $\xi_i = \int_0^1 f_i(t)X_t dt$ are eigenvectors of the Escoufier operator, $\mathbf{W}^X$, defined by

$$\mathbf{W}^X Z = \int_0^1 E(X_t Z)X_t dt, \quad Z \in L_2(\Omega). \tag{4}$$

The process $\{X_t\}_{t\in[0,1]}$ and the set of its principal components, $\{\xi_k\}_{k\geq 1}$, span the same linear space. Thus, the regression of Y on X is equivalent to

the regression on $\{\xi_k\}_{k\geq 1}$ and we have $\hat{Y} = \sum_{k\geq 1} \frac{E(Y\xi_k)}{\lambda_k}\xi_k$.
In practice we need to choose an approximation of order q, $q \geq\ 1$:

$$\hat{Y}_{PCR(q)} = \sum_{k=1}^{q} \frac{E(Y\xi_k)}{\lambda_k}\xi_k = \int_0^1 \hat{\beta}_{PCR(q)}(t)X_t dt. \tag{5}$$

But the use of principal components for prediction is heuristic because they are computed independently of the response. One alternative is the PLS approach which builds directions for regression (PLS components) taking into account the response variable Y.

2.2 PLS regression on a stochastic process

The PLS (Partial Least Squares) approach offers a good alternative to the PCR method by replacing the least squares criterion with that of maximal covariance between $(X_t)_{t\in[0,1]}$ and Y ([Preda and Saporta, 2002]).

The PLS regression is an iterative method. Let $X_{0,t} = X_t$, $\forall t \in [0,1]$ and $Y_0 = Y$. At step q, $q \geq 1$, of the PLS regression of Y on X, we define the q^{th} PLS component, t_q, by the eigenvector associated to the largest eigenvalue of the operator $\mathbf{W}^X_{q-1}\mathbf{W}^Y_{q-1}$, where $\mathbf{W}^X_{q-1}$, respectively $\mathbf{W}^Y_{q-1}$, are the Escoufier's operators associated to X, respectively to Y_{q-1}. The PLS step is completed by the ordinary linear regression of $X_{q-1,t}$ and Y_{q-1} on t_q. Let $X_{q,t}$, $t \in [0,1]$ and Y_q be the random variables which represent the residual of these regressions : $X_{q,t} = X_{q-1,t} - p_q(t)t_q$ and $Y_q = Y_{q-1} - c_q t_q$.

Then, for each $q \geq 1$, $\{t_q\}_{q\geq 1}$ forms an orthogonal system in $L_2(X)$ and the following decomposition formulas hold :

$$Y = c_1 t_1 + c_2 t_2 + \ldots + c_q t_q + Y_q,$$
$$X_t = p_1(t)t_1 + p_2(t)t_2 + \ldots + p_q(t)t_q + X_{q,t}, \quad t \in [0,1].$$

The PLS approximation of Y by $\{X_t\}_{t\in[0,1]}$ at step q, $q \geq 1$, is given by :

$$\hat{Y}_{PLS(q)} = c_1 t_1 + \ldots + c_q t_q = \int_0^1 \hat{\beta}_{PLS(q)}(t)X_t dt. \tag{6}$$

[de Jong, 1993] and [Phatak and De Hoog, 2001] show that for a fixed q, the PLS regression fits closer than PCR, that is,

$$R^2(Y, \hat{Y}_{PCR(q)}) \leq R^2(Y, \hat{Y}_{PLS(q)}). \tag{7}$$

In [Preda and Saporta, 2002] we show the convergence of the PLS approximation to the approximation given by the classical linear regression :

$$\lim_{q\to\infty} E(|\hat{Y}_{PLS(q)} - \hat{Y}|^2) = 0. \tag{8}$$

In practice, the number of PLS components used for regression is determined by cross-validation ([Tenenhaus, 1998]).

3 LDA and linear regression for functional data

Let us denote by

$$p_0 = \mathrm{P}(Y=0),\ p_1 = 1-p_0 = \mathrm{P}(Y=1),$$
$$\mu_0(t) = \mathbb{E}(X_t|Y=0),\ \mu_1(t) = \mathbb{E}(X_t|Y=1), t \in [0,1].$$

Since $\mathbb{E}(X_t)=0$, it follows that $p_0\mu_0(t)+p_1\mu_1(t)=0$, $\forall t \in [0,1]$.
Let also denote by $\mathbf{C}$ the covariance operator associated to the process X defined on $L_2[0,1]$ by

$$f \xmapsto{\mathbf{C}} g, \quad g(t) = \int_0^1 \mathbb{E}(X_t X_s) f(s) ds,$$

and by $\mathbf{B}$ the operator on $L_2[0,1]$ defined by

$$f \xmapsto{\mathbf{B}} g, \quad g(t) = \int_0^1 B(t,s) f(s) ds,$$

where $B(t,s) = p_0\mu_0(t)\mu_0(s) + p_1\mu_1(s)\mu_1(t) = p_0p_1(\mu_0(t)-\mu_1(t))(\mu_0(s)-\mu_1(s))$. Denoting by $\phi = \sqrt{p_0p_1}(\mu_0-\mu_1)$, it follows that

$$\mathbf{B} = \phi \otimes \phi,$$

where $\phi \otimes \phi(g) = \phi\langle \phi, g\rangle_{L_2[0,1]}$, $g \in L_2[0,1]$.

As in the classical setting, the discriminant coefficient function, $\beta \in L_2[0,1]$, which satisfies the criterion given in (1), corresponds to the largest λ, $\lambda \in \mathbb{R}$, such that

$$\mathbf{B}\beta = \lambda \mathbf{C}\beta, \tag{9}$$

with $\langle \beta, \mathbf{C}\beta\rangle_{L_2[0,1]} = 1$.

Without loss of generality, let us recode Y by : $0 \rightsquigarrow \sqrt{\frac{p_1}{p_0}}$ and $1 \rightsquigarrow -\sqrt{\frac{p_0}{p_1}}$. If β is a solution of (9) then $\lambda = \langle \phi, \beta\rangle^2_{L_2[0,1]}$ and β is solution of the Wiener-Hopf equation

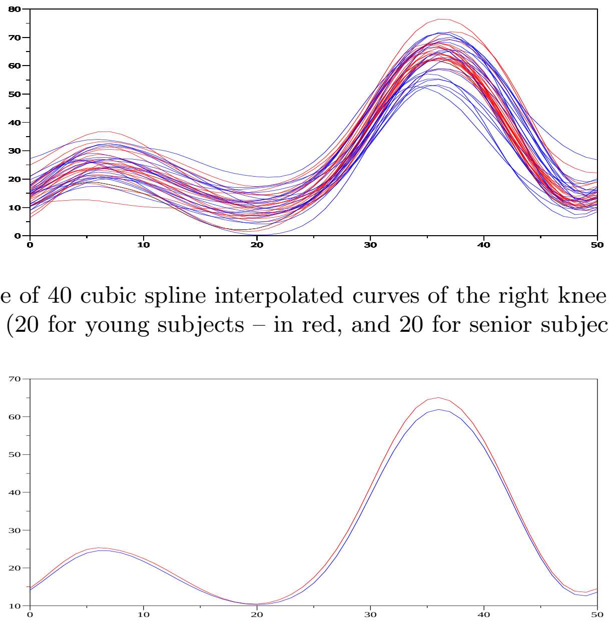
$$\mathbb{E}(YZ_t) = \int_0^1 \mathbb{E}(Z_t Z_s)\beta(s) ds, \tag{10}$$

where $Z_t = \langle \phi, \beta\rangle_{L_2[0,1]} X_t$, $t \in [0,1]$. The function β given by equation (10) is the regression coefficient function of the linear regression of Y on $Z = \{Z_t\}_{t\in[0,1]}$. Equation (10) has an unique solution under conditions of convergence of series implying the eigenvalues and eigenvectors of the covariance operator of the process X [Saporta, 1981]. These conditions are rarely satisfied. Thus, in practice, the problem to find β is generally an ill-posed problem.

However, if the aim is to find the discriminant variable (scores), then one can use the above relationship between LDA and linear regression. The regularized linear methods proposed in Section 2 provides good approximations by using (5) and (6) with Y recoded as above. Then $\hat{\beta}_{PCR_{(q)}}$ and $\hat{\beta}_{PLS_{(q)}}$ can be used to compute the discriminant score for a new observation for which one has only the observation of X. The prediction for a new observation is given with respect to a reference score value which is determined on a test sample such that the classification error rate is minimum.

4 Application to gait data

The application deals with data provided by the Department of Movement Disorders, Lille University Medical Center (France). This data is described by a set of curves representing the knee flexion angle evolution over one complete gait cycle and characterizes patients from two classes of age ([Duhamel *et al.*, 2004]). We are interested in predicting the class of age from the knee curve.



a) A sample of 40 cubic spline interpolated curves of the right knee angular rotation (20 for young subjects – in red, and 20 for senior subjects – in blue).

b) Mean estimation of angular rotation of the right knee during a complete cycle for each group.

Fig. 2. Knee flexion angular data

Two groups of 30 subjects were studied : 30 young students (mean age 27 years and standard deviation 4 years) and 30 healthy senior citizens (mean age 64 years and standard deviation 6 years). For each subject the observed data represent the flexion angle for the right knee measured during one complete gait cycle. Each curve represents a gait cycle and is given by a set $\{(x_{t_i}, t_i)\}_{i=1,\ldots,50}$ of 50 values corresponding to an equidistant discretisation of the cycle.

We assume that data represent sample paths of a stochastic process $\{X_t\}_{t \in T}$ of second order and L_2 continuous. Also, it is natural to consider that the paths are derivable functions of time (percent of gait cycle) and therefore, cubic spline interpolation is performed for each curve.

Data is randomly divided into two samples, a learning sample of 40 subjects (Figure 2a) and a test sample of 20 patients. Each sample contains the same number of young and senior subjects.

In order to approximate the discriminant variable $\Phi(X) = \int_0^1 X_t \beta(t) dt$, we use the PLS regression ([Preda and Saporta, 2002]) for binary response. The number of PLS components in the model is given by cross validation [Tenenhaus, 1998]. A PLS model with q components is quoted by $LDA\_PLS(q)$. In our example $q = 3$ and the proportion of inertia of X explained by $\{t_1, t_2, t_3\}$ is 0.825. The PLS approach is compared with linear discriminant analysis and logistic regression using the principal components of $X = \{X_t\}_{t \in [0,1]}$ as predictors (the four first principal components explain 94.64% of the total inertia of X). Let us quote by $LDA\_PCR(q)$ and $LogPCR(q)$ these models using the q first principal components. The logistic regression using q PLS components is quoted by $LogPLS(q)$. The comparison criterion is the area under the ROC (Receiver Operating Characteristic) curve (Figure 3) estimated on the test sample.

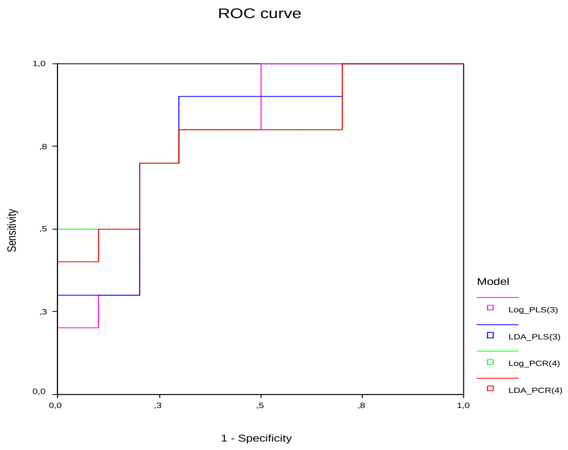


Fig. 3. ROC curves for each discriminant function.

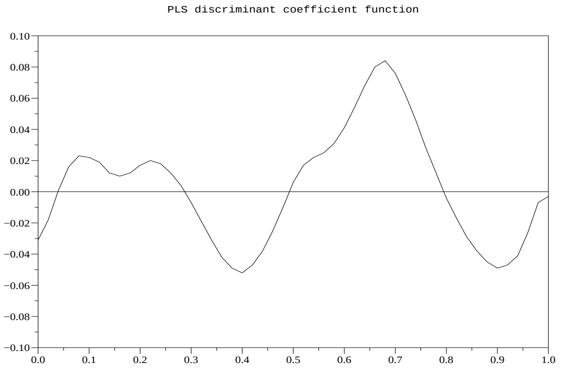


Fig. 4. Discriminant coefficient function $\hat{\beta}_{PLS(3)}$ for LDA_PLS(3)

Model	LDA_PLS(3)	LDA_PCR(4)	Log_PCR(4)	Log_PLS (3)
Area	0.790	0.780	0.790	0.780

Table 1. Area under the ROC curve. Sample test estimation.

5 Conclusion

PLS regression on functional data is used for linear discriminant analysis with binary response. It is an interesting alternative to classical linear methods based on principal components of predictors. Our intuition that similar or better results may be obtained with less PLS components than principal components is confirmed by an example on medical data.

Acknowledgements

We are grateful to Department of Movement Disorders, Roger Salengro Hospital, Lille University Medical Center (France) for providing us with data for testing our methods.

References

[Aguilera *et al.*, 1998]A.M. Aguilera, F. Ocaña, and M.J. Valderrama. An approximated principal component prediction model for continous-time stochastic process. *Applied Stochastic Models and Data Analysis*, pages 61–72, 1998.

[Araki and Sadanori, 2004]Y. Araki and K. Sadanori. Functional regression models via regularized radial basis function networks. In *The 2004 Hawaii International Conference on Statistics and Related Fields*, 2004.

[Biau *et al.*, 2004]G. Biau, F. Bunea, and M. Wegkamp. Function classification in Hilbert spaces. *Submitted*, 2004.

[Cardot *et al.*, 1999]H. Cardot, F. Ferraty, and P. Sarda. Functional linear model. *Statist. Probab. Lett.*, pages 11–22, 1999.

[de Jong, 1993]S. de Jong. PLS fits closer than PCR. *Journal of Chemometrics*, pages 551–557, 1993.

[Duhamel *et al.*, 2004]A. Duhamel, J.L. Bourriez, P. Devos, P. Krystkowiak, A. Destée, P. Derambure, and L. Defebvre. Statistical tools for clinical gait analysis. *Gait and Posture*, pages 204–212, 2004.

[Escabias *et al.*, 2004]M. Escabias, A.M. Aguilera, and M.J. Valderrama. Modeling environmental data by functional principal component logistic regression. *Environmetrics*, 2004.

[Ferraty and Vieu, 2003]F. Ferraty and P. Vieu. Curves discrimination: a nonparametric approach. *Computational Statistics & Data Analysis*, pages 161–173, 2003.

[Hastie *et al.*, 2001]T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning. Data mining, Inference and Prediction.* Springer, 2001.

[Phatak and De Hoog, 2001]A. Phatak and F. De Hoog. PLSR, Lanczos, and conjugate gradients. *CSIRO Mathematical & Information Sciences*, pages 551–557, 2001.

[Preda and Saporta, 2002]C. Preda and G. Saporta. Régression PLS sur un processus stochastique. *Revue de Statistique Appliquée*, pages 27–45, 2002.

[Ramsay and Silverman, 1997]J. O. Ramsay and B.W. Silverman. *Functional Data Analysis.* Springer, 1997.

[Ramsay and Silverman, 2002]J. O. Ramsay and B.W. Silverman. *Applied Functional Data Analysis : Methods and Case Studies.* Springer, 2002.

[Saporta, 1981]G. Saporta. *Méthodes exploratoires d'analyse de données temporelles.* Cahiers du B.U.R.O., Université Pierre et Marie Curie, Paris, 1981.

[Tenenhaus, 1998]M. Tenenhaus. *La régression PLS. Théorie et pratique.* Editions Technip, Paris, 1998.

S-Class, A Divisive Clustering Method, and Possible "Dual" Alternatives

Jean-Paul Rasson, François Roland, Jean-Yves Pirçon, Séverine Adans, and Pascale Lallemand

Facultés Universitaires Notre-Dame de la Paix
Unité de Statistiques
5000 Namur, Belgique
(e-mail: jean-paul.rasson@fundp.ac.be, francois.roland@fundp.ac.be, severine.adans@fundp.ac.be)

Abstract. A new partitioning method based on the non-homogeneous Poisson processes is presented. The principle of construction is of hierarchical divisive monothetic type. A variable is selected at each stage to cut a group into two subsets in a recursive way. The criterion consists in maximizing the 'gap' between the data. This last-one is deduced from the maximum likelihood criterion. A pruning phase, that is a simplification of the tree structure, based on the Gap test is then performed. An application of this algorithm on the well-know Ichino's oils dataset (interval data) is described.

Keywords: Clustering trees, Non-homogeneous Poisson processes, Gap test, Symbolic data.

1 Introduction

One of the most common tasks in data analysis is the detection and construction of groups of objects in a population E such that objects from the same group show a high similarity whereas objects from different groups are typically more dissimilar. Such groups are usually called 'clusters' and must be constructed on the basis of the data which were recorded for the objects. This problem is know as clustering.

The present method is a divisive monothetic clustering method for a symbolic $n \times p$ data array $\underline{X}$.

The resulting classification structure is a k-partition.

2 Input Data: Interval Data

This algorithm studies the case where n symbolic objects are described by p interval variables $Y_1, \ldots, Y_p$.

The interval-valued variable $Y_j (j = 1, \ldots, p)$ is measured for each element of the basic set $E = \{1, \ldots, n\}$. For each element $x \in E$, we denote the interval $Y_j(x)$ by $[\underline{y}_{jx}, \bar{y}_{jx}]$, thus $\underline{y}_{jx}$ (resp. $\bar{y}_{jx}$) is the lower (resp. the upper) bound of the interval $Y_j(x) \subseteq \mathcal{R}$.

An example is given by table 1.

3 The Clustering Tree Method

The proposed algorithm is a recursive one intended to divide a given population of symbolic objects into classes. According to the clustering tree method, nodes are split recursively by choosing the best interval variable.

The original contribution of this method lies in the way of splitting a node. The cut will be based on the only assumption that the distributions of points can be modeled by non-homogeneous Poisson process, where the intensity will be estimated by the kernel method. The cut will be made in order to maximize the likelihood function.

3.1 General Hypothesis: Non-Homogeneous Poisson Process

The only assumption on which the clustering problem rests is that the observed points are generated by a non-homogeneous Poisson process with intensity $q(.)$ and are observed in E, where E is the union of k disjoint convex fields.

The likelihood function, for the observations $\underline{x} = (x_1, x_2, \ldots, x_n)$ with $x_i \in R^d, i = 1, \ldots, n$ is:

$$f_E(\underline{x}) = \frac{1}{(\rho(E))^n} \prod_{i=1}^{n} 1\!\!1_E(x_i).q(x_i)$$

where

- $\rho(E) = \int_E q(x)dx$ is the integrated intensity;
- $q(.)$ is the process intensity ($q(x) > 0 \ \forall x$).

Consequently, if the intensity of the process is known, the solution of the maximum likelihood will correspond to k disjoint convex fields containing all the points and for which the sum of the integrated intensities is minimal. For an homogenous Poisson process on the line, this gives exactly the N-N rule. When the intensity is unknown, it will be estimated.

3.2 Kernel Method

To estimate the intensity of a non-homogeneous Poisson process, the non-parametric kernel method is used. Because this algorithm proceeds in a monothetic way, formulas dont't need to be extended beyond one dimension. The kernel estimator, which is a sum of 'bumps', each of these centered on an observation, is defined by:

$$\hat{q}(x) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{h} K\Big(\frac{x - X_i}{h}\Big)$$

where

- K is the kernel and is a positive continuous symmetric function satisfying $\int K(x)dx = 1$. The kernel determines the shapes of the bumps.
- h is the window width, also called the smoothing parameter and determines the width of the bumps.

The choice of the smoothing parameter is important. If it is too small, the estimator degenerates into a succession of peaks located at each point of the sample. If it's too large, the estimation approaches an uniform law and then we will have a loss of details.

3.3 Bumps and Multi-modalities

Within the clustering context, Silverman ([Silverman, 1981], [Silverman, 1986]) defined a mode in a density f as a local maximum while a bump is characterized by an interval, in such way that the density is concave on this interval but not on a larger interval.

In the framework of density estimation by the kernel method, the number of modes will be determined by the smoothing parameter, following Silverman's assertion : the number of modes is a decreasing function of the smoothing parameter h ([Silverman, 1981],[Silverman, 1986]).

This has been proved at least for the normal kernel defined by :

$$K_{\mathcal{N}}(t) = \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}}.$$

Consequently, this one was prefered to perform estimation of the intensity of the non-homogeneous Poisson process.

Because of this choice, there is a critical value h_{crit} of the smoothing parameter for which the estimation changes from unimodality to multimodality. The split criterion will seek this value.

3.4 Splitting Criteria

For each variable, a dichotomic process computes the highest value of parameter h, giving a number of modes of the associated intensities strictly larger than 1. Once this h determined, E is split into two convex disjoint fields E_1 and E_2, such that $E = E_1 \cup E_2$, for which the likelihood function

$$f_{E_1,E_2}(\underline{x}) = \frac{1}{(\rho(E_1) + \rho(E_2))^n} \prod_{i=1}^{n} 1\!\!1_{E_1 \cup E_2} . \hat{q}(x_i)$$

is maximum, i.e. for which the integrated density $\rho(E_1) + \rho(E_2)$ is smallest.

Since the algorithm proceeds variable by variable, the best variable, i.e. the one which generates the "largest gap" (the density integrated on this gap is the largest), is selected.

This procedure is recursively performed until some stopping rule is fulfilled: the number of points in a node must be under a cut-off value.

3.5 Pruning Method

At the end of the splitting process, a large tree is obtained. A pruning method to select the best subtree was then developped. This pruning method takes the form of a classical hypothesis test: the Gap test ([Kubushishi, 1996], [Rasson and Kubushishi, 1994]).

The principle is the following: each cut is examined to determine if it is a good one (Gap Test satisfied) or a bad one (Gap Test unsatisfied). In the case of two classes D_1 and D_2, with $D_1 \cup D_2 = D$, the hypotheses are:

$$H_0\text{: there are } n = n_1 + n_2 \text{ points in } D_1 \cup D_2$$
$$\text{VS}$$
$$H_1\text{: there are } n_1 \text{ points in } D_1 \text{ and } n_2 \text{ points in } D_2\text{, with } D_1 \cap D_2 = \emptyset.$$

This pruning method crosses the tree branch by branch, from its root to its leaves, in order to index the good cuts and the bad cuts. The ends of the branches for which there are only bad cuts are pruned.

3.6 Application to Interval Data

The current problem is to apply this new method to symbolic data of interval type. Let an interval set

$$I = \{[a_i, b_i],\ i = 1, \dots, n,\ a_i \leq b_i\}.$$

The usual distance used for interval variables is the Hausdorff distance:

$$d_H([a_1, b_1], [a_2, b_2]) = Max(|a_1 - a_2|, |b_1 - b_2|)$$

or ([Chavent and Lechevallier, 2002], [Chavent, 1997])

$$d([a_1, b_1], [a_2, b_2]) = |M_1 - M_2| + |L_1 - L_2|$$

where $M_i = \frac{a_i + b_i}{2}$ is the middle point of the interval $[a_i, b_i]$ and $L_i = \frac{b_i - a_i}{2}$ is its half-length.

So each interval can be represented by its coordinates *(middle,half-length)*, on the space $(M, L) \subseteq I\!R \times I\!R^+$.

It is clear that separations must respect the order of the classes centers and thus, in the half-plane $I\!R \times I\!R^+$, only partitions invariant in relation to M are considered.

In the most general case of a non-homogeneous Poisson process, the integrated intensity has to be minimized:

$$\int_{M_i}^{M_{i+1}} \rho_1(m)dm + \int_{Min(L_i, L_{i+1})}^{Max(L_i, L_{i+1})} \rho_2(l)dl. \tag{1}$$

Any bipartition generated by a point being located inside the interval which maximizes (1) is appropriate.

3.7 Set of Binary Questions for Interval Data

In the framework of the divisive clustering method, the split of a node C is performed on the basis of one single variable (suitably chosen) and answers ([Chavent and Lechevallier, 2002], [Chavent, 1997]) to a specific binary question of type *'Is $Y_j \leq c$?'*, where c is called the cut value.

To the binary question *'Is $Y_j \leq c$?'*, an object $x \in C$ answers 'yes' or 'no' according to a binary function $q_c : E \rightarrow \{true, false\}$. The bipartition (C_1, C_2) of C induced by the binary question is as follows :

- $C_1 = \{x \in C \mid q_c(x) = true\}$
- $C_2 = \{x \in C \mid q_c(x) = false\}$

Consider the case of interval variables: Let $Y_j(x) = [\alpha, \beta]$, the middle of $[\alpha, \beta]$ is $m_x = \frac{\alpha+\beta}{2}$.

1. The binary question is "Is $m_x \leq c$?".
2. The function q_c is defined by:
   - $q_c(x) = true$ if $m_x \leq c$
   - $q_c(x) = false$ if $m_x > c$
3. The bipartition (C_1, C_2) of C induced by the binary question is :
   - $C_1 = \{x \in C \mid q_c(x) = true\}$
   - $C_2 = \{x \in C \mid q_c(x) = false\}$

3.8 Output Data and Results

After the tree-growing algorithm and the pruning procedure, the final clustering tree is obtained.

The nodes of the tree represent the binary questions selected by the algorithm and the k leaves of the tree define the k-partition. Each cluster is characterized by a rule, i.e, the path in the tree which provided it. The clusters therefore become new symbolic objects defined according to the binary questions leading from the root to the corresponding leaves.

4 Example on the Oils and Fats Data

The above clustering method has been examined with the well-known Ichino's oils dataset. The data set (Table 1) is composed of 8 oils described in terms of four interval variables.

This divisive algorithm yields the 3-cluster partition represented in the tree given in figure 1.

Two binary questions correspond to two binary functions $E \rightarrow \{true, false\}$, given by q_1 = [Spec. Grav.$(x) \leq 0.89075$] and q_2 = [Iod. Val.$(x) \leq 148.5$].

Each cluster corresponds to a symbolic object, e.g. a query assertion:

Sample	Specific Gravity	Freezing point	Iodine Value	Saponification Value
linseed oil	[0.930;0.935]	[-27;-18]	[170;204]	[118;196]
perilla oil	[0.930;0.937]	[-5;-4]	[192;208]	[188;197]
cottonseed oil	[0.916;0.918]	[-6;-1]	[99;113]	[189;198]
sesam oil	[0.920;0.926]	[-6;-4]	[104;116]	[187;193]
camelia oil	[0.916;0.917]	[-21;-15]	[80;82]	[189;193]
olive oil	[0.914;0.919]	[0;6]	[79;90]	[187;196]
beef tallow	[0.860;0.870]	[30;38]	[40;48]	[190;199]
hog fat	[0.858;0.864]	[22;32]	[53;77]	[190;202]

Table 1. *Table of oils and fats*

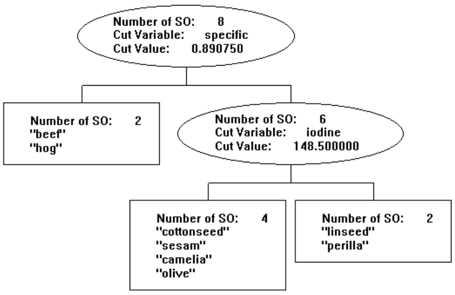


Fig. 1. Clustering tree obtained on the Ichino'oils dataset.

- $C_1 =$ [Spec. Grav.$(x) \leq 0.89075$],
- $C_2 =$ [Spec. Grav.$(x) > 0.89075$] $\wedge$ [Iod. Val.$(x) \leq 148.5$],
- $C_3 =$ [Spec. Grav.$(x) > 0.89075$] $\wedge$ [Iod. Val.$(x) > 148.5$].

Then, the resulting 3-cluster partition is: $C_1 = \{$beef, hog$\}$, $C_2 = \{$cottonseed, sesam, camelia, olive$\}$, $C_3 = \{$linseed, perilla$\}$.

5 Further Works and Conclusions

Following that work, a new clustering method was conceived. It's also a hierarchical clustering method but a multivariate agglomerative one. The basic idea was to find a merging criterion which would have been dual and complementary to the splitting one. But the strictly dual criterion, consisting in measuring the area sustended by the density between 2 points (or groups of points) and then merging the 2 points (or groups) which are the closest in that sense, presents a risk: gathering 2 points (or groups) which are obviously in different groups.

If a model in dimension d is used, the real criterion (the maximum likelihood criterion) for the divisive method, e.g. between two convex clusters consists in finding the two clusters such that the difference of the hypervolumes sustended by the density between the global convex hulls of the two clusters is the largest. In an agglomerative way, this difference should be the smallest.

Computing hypervolumes causes computational problems. But, if all the sustended areas (on each axis) between the respective coordinates of the two points are small, then the hypervolume in dimension d will be small (This implication is not reversible).

Therefore for each couple of points $x_i = (x_{i1}, \cdots, x_{id})$ and $x_j = (x_{j1}, \cdots, x_{jd})$, the following quantities are computed

$$\text{diss}(x_i, x_j) = \max_{1 \le k \le d} | \int_{x_{ik}}^{x_{jk}} \hat{f}_k(x) dx | \qquad (2)$$

where $\hat{f}(\cdot)$ is an estimation of the density function for the variable k:

$$\hat{f}_k(x) = \frac{1}{nh_k} \sum_{i=1}^{n} \frac{1}{\sqrt{2\pi}} e^{-\frac{(x-x_i)^2}{2h_k^2}}.$$

The value of h_k, the smooting parameter is chosen following Silverman ([Silverman, 1986]) as $h_k = 1,06 \cdot \min(\sigma_k, R_k/1,34) \cdot n^{-0,2}$, where σ_k (respectively R_k) is the standard deviation (respectively the interquartil range) of the n values $x_{1k}, \cdots, x_{nk}$.

It can be shown easily that (2) is a dissimilarity measure. For two clusters C_i and C_j, there exist many ways to define $\text{diss}(C_i, C_j)$. For example:

- the single linkage method where $\text{diss}(C_i, C_j) = \min_{x \in C_i, y \in C_j} \text{diss}(x, y)$,
- the complete linkage method where $\text{diss}(C_i, C_j) = \max_{x \in C_i, y \in C_j} \text{diss}(x, y)$.

Based on these definitions, the merging criterion consists in grouping the two objects X and Y (either points or clusters) for which $\text{diss}(X, Y)$ is the smallest.

The method proceeds from the situation where all the points are in separate clusters until they all form a unique cluster. Consecutive merging can be represented by a dendrogram (figure 2).

The resulting algorithm based on these concepts was implemented and seems to be very powerful. The first results obtained are promising. For example, the structure of the dendrogram (figure 2) constructed by the method on the Ichino's Oils dataset is very good if compared with the tree obtained with the first method or those presented in ([Chavent, 1997], page 139).

A new hierarchical divisive monothetic method was first developped. The only hypothese needed was that the observed points are generated by a non-homogeneous Poisson process. The algorithm performed in two steps : splitting and pruning. The splitting rule was deduced from a maximum likelihood criterion; the pruning method was based on the Gap test. An application of this algorithm was presented on a well-known interval dataset. The splitting criterion also gave the idea to develop a new dissimilarity measure for

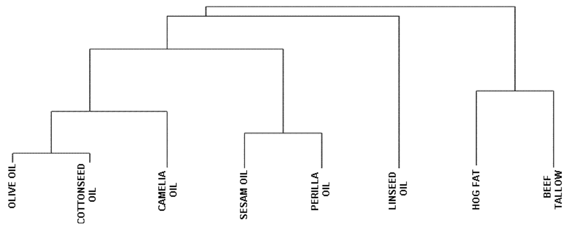


Fig. 2. Dendrogram obtained on the Ichino'oils dataset, complete linkage method.

hierarchical agglomerative clustering. The resulting algorithm was briefly described. Applied on the same dataset, it produced very interesting results. All these ways will be thorough in the future.

References

[Chavent and Lechevallier, 2002]M. Chavent and Y. Lechevallier. Dynamical clustering of interval data: Optimization of an adequacy criterion based on Hausdorff distance. In K. Jujuga, A. Sokolowski, and H.H. Bock, editors, *Classification, Clustering, and Data Analysis*, pages 53–60, 2002.

[Chavent, 1997]M. Chavent. *Analyse des Données Symboliques: Une méthode divisive de classification.* PhD thesis, Université Paris IX-Dauphine, 1997.

[Kubushishi, 1996]T. Kubushishi. *On some Applications of the Point Process Theory in Cluster Analysis and Pattern Recognition.* PhD thesis, Facultés Universitaires Notre-Dame de la Paix, Namur, 1996.

[Rasson and Kubushishi, 1994]J.-P. Rasson and T. Kubushishi. The gap test: an optimal method for determining the number of natural classes in cluster analysis. In E. Diday, Y. Lechevallier, M. Shader, P. Bertrand, and B. Burtschy, editors, *New approaches in Classification and Data Analysis*, pages 186–193, 1994.

[Silverman, 1981]B. W. Silverman. Using kernel density estimates to investigate multimodality. *Journal of The Royal Statistic Society, B*, pages 97–99, 1981.

[Silverman, 1986]B.W. Silverman. *Density Estimation for Statistics and Data Analysis.* Chapman and Hall, 1986.

Normalized k-means clustering of hyper-rectangles

Marie Chavent

Mathématiques Appliquées de Bordeaux, UMR 5466 CNRS,
Université Bordeaux 1 - 351, Cours de la libération,
33405 Talence Cedex, France
(e-mail: chavent@math.u-bordeaux1.fr)

Abstract. Interval variables can be measured on very different scales. We first remind a general methodology used for measuring the dispersion of a variable from an optimal center and we define two measures of dispersions associated to two optimal "centers" for interval variables. Then we study the relations between the standardization of a data table and the use in clustering of a normalized distance. Finally we define two normalized distances between hyper-rectangles and their use in two normalized k-means clustering algorithms.

Keywords: Interval data, Standardization, Normalized Hausdorff distance, Clustering.

1 Introduction

A classical quantitative data table $(x_i^j)_{n \times p}$ describes n objects $\{1, ..., i, ..., n\}$ by p quantitative variables $\{1, ..., j, ..., p\}$ which may be defined on different scales. This phenomenon is measured by the dispersion (standard deviation, range, percentile ranges...) of each variable.

Dealing with variables measured on very different scales is a problem when comparing two objects globally on all the variables. For instance the Euclidean distance or more generaly the L_p-distance will give implicitly more importance to variables of strong dispersion and the comparison between objects will only reflect their differences on those variables. This phenomenon has then an incidence on the clustering into classes of homogeneous objects (i.e. objects highly similar to each other): only variables with strong dispersion will have an important contribution in the construction of clusters. A natural way to avoid this effect is either to normalize the data table or to use normalized distances.

Recently, several clustering methods have been proposed in the field of symbolic data analysis [Diday, 1988], [Bock and Diday, 2000]. Several works on k-means clustering of interval data sets have been published [Bock, 2001], [Chavent and Lechevallier, 2002], [De Carvalho *et al.*, 2003], [Chavent *et al.*, 2003], [De Souza and De Carvalho, 2004] and [Chavent, 2004].

The problem of the standardization of this new type of data is now naturally arising. In [Chavent, 1997], the symbolic data set was not directly normalized but normalized distances between symbolic objects were used:

$$d(i,i') = (\sum_{j=1}^{p} \frac{1}{(\sigma^j)^\alpha} d(x_i^j, x_{i'}^j)^\alpha)^{1/\alpha} \tag{1}$$

where d was a measure of comparison between two symbolic descriptions (two intervals for instance) and σ^j a measure of dispersion of a variable j defined by:

$$\sigma^j = \frac{1}{2n} \sum_{i=1}^{n} \sum_{i'=1}^{n} d^2(x_i^j, x_{i'}^j) \tag{2}$$

The use of a double sum in [2] was not really appropiate for computing σ^j on voluminous data sets.

This question of the standarization of symbolic data has also been clearly raised for interval data in [De Carvalho *et al.*, 2003] where the authors proposed measures of dispersion based on the dispersion of the centers, the lower bounds or the upper bounds of the intervals.

In [Chavent and Lechevallier, 2002] and [Chavent, 2004], two k-means clustering algorithms of hyper-rectangles with Hausdorff distances where proposed. The idea here is to use the explicit formula of the optimum class prototype given in those two papers in order to define two "mean" intervals optimizing two measures of dispersion (see section 2). Those two measures of dispersion are called the "star" and the "radius" of an interval variable (see sections 2.1 and 2.2). After a few words on the relation between standardizing an interval data table and using a normalized distance between hyper-rectangles (see section 3), the two k-means algorithms given in [Chavent and Lechevallier, 2002] and in [Chavent, 2004] are "normalized" (see section 4).

In the rest of this paper we will consider an interval data table $(x_i^j)_{n\times p}$ where each object i is described for each variable j by an interval

$$x_i^j = [a_i^j, b_i^j] \in I = \{[a,b] \mid a, b \in \Re \ , \ a \leq b\}$$

Each object i is then described by an hyper-rectangle of $\Re^p$:

$$x_i = \prod_{j=1}^{p} [a_i^j, b_i^j]$$

2 Measure of centrality and dispersion

For a classical quantitative variable j the mean squared deviation measures the dispersion from the mean $\bar{x}^j$ which is the optimal solution $\hat{y}$ of the following minimization problem:

$$\min_{y\in\Re} \sum_{i=1}^{n} (x_i^j - y)^2 = \min_{y\in\Re} \underbrace{\sum_{i=1}^{n} d^2(x_i^j, y)}_{f(y)} \tag{3}$$

In the same way the mean absolute deviation measures the dispersion from the median x^j_M which is the optimal solution $\hat{y}$ of the following minimization problem:

$$\min_{y\in\Re}\sum_{i=1}^{n}|x^j_i - y| = \underbrace{\min_{y\in\Re}\sum_{i=1}^{n} d(x^j_i, y)}_{f(y)} \tag{4}$$

In both cases, $f(\hat{y})$ is a measure of dispersion.

For an interval variable j we have $x^j_i = [a^j_i, b^j_i]$ and the "measures" of centrality are not real values like the mean or the median values but an interval of values noted $y = [\alpha, \beta]$. We have seen that the mean and the median are optimal centers of two different dispersion measures f. Our aim is then to define optimal centers $\hat{y} = [\hat{\alpha}, \hat{\beta}]$ for functions f chosen to measure the dispersion. Those functions are based on a distance d between intervals.

The distance chosen here to compare two intervals is the Hausdorff distance. This set-distance d_H is simplified in the particular case of two intervals to:

$$d_H([a^j_i, b^j_i], [a^j_{i'}, b^j_{i'}]) = \max(|a^j_i - a^j_{i'}|, |b^j_i - b^j_{i'}|) \tag{5}$$

In the next sections we will define two different optimal "centers" $\hat{y} = [\hat{\alpha}, \hat{\beta}]$ and two different measures of dispersion $f(\hat{y})$.

2.1 The "star"

We consider the following measure of dispersion from $\hat{y}$:

$$f(\hat{y}) = \sum_{i=1}^{n} d_H(x^j_i, \hat{y}) \tag{6}$$

where d_H is the Hausdorff distance between the intervals x^j_i and $\hat{y}$ and where $\hat{y}$ is defined by:

$$\hat{y} = arg\min_{y\in I}\sum_{i=1}^{n} d_H(x^j_i, y) \tag{7}$$

We use a result of [Chavent and Lechevallier, 2002] to define an explicit formula for the optimal "central" interval $\hat{y} = [\hat{\alpha}, \hat{\beta}]$: by a simple rewriting of the intervals $x^j_i = [a^j_i, b^j_i]$ according to their middle point m^j_i and their half-length l^j_i, the authors proved that the middle point $\hat{\mu}$ and the half-length $\hat{\lambda}$ of the interval $\hat{y}$ minimizing $\sum_{i=1}^{n} d_H(x^j_i, y)$ is:

$$\hat{\mu} = median\{m^j_i \mid i = 1, ..., n\} \tag{8}$$

$$\hat{\lambda} = median\{l^j_i \mid i = 1, ..., n\} \tag{9}$$

The following measure of dispersion σ^j is defined:

$$\sigma^j = \sum_{i=1}^{n} \max(|a_i^j - \hat{\mu} + \hat{\lambda}|, |b_i^j - \hat{\mu} - \hat{\lambda}|) \tag{10}$$

Because the formulation of f given in (6) is close to the measure of homogeneity of a cluster C called the "star":

$$\min_{i \in C} \sum_{j \in C} d_{ij}$$

we will call σ^j defined in (10) the "star" of the interval variable j.

2.2 The "radius"

We consider the following measure of dispersion from $\hat{y}$:

$$f(\hat{y}) = \max_{i=1\ldots n} d_H(x_i^j, \hat{y}) \tag{11}$$

where d_H is once again the Hausdorff distance between the intervals x_i^j and y and where $\hat{y}$ is defined by:

$$\hat{y} = arg \min_{y \in I} \max_{i=1\ldots n} d_H(x_i^j, y) \tag{12}$$

We use here a result of [Chavent, 2004] to define an explicit formula for the optimal "central" interval $\hat{y} = [\hat{\alpha}, \hat{\beta}]$: the author proved that the lower and upper bounds of interval $\hat{y}$ minimizing $\max_{i=1\ldots n} d_H(x_i^j, y)$ are:

$$\hat{\alpha}^j = \frac{\max_{i=1\ldots n} a_i^j + \min_{i=1\ldots n} a_i^j}{2} \tag{13}$$

$$\hat{\beta}^j = \frac{\max_{i=1,\ldots,n} b_i^j + \min_{i=1,\ldots,n} b_i^j}{2} \tag{14}$$

The following measure of dispersion σ^j can then be defined:

$$\sigma^j = \max_{i=1\ldots n} \max(|a_i^j - \hat{\alpha}^j|, |b_i^j - \hat{\beta}^j|) \tag{15}$$

Because the formulation of f given in (11) is close to the measure of homogeneity of a cluster C called the "radius":

$$\min_{i \in C} \max_{j \in C} d_{ij}$$

we will call σ^j defined in (15) the "radius" of the interval variable j.

3 Standardization, distance and clustering

For a classical quantitative data table $(x_i^j)_{n \times p}$, standardizing is a technique for removing location and scale attributes. The standardized variables z^j have mean equal to 0 and standard deviation equal to 1 when the variables x^j are centered by their mean $\bar{x}^j$ and normalized (reduced) by their standard deviation σ^j. The Euclidean distance between two objects i and i' of the standardized matrix $(z_i^j)_{n \times p}$ is then:

$$d(z_i, z_{i'}) = \sqrt{\sum_{j=1}^{p} (\frac{x_i^j - \bar{x}^j}{\sigma^j} - \frac{x_{i'}^j - \bar{x}^j}{\sigma^j})^2} \tag{16}$$

$$= \sqrt{\sum_{j=1}^{p} \frac{1}{(\sigma^j)^2} (x_i^j - x_{i'}^j)^2} \tag{17}$$

$$= d_M(x_i, x_{i'}) \tag{18}$$

where d_M is the weighed Euclidean distance and $M = D_{1/\sigma^2}$. This weighed distance is also sometimes called the normalized Euclidean distance.

We can then notice that:

- the clustering obtained from the initial data table $(x_i^j)_{n \times p}$ is similar to the clustering obtained from the centered data table $(x_i^j - \bar{x}^j)_{n \times p}$ (because the distances are equal). Indeed we are not directly concerned in this article with the problem of centering interval data even if we have defined a “central” interval previously in this article.
- the clustering performed with the initial data table $(x_i^j)_{n \times p}$ and the normalized Euclidean distance d_M is similar to the clustering performed with the standardized (or simply normalized) data table $(z_i^j)_{n \times p}$ and the ”simple” Euclidean distance.

We have of course the same kind of results with the Minkowsky distance.

The questions are now: do we have the same kind of results for interval data ? Is it equivalent to ”normalize” the intervals $x_i^j = [a_i^j, b_i^j]$ and to use a ”normalized” distance ? What does “normalizing” an interval or “normalizing” a distance between hyper-rectangles mean ?

Here we will try to answer those questions in the particular case of two distances between hyper-rectangles of $\Re^p$ used in [Chavent and Lechevallier, 2002] and [Chavent, 2004]. We consider

$$x_i = \prod_{j=1}^{p} \underbrace{[a_i^j, b_i^j]}_{x_i^j}$$

and

$$x_{i'} = \prod_{j=1}^{p} \underbrace{[a_{i'}^j, b_{i'}^j]}_{x_{i'}^j}$$

The first distance d_1 is not a real $\Re^p$-set Hausdorff distance but a sum of Hausdorff distances d_H between intervals:

$$d_1(x_i, x_{i'}) = \sum_{j=1}^{p} d_H(x_i^j, x_{i'}^j) \tag{19}$$

The second distance d_2 is a real $\Re^p$-set Hausdorff distance called the L_∞-Hausdorff distance which can be written in the particular case of hyper-rectangles as a maximum of Hausdorff distances d_H between intervals:

$$d_2(x_i, x_{i'}) = \max_{j=1\dots p} d_H(x_i^j, x_{i'}^j) \tag{20}$$

If we consider now that "normalizing" an interval $x_i^j = [a_i^j, b_i^j]$ consists in dividing its lower and upper bounds by the same measure of dispersion σ^j, the "normalized" interval of x_i^j is $z_i^j = [\frac{a_i^j}{\sigma^j}, \frac{b_i^j}{\sigma^j}]$.

The Hausdorff distance between two "normalized" intervals is then:

$$d_H(z_i^j, z_{i'}^j) = \max(|\frac{a_i^j}{\sigma^j} - \frac{a_{i'}^j}{\sigma^j}|, |\frac{b_i^j}{\sigma^j} - \frac{b_{i'}^j}{\sigma^j}|) = \frac{1}{\sigma^j} d_H(x_i^j, x_{i'}^j) \tag{21}$$

and the distances d_1 and d_2 between the two "normalized" hyper-rectangles z_i and $z_{i'}$ can then be written as:

$$d_1(z_i, z_{i'}) = \sum_{j=1}^{p} \frac{1}{\sigma^j} d_H(x_i^j, x_{i'}^j) \tag{22}$$

and

$$d_2(z_i, z_{i'}) = \max_{j=1\dots p} \frac{1}{\sigma^j} d_H(x_i^j, x_{i'}^j) \tag{23}$$

The normalized distance is then defined for d_1 by:

$$d_1(x_i, x_{i'}) = || \frac{(d_H(x_i^j, x_{i'}^j)_{j=1,\dots,p}}{\sigma^j} ||_{L_1} \tag{24}$$

and for d_2 by:

$$d_2(x_i, x_{i'}) = || \frac{(d_H(x_i^j, x_{i'}^j)_{j=1,\dots,p}}{\sigma^j} ||_{L_\infty} \tag{25}$$

Finally, we have once again the result that the clustering performed with the initial interval data table $(x_i^j)_{n\times p}$ and the normalized distances d_1 or d_2 (given in (24) and 25)) is similar to the clustering performed with the "normalized" interval data table $(z_i^j)_{n\times p}$ and the "simple" distances d_1 or d_2 (given in (19) and (20)).

4 Normalized k-means of hyper-rectangles

Dynamical clustering [Diday and Simon, 1976] called here for simplification k-means clustering, proceeds by iteratively determining K class prototypes and then reassigning all objects to the closest class prototype. If the prototype $\hat{y}$ of a cluster C is properly defined by optimization of an adequacy criterion f (measuring the “dissimilarity” between the prototype and the cluster), the algorithm converges and the partitioning criterion decreases at each iteration.

For classical quantitative data, when the prototype $\hat{y}$ of a cluster C is the mean-vector, the adequacy criterion minimized is:

$$f(y) = \sum_{i \in C} d^2(x_i, y) = \sum_{i \in C} \sum_{j=1}^{p} (x_i^j - y^j)^2 \tag{26}$$

When a standardization is necessary, the columns x^j are usually normalized by $\sigma^j = \sqrt{\sum_{i=1}^{n}(x_i^j - \bar{x^j})^2}$ or the normalized Euclidean distance d_M with $M = D_{1/\sigma^2}$ is used. The adequacy criterion measured on $\Omega = \{1, ..., n\}$ is then equal to p, the number of variables.

In the same way when the prototype $\hat{y}$ of a cluster C is the median-vector x_m, the adequacy criterion minimized is:

$$f(y) = \sum_{i \in C} d(x_i, y) = \sum_{i \in C} \sum_{j=1}^{p} |x_i^j - y^j| \tag{27}$$

When a standardization is necessary, the columns x^j are normalized by $\sigma^j = \sum_{i=1}^{n} |x_i^j - x_m^j|$ or the normalized Euclidean distance d_M with $M = D_{1/\sigma}$ is used. The adequacy criterion measured on $\Omega = \{1, ..., n\}$ is then once again equal to p, the number of variables.

In the particular case of interval data the optimal prototype of a cluster is an hyper-rectangle. We can repeat the previous reasoning for ”normalizing” any k-means clustering algorithm of hyper-rectangles when the prototypes are properly defined by optimization of an adequacy criterion. Here we use:

- the normalized distance (24) with σ^j the “star” defined in (10) for ”normalizing” the k-means method of [Chavent and Lechevallier, 2002]
- the normalized distance (25) with σ^j the “radius” defined in (15) for ”normalizing” the k-means method of [Chavent, 2004]

5 Conclusion

In this paper we have proposed a general approach for the “normalization” of dynamical clustering algorithms. We have seen that if the prototype of a cluster is properly defined by optimization of an homogeneity criterion, this result can also be used to define a measure of dispersion and then to normalize either the data or the distances. We have applied this methodology in the particular case of two k-means clustering algorithms of hyper-rectangles. The first one uses a "star" homogeneity criterion and a distance between hyper-rectangles which is a sum of Hausdorff distances between intervals. The second one uses a "radius" homogeneity criterion and the L_∞ Hausdorff distance between hyper-rectangles. The two corresponding dispersion measures of interval variables called here the "star" and the "radius" are then simply used to "normalize" those two algorithms.

References

[Bock and Diday, 2000]H.-H. Bock and E. Diday, editors. *Analysis of symbolic data. Exploratory methods for extracting statistical information from complex data.* Studies in classification, data analysis and knowledge organisation. Springer Verlag, Heidelberg, 2000.

[Bock, 2001]H.-H. Bock. Clustering algorithms and Kohonen maps for symbolic data. In *ICNCB Proceedings*, pages 203–215, Osaka, 2001.

[Chavent and Lechevallier, 2002]M. Chavent and Y. Lechevallier. Dynamical clustering of interval data. Optimization of an adequacy criterion based on Hausdorff distance. In K. Jajuga, A. Sokolowski, and H.-H. Bock, editors, *Classification, Clustering, and Data Analysis*, pages 53–60, Berlin, 2002. Springer Verlag.

[Chavent *et al.*, 2003]M. Chavent, F.A.T. De Carvalho, Y. Lechevallier, and R. Verde. Trois nouvelles méthodes de classification automatique de données symboliques de type intervalle. *Revue de Statistiques Appliquées*, LI(4), 2003.

[Chavent, 1997]M. Chavent. *Analyse des données symboliques. Une méthode divisive de classification.* PhD thesis, Université Paris-IX Dauphine, 1997.

[Chavent, 2004]M. Chavent. An Hausdorff distance between hyper-rectangles for clustering interval data. In D. Banks and al., editors, *Classification, Clustering and Data Mining Applications*, pages 333–340. Springer, 2004.

[De Carvalho *et al.*, 2003]F.A.T. De Carvalho, P. Brito, and H.-H. Bock. Une méthode type nuées dynamiques pour les données symboliques quantitatives. In Y. Dodge and G. Melfi, editors, *Méthodes et perspectives en Classification*, pages 79–81. Presses Académiques Neuchâtel, 2003.

[De Souza and De Carvalho, 2004]R.M.C.R. De Souza and F.A.T. De Carvalho. Clustering of interval data based on city-block distances. *Pattern Recognition Letters*, 25:353–365, 2004.

[Diday and Simon, 1976]E. Diday and J. C. Simon. Clustering analysis. In K. S. Fu, editor, *Digital Pattern Classification*, pages 47–94. Springer Verlag, 1976.

[Diday, 1988]E. Diday. The symbolic approach in clustering and related methods of data analysis: The basic choices. In H.-H. Bock, editor, *Classification and related methods of data analysis*, pages 673–684, Amsterdam, 1988. North Holland.

Generalized Symbolic Marking Of Complex Objects Through Intelligent Complex Miner Software CRM Applications

Mireille Gettler Summa[1], Frederick Vautrain[2], and Matthieu Barrault[2]

[1] Centre de Recherche de Mathematique de la Decision
University of Paris Dauphine
1 Pl. du Ml. De Lattre de Tassigny
75016 - Paris France
(e-mail : summa@ceremade.dauphine.fr)

[2] ISTHMA Ltd
6 rue du Soleillet
75020 - Paris France
(e-mail: barrault@isthma.com, vautrain@isthma.com)

Abstract. In this paper we propose an automatic method of describing classes of complex objects (lists, diagrams, intervals, histograms, time series). The approach simultaneously generalizes a class and discriminates it from the others. This method belongs to a family of algorithms called MGS (Marking and Generalization by Symbolic objects) which were already applied on classical inputs, either to Factorial Analysis interpretation in [Gettler Summa, 1992] [Giordano *et al.*, 2000] or to the interpretation of partitions [Gettler Summa *et al.*, 1994]. It was also used for summarizing huge databases in [Massrali *et al.*, 1998]. For Customer Relationship Management, MGS provides sets of client profiles which target shops, brands or couponing analysis. An application through Intelligent Complex Miner software is presented on jointed data bases of sells, couponing information, client socio-demographic elements, and geo-marketing data.

Keywords: discriminant description, generalization, symbolic marking, generalized V-test, CRM.

1 Introduction

Data analysis on classes of statistical units is a crucial issue because huge data bases are stored and results interpretation takes more and more time. Furthermore, to take into account multiple arrays, distribution values, time series or continuous functions appear to be the very appropriate inputs for summarizing the data without loss of information, towards knowledge extraction. Very few approaches face such complex data analysis: Symbolic Data Analysis brings for example a theoretical framework [Diday, 1988] for such a challenge. Discrimination and generalisation are to be redefined in that new context: Marking and Generalisation by Symbolic objects generalises to symbolic inputs some Machine Learning approaches [Stepp, 1984] [Ho tu *et al.*, 1988] [Ganascia, 2000], or supervised classification algorithms that are generally not able to treat complex matrices [Gordon, 1999]. MGS provides discriminant generalizing objects that describe subsets of the power set of an initial classical data set. Some other symbolic approaches have recently been published [Vrac and Diday, 2001] for similar purposes. As in recursive partition algorithms [Périnel *et al.*, 2003], the results could be written as complex production rules but the inference validation phase is not included in this paper.

2 The input matrix

We consider a set $\Omega = \{\omega_1, ..., \omega_n\}$ of n symbolic objects for p variables Y_j :

$$\begin{aligned} Y_j : \Omega &\longrightarrow \Upsilon_j \\ \omega &\longrightarrow Y_j(\omega) \end{aligned}$$

$Y_j(i)$ may be :

- a single real number or a single category (classical data);
- a finite set of real numbers or categories (multi-valued variable);
- a discrete finite frequency distribution (diagram variable); frequencies are in this paper the frequencies of a statistical distribution, as it happens for example when symbolic objects come from a query on a classical categorical data base;
- an interval;
- a continuous frequency distribution on a finite number of intervals (histogram variable); an hypothesis of uniform distribution along the intervals allows linear interpolation to calculate frequencies on sub intervals.

Let E and $\bar{E}$ be two classes of a binary partition on Ω (if the given partition is not binary, E is the class to be marked and $\bar{E}$ its complementary part, union of the other classes of the partition).

$Y(\omega) = \{y_j(\omega),\ j \in \{1, ..., p\}\}$ is the description of ω, denoted also d_ω. A 'partial description' has fewer variables than in initial individuals. A symbolic object s, is a triplet (a, R, d) where a is a mapping $E \rightarrow \{0, 1\}$ which measures the fit between d_ω and d, R is a relation which associates (in this paper) to a couple of descriptions a Boolean value (for example R is the inclusion operator). The extent of a symbolic object s in E is defined in this paper in the Boolean case by : $Ext_E(s) = \{\omega \in E / a(\omega) = 1\}$.

Let S_E be the set of symbolic objects belonging to E.

3 GV-TEST criterion

Marking is a process which builds discriminant descriptions of S_E. Several trees are simultaneously explored top down from initial nodes. Depending on some parameter values, final descriptions:

1. may be totally discriminant or only partially;
2. may have overlapping extents or not;
3. may include in their extent all E elements or not.

The first and the second points are simultaneously taken into account in the descending process through a threshold Thr_{Cr} for some criterion Cr which measures the link between any subset of Ω and E.

Let M_g be an intent of a subset in the case of a classical initial data set (for example: 'colour = yellow and length = short', partial description for all 'yellow and short' units).

Almost all of them use for classical data the following quantities:

$$n_g = Card[ext_{\Omega}(M_g)],\ n_{E,g} = Card[ext_E(M_g)],\ n_g - n_{E,g} = Card[ext_{\bar{E}}(M_g)]$$

	E	$\bar{E}$	Ω
M_g	$n_{E,g}$	$n_g - n_{E,g}$	n_g
Ω	n_E	$n - n_E$	n

Table 1. GV-TEST criterion

We propose the GV-test criterion which is a generalization of the V-test [Alevizos and Morineau, 1992] for symbolic objects.

The V-test is based on the hyper geometric distribution hypothesis; for its 5% upper point, its value, which can be calculated by a Laplace Gauss approximation, is greater or equal to 1.96. Explicit formula is the following :

$$T_H = \frac{n_{E,g} - n\dfrac{n_g}{n_E}}{\left[\dfrac{n_E n_g (n - n_E)}{(n-1)n(1 - \dfrac{n_g}{n})}\right]^{\frac{1}{2}}}$$

In the case of categorical variables, the V-test criterion can be used on not too small data sets [Gettler Summa, 2000]. For non classical objects, some other calculations (extents cardinalities) are to be considered depending on the situation:

- for an interval variable Y, the frequency of an interval I_1 among all possible intervals for Y values (in E, in $\bar{E}$, in Ω, in M_g) is equal to the number of intervals that include I;
- for a multi-valued variable, the frequency of a list L among all possible lists for Y values (in E, in $\bar{E}$, in Ω, in M_g) is equal to the number of lists which include L;

- for a diagram variable Y, the frequency which occurs is that of a class of bars of the following type(see 'initial nodes for the sets of diagrams variables'): $\{m_k, w \geq w_{min,m_k}\}$; this cardinality is equal to the number of bars among all Y diagrams (in E, in $\bar{E}$, in Ω, in M_g) $\{m_{ki}, w_{ki}\}$ where $m_{ki} = m_k$ and $w \geq w_{min,m_k}$;
- for a histogram variable Y, the frequency is the one of a class of rectangles of the following type(see 'initial nodes for the sets of histograms variables') : $\{I_k, w \geq w_{min,I_k}\}$; this cardinality is equal to the number of rectangles among all Y histograms (in E, in $\bar{E}$, in Ω, in M_g) $\{I_{ki}, w_{ki}\}$ where $I_k \subseteq I_{ki}$ and $(I_k/I_{ki})w_{ki} \geq w_{min,I_k}$.

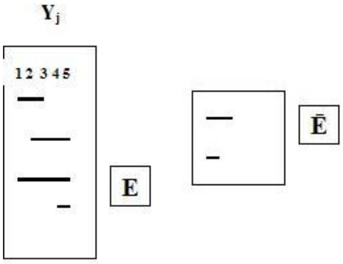


Fig. 1. Example of GV-test computation for an interval variable.

The GV-test of a variable value is used for ranking the initial nodes in order to begin the descending process in the MGS algorithm.

For example, one can easily calculate from figure 1 the elements which occur for the GV-Test computation, for variable Y_j taking its value in the interval [4 5] : Let $s_{[4,5]}$ be the symbolic object associated to the partial description $d = (Y_j = [4, 5])$. $n_g = Card[ext_\Omega(s_{[4,5]})] = 3, n_{E,g} = Card[ext_E(s_{[4,5]})] = 3, n_g - n_{E,g} = Card[ext_{\bar{E}}(s_{[4,5]})] = 0, n_E = 4, n = 6$.

4 Initial Nodes

The choice of the initial nodes depends on the variables nature. Details are given for diagram and histogram variables.

4.1 Initial nodes for the sets of categorical or continuous classical variables

Continuous variables are discretized into optimized classes by a supervised method in order to take into account the binary partition: for example the supervised Fisher algorithm or any decision tree one-to-one variable [Zighed *et al.*, 1997]. Initial nodes are then built by the same method as for categorical variables, each class being considered as a category. This approach on classical data is fully described in [Gettler Summa, 2000]. Let IN_C be the subset of the initial nodes.

4.2 Initial nodes for the set of diagram variables

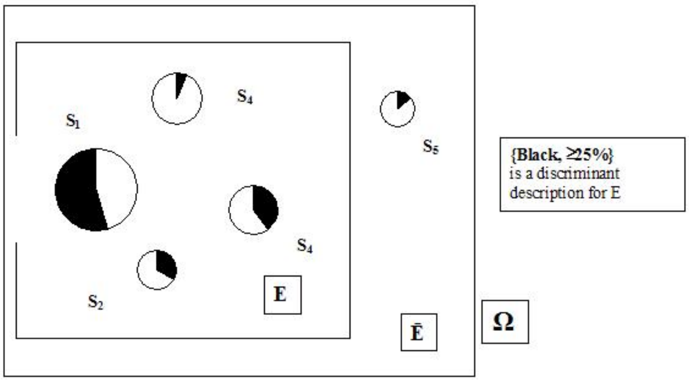


Fig. 2. Initial nodes for the set of diagram variables.

Let $S_{w_c,E}$ be the set of all weighted categories belonging to the variables describing S_E and p_{w_c} be the number of the distinct categories in all diagram variables.
Let's call IN_D the subset of $S_{w_c,E}$ which belong to the initial nodes set.
Let w_1 and w_2 be two weights of a same category, m_k; suppose $w_1 > w_2$. Taking into account $\{m_k, w \geq w_2\}$ implies taking into account $\{m_k, w = w_1\}$ because weights have a statistical frequency semantic.
Let then w_{min,m_k} be the minimum of m_k weights that correspond to GV-test values greater than the a priori threshold Thr_{GV}. It may happen for some category that no weight provides a good GV-test. Let p'_{w_c} be the number of those categories m_k for which w_{min,m_k} does exist ($p'_{w_c} \leq p_{w_c}$).
IN_D is then built of all $\{m_k, w \geq w_{min,mk}\} k \in 1, ..., p'_{w_c}$.

4.3 Initial nodes for the set of histogram variables

A simple situation is the one of interval variables as it is shown in the GV-test paragraph. It will thus not be detailed furthermore so let us present directly the case of histogram variables.
Let $S_{H,E}$ be the set of all weighted intervals (I_k, w_k), belonging to S_E.
Let IN_H (IN_I for interval variables) be the subset of $S_{H,E}$ which will belong to the initial nodes set.
Let p_H be the cardinality of $S_{H,E}$.

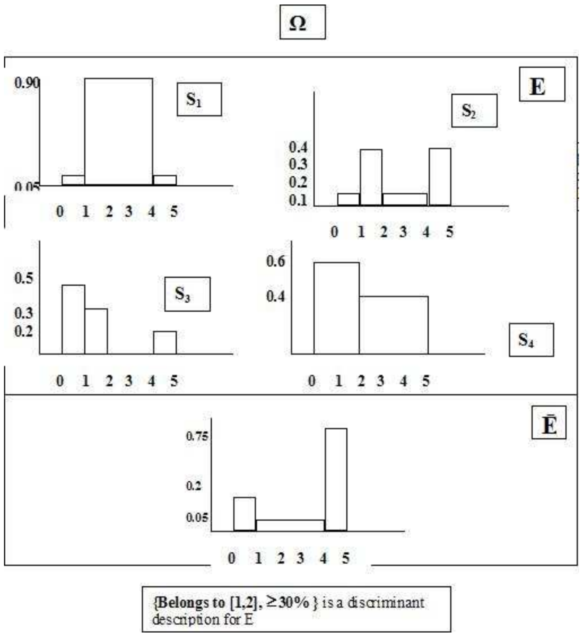


Fig. 3. Initial nodes for the set of histogram variables.

Let us consider the ordered list of the ordered bounds of the intervals $\{(I_k, w_k), k \in 1, ..., p_H\}$.
That list defines $p'_H = p_H$ non overlapping intervals $\{I'_k, k \in 1, ..., p'_H\}$.
Let S_{I_k} be the set of weighted intervals including I'_k. For each of SI_k intervals, a new weight is calculated by linear interpolation (because of the uniform distribution hypothesis).
Let w_1 and w_2 be two weights of a same interval, I'_k, depending of two corresponding $S_{H,E}$ intervals the intersection of which is I'_k ; let suppose $w_1 > w_2$.
Considering $\{I'_k, w \geq w_2\}$ implies considering $\{I'_k, w = w_1\}$ because weights have a statistical frequency semantic.
Let $wmin, I'_k$ be the minimum of SI_k weights that correspond to GV-test values greater than the a priori threshold ThrGV. Note that it may happen for some interval that no weight provides a good GV-test. Let p''_H be the number of those intervals for which w_{min,I'_k} does exist $(p''_H \leq p'_H)$.
IN_H is then built of all $\{(I'_k, w \geq w_{min,I'_k}), k \in 1, ..., p''_H\}$.

5 Generalized MGS Algorithm

In the proposed method a set of initial nodes is built by first exploring a particular lattice, and then by the conjunction of some vertices, according to various chosen criteria, further steps build the final Markings set.

'The output of the marking process consists of a set of partial descriptions ('Markings'), the number of which is inferior to the number of initial individuals.' Let now call a partial description d and the symbolic object, triplet (a, R, d) that is associated to it by the same notation .

Let's denote M_g a generic marking for E. A threshold Thr_{GV} for the GV-test is to be chosen as an input parameter in order to select the initial nodes for each type of variable. Let us denote L the union of initial nodes sets of different types (see 'initial node'):

$$L = IN_C \cup IN_I \cup IN_D \cup IN_H$$
$$L = \{l_g, 1 \leq g \leq v\}$$

Each of L elements has a GV-test value (see GV-test) with respect to E, such that GV-test values are not metric but ordered values, i.e. the greater the absolute value is, the stronger is the link which is measured.
L elements are thus ordered according to their V-test values.
Various heuristics have already been proposed to construct Markings. The main differences are, whether it is top down or bottom up [Gettler Summa *et al.*, 1995], greedy [Ho tu *et al.*, 1988] or not, depth first or breadth first, allowing overlapping branches or not etc.
Let's denote S_M a set of Makings.
Let's denote $Cov(l_g) = \dfrac{Card[ext_E(l_g)]}{Card(E)}$.
Let's denote $Err(l_g) = \dfrac{Card[ext_{\bar{E}}(l_g)]}{Card(\Omega)}$.
Two a priori thresholds are to be chosen:

- The final degree in which E is covered by the union of the markings, R_{Cov}; a final marking set S_M should be such that:

$$R_{Cov} \leq \frac{Card(\cup ext_E[M_g, M_g \in S_M])}{Card(E)} \tag{1}$$

- The error ratio made by the markings by covering elements out of E, R_{Err}; each marking should be such that:

$$\forall M_g \in S_M, R_{Err} \geq \frac{Card[ext_{\bar{E}}(M_g)]}{Card(\Omega)} \tag{2}$$

Step 1 : All initial nodes build a first set of markings. Criteria (1) and (2) are calculated for each marking. If any node does not respect Criterion (2), it is deleted from the markings. A first set of markings is thus constructed:

$$S_M^1 = \{M_g^1, M_g^1 = l_g, 1 \le g \le v_1 \le v\}$$
$$Card(S_M^1) = v_1$$
$$\forall M_g^1 \in S_M^1, \ \frac{Err(M_g^1)}{Card(\Omega)} \le R_{Err}$$

The two following quantities are also calculated:

$$Cov(S_M^1) = \frac{Card[\cup ext_E(M_g, M_g \in S_M^1)]}{Card(E)}$$
$$Err(S_M^1) = \frac{Card[\cup ext_{\bar{E}}(M_g, M_g \in S_M^1)]}{Card(\Omega)}$$

Step 2 : Each element of S1M will be a root for descending branches built as follows :

- the constituents of S1M are ordered by their corresponding GV-test values;
- the greatest GV-test value corresponds to the root which is processed at first and so on;
- branches are constructed from each node by choosing the L elements according to the above defined order;
- for each branch, one has to check if it has not yet been constructed to avoid redundancy;
- for each branch, the error ratio is calculated; if it is greater than RErr, the branch is abandoned;
- for each branch, the GV-test is calculated; if it is smaller than ThrGV, the branch is abandoned;
- each remaining branch as a whole is a new marking.

A second set of markings is thus substituted to the first one:

$$S_M^2 = \{M_g^2, 1 \le g \le v_2\}$$
$$Card(S_M^2) = v_2$$
$$\forall M_g^2 \in S_M^2, \ \frac{Err(M_g^2)}{Card(\Omega)} \le R_{Err}$$

The two following quantities are also calculated:

$$Cov(S_M^2) = \frac{Card[\cup ext_E(M_g, M_g \in S_M^2)]}{Card(E)}$$
$$Err(S_M^2) = \frac{Card[\cup ext_{\bar{E}}(M_g, M_g \in S_M^2)]}{Card(\Omega)}$$

Further steps : Step 2 procedure is iterated; as the number of nodes is limited by the GV-test criterion and redundancy of branches is avoided, the algorithm is not fully combinatory and comes to an end according to some stopping rules which are described in the following paragraph:

- a step f is the last one if $Cov(S_M^f) \ge R_{Cov}$ i.e. E is sufficiently marked;

- if one does not want long branches (for example for providing a quick decision aid rule in an application), a parameter is proposed in input to fix a maximum hmax for the number of nodes in a branch. The h^{th} step will thus be at the most, the last one;
- if a marking M_f is such that $Err(M_f) \geq R_{Err}$, it can be cancelled, as an option of the algorithm, from the results.

6 Applications

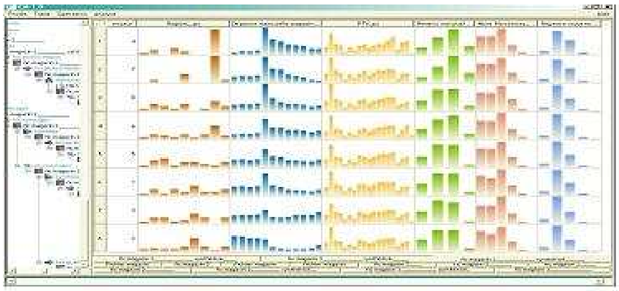


Fig. 4. Complex data editor

MGS is one of the statistical methods implemented in the Interactive Complex Miner software. Based on a collaboration between Ceremade laboratory from Dauphine University and Isthma Company, this software can be used to manage and analyse complex and "multivalued" data as curves, distributions, intervals, sets as well as classical data. Temporal and geographical data are the most frequent complex data types which are analysed in the applications.

Current application consists in three related data bases: 200 shops described by their monthly turnover on several years, 800000 households described by sociogeographic variables and one shop variable, and 4M coupon data base with household, shop and time variables. A complex shop database is generated by merging the three databases. The shops are thus symbolic descriptions (see fig. 4).

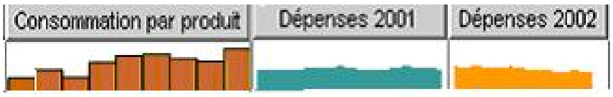


Fig. 5. Client profile 1

A symbolic hierarchical clustering is carried on these shops. Each class is then characterised by 'shops profiles' through MGS. Each profile is graphically displayed (see fig. 5) through ICM editor. Each profile is a vector providing a 'partial symbolic description' (some variables don't appear in the description) called a 'marking core' in the case of Marking approach.

7 Conclusion

Generalized MGS provides sets of descriptions for a chosen subset, depending on initial discrimination quality request. It can be just a generalization of the whole subset if this quality is null; it may also produce lots of specified descriptions with little extents if the quality is high. MGS could also be used for inference to provide rules if a validation step was added to the supervised learning phase. But its best purpose remains a descriptive process of the data, with generalizing and discriminating potential.

References

[Alevizos and Morineau, 1992]P. Alevizos and A. Morineau. Tests et valeurs tests. *RSA*, pages 27–43, 1992.

[Diday, 1988]E. Diday. The symbolic approach in clustering and relative methods of data analysis ; the basic choices. *IFCS*, pages 673–684, 1988.

[Ganascia, 2000]J.G. Ganascia. Charade et fils : évolution, applications et extensions. *Induction symbolique numérique à partir de données*, pages 303–318, 2000.

[Gettler Summa *et al.*, 1994]M. Gettler Summa, E. Périnel, and J. Ferraris. Automatic aid to symbolic cluster interpretation. In Springer Ed., editor, *New approaches in Classificationi and Data Analysis*, pages 405–413, 1994.

[Gettler Summa *et al.*, 1995]M. Gettler Summa, A. Morineau, and H. Pham ti tong. Marquage des axes et des classes. In *ASU proceedings*, pages 468–472, 1995.

[Gettler Summa, 1992]M. Gettler Summa. Factorial axis interpretation by symbolic objects. In CEREMADE Ed., editor, *3ème journées numérique- symbolique*, 1992.

[Gettler Summa, 2000]M. Gettler Summa. Marquages de sous ensembles de données. *Induction symbolique numérique à partir de données*, pages 339–362, 2000.

[Giordano *et al.*, 2000]G. Giordano, M. Gettler-Summa, and R. Verde. Symbolic interpretation in a clustering strategy on multiatribute preference data. In *Statistica Applicata*, pages 473–495, 2000.

[Gordon, 1999]A.D. Gordon. *Classification (2nd edition)*. Chapman and Hall/CRC, Boca Raton, 1999.

[Ho tu *et al.*, 1988]B. Ho tu, E. Diday, and M. Gettler Summa. Generating rules for expert systems from observations. In *Pattern Recognition Letters, Volume 7*, 1988.

[Massrali *et al.*, 1998]M. Massrali, E. Diday, and M. Gettler Summa. Knowledge extraction from data tables through symbolic data analysis. In Eurostat, editor, *KESDA proceeding*, 1998.

[Périnel *et al.*, 2003]E. Périnel, A. Ciampi, J. Lebbe, R. Vignes, and E. Diday. Tree growing with imprecise data. In *Pattern Recognition Letters*, pages 787–803, 2003.

[Stepp, 1984]R. Stepp. A description and user's guide for cluster/2 a program for conjunctive conceptual clustering. *Report n^o UIUCDCS-R61084*, 1984.

[Vrac and Diday, 2001]M. Vrac and E. Diday. Description symbolique de classes. *Cahier du CEREMADE n^o 0115*, 2001.

[Zighed *et al.*, 1997]D. Zighed, R. Rakotomala, and F. Feschet. Optimal multiple intervals discretization of continuous attributes for supervised learning. In *Proceedings of the Thirteenth International Conference on Machine Learning*, pages 295–298, 1997.

Multidimensional Interval-Data: Metrics and Factorial Analysis

Francesco Palumbo[1]* and Antonio Irpino[2]

[1] Department of Economics and Finance - Università di Macerata
Via Crescimbeni, 20
I-62100 Macerata, Italy
(e-mail: francesco.palumbo@unimc.it)

[2] Department of Business Strategies and Quantitative Methods - Seconda Università di Napoli
Corso Gran Priorato di Malta
I-80126 Capua, Italy
(e-mail: irpino@unina.it)

Abstract. Statistical units described by interval-valued variables represent a special case of Symbolic Objects, where all descriptors are quantitative variables. In this context, the paper presents two different metrics in $\mathbb{R}^p$ for interval-valued data that are based on the definition of the Hausdorff distance in $\mathbb{R}$. Hausdorff distance in $\mathbb{R}^p$ (for any $p \geq 1$) is a L_∞ norm between pairs of closed sets. However, when $p > 1$ the problem complexity leads towards the definition of L_2 norms approximating as well as possible the Hausdorff distance. Given a set of n units described by p interval-valued variables, we compute and represent the distances over factorial planes that are defined by factorial analyses that are consistent with the two distance measure definitions.

Keywords: Factorial Analysis, Hausdorff distance, Interval Data.

1 Introduction

Let $\mathbf{\Omega} = \{\omega_1, \omega_2, \ldots, \omega_n\}$ be a set of individuals with description in the space $\mathbb{IR}^p$, where $\mathbb{IR}^p$ indicates the p-dimensional space of the closed subsets in $\mathbb{R}$. The individuals can be modeled as Symbolic Objects (SO) described by interval descriptors. Interval data represent a special case of set-valued data, where the sets are compact and identified by ordered couples of values: $[a] = [\underline{a}, \overline{a}] \subset \mathbb{R}$, which correspond to the interval bound values [Hickey *et al.*, 2001]. The generic $n \times p$ interval data matrix $[\mathbf{A}]$ has general term $[a]_{i,j}$, where $i = 1, \ldots, n$ and $j = 1, \ldots, p$ indicate the generic statistical unit and the generic descriptor, respectively. The general term $[a]_{i,j}$ can also be represented as the *midpoint* $a^c_{i,j}$ and *range* (or *radius*) $a^r_{i,j}$ notation: $[a]_{i,j} =$

* *Acknownledgements:* The paper was financially supported by PRIN2003 grant: "*Multivariate Statistical and Visualization Methods to Analyze, to Summarize, and to Evaluate Performance Indicators*" research unit of University of Macerata, responsible F. Palumbo.

$[\underline{a}_{i,j}, \overline{a}_{i,j}] = [a^c_{i,j} - a^r_{i,j}, a^c_{i,j} + a^r_{i,j}]$. Midpoints and ranges are respectively defined by:

$$a^c = \frac{1}{2}(\overline{a} + \underline{a}), \qquad a^r = \frac{1}{2}(\overline{a} - \underline{a}).$$

In the midpoints/ranges notation, the matrix $[\mathbf{A}]$ is split in the matrices $\mathbf{A}^c$ and $\mathbf{A}^r$ that are called center and range matrix, respectively.

The interval (data) arithmetic has a wide specialized literature, see [Alefeld and Mayer, 2000] for an exhaustive survey. However, the direct treatment of interval-variables in statistics is limited to very few cases, this occurs because the computation of the variance-bounds is an *NP*-hard problem and does not have approximate solutions [Xiang *et al.*, 2004]. It is worth noticing that this aspect involves also the Principal Component Analysis (PCA) and factorial analysis, more generally.

Facing the problem from a geometric point of view and starting from different definitions of distance between intervals, many authors have proposed different approaches to the factorial analysis for interval data (see [Cazes *et al.*, 1997], [Lauro and Palumbo, 2000], [Lauro and Palumbo, 2005], [Giordani and Kiers, 2004]). Generally, a distance between intervals takes into account only some representative points. Cazes et *al.* and Giordani and Kiers based their analysis on the distance between the interval bounds (*vertices*); Lauro and Palumbo proposed a distance measure based on the interval centers and *radii* (or ranges). However, there exist many distance definitions for interval data and more generally for set-valued data; given any general function of distance or proximity, it is possible to arrange a $n \times n$ matrix on which to perform a MultiDimensional Scaling (MDS) analysis and to represent the SO as points in the reduced space.

Dealing with punctual data, a statistical unit is represented by a dimensionless point in $\mathbb{R}^p$ $\forall p$; whereas, the geometric nature of a closed subset ω_i in $\mathbb{R}^p$ varies according to p; it is a segment if $p = 1$, a parallelogram if $p = 2$ a parallelepiped when $p = 3$ and, more generally, a parallelotope when $p > 3$, where $\omega_i = ([a]_{i,1}, [a]_{i,2}, \ldots, [a]_{i,p})$ indicates the generic subsets in $\mathbb{R}^p$, $\forall p \geq 1$.

Differently from the MDS, our aim is to represent the distances but also the *size* and *shape* of the SO [Lauro and Palumbo, 2005].

In section 2 we shall introduce the Hausdorff metric and in section 3 we shall present two distances for interval valued data in $\mathbb{IR}^p$, both of them are derived from the Hausdorff notion of distance. Section 4 presents an application of the two distances on the *Italian peppers* data set. Distances and SO sizes and shapes are represented over factorial planes by means of two factorial analyses; section 5 closes the paper.

2 Distance measures in $\mathbb{IR}^p$

In this section we present the Hausdorff metric for interval data and we introduce two different generalizations in the $\mathbb{IR}^p$ space. We shall show that these distances are good approximations of the Hausdorff distance in $\mathbb{R}^p$ and can be easily decomposed in suitable factorial models.

The Hausdorff metric was proposed by Felix Hausdorff in the early of 20^{th} century as a measure of distance between compact subsets in $\mathbb{R}^p$.

Given a metric $d(\cdot)$, the distance *from* a generic point $x \in \mathbb{R}^p$ *to* a closed subset $A \subset \mathbb{R}^p$ is defined as:

$$d(x, A) = \min_{\tilde{a} \in A} d(x, \tilde{a}).$$

Let $\mathcal{H}(X)$ be the space of all non-empty compact subsets of X, the Hausdorff metric on $\mathcal{H}(X)$ is defined on the basis of the following quantities:

$$h(A, B) = \max_{\tilde{a} \in A} d(\tilde{a}, B), \qquad h(B, A) = \max_{\tilde{b} \in B} d(\tilde{b}, A),$$

where $\{A, B\} \in \mathcal{H}(X)$ and $\{\tilde{a} \in A,\ \tilde{b} \in B\}$.

The Hausdorff distance $H(A, B)$ is defined by:

$$\begin{aligned} H(A, B) &= \max\{\max\{d(\tilde{a}, B)\}, \max\{d(\tilde{b}, A)\} = \\ &= \max\left(h(A, B), h(B, A)\right). \end{aligned} \tag{1}$$

In the special case of $\mathbb{R}$, the Hausdorff distance between two generic intervals is given by: $H(A, B) = \max\{\mid \overline{a} - \overline{b} \mid, \mid \underline{a} - \underline{b} \mid\} = \mid a^c - b^c \mid + \mid a^r - b^r \mid$. It is easy to show that $H(A, B) \geq 0$ and $H(A, B) = H(B, A)$. Moreover, let C be a generic compact subset in $\mathbb{R}$, the triangular inequality $H(A, C) \leq H(A, B) + H(B, C)$ can be easily proved taking into account the definition of distance in (1) [Neumaier, 1990].

3 Generalization of $H(A, B)$ in $\mathbb{R}^p$

The generalization of the Hausdorff distance in $\mathrm{I\!R}^p$ tends to be very complex as p tends to be large. Readers interested in the properties of the Hausdorff metric in $\mathbb{R}^p$ space may refer to [Braun *et al.*, 2003]. However, when the compact subsets in $\mathrm{I\!R}^p$ are restricted to some special cases, the Hausdorff metric can be easily generalized. This paper will focus the attention on two special cases: *boxes* and hyperspheres.

3.1 Distance between *boxes*

In order to have a distance measure easy to handle in $\mathrm{I\!R}^p$, we introduce a measure of distance that generalizes the Minkowski metric.

In the $p-$dimensional space $\mathbb{R}^p$, $\mathcal{H}(X_1, X_2, \ldots, X_p)$ indicates the set of all possible bounded *boxes* (or *parallelotopes*) in the space $\mathrm{I\!R}^p$.

Given two *boxes* $\{A, B\} \in \mathcal{H}(X_1, X_2, \ldots, X_p)$, the quantity:

$$H(A, B) = \left\{ \textstyle\sum_{j=1}^{p} \mid H(A_j, B_j) \mid^{\alpha} \right\}^{\frac{1}{\alpha}} \geq 0, \tag{2}$$

for any $\alpha \geq 1$, is a metric. It is obvious that $H(A, A) = 0 \Leftrightarrow A = A$, $\forall A \in \mathcal{H}(X_1, X_2, \ldots, X_p)$, being $H(A_j, A_j) = 0, \forall j = 1, \ldots, p$. The two following properties of (2) can be easily demonstrated:

i) $H(A, B) = H(B, A)$ (*Symmetry*):
For any $(A, B) \in \mathcal{H}(X_1, X_2, \ldots, X_p)$ is:

$$\begin{aligned} H(A,B) &= \left\{\textstyle\sum_{j=1}^{p} \left[H(A_j, B_j)\right]^{\alpha}\right\}^{\frac{1}{\alpha}} = \\ &= \left\{\textstyle\sum_{j=1}^{p} \left[H(B_j, A_j)\right]^{\alpha}\right\}^{\frac{1}{\alpha}} = H(B,A) \end{aligned} \quad (3)$$

given the symmetry of $H(A_j, B_j)$ for any $j = 1, \ldots, p$.
For any $A \in \mathcal{H}(X_1, X_2, \ldots, X_p)$ is:

$$H(A,A) = \left\{\textstyle\sum_{j=1}^{p} \left[H(A_j, A_j)\right]^{\alpha}\right\}^{\frac{1}{\alpha}} = 0 \quad (4)$$

ii) $H(A, B) + H(A, C) \geq H(B, C)$ (*Triangular inequality*): For any $(A, B, C) \in \mathcal{H}(X_1, X_2, \ldots, X_p)$ under the hypothesis that the distance $H(A_j, B_j)$ satisfies the triangular inequality for any $j = 1, \ldots, p$, this follows from equation (1) (see [Neumaier, 1990] for a complete specification of the metric properties in the $\mathbb{IR}$ space). The following proves the inequality $H(A, B) + H(A, C) \geq H(B, C)$:

$$\begin{aligned} H(A,B) + H(A,C) &= \left\{\textstyle\sum_{j=1}^{p} \left[H(A_j, B_j)\right]^{\alpha}\right\}^{\frac{1}{\alpha}} + \left\{\textstyle\sum_{j=1}^{p} \left[H(A_j, C_j)\right]^{\alpha}\right\}^{\frac{1}{\alpha}} \geq \\ &\geq \left\{\textstyle\sum_{j=1}^{p} \left[H(A_j, B_j) + H(A_j, C_j)\right]^{\alpha}\right\}^{\frac{1}{\alpha}} \geq \\ &\geq \left\{\textstyle\sum_{j=1}^{p} \left[H(B_j, C_j)\right]^{\alpha}\right\}^{\frac{1}{\alpha}} = H(B,C), \end{aligned} \quad (5)$$

being $H(A_j, B_j) + H(A_j, C_j) \geq H(B_j, C_j)$ satisfied for any j, according to the Hausdorff metric definition in $\mathbb{R}$.

The distance in $\mathbb{IR}^p$ introduced in (2), for $\alpha = 2$ can also be expressed in terms of centers and *radii*:

$$H(A,B) = \sqrt{\sum_{j=1}^{p} \left[(a_j^c - b_j^c)^2 + (a_j^r - b_j^r)^2 + 2 \mid a_j^c - b_j^c \mid\mid a_j^r - b_j^r \mid\right]}. \quad (6)$$

This notation will be useful when we shall present the factorial model.

3.2 Hausdorff distance between two spheres in $\mathbb{R}^p$

Another distance in $\mathbb{R}^p$, which derives from the Hausdorff metric in $\mathbb{R}$, is given by the distance defined between the spheres inscribing the *parallelotopes*. This distances coincides with the Hausdorff metric when the SO are hypercubes (equal edges).

Before illustrating the distance we prove the following theorem that defines the Hausdorff distance between spheres in the $\mathbb{R}^p$ space.

In this section capital letters $\{A, B, \ldots\}$ indicate spheres in the $\mathbb{R}^p$ space; the general sphere A has center in $A^c = [a_j^c]$ $(j = 1, \ldots, p)$ and *radius* $A^r \geq 0$.

Theorem 1 *Given two spheres $\{A, B\}$ in the $\mathbb{R}^p$ space, the Hausdorff distance between A and B is given by:*

$$H(A,B) = \sqrt{\sum_{j=1}^{p} \left(a_j^c - b_j^c\right)^2} + \mid A^r - B^r \mid \tag{7}$$

Proof. We remind that the equation of the sphere A is: $\sum_{j=1}^{p} \left(x_j - a_j^c\right)^2 = \left(A^r\right)^2$. The minimum and the maximum Euclidean distance from a point $\mathsf{O} = [x_j]$ with $(j = 1, \ldots, p)$ to the sphere A are the radii of the spheres, having centers in O, external to A and containing A, respectively. So that, the Euclidean Hausdorff distance between O and A is the minimum one.

Two spheres A and B are tangent if:

$$\sum_{j=1}^{p} \left(a_j^c - b_j^c\right)^2 = (A^r \pm B^r)^2. \tag{8}$$

If A does not intersect B, we have the sign +; if A is inside B, we have the sign −. Let us suppose that O represents the center of the sphere B. If O belongs to A, the minimum distance between O and A is 0, obviously. If O is external to A, we need to solve the following equation for $r^\diamond$:

$$\sum_{j=1}^{p} \left(a_j^c - x_j\right)^2 = (A^r + r^\diamond)^2. \tag{9}$$

Solving with respect to $r^\diamond$ we have:

$$r^\diamond = \sqrt{\sum_{j=1}^{p} \left(a_j^c - x_j\right)^2} - A^r = \min_{\tilde{a} \in A} d(\mathsf{O}, A). \tag{10}$$

Let us assume that O belongs to B. The $\max\min\{d(B, A)\}$ is given by:

$$\max_{\tilde{b} \in B} \min_{\tilde{a} \in A} \left(d(\tilde{a}, \tilde{b})\right) = \max_{\tilde{b} \in B} \left(\sqrt{\sum_{j=1}^{p} \left(a_j^c - \tilde{b}_j\right)^2}\right) - A^r. \tag{11}$$

Equivalently, the minimum *radius* sphere with the same center as A that both contains and is tangent to B.

According to (8) we have to solve the following equation in r^*:

$$\sum_{i=1}^{p} \left(a_j^c - b_j^c\right)^2 = \left(r^* - B^r\right)^2$$

$$r^* = \sqrt{\sum_{j=1}^{p} \left(a_j^c - b_j^c\right)^2} + B^r. \tag{12}$$

Then the Hausdorff distance based on the Euclidean distance from a sphere B to a sphere A is:

$$h(B,A) = \max_{\tilde{b}\in B} \min_{\tilde{a}\in A} d(\tilde{b},\tilde{a}) = \sqrt{\sum_{j=1}^{p} \left(a_j^c - b_j^c\right)^2} + (B^r - A^r) \qquad (13)$$

then $H(A,B) = \max(h(A,B), h(B,A)) = \sqrt{\sum_{j=1}^{p} \left(a_j^c - b_j^c\right)^2} + |A^r - B^r|$ and the proof is complete. □

Given two spheres $\{A, B\}$ in the $\mathbb{R}^p$ space, $H(A,B)$ is a metric. Reflexive and symmetric properties are intuitive. We need to prove that $H(A,B) + H(B,C) \geq H(A,C)$ is true (*triangular inequality*).
For the triangular property of Euclidean (for centers) and Manhattan (for radii) distance we may assert that:

$$\sqrt{\sum_{j=1}^{p} \left(a_j^c - b_j^c\right)^2} + \sqrt{\sum_{j=1}^{p} \left(b_j^c - c_j^c\right)^2} \geq \sqrt{\sum_{j=1}^{p} \left(a_j^c - c_j^c\right)^2}$$
$$|A^r - B^r| + |B^r - C^r| \geq |A^r - C^r| \,.$$

Then it follows that:

$$\sqrt{\sum_{j=1}^{p} \left(a_j^c - b_j^c\right)^2} + |A^r - B^r| + \sqrt{\sum_{j=1}^{p} \left(b_j^c - c_j^c\right)^2} + |B^r - C^r| \geq$$
$$\geq \sqrt{\sum_{j=1}^{p} \left(a_j^c - c_j^c\right)^2} + |A^r - C^r|$$

4 A comparison between two metrics

This section presents a comparison between the factorial analysis based on the two proposed measures of distance for interval valued variables. The example shows the results obtained on the "*Italian Peppers*" dataset; these data are a good example of native interval variables, they describe some chemio-physical properties (*H_2O, Glicide, Lipid, Protein*) of eight different species of Italian peppers: (*Cuban, Cuban Nano, Corno di Bue, Grosso di Nocera, Pimento, Quadrato d'Asti, Sunnybrook, Yolo wonder*). [Lauro and Palumbo, 2005]

Each factorial approach has been chosen to ensure the maximum degree of consistency with respect to the distance measure. We remind that statistical factorial analysis for interval variables does not limit itself to the study of proximities among dimensionless points but, it must take into account the size and shape of the compact subsets in $\mathbb{R}^p$

Let $[\mathbf{X}]$ be a generic $n \times p$ interval data matrix. In order to simplify the notation, we define the centers matrix $\mathbf{C} = \frac{1}{2}(\overline{\mathbf{X}} + \underline{\mathbf{X}})$ and the ranges matrix $\mathbf{R} = \frac{1}{2}(\overline{\mathbf{X}} - \underline{\mathbf{X}})$ where, $\underline{\mathbf{X}}$ and $\overline{\mathbf{X}}$ are the minimum and the maximum values matrices, respectively. All these matrices are of $n \times p$ order.

The arithmetic mean of the generic interval-valued variables $[\mathbf{x}]_j$, according to the the basic principles of the interval arithmetic [Hickey *et al.*, 2001] is defined as:

$$[\bar{\mathbf{x}}]_j = \frac{1}{n}\sum_{i=1}^{n}[x]_{i,j} = \frac{1}{n}\left[\sum_{i=1}^{n}\underline{x}_{i,j}, \sum_{i=1}^{n}\overline{x}_{i,j}\right] = \left\{\frac{1}{n}\sum_{i=1}^{n}x^c_{i,j}, \frac{1}{n}\sum_{i=1}^{n}x^r_{i,j}\right\}. \quad (14)$$

Whereas dealing with single valued variables, in $\mathbb{R}$ space, difference and distance measures are equivalent apart the sign; this is not true when variable is interval-valued. Lauro and Palumbo (2005) defined the following measure of variability for interval-valued variables based on the Hausdorff distance:

$$\text{var}([x]_j) = \frac{1}{n}\sum_{i=1}^{n}\left[\mid x^c_{i,j} - \bar{x}^c_j \mid + \mid x^r_{i,j} - \bar{x}^r_j \mid\right]^2, \quad (15)$$

where $\bar{x}^c_j$ and $\bar{x}^r_j$ represent, respectively, the mean midpoint and the mean range of the generic interval variable $[X]_j$. We call centered and reduced the interval valued variable:

$$[z]_j = \{z^c_j, z^r_j\} = \left\{\frac{(x^c_j - \bar{x}^c_j)}{\sqrt{\text{var}[x]_j}}, \frac{\mid x^r_j - \bar{x}^r_j \mid}{\sqrt{\text{var}[x]_j}}\right\}.$$

The distance presented in equation (6) can be rewritten in matrix notation as: $H^2 = \mathbf{C}\mathbf{C}^\mathsf{T} + \mathbf{R}\mathbf{R}^\mathsf{T} + \mid \mathbf{C} \mid\mid \mathbf{R}^\mathsf{T} \mid + \mid \mathbf{R} \mid\mid \mathbf{C}^\mathsf{T} \mid$, where we assume that interval variables have been centered and reduced. The quantity $trace(H^2)$ is the sum of the n squared distances from the mean. However, in the PCA practice it is preferred to apply the SVD to the $p\times p$ correlation (or covariance) matrix, in this case we will apply the SVD to the matrices $\mathbf{C}^\mathsf{T}\mathbf{C}$ and $\mathbf{R}^\mathsf{T}\mathbf{R}$. The MR-PCA of Lauro and Palumbo performs two separate PCA's on the matrices $\mathbf{C}^\mathsf{T}\mathbf{C}$ and $\mathbf{R}^\mathsf{T}\mathbf{R}$ and permits to recover the intervals on the factorial plan by adding and subtracting the rotated and translated *radii* into the space of the centers coordinates in their own space. The rotation matrix $\mathbf{T}$ is defined maximizing the quantity $\mathbf{C}^\mathsf{T}\mathbf{R}$. Notice that the square matrix $\mathbf{\Sigma}$:

$$\mathbf{\Sigma} = \mathbf{C}^\mathsf{T}\mathbf{C} + \mathbf{R}^\mathsf{T}\mathbf{R} + \mid \mathbf{C}^\mathsf{T} \mid\mid \mathbf{R} \mid + \mid \mathbf{R}^\mathsf{T} \mid\mid \mathbf{C} \mid$$

is symmetric; the extra-diagonal terms vary in $[-1, 1]$ and the diagonal terms are equal to 1. It represents a sort of correlation matrix for interval valued variables, where the total correlation is the sum of three different components: midpoints association, ranges association and the midpoints/ranges congruence.

The output of the method of Palumbo and Lauro consists in a representation of the centers and of the *radii* taken into account singly, and of a joint representation where interval objects are represented by rectangles having sides parallel to the axes. However, here we propose only the midpoints and *radii* joint representation.

In order to present the second factorial approach based on the definition of the distance in (7), differently from the previous approach, we consider that both center and *radii* variables, respectively in the matrices $\mathbf{C}$ and $\mathbf{R}$, are reduced with respect to standard deviations of the centers (see [Giordani and Kiers, 2004]). Notice that the matrix B in (7) has a constant value over the main diagonal, it corresponds to the norm of the average units *radius.* The matrix notation of the distance in (7) is equal to: $\sum_{i=1}^{n} H(A_i, \bar{A})^2 = tr\, \mathbf{C}^\mathsf{T}\mathbf{C} + tr\, \mathbf{R}^\mathsf{T}\mathbf{R}$. The symbol $\bar{A}$ indicates the mean SO that is obtained by applying the formula (14). The problem consists in finding the orthogonal subspace the maximizes $tr\, \mathbf{C}^\mathsf{T}\mathbf{C} + tr\, \mathbf{R}^\mathsf{T}\mathbf{R}$ simultaneously. We introduce the super matrix $\mathbf{Y}$:

$$\mathbf{Y} = \begin{bmatrix} \mathbf{C} \\ \mathbf{R} \end{bmatrix}.$$

The projection of $\mathbf{Y}$ on a common orthogonal subspace can be obtained by means of the extraction of the principal components of $\mathbf{C}^\mathsf{T}\mathbf{C}$ denoted as $\mathbf{D_{CC}}$. Considering the projection of $\mathbf{Y}$ on the space spanned by the centers using the projector $\mathbf{P_C}$ and on the orthogonal projection using $\mathbf{P_C^\perp}$ we have:

$$\begin{aligned} \mathbf{Y} &= \mathbf{P_C}\mathbf{Y} + \mathbf{P_C^\perp}\mathbf{Y} = (\mathbf{P_C}\mathbf{C}, \mathbf{P_C}\mathbf{R}) + (\mathbf{P_C^\perp}\mathbf{C}, \mathbf{P_C^\perp}\mathbf{R}) = \\ &= (\mathbf{C}, \mathbf{P_C^\perp}\mathbf{R}) + (\mathbf{0}, \mathbf{P_C^\perp}\mathbf{R}) \end{aligned} \tag{16}$$

that leads to the following decomposition:

$$\mathbf{D_{YY}} = \begin{bmatrix} \mathbf{D_{CC}} & \mathbf{D_{CR}} \\ \mathbf{D_{RC}} & \mathbf{D_{RR}} \end{bmatrix} = \begin{bmatrix} \mathbf{D_{CC}} & \mathbf{D_{CR}} \\ \mathbf{D_{RC}} & \mathbf{D_{RRC}} \end{bmatrix} + \begin{bmatrix} 0 & 0 \\ 0 & \mathbf{D_{RR.\frac{\perp}{C}}} \end{bmatrix} \tag{17}$$

where $\mathbf{D_{RR.\frac{\perp}{C}}}$ can be obtained computing the first principal components of $\mathbf{C}$ and then obtaining the structure matrix of $\mathbf{D_{CC}}$ on the base of the set of the principal components of $\mathbf{D_{CC}}$. For further details see [Takeuchi *et al.*, 1982]. Taking into account these results, there are several possible approaches to the analysis [Lebart *et al.*, 1995]; for sake of space, here we do not discuss the choices and their motivations. Figure 1 shows the results based on the distance between *boxes*: ranges are rotated and projected into the space of the midpoints as supplementary points. The total variability associated to the first factorial plan is 65.76%. Figure 2 shows the results obtained representing the SO with respect to the distance between hyperspheres. Here the variability associated to the first factorial plan is equal to 76.48%.

Looking at the outputs we notice that the SO can be distinguished according to their position, size and shape. It is worth noticing that, with respect to the positions, results of the two analyses are consistent. SO have the same order on the first factor in both analyses. The size interpretation is quite intuitive. Few words are necessary to correctly interpret the shapes: it is necessary to take into account the shape itself and also the range orientation. Looking at figure 1, we notice that *Sunnybrook* and *Quadrato d'Asti* have similar shapes but they have opposite range orientation. This is confirmed

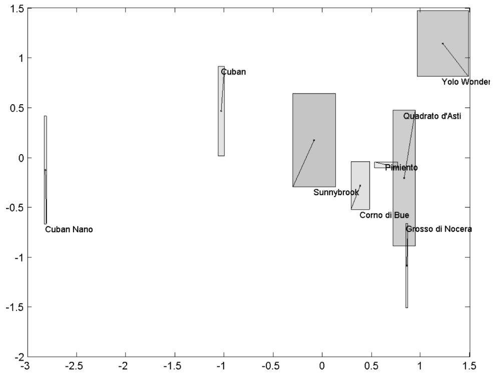


Fig. 1. Distance between *boxes*: first factorial plan (65.76%)

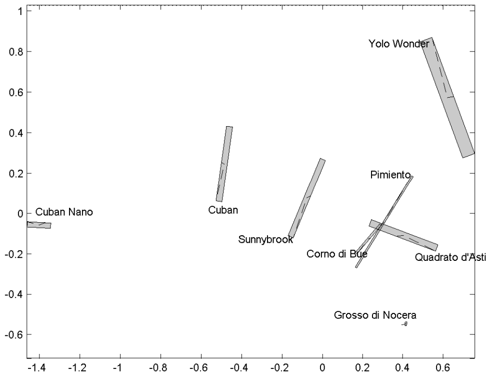


Fig. 2. Distance between hyperspheres: first factorial plane (76.48%)

also in the other analysis: in figure 2 we see that *Sunnybrook* and *Quadrato d'Asti* appear orthogonal.

To understand which variables have mainly characterized the positioning and the size and shape of the SO it is necessary to look at the variables representations on the same factorial plans.

5 Conclusion and future work

Since Edwin Diday introduced Symbolic Data Analysis [Diday, 1989] we have noticed a growing interest for the analysis of complex data structures. The first book entirely dedicated to Symbolic Data Analysis appeared five years ago [Bock and Diday, 2000]. These new statistical data need new concepts not having a counterpart in the "classical" data analysis, necessarily. At the beginning, many have proposed special data-codings to make data tractable by the traditional methods; so that, most of the big effort done up to now allowed us to treat complex data with *suitably adapted* methods for single-valued data. We believe that the next challenge is to setup numerical and statistical methods that are specifically designed for the complex-data structures. We see two main research directions: *i*) definition of new statistical indexes (measures of central tendency, variability, etc.) that take into account the innovative nature of the data; *ii*) development of analytical and numerical methods allowing to treat intervals as mathematical structures. Interval arithmetic has been mainly developed to treat data imprecision caused by the "old" fix-point CPU (the round-off error) and its generalization to the statistical interval-valued data requires a big effort.

Nevertheless, the treatment of set-valued variables is a field with very high potential for further developments.

References

[Alefeld and Mayer, 2000]G. Alefeld and G. Mayer. Interval analysis: theory and applications. *Jour. of Comp. and Appl. Mathematics*, 121:421–464, 2000.

[Bock and Diday, 2000]H. H. Bock and E. Diday, eds. *Analysis of Symbolic Data. Exploratory methods for extracting statistical information from complex data.* Springer Verlag, Heidelberg, 2000.

[Braun *et al.*, 2003]D. Braun, J. Mayberry, A. Powers, and S. Schlicker. The geometry of the Hausdorff metric. Available at: `http://faculty.gvsu.edu/schlicks/Hausdorff_ paper.pdf`, August 2003.

[Cazes *et al.*, 1997]P. Cazes, A. Chouakria, E. Diday, and Y. Schektman. Extension de l'analyse en composantes principales à des données de type intervalle. *Revue de Statistique Appliquée*, XIV(3):5–24, 1997.

[Diday, 1989]E. Diday. Introduction à l'approche symbolique en analyse des données. *Rev. d'Aut., d'Informatique et de Rec. Opérationnelle*, 32(2), 1989.

[Giordani and Kiers, 2004]P. Giordani and H.A.L. Kiers. Principal component analysis of symmetric fuzzy data. *Comp. Stat. Data Anal.*, 45:519–548, 2004.

[Hickey *et al.*, 2001]T. Hickey, Q. Ju, and M. H. Van Emden. Interval arithmetic: From principles to implementation. *Jour. of the ACM*, 48(5):1038–1068, 2001.

[Lauro and Palumbo, 2000]C. N. Lauro and F. Palumbo. Principal component analysis of interval data: A symbolic data analysis approach. *Computational Statistics*, 15(1):73–87, 2000.

[Lauro and Palumbo, 2005]C. N. Lauro and F. Palumbo. Principal component analysis for non-precise data. In M. Vichi *et al.*, eds, *New Developments in Classification and Data Analysis*, pages 173–184. Springer, 2005.

[Lebart *et al.*, 1995]Ludovic Lebart, Alain Morineau, and M. Piron. *Statistique exploratorie multidimensionelle.* Dunod, Paris, 1995.

[Neumaier, 1990]A. Neumaier. *Interval methods for systems of Equations.* Cambridge University Press, Cambridge, 1990.

[Takeuchi *et al.*, 1982]K. Takeuchi, H. Yanai, and B. N. Mukherjee. *The Fundations of Multivariate Analysis.* Wiley Eastern Ltd., New Delhi, 1982.

[Xiang *et al.*, 2004]G. Xiang, S. A. Starks, V. Kreinovich, and L. Longpre. New algorithms for statistical analysis of interval data. Utep-cs-04-04, NASA PACES, El Paso, TX 79968, USA, 2004. at: `http://www.cs.utep.edu/vladik/2004/list04.html`.

Clayton copula and mixture decomposition

Etienne Cuvelier and Monique Noirhomme-Fraiture

FUNDP(Facultés Universitaires Notre-Dame de La Paix)
Institut d'Informatique
B-5000 Namur, Belgium
(e-mail: cuvelier.etienne@info.fundp.ac.be,
noirhomme.monique@info.fundp.ac.be)

Abstract. A symbolic variable is often described by a histogram. More generally, it can be provided in the form of a continuous distribution. In this case, the problem is to solve the most frequent problem in data mining, namely: to classify the objects starting from the description of the variables in the form of continuous distributions. A solution is to sample each distribution in a number N of points, and to evaluate the joint distribution of these values using the copulas, and also to adapt the dynamical clustering (nuées dynamiques) method to these joint densities. In this paper we compare the Clayton copula and the Normal copula for more than 2 dimensions, and we compare results of clustering by using on the one hand the method based on the Clayton copula and traditional methods (MCLUST, and K-means). Our comparison is based on 2 well-known classical data files.

Keywords: symbolic data analysis, mixture decomposition, Clayton copula, clustering.

1 Introduction

The mixture decompostion is a classical tool used in clustering. The method consists in estimating a probability density function from a given sample in R^q, considering that the reached function f is a finite mixture of K densities:

$$f(x_1, ..., x_q) = \sum_{i=1}^{K} p_i \cdot f(x_1, ..., x_q, \beta_i) \tag{1}$$

with $\forall\ i \in \{1, ..., K\}$, $0 < p_i < 1$, and $\sum_{i=1}^{K} p_i = 1$. The function $f(., \beta)$ is a density function with parameter β belonging to R^d and p_i is the probability that one element of the sample get the density $f(., \beta_i)$. In this clustering approach each component of the mixture corresponds to a cluster.

To find the partition $P = (P_1, ..., P_K)$, which is the best adapted to the data two main algorithms were proposed : the EM algorithm (Estimation, Maximisation) [Dempster *et al.*, 1977] and the dynamical clustering algorithm [Diday *et al.*, 1974].

A use of the dynamical clustering algorithm in the symbolic data analysis framework when the data are distribution probabilities was proposed in [Diday, 2002]. In a symbolic data table, a statistical unit can be described by

numbers, intervals, histograms and probability distributions. We suppose to have a table T with n lines and p columns, and that the j^{th} column contains probability distributions, i.e. if we note Y^j the j^{th} variable then Y_i^j is a distribution $F_i(.)$ for all $i \in \{1, ..., n\}$. To cluster this last type of data two main ideas were proposed in [Diday, 2002]. The first idea is to use as sample the values of the distributions found in table T in q quite selected values $T_1, ..., T_q$: $\{(F_i(T_1), ..., F_i(T_q)) : i \in \{1, ..., n\}\}$. The second idea is to estimate the margins of $f(., \beta_i)$ in a first step , and to join them in a second step using copulas.
[Vrac *et al.*, 2001] used this approach with success to cluster atmospheric data with the Franck copula of dimension 2 (i.e. with only two real values T_1 and T_2 where distributions are computed). The starting point of our work is to extend this approach with copulas with a higher number of dimensions with the Clayton n-copula.
The organization of the paper is as follows. In section 2 we set the symbolic data analysis framework for the mixture decomposition when data are probability distributions. A general presentation of the copulas is made in section 3, and we focus in section 4 on the Clayton copula. In the following section we show the implementation and results, and we conclude with perspectives and future work in the last section.

2 The symbolic data analysis framework

2.1 Distributions of distributions

We suppose to have a table T with n lines and p columns, and that the j^{th} column contains probability distributions, i.e. if we note Y^j the j^{th} variable then Y_i^j is a distribution $F_i(.)$ for all $i \in \{1, ..., n\}$. In the following we note ω_i the concept described by the i^{th} row, and $F_{\omega_i}(.)$ the associated distribution. We choose q real values $T_1, ..., T_q$ (we don't discuss of the choice of this values here), and for each $i \in \{1, ..., n\}$ we compute $F_{\omega_i}(T_1), ..., F_{\omega_i}(T_q)$. Then, if we call Ω the set of all concepts, the joint distribution of the $F_i(T_j)$ values is defined by:

$$H_{T_1,...,T_q}(x_1, ..., x_q) = P\left(\omega \in \Omega : \{F_\omega(T_1) \le x_1\} \cap ... \cap \{F_\omega(T_q) \le x_q\}\right) \quad (2)$$

which is called distribution of distributions. The classical classification method consists in considering this distribution as the result of a finite mixture distributions:

$$H_{T_1,...,T_q}(x_1, ..., x_q) = \sum_{i=1}^{K} p_i \cdot H^i_{T_1,...,T_q}(x_1, ..., x_q; \beta_i) \quad (3)$$

with $\forall\ i \in \{1, ..., K\} : 0 < p_i < 1$ and $\sum_{i=1}^{K} p_i = 1$.
The distribution of $\mathrm{i^{th}}$ cluster is given by $H^i_{T_1,...,T_q}(x_1, ..., x_q; \beta_i)$, where parameter $\beta_i \in R^d$, and p_i is the probability that one element is in this cluster.

If we take a look at the densities, then the probability density of H is

$$h(x_1, ..., x_q) = \frac{\partial^q}{\partial x_1 ... \partial x_q} H(x_1, ..., x_q) \tag{4}$$

And the mixture densities is given by:

$$h_{T_1,...,T_q}(x_1, ..., x_q) = \sum_{i=1}^{K} p_i \cdot h^i_{T_1,...,T_q}(x_1, ..., x_q; \beta_i) \tag{5}$$

2.2 Clustering algorithm

The clustering algorithm proposed by [Diday, 2002] is an extension of the dynamical clustering method [Diday *et al.*, 1974] for density mixtures. The main idea is to estimate at each step, the density which describes at best the clusters of the current partition P, according to a given quality criterion. We considered the classifier log-likelihood :

$$lvc(P, \beta) = \sum_{i}^{K} \sum_{\omega \in P_i} log(h(w)) \tag{6}$$

where

$$h(w) = h_{T_1,...,T_q}(F_\omega(T_1), ..., F_\omega(T_q)) \tag{7}$$

The classification starts with a random partition, then the two following steps are repeated:

- **Step 1 : Parameters estimation**
  Find the vector $(\beta_1, ..., \beta_K)$ which maximizes the chosen criterion;
- **Step 2 : Distribution of units in new classes**
  Build new classes $(P_i)_{i=1,...,K}$ with parameters found at Step 1 :

$$P_i = \{\omega : p_i \cdot h(\omega, \beta_i) \geq p_m \cdot h(\omega, \beta_m) \forall m\} \tag{8}$$

until the stabilization of partition.

2.3 Estimation

Before using this algorithm we must know how to estimate the density of each cluster.
For univariate distributions we may use :

- a parametric approach, and use well-known laws as the Beta law (Dirichlet's law in one dimension) or the Normal law,

- a non-parametric approach, as the kernel density estimation :

$$\hat{f}(x) = \frac{1}{n \cdot h} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right) \tag{9}$$

where
  - $(X_1, ..., X_n)$, is the sample over which the estimation is made,
  - K is the kernel density function (many possible choices...)
  - h is the window width, and can be automatically estimated with Mean Integrated Square Error(MISE) formulae $h = 1.06\sigma N^{-1/5}$ [Silverman, 1986], where σ is the standart deviation of the sample.

For multivariate distributions, we can also use parametric estimation, with a Normal multivariate distribution for example like in [Fraley and Raftery, 2002] or a non parametric approach (the kernel estimation exists also in higher dimensions, but is heavier in calculations), but we can also attempt to re-build the joint distributions H with marginals coupling, by using copula, and at the same time have a model of the dependence structure of the data.

3 Multivariate copulas

A multivariate copula, also called n-copula, is a function C from $[0,1]^n$ to $[0,1]$ with the following properties :

- $\forall$ $\mathbf{u} \in [0,1]^n$,
  - $C(\mathbf{u}) = 0$, if at least one coordinate of $\mathbf{u}$ is 0,
  - $C(\mathbf{u}) = u_k$, if all coordinates of $\mathbf{u}$ are 1 except u_k

- $\forall$ $\mathbf{a}, \mathbf{b} \in [0,1]^n$, such that $\mathbf{a_i} \leq \mathbf{b_i}, \forall\ 1 \leq i \leq n$,

$$V_C([\mathbf{a}, \mathbf{b}]) \geq 0, \tag{10}$$

where $[\mathbf{a}, \mathbf{b}] = [\mathbf{a_1}, \mathbf{b_1}] \times ... \times [\mathbf{a_n}, \mathbf{b_n}]$, and $V_C([\mathbf{a}, \mathbf{b}])$ is the nth order difference of H on $[\mathbf{a}, \mathbf{b}]$:

$$V_C([\mathbf{a}, \mathbf{b}]) = \Delta_{\mathbf{a}}^{\mathbf{b}} C(\mathbf{t}) = \Delta_{a_n}^{b_n} \Delta_{a_{n-1}}^{b_{n-1}} ... \Delta_{a_2}^{b_2} \Delta_{a_1}^{b_1} C(\mathbf{t}) \tag{11}$$

with

$$\Delta_{a_k}^{b_k} C(\mathbf{t}) = C(..., t_{k-1}, b_k, t_{k+1}, ...) - C(..., t_{k-1}, a_k, t_{k+1}, ...) \tag{12}$$

The copulas are powerfull tools in modeling dependences since Abe Sklar stated the following theorem [Sklar, 1959]:
Let H be an n-dimensional distribution function with margins $F_1, ..., F_n$. Then there exists an n-copula C such that for all $\mathbf{x}$ in $\bar{R}^n$,

$$H(x_1, ..., x_n) = C(F_1(x_1), ..., F_n(x_n)). \tag{13}$$

If $F_1, ..., F_n$ are all continuous, then C is unique; otherwise, C is uniquely determined on Range of $F_1 \times ... \times$Range of F_n.

In fact the copula captures the dependence structure of the distribution. In our case, if we note a univariate margin :

$$G_T(x) = Pr\left(\omega \in \Omega : \{F_\omega(T) \leq x\}\right) \tag{14}$$

then the mixture can be written as follows

$$H_{T_1,...,T_q}(x_1, ..., x_q) = \sum_{i=1}^{K} p_i \cdot C^i(G^i_{T_1}(x_1), ..., G^i_{T_q}(x_q); \beta_i) \tag{15}$$

and in terms of densities

$$h^i_{T_1,...,T_q}(x_1, ..., x_q; \beta_i) = \prod_{i=1}^{q} \frac{dG^i_{T_i}}{dx}(x_i) \times \frac{\partial^q}{\partial u_1 ... \partial u_q} C^i(G^i_{T_1}(x_1), ..., G^i_{T_q}(x_q); \beta_i) \tag{16}$$

The use of copulas allows us to estimate all the marginals first, and in a second time to estimate the parameters of each copula. The copula modelises the dependences of the $F_\omega(T_i)$ values inside each cluster. Note well that this use of copulas can be made, not only when the original data are symbolic data described by a continuous distribution, but also with quantitative unspecified variables.

4 Clayton copula

In the following we present the Clayton's copula we use for our implementation, and the Normal copula for comparison.

The Clayton copula is an Archimedean copula. These copulas are generated by a function ϕ, called the generator:

$$C(u_1, ..., u_n) = \phi^{-1}\left(\sum_{i=1}^{n} \phi(u_i)\right) \tag{17}$$

where ϕ is a function from $[0, 1]$ to $[0, \infty]$ such that:

- ϕ is a continuous strictly decreasing function
- $\phi(0) = \infty$
- $\phi(1) = 0$
- ϕ^{-1} is completely monotonic on $[0, \infty[$ i.e.

$$(-1)^k \frac{d^k}{dt^k} \phi^{-1}(t) \geq 0 \tag{18}$$

for all t in $[0, \infty[$ and for all k.

If we use $\phi_\theta(t) = t^\theta - 1$ as generator, then we get the Clayton's copula

$$C(u_1, ..., u_n) = \left(1 - n + \sum_{i=1}^{n} u_i^{-\theta}\right)^{-1/\theta} \tag{19}$$

which is a copula only if $\theta > 0$.

We choose this copula in the set of the multivariate Archimedean copulas because as showed in [Cuvelier and Noirhomme-Fraiture, 2003], the density is easy to compute:

$$c(u_1, ..., u_q) = \left(1 - q + \sum_{i=1}^{q} u_i^{-\theta}\right)^{-q-\frac{1}{\theta}} \prod_{j=1}^{q} \left(u_j^{-\theta-1} \{(j-1)\theta + 1\}\right). \tag{20}$$

It is important to notice that all the k-margins of an Archimedean copula are identical: $C(u_1, ..., u_{n-1}, 1) = \phi^{-1}\left(\sum_{i=1}^{n-1} \phi(u_i)\right)$. This fact limits the nature of dependence structure in these families because it introduces a certain symmetry.

The Normal copula is built by the most obvious process: the inversion method. If we have a multivariate distribution H, with margins $F_1, ..., F_n$, then for any $\mathbf{u}$ in $[0, 1]^n$:

$$C(u_1, ..., u_n) = H(F_1^{(-1)}(u_1), ..., F_n^{(-1)}(u_n)) \tag{21}$$

is a copula. Let ρ be a positive correlation matrix, Φ_ρ the Normal multivariate distribution defined with this matrix, and Φ the standard Gaussian distribution. The Normal copula is then defined by:

$$C(u_1, ..., u_n) = \Phi_\rho(\Phi^{-1}(u_1), ..., \Phi^{-1}(u_n)). \tag{22}$$

and its density is given by

$$c(u_1, ..., u_n) = \frac{1}{|\rho|^{\frac{1}{2}}} \; exp\left(-\frac{1}{2} \, \varsigma^\tau (\rho^{-1} - I) \, \varsigma\right) \tag{23}$$

where $\varsigma = \Phi^{-1}(u_i)$, and $\mathbf{I}$ is the $(n \times n)$ unity matrix. This copula has two main advantages : there is a formula to calculate its density in any dimension and, more significantly, a large set of parameters ($\frac{n \cdot (n-1)}{2}$) which indicates that one can have a very flexible modelisation of the dependence.

To show the difference between these two copulas, we generated 1000 random couples of numbers, once with Clayton copula ($\theta = 5$, figure 1), and then with the Normal copula (with a correlation of 0.5 between the two variables, figure 2).

As we can see the spatial distributions of the generated points have radically different forms. That implies that the choice of one of these two copulas

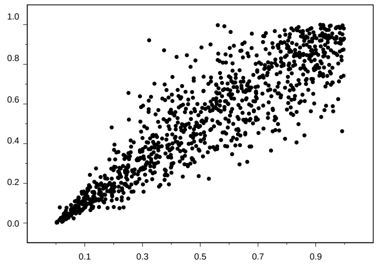


Fig. 1. Dependence structure of Clayton copula

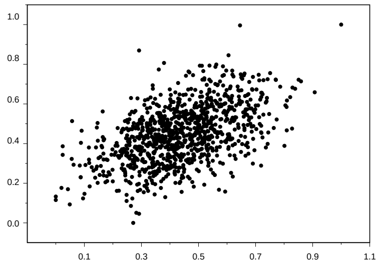


Fig. 2. Dependence structure of normal copula

will influence the shape of the clusters we can retrieve in the data. The Normal copula, and more generally the Normal distribution, tends to form elliptic groups whereas, as we can see, the copula of Clayton will tend to form groups "with pear shape".
In fact the "pear shape" shown in figure 1 is due to a property of the Clayton copula called **lower tail dependence**: a copula C has lower tail dependence if

$$lim_{u \to 0} \frac{C(u,u)}{u} > 0 \tag{24}$$

Of course the use of the Normal copula, in addition with Normal margins, corresponds to the use of the Normal multivariate distribution which was already largely studied and used in clustering methods. We will compare our results to the results of MCLUST [Fraley and Raftery, 2002] on the same data set.

5 Implementation of the algorithm and results

In this section we call our clustering algorithm (i.e. the dynamical clustering algorithm, with Clayon copula): Clayton Copule-Based clustering (CCBC). We compare the results of CCBC to the k-means implemented in S-Plus, and to the Model-Based clustering (MCLUST, [Fraley and Raftery, 2002] and [Fraley and Raftery, 1999]).
Our implementation of CCBC was made in the statistical language S, using the S-plus software. To estimate the unidimensionnal margins, we used kernel density estimation for margins (with Normal kernel).

To test our implementation we used two classical data sets. We used first the very well known Iris database from Fisher. The data set contains 3 classes of 50 instances each, where each class refers to a type of Iris plant (Iris Setosa, Iris Versicolour and Iris Virginica). The 4 numerical attributes are : sepal length, sepal width, petal length and petal width. We found the same clusters with few misclassified individuals as it can be seen in table 1. The results are encouraging, especially taking into account the fact that MCLUST uses multivariate Normal laws, and so uses 6 parameters for each law, which supposes a greater flexibility to adapt it to various dependence structures.
After this we used the UCI Wisconsin diagnostic breast cancer data. In a widely publicized work [Mangasarian *et al.*, 1995], 176 consecutive future cases were successfully diagnosed from 569 instances through the use of linear programming techniques to locate planes separating classes of data. Their results were based on 3 out of 30 attributes: `extreme area`, `extreme smoothness` and `mean texture`. The three explanatory variables were chosen via cross-validation comparing methods using all subsets of 2, 3, and 4 features and 1 or 2 linear separating planes. The data is avalaible from the UCI Machine Learning Repository (http://kdd.ics.uci.edu/). The three variables of interest are shown in figure 3, and we can see that, if the joint distribution of variables `extreme smoothness` and `mean texture` seems Normal, on the other hand, the two other joint distributions are closer to Clayton copula.

-	**CCBC**	MCLUST	k-means
Misclass. Numb.	**9**	5	17
Percent.	**6%**	3.33%	11.33%

Table 1. Misclassified data from Fisher's Iris

We can see in table 2 that the mixture model with the Clayton copula captures the structure dependence of the breast cancer data better than the multivariate Normal distribution, in spite of the fact that all the k-margins of the copula of Clayton are identical, i.e. that this one seeks clusters necessarily presenting a certain symmetry.

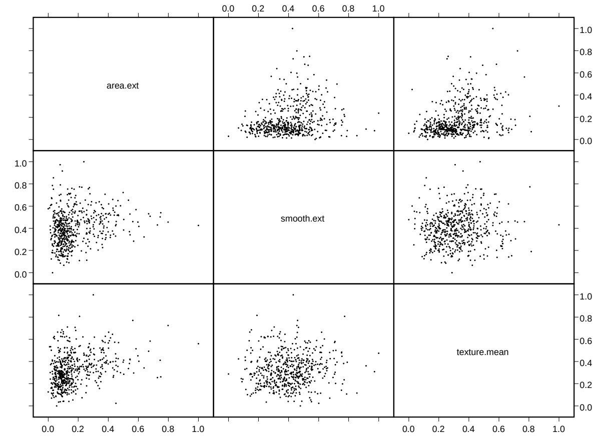


Fig. 3. Pair plots of Wisconsin Diagnostic Breast Cancer Data

-	**CCBC**	MCLUST	k-means
Misclass. Numb.	**27**	29	62
Percent.	**4.7%**	5%	10.89%

Table 2. Misclassified data from Wisconsin Diagnostic Breast Cancer Data

6 Conclusions

Mixture decomposition is a tool for classification which has already largely proved its reliability. In the same way the interest of the Copulas in the study of the dependence structures is well-known. One of the main interest of the copulas is to escape to the normality assumption and to the linear correlation.

We have shown that we can obtain equivalent or better results for clustering as with other methods, even if Clayton copula shares the weakness of all the Archimedean copulas [Nelsen, 1999]: first, in general all the k-magins of an archimedean n-copula are identical, secondly, the fact that there are only one or two parameters limits the nature of the dependence structure in these families. To overcome this weakness, in future work we intend to use other copulas with more flexible dependence structures. Now we can start to test CCBC on symbolic data.

References

[Cuvelier and Noirhomme-Fraiture, 2003]Etienne Cuvelier and Monique Noirhomme-Fraiture. Mélange de distributions de distributions, décomposition de mélange avec la copule de clayton. In *XXXV èmes Journées de Statistiques*, 2003.

[Dempster *et al.*, 1977]A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society (Series B)*, 39(1):1–38, 1977.

[Diday *et al.*, 1974]Edwin Diday, A. Schroeder, and Y. Ok. The dynamic clusters method in pattern recognition. In *IFIP Congress*, pages 691–697, 1974.

[Diday, 2002]E. Diday. Mixture decomposition of distributions by copulas. In *Classification, Clustering and Data Analysis*, pages 297–310, 2002.

[Fraley and Raftery, 1999]C. Fraley and A. E. Raftery. Mclust: Software for model-based cluster analysis. Technical report, Department of Statistics, University of Washington, 1999.

[Fraley and Raftery, 2002]C. Fraley and A. E. Raftery. Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association*, 97:611–631, 2002.

[Mangasarian *et al.*, 1995]Olvi L. Mangasarian, W. Nick Street, and William H. Wolberg. Breast cancer diagnosis and prognosis via linear programming. *Operations Research*, 43:570–577, 1995.

[Nelsen, 1999]R.B. Nelsen. *An introduction to copulas.* Springer, London, 1999.

[Silverman, 1986]B. W. Silverman. *Density estimation for statistics and data analysis.* Chapman and Hall, London, 1986.

[Sklar, 1959]Abe Sklar. Fonctions de répartition à n dimensions et leurs marges. *Publications Statistiques Université de Paris*, 8:229–231, 1959.

[Vrac *et al.*, 2001]Mathieu Vrac, Edwin Diday, Alain Chédin, and Philippe Naveau. Mélange de distributions de distributions, décomposition de mélange de copules et application à la climatologie. In *Actes du VIIIème congrés de la Société Francophone de Classification*, pages 348–355, 2001.

Speeding up the self organizing map for dissimilarity data

Aïcha El Golli

Projet AxIS
INRIA-Rocquencourt,
Domaine De Voluceau, BP 105,
78153 Le Chesnay Cedex, France
(e-mail: aicha.elgolli@inria.fr)

Abstract. This paper presents an optimization of the self organizing map for dissimilarity data. In fact, practical clustering algorithms for dissimilarity data are extremely costly because of the calculation of the dissimilarity table and require multiple data scans to achieve convergence. Therefore, we propose working on sample set data to speed up the training process and also to handle large data set.

Keywords: Self organizing map, dissimilarity, random sampling.

1 Introduction

The self organizing map (SOM) [Kohonen, 1982a], [Kohonen, 1982b] and [Kohonen, 1997] is considered as a clustering method and also a projection method. It can be used at the same time both to reduce the amount data by clustering, and for projecting the data nonlinearly onto a lower dimensional display. Due to its unsupervised learning and topology preserving proprieties it has proven to be especially suitable in analysis of complex systems. The SOM algorithm implements a nonlinear topology preserving mapping from a high-dimensional input metric vector data space, $\mathbb{R}^p$, into a two-dimensional network or grid of neurons. To understand what the SOM really shows, it is important to understand that it actually performs two tasks: vector quantization and vector projection. Vector quantization creates from the original data a smaller, but still representative, data set to be worked with. The set of prototype vectors reflects the properties of the data space. The projection performed by the SOM is nonlinear and restricted to a regular grid (the map grid). The SOM tries to preserve the topology of the data space rather than relative distances.

The Kohonen's SOM is based on the notion of center of gravity and unfortunately, this concept is not applicable to many kinds of complex data. The extension of the self organizing map to dissimilarity data [El Golli *et al.*, 2004] is an alternative solution for new forms of complex data and so allows its process on dissimilarity measures rather than on raw data. With this alternative only the definition of a dissimilarity for each type of data is necessary to apply the method and so treat complex data. This extension is an adaptation of the batch-learning version of the SOM to dissimilarity data. At each stage, the learning is performed by alternating an assignment step and a representation step.

We focus on the problem of clustering large data set. In fact, when we work with this kind of data this extension of the SOM to complex data is extremely costly because of the calculation of the dissimilarity table. In order to solve this problem we propose to work on a sample set either on the whole learning set.

The paper is organized as follows: we first recall our adaptation of the SOM algorithm in its batch version for the dissimilarity data. Then we describe the algorithm working with a sample set.

2 Batch self-organizing map for dissimilarity data

The SOM can be considered as carrying out vector quantization and/or clustering while preserving the spatial ordering of the prototype vectors (also called referent vectors) in one or two dimensional output space. The SOM consists of neurons organized on a regular low-dimensional map. More formally, the map is described by a graph $(C, \mathit{\Gamma})$. C is a set of m interconnected neurons having a discrete topology defined by $\mathit{\Gamma}$.

For each pair of neurons (c, r) on the map, the distance $\delta(c, r)$, is defined as the shortest path between c and r on the graph. This distance imposes a neighborhood relation between neurons. The batch training algorithm is an iterative algorithm in which the whole data set (noted $\mathit{\Omega}$) is presented to the map before any adjustments are made. We note z_i an element of $\mathit{\Omega}$ and $\mathbf{z_i}$ the representation of this element in the space D called representation space of $\mathit{\Omega}$. In our case, the main difference with the classical batch algorithm is that the representation space is not $\mathbb{R}^p$ but an arbitrary set on which dissimilarity (denoted d) is defined.

Each neuron c is represented by a set $A_c = z_1, ..., z_q$ of elements of $\mathit{\Omega}$ with a fixed cardinality q, where z_i belongs to $\mathit{\Omega}$. A_c is called an individual referent. We denote A the set of all individual referents, i.e. the list $A = A_1, ..., A_m$. In our approach each neuron has a finite number of representations. We define a new adequacy function d^T from $\mathit{\Omega} \times P(\mathit{\Omega})$ to $\mathbb{R}^+$ by:

$$d^T(z_i, A_c) = \sum_{r \in C} K^T(\delta(r, c)) \sum_{z_j \in A_r} d^2(\mathbf{z_i}, \mathbf{z_j}) \tag{1}$$

d^T is based on the kernel positive function K. $K^T(\delta(c, r))$ is the neighborhood kernel around the neuron r. This function is such that $\lim\limits_{|\delta| \longrightarrow \infty} K(\delta) = 0$ and allows us to transform the sharp graph distance between two neurons on the map ($\delta(c, r)$) into a smooth distance. K is used to define a family of functions K^T parameterized by T, with $k^T(\delta) = K(\frac{\delta}{T})$. T is used to control the size of the neighborhood [Anouar *et al.*, 1997], [Dreyfus *et al.*, 2002]; when the parameter T is small, there are few neurons in the neighborhood. A simple example of K^T is defined by $K^T(\delta) = e^{-\frac{\delta^2}{T^2}}$.

During the learning, we minimize a cost function E by alternating an assignment step and a representation step. During the assignment step, the assignment function f assigns each individual z_i to the nearest neuron, here in terms of the function d^T:

$$f(z_i) = arg \min_{c \in C} d^T(z_i, A_c) \tag{2}$$

If there is equality, we assign the individual z_i to the neuron with the smallest label.

During the representation step, we have to find the new individual referents A^* that represent the set of observations in the best way in terms of the following cost function E:

$$E(f,A) = \sum_{z_i \in \Omega} d^T(z_i, A_{f(z_i)}) = \sum_{z_i \in \Omega} \sum_{r \in C} K^T(\delta(f(z_i), r)) \sum_{z_j \in A_r} d^2(\mathbf{z_i}, \mathbf{z_j}) \tag{3}$$

This function calculates the adequacy between the induced partition by the assignment function and the map referents A.

The criterion E is additive so this optimization step can be carried out independently for each neuron. Indeed, we minimize the m following functions:

$$E_r = \sum_{z_i \in \Omega} K^T(\delta(f(z_i), r)) \sum_{z_j \in A_r} d^2(\mathbf{z_i}, \mathbf{z_j}) \tag{4}$$

In the classical batch version, this minimization of E function is immediate because the positions of the referent vectors are the averages of the data samples weighted by the kernel function.
Here is the algorithm:

Initialization: iteration $k = 0$, choose an initial codebook A^0. Fix $T = T_{max}$ and the total number of iterations N_{iter}

Iteration: At iteration k, the set of individual referents of the previous iteration A^{k-1} is known. Calculate the new value of T:

$$T = T_{max} * (\frac{T_{min}}{T_{max}})^{\frac{k}{N_{iter-1}}}$$

▶ **affectation step:** up date the affectation function f_{A^k} associated to the A^{k-1} referent. Affecting each individual z_i to the referent as defined in equation (2).

▶ **representation step:** determine the new codebook A^{k*} that minimizes the $E(f_{A^k}, A)$ function (with respect to A) A_c^{k*} is defined from equation (4).

Repeat **Iteration** until $T = T_{min}$

3 Incorporating sampling

The extension of the SOM to dissimilarity data (DisSom) is a solution for different kind of complex data since we can define a dissimilarity but the computational complexity constitute a problem when we have a large data sets. In order to handle large data sets, we need an efficient mechanism for reducing the size of the learning set of the DisSom. One approach to achieving this is via random sampling ($S \subset \Omega$), the key idea is to apply DisSom's clustering algorithm to the new learning set S drawn from the data set rather than the entire data set. Typically, the random sample S will fit in main memory and will be much smaller than the original data set. Consequently, significant improvements in execution times for DisSom can be realized. When we choose a random samples S of moderate sizes we preserve information about the geometry of clusters fairly accurately, thus enabling DisSom to correctly cluster the input. We propose to use the algorithm [**?**] for drawing a sample randomly from data using constant space.

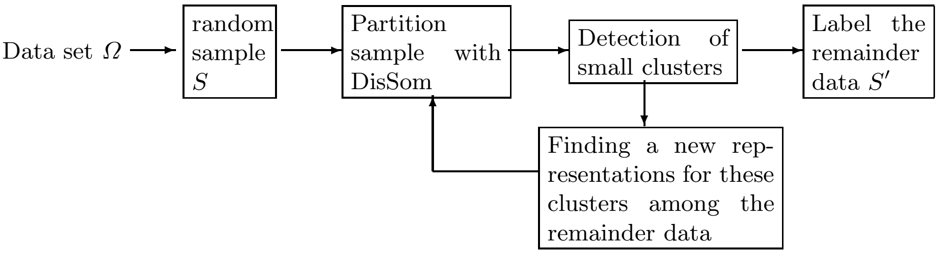

Data set Ω
random sample S
Partition sample with DisSom
Detection of small clusters
Label the remainder data S'
Finding a new representations for these clusters among the remainder data

Fig. 1. Overview of the steps of optimized DisSom

Once clustering of the random sample S is completed, the individual referent of each cluster is used to label the remainder of the data set (S'), where $\Omega = S \cup S'$ and $S \cap S' = \emptyset$. Each data point $z_i \in S'$ will be assigned to the closest individual referent of the map using the assignment function f (equation 2). But since we do not consider the entire data set, information about certain clusters may be missing in the input. As a result, our clustering algorithm may miss out certain clusters or incorrectly identify certain clusters. To this end, before labelling the remainder of the learning set we detect small clusters (c) and we have to find new individuals referent from the remainder data S' that minimizing the inertia criterion of the clusters c:

$$arg \min_{z_i \in S'} \sum_{z_j \in c} d^2(z_i, z_j)$$

To detect outliers we can use the Chernoff bounds. In fact, assuming that each cluster has a certain minimum size, we can use Chernoff bounds

[Motwani and Raghavan, 1995] to calculate the minimum sample size for which the sample contains, with high probability, at least a fraction *fr* of every clusters [Guha *et al.*, 1998].

The steps involved in clustering with DisSom are described in Figure 1. Since the learning set of the DisSom clustering algorithm is a set of randomly sampled points from original data set, the final k clusters involve only a subset of the entire set of points. In DisSom, the algorithm for assigning the appropriate cluster labels to the remaining data points employs a fraction of randomly selected individuals referent for each of the final k clusters. Each data point is assigned to the cluster containing the individual referent closest to it.

4 Conclusion

In this paper, we propose to speed up the self organizing map on dissimilarity data for large data sets. In fact, we propose to employ random sampling that allows to handle large data sets efficiently.

References

[Anouar *et al.*, 1997]F. Anouar, F. Badran, and S. Thiria. Self organized map, a probabilistic approach. 1997.

[Dreyfus *et al.*, 2002]G. Dreyfus, J.M. Martinez, M. Samuelides, M. Gordon, F. Badran, S. Thiria, and L. Hérault. *Réseaux de neurones méthodologie et applications*. Eyrolles, Paris, 2002.

[El Golli *et al.*, 2004]A. El Golli, B. Conan-Guez, and F. Rossi. a self-organizing map for dissimilarity data. In D. Banks, L. House, F. R. McMorris, P. Arabie, and W. Gaul, editors, *Classification, Clustering and Data Mining Application (Proceeding of IFCS)*, pages 61–68, Chicago, Illinois, 2004. Springer.

[Guha *et al.*, 1998]S. Guha, R. Rastogi, and K. Shim. Cure: An efficient clustering algorithm for large databases. In *In ACM SIGMOD Conf.*, pages 73–84, 1998.

[Kohonen, 1982a]T. Kohonen. Analysis of a simple self-organizing process. *Biol. cybern.*, 44:135–140, 1982.

[Kohonen, 1982b]T. Kohonen. Self-organized formation of topologically correct feature map. *Biol. Cybern*, 43:59–69, 1982.

[Kohonen, 1997]T. Kohonen. *Self-Organizing Maps*. Springer Verlag, New York, 1997.

[Motwani and Raghavan, 1995]R. Motwani and P. Raghavan. *Randomized algorithms*. In Cambridge university press, 1995.

Characterization of Galois closed sets using multiway dissimilarities

Jean Diatta

IREMIA
Université de La Réunion
15 avenue René Cassin - BP 7151,
97715 Saint-Denis messag cedex 9, France
(e-mail: jdiatta@univ-reunion.fr)

Abstract. We place ourselves in a so-called meet-closed description context; that is a context consisting of a finite nonempty entity set E whose elements are described in a complete meet-semilattice $\underline{D}$, by means of a descriptor δ. Then we consider multiway quasi-ultrametric dissimilarities on E, a class of multiway dissimilarities that, with their relative k-balls, extend the fundamental in classification bijection between ultrametric dissimilarities and indexed hierarchies. We also consider multiway dissimilarities agreeing with entity descriptions in a quite natural sense called δ-meet compatibility. It turns out that there exists an integer k such that any strictly δ-meet compatible k-way dissimilarity is quasi-ultrametric. On the other hand, the descriptor δ induces a Galois connection between the powerset $\mathcal{P}(E)$ and $\underline{D}$, which, in turn, induces a closure operator, say ϕ_δ, on $\mathcal{P}(E)$. then it is proved that nonempty ϕ_δ-closed subsets of E coincide with k-balls relative to some strictly δ-meet compatible multiway dissimilarities on E.

Keywords: Galois connection, Multiway dissimilarity, Closure operator, Description-meet compatibility, Quasi-ultrametric.

1 Introduction

Multiway dissimilarities are natural extensions of classical pairwise dissimilarities, that allow global comparison of more than two entities. In the last decade, they have been investigated or considered from different approaches in many works among which we just mention [Bandelt and Dress, 1994], [Joly and Le Calvé, 1995], [Daws, 1996] and [Bennani and Heiser, 1997]. In this paper, these approaches are extended onto the so-called meet-closed data description context. A meet-closed description context represents a finite entity set E using a complete meet-semilattice $\underline{D}$. Then we consider multiway quasi-ultrametric dissimilarities on E [Bandelt and Dress, 1994], [Diatta, 1997], a class of multiway dissimilarities that, with their relative k-balls, extend the fundamental in classification bijection between ultrametric dissimilarities and indexed hierarchies [Johnson, 1967]. We also consider multiway dissimilarities agreeing with entity descriptions in a quite natural sense called δ-meet compatibility. It turns out that there exists an integer k such that any strictly δ-meet compatible k-way dissimilarity is quasi-ultrametric.

On the other hand, any descriptor δ induces a Galois connection between the powerset $\mathcal{P}(E)$ and $\underline{D}$, which, in turn, induces a closure operator, say ϕ_δ, on $\mathcal{P}(E)$ [Birkhoff, 1967]. It is proved that nonempty ϕ_δ-closed subsets of E are the k-balls of some strictly δ-meet compatible multiway dissimilarities on E.

2 Multiway dissimilarities

Before introducing multiway dissimilarities, let us first recall the classical pairwise ones. Let E be a finite nonempty set.

A (pairwise) dissimilarity on E is a map $d : E \times E \to \mathbb{R}$ satisfying reflexivity ((R2) $d(x,x) = 0$), non-negativity ((N2) $d(x,y) \geq 0$) and symmetry ((S2) $d(x,y) = d(y,x)$).

Considering maps on $E^3, E^4, \dots, E^k$, with similar properties, naturally leads to the notion of 3-way, 4-way,..., k-way dissimilarity. For instance, a 3-way dissimilarity on E will be any map $d : E^3 \to \mathbb{R}$ satisfying: (R3) $d(x,x,x) = 0$, (N3) $d(x,y,z) \geq 0$ and (S3) $d(x,y,z) = d(x,z,y) = d(y,x,z) = d(y,z,x) = d(z,x,y) = d(z,y,x)$. The term *multiway* dissimilarity will be used to mean a k-way dissimilarity, for some $k \geq 2$.

Of course, due to the tuple-based definition above, the complexity of expressions related to k-way dissimilarities increases when k grows. Then, for the sake of simplicity, we adopt in the present paper a set-based definition based on the following observation: according to (R2) and (N2), $d(x,x) \leq d(x,y)$ for all x, y. Thus, a dissimilarity on E can be defined as being a nonnegative real valued map d on the set of singletons and pairs of E, satisfying $d(\{x\}) = 0$ and $d(\{x\}) \leq d(\{x,y\})$. This set-based definition makes the symmetry condition implicit. Moreover, for $k \geq 2$, its generalization to k-way dissimilarities involves shortest expressions.

For reasons explained in Remark 3 below, we will drop out the reflexivity condition and thus be rather concerned with so-called (multiway) pseudo-dissimilarities. However, we will still use the term dissimilarity, keeping in mind that the condition $d(\{x\}) = 0$ is not required.

For any set S and any integer $k \geq 1$, $S^*_{\leq k}$ will denote the set of all nonempty subsets of S with at most k elements. Then we formally define multiway dissimilarities as follows.

Definition 1 *A* k-way dissimilarity *on* E *will be any nonnegative real valued and isotone map defined on the set of all nonempty subsets of* E *with at most* k *elements, i.e., any map* $d : E^*_{\leq k} \to \mathbb{R}_+$ *such that* $d(X) \leq d(Y)$ *when* $X \subseteq Y$*.*

Example 1 *Table 2 presents a dataset, say* $\mathcal{D}$*, about seven market baskets and five items: bread (brd), butter (btr), cheese (chs), eggs (egg), milk (mlk); for instance, the market basket labeled 1 contains bread and cheese. For any* k *such that* $2 \leq k \leq 5$*, a* k*-way dissimilarity on the set of items, can be* *defined by letting* $\mathrm{dis}_k(X)$ *be seven minus the number of baskets that contain each of the items in* X*. Then, for instance,* $\mathrm{dis}_3(\{brd, chs\}) = 4$ *and* $\mathrm{dis}_3(\{brd, btr, chs\}) = 7$*.*

	brd	btr	chs	egg	mlk
1	x		x		
2		x	x		x
3	x		x	x	
4		x		x	x
5	x	x			x
6	x	x		x	
7	x		x		x

Table 1. Example dataset

Remark 1 *For* $\{x, y, z\} \subseteq E$*, we will simply write* $d(x)$ *or* $d(x, y)$ *or* $d(x, y, z)$ *instead of* $d(\{x\})$ *or* $d(\{x, y\})$ *or* $d(\{x, y, z\})$*, respectively. Moreover, as in the tuple-based setting, the notation* $d(x, y)$ *or* $d(x, y, z)$ *will not require* x*,* y *and* z *be distinct.*

3 Quasi-ultrametric multiway dissimilarities

Key notions in the definition of quasi-ultrametrics given below are those of a d-ball, (d, k)-ball and d-diameter, where d is a k-way dissimilarity. To catch their meaning, let us first cast them in the case of a pairwise dissimilarity, say d_2.

The d_2-diameter of a nonempty subset Z of E is the maximum d_2-dissimilarity between elements of Z, i.e.: $\mathrm{diam}_{d_2}(Z) = \max\{d_2(x, y) : x, y \in Z\}$.

Let now x and y be two distinct elements of E and r a nonnegative real number. The d_2-ball of center x and radius r is the set $B^{d_2}(x, r)$ of elements of E whose d_2-dissimilarity degree from x is at most r, i.e., formally, $B^{d_2}(x, r) = \{z \in E : d_2(x, z) \leq r\}$; the $(d_2, 2)$-ball generated by $\{x\}$ is the set $B^{d_2}_x = B^{d_2}(x, d_2(x))$, and the $(d_2, 2)$-ball generated by $\{x, y\}$ is the set $B^{d_2}_{xy} = B^{d_2}(x, d_2(x, y)) \cap B^{d_2}(y, d_2(x, y))$. If $x = y$, $B^{d_2}_{xy} = B^{d_2}_x$.

All these notions have been naturally generalized to multiway dissimilarities in [Diatta, 1997]. For $k \geq 2$, let d_k denote a k-way dissimilarity on E.

The d_k*-diameter* (or, simply, *diameter*) of a nonempty subset Z of E is the maximum d_k-dissimilarity degree between elements of Z, i.e.: $\mathrm{diam}_{d_k}(Z) = \max\{d_k(T) : T \in Z^*_{\leq k}\}$.

Let $X \in E^*_{\leq k-1}$. The d_k*-ball* (or, simply, *ball*) of center X and radius r is the set $B^{d_k}(X, r)$ defined by $B^{d_k}(X, r) = \{y \in E : d_k(X \cup \{y\}) \leq r\}$. If

$X \in E^*_{\leq k}$, then the (d_k, k)-*ball* (or, simply, k-*ball* relative to d_k) generated by X is the set $B_X^{d_k}$ defined by $B_X^{d_k} = B^{d_k}(X, d_k(X))$ when $|X| \leq k-1$, and $B_X^{d_k} = \bigcap_{x \in X} B^{d_k}(X \setminus \{x\}, d_k(X))$ otherwise. The superscript d_k may be omitted if there is no risk of confusion.

Before defining quasi-ultrametrics, let us recall a well-known particular case of them, namely ultrametrics. A (2-way) dissimilarity d_2 is said to be *ultrametric* if for all x, y, z:

$$d_2(x,y) \leq \max\{d_2(x,z), d_2(y,z)\}.$$

Next are some characterizations of ultrametric 2-way dissimilarities, which may help in understanding the definition of quasi-ultrametrics given below.

Proposition 1 *[Diatta and Fichet, 1998] For a* 2*-way dissimilarity d_2 on E, the following assertions are equivalent.*

(i) *d_2 is ultrametric.*
(ii) *for all x, y, z: the greatest two values among $d_2(x,y)$, $d_2(x,z)$ and $d_2(y,z)$ are equal.*
(iii) *for all x, y:* $\mathrm{diam}_{d_2}(B(x, d_2(x,y))) = d_2(x,y)$ *(diameter condition).*
(iv) *for all x, y, u, v: $u, v \in B(x, d_2(x,y))$ implies $B(u, d_2(u,v)) \subseteq B(x, d_2(x,y))$ (inclusion condition).*

Example 2 *Figure 1 presents three dissimilarities d_1, d_1' and d_1'' on the set $\{i,j,k,l\}$. It is easily checked that d_1 satisfies the diameter condition; but d_1 does not satisfy the inclusion condition because $j, k \in B_{jl}^{d_1}$ whereas $i \in B_{jk}^{d_1}$ and $i \notin B_{jl}^{d_1}$. It is also easily checked that d_1' satisfies the inclusion; but d_1' does not satisfy the diameter condition because $i, j \in B_{kl}^{d_1'}$ so that* $\mathrm{diam}_{d_1'}(B_{kl}^{d_1'}) > d_1'(k,l)$*. The dissimilarity d_1'' is clearly quasi-ultrametric since $B_i^{d_1''} = B_j^{d_1''} = B_{ij}^{d_1''} = \{i,j\}$, for $x \neq i,j$, $B_x^{d_1''} = \{x\}$, and for $\{x,y\} \neq \{i,j\}$, $B_{xy}^{d_1''} = \{i,j,k,l\}$.*

i	0			
j	1	0		
k	1	1	0	
l	3	2	1	0
	i	j	k	l

d_1

i	0			
j	3	0		
k	1	1	0	
l	1	1	2	0
	i	j	k	l

d_1'

i	0			
j	0	0		
k	1	1	0	
l	1	1	1	0
	i	j	k	l

d_1''

Fig. 1. Three pairwise dissimilarities on the set $\{i,j,k,l\}$: d_1 satisfies the diameter but not the inclusion condition; d_1' satisfies the inclusion but not the diameter condition; d_1'' is quasi-ultrametric.

Conditions (iii) and (iv) of Proposition 1 above can be extended to the case of multiway dissimilarities by replacing balls with k-balls. The two extended conditions define what we call the quasi-ultrametric multiway dissimilarities [Diatta, 1997]:

Definition 2 *A k-way dissimilarity d_k on E is said to*

(i) satisfy the inclusion condition *if for all $X, Y \in E^*_{\leq k}$, $Y \subseteq B^{d_k}_X$ implies $B^{d_k}_Y \subseteq B^{d_k}_X$;*

(ii) satisfy the diameter condition *if for all $X \in E^*_{\leq k}$, $\mathrm{diam}_{d_k}(B^{d_k}_X) = d_k(X)$;*

(iii) be quasi-ultrametric *if it satisfies both of the inclusion and the diameter conditions.*

Example 3 *The reader may check that the* 3*-way dissimilarity* dis_3 *defined in Example 1 is quasi-ultrametric. This can also be derived from Theorem 1 below (see Remark 4).*

4 Description-meet compatibility

In this section, we place ourselves in a so-called *meet-closed description context.* That is a context consisting of a finite nonempty entity set E whose elements are described in a complete meet-semilattice $\underline{D}$, by means of a descriptor δ. We will denote such a context as a triple $\mathbb{K} = (E, \underline{D}, \delta)$ where E stands for the entity set, $\underline{D} := (D, \leq)$ the entity description space, and δ the descriptor that associates to each entity $x \in E$ its description $\delta(x)$ in $\underline{D}$.

In all what follows, E will denote a finite nonempty entity set, $\underline{D}$ a complete meet-semilattice, δ a descriptor that maps E into $\underline{D}$, and $\mathbb{K}$ the meet-closed description context $(E, \underline{D}, \delta)$.

Example 4 *Consider Table 4 presenting five visitors of a given Web site, described by three attributes:* LiLo, NoLi, ReSu*, where* $\mathrm{LiLo}(x)$ *is the login-logout time interval of visitor x within the interval* $[0, 24]$*,* $\mathrm{NoLi}(x)$ *is the number of times visitor x logs in at* $\mathrm{LiLo}(x)$ *interval during a given fixed period, and* $\mathrm{ReSu}(x)$ *is the subjects requested by x during a session; requested subjects are sets of subjects from: Arts & Humanities (AH), Business & Economy (BE), Computers & Internet (CI), News & Media (NM), Recreation & Sports (RS), Science & Health (SH), Society & Culture (SC).*
Then Table 4 can be seen as representing a meet-closed description context $\mathbb{K}_2 := (E_2, \underline{D}_2, \delta_2)$ *where E_2 is the set* $\{1, 2, 3, 4, 5\}$*, $\underline{D}_2$ the direct product of three partially ordered sets (posets): the set* $(\mathrm{FUCI}([0, 24]), \subseteq)$ *of finite unions of closed intervals of* $[0, 24]$ *endowed with the set inclusion order, the set* $([\![30; 40]\!], \leq)$ *of integers from 30 to 40, endowed with the integer usual order, and the powerset $(\mathcal{P}(S), \subseteq)$ of the set $S = \{AH, BE, CI, NM, RS, SC\}$, endowed with the set inclusion order, and* $\delta_2(x) = (\mathrm{LiLo}(x), \mathrm{NoLi}(x), \mathrm{ReSu}(x))$.

	LiLo	NoLi	ReSu
1	0-2	30	CI,RS
2	21-24	35	AH,NM,SC
3	0-3	40	AH,BE,CI,RS
4	22-24	35	AH,SC
5	12-14	30	BE,NM

Table 2. Example meet-closed description context

The description-meet compatibility defined below has been introduced in [Diatta and Ralambondrainy, 2002] in the case of pairwise dissimilarities. It uses the notion of valuation on a poset.

A *valuation* on a poset $(P, \leq)$ is a map $h : P \to \mathbb{R}_+$ such that $h(x) \leq h(y)$ when $x \leq y$. A *strict* valuation will then be a valuation h such that $x < y$ implies $h(x) < h(y)$.

Before defining the description-meet compatibility, let us introduce a further notation: for any $X \subseteq E$, $\delta(X)$ will denote the set of descriptions of entities belonging to X.

A multiway dissimilarity d on E will be said to be *δ-meet compatible* if there exists a valuation h on $\underline{D}$ with which it is δ-meet compatible, i.e., such that

$$d(X) \leq d(Y) \iff h(\inf \delta(X)) \geq h(\inf \delta(Y)),$$

for $X, Y \subseteq E$. If h is a strict valuation, d will be said to be *strictly* δ-meet compatible.

Remark 2 *The reader may observe that when $\underline{D}$ is a complete join-semilattice, a dual compatibility condition, say δ-join compatibility, can be defined by reversing the right-hand side inequality in the above equivalence and replacing meets by joins.*

Description-meet compatibility is a kind of natural agreement expressing the following fact: the lower the meet of descriptions of entities in X, the larger the dissimilarity degree of X.

Remark 3 *If d is a strictly δ-meet compatible (multiway) dissimilarity, then $\delta(x) < \delta(y)$ implies $d(y) < d(x)$. This is why we drop out the condition $d(x) = 0$, since it is very likely to happen that two entities x and y satisfy $\delta(x) < \delta(y)$.*

Example 5 *Consider the meet-closed description context $\mathbb{K}_2$ defined in Example 4. Define a multiway dissimilarity on E_2 by*

$$\mathrm{dis}'(X) = 47 - (\lambda(\cap_{x \in X} \mathrm{LiLo}(x)) + \min_{x \in X} \mathrm{NoLi}(x) + |\cap_{x \in X} \mathrm{ReSu}(x)|),$$

where $\lambda([\alpha, \beta]) = \beta - \alpha$. For instance, $\mathrm{dis}'(1, 2, 3) = 47 - (\lambda([0, 2] \cap [21, 24] \cap [0, 3]) + \min\{30, 35, 40\} + |\{CI, RS\} \cap \{AH, NM, SC\} \cap \{AH, BE, CI, RS\}|) =$

$47 - (\lambda(\varnothing) + 30 + |\varnothing|) = 47 - (0 + 30 + 0) = 17$. *Then* dis' *is strictly* δ_2*-meet compatible. Indeed,* λ, $x \mapsto x$ *and* $Y \mapsto |Y|$ *are strict valuations on* $(\mathrm{FUCI}([0,24]), \subseteq)$, $([\![30; 40]\!], \leq)$ *and* $(\mathcal{P}, \subseteq)$*, respectively. Thus* h_2 *defined by*

$$h_2(u, v, w) = \lambda(u) + v + |w|$$

is a strict valuation on $\underline{D}_2$*, and the fact that* dis' *is* δ_2*-meet compatible with* h_2 *follows from the fact that* $\mathrm{dis}'(X)$ *is decreasing w.r.t.* $h_2(\inf \delta_2(X))$*.*

Before outlining the relationship between quasi-ultrametricity and description-meet compatibility, let us recall the following technical notion: the *breadth* of a meet-semilattice $(P, \leq)$ is the least positive integer k such that the meet of any $(k+1)$ elements of P is always the meet of k elements among these $k+1$ [Birkhoff, 1967]. Having noticed this, we agree to say that a subset Q of a meet-semilattice is of breadth k if k is the least positive integer such that for any $(k+1)$-element subset W of Q there is $w \in W$ such that $\inf(W \setminus \{w\}) \leq w$.

Example 6 *Consider the dataset* $\mathcal{D}$ *given in Table 2 as presenting a meet-closed description context* $\mathbb{K}_1 := (E_1, \underline{D}_1, \delta_1)$*, where* E_1 *is the set of five items and* $\underline{D}_1$ *the boolean lattice* $\{0,1\}^7$*; for instance* $\delta_1(brd) = (1,0,1,0,1,1,1)$*. Then* $\delta_1(E_1)$ *is of breadth at least* 3 *since*

$$\inf \delta_1(\{brd, chs, mlk\}) = (0,0,0,0,0,0,1),$$

which is different from either of $\delta_1(brd) \wedge \delta_1(chs) = (1,0,1,0,0,0,1)$, $\delta_1(brd) \wedge \delta_1(mlk) = (0,0,0,0,1,0,1)$ *and* $\delta_1(chs) \wedge \delta_1(mlk) = (0,1,0,0,0,0,1)$*. Moreover,*

$$\begin{aligned}\inf \delta_1(\{brd, btr, chs, egg\}) &= \inf \delta_1(\{brd, btr, chs, mlk\}) \\ &= \inf \delta_1(\{brd, btr, chs\}),\end{aligned}$$

$$\begin{aligned}\inf \delta_1(\{brd, btr, egg, mlk\}) &= \inf \delta_1(\{brd, chs, egg, mlk\}) \\ &= \inf \delta_1(\{brd, egg, mlk\}),\end{aligned}$$

and $\inf \delta_1(\{btr, chs, egg, mlk\}) = \inf \delta_1(\{btr, chs, egg\})$*, so that* $\delta_1(E_1)$ *is of breadth* 3*.*

We now go on stating the result showing the existence of an integer $k \geq 2$ such that any strictly δ-meet compatible k-way dissimilarity on E is quasi-ultrametric.

Theorem 1 *(i) If* $\delta(E)$ *is of breadth one, then every strictly* δ*-meet compatible* 2*-way dissimilarity on* E *is ultrametric.*

(ii) If $\delta(E)$ *is of breadth* $k \geq 2$*, then every strictly* δ*-meet compatible* k*-way dissimilarity on* E *is quasi-ultrametric.*

The converse of Theorem 1 does clearly not hold since, for $k \geq 2$, every constant k-way dissimilarity on E is quasi-ultrametric but never strictly δ-meet compatible, regardless of the descriptor δ. Indeed, otherwise, we would have, for all $x, y \in E$, $\delta(x) = \delta(y)$ so that $\delta(E)$ would be a singleton, hence of breadth one.

Remark 4 *As claimed in Example 3, it follows from Theorem 1 that the* 3*-way dissimilarity* dis_3 *defined in Example 1 is quasi-ultrametric. Indeed, on the one hand, as observed in Example 6,* $\delta_1(E_1)$ *is of breadth* 3*. On the other hand, for each* k *such that* $2 \leq k \leq 5$*,* dis_k *is strictly* δ_1*-meet compatible with the valuation* h_1 *defined on* $\underline{D}_1$ *by letting* $h_1(x)$ *be the number of ones occurring in* x*.*

The entity set E being finite, there is an integer $k \geq 1$ such that k is the breadth of $\delta(E)$. Moreover, as any pairwise ultrametric dissimilarity is quasi-ultrametric, we derive the following from Theorem 1.

Corollary 1 *There is an integer* $k \geq 2$ *such that any strictly* δ*-meet compatible* k*-way dissimilarity on* E *is quasi-ultrametric.*

Following [Diatta, 1997], a k-way dissimilarity d will be said to be ultrametric if for all $X \in E^*_{\leq k}$ and $x \in E$:

$$d(X) \leq \max_{Y \in X^*_{\leq k-1}} d(Y + x).$$

When $\delta(E)$ is of breadth one, Theorem 1 (i) extends to ultrametric multiway dissimilarities:

Theorem 2 *If* $\delta(E)$ *is of breadth one, then for* $k \geq 2$*, every strictly* δ*-meet compatible* k*-way dissimilarity on* E *is ultrametric.*

5 Characterization of Galois closed entity sets

Given the meet-closed description context $\mathbb{K} = (E, \underline{D}, \delta)$, the descriptor δ induces a Galois connection between $(\mathcal{P}(E), \subseteq)$ and $\underline{D}$ by means of the maps

$$f : X \mapsto \inf \{\delta(x) \,:\, x \in X\}$$

and

$$g : I \mapsto \{x \in E \,:\, I \leq \delta(x)\},$$

for $X \subseteq E$ and $I \in \underline{D}$. Then it is well known that, in these conditions, the map $\phi_\delta := g \circ f$ is a closure operator on $\mathcal{P}(E)$ [Birkhoff, 1967]. That is ϕ_δ is *extensive* ($X \subseteq \phi_\delta(X)$), *isotone* ($X \subseteq Y$ implies $\phi_\delta(X) \subseteq \phi_\delta(Y)$) and *idempotent* ($\phi_\delta(\phi_\delta(X)) = \phi_\delta(X)$). A subset X of E is said to be ϕ_δ-*closed* (or a *Galois closed entity set* of $\mathbb{K}$, relative to ϕ_δ) when $\phi_\delta(X) = X$.

Galois closed entity sets play an important role in classification because they provide easy-to-interpret clusters [Domenach and Leclerc, 2002]. Indeed, if X is a Galois closed entity set, then $f(X)$ is the description of X.

When $\underline{D}$ is a complete join-semilattice, the descriptor δ induces a Galois connection between $(\mathcal{P}(E), \subseteq)$ and the order-dual of $\underline{D}$ by means of the maps

$$f^{\partial} : X \mapsto \sup \{\delta(x) \,:\, x \in X\}$$

and

$$g^{\partial} : I \mapsto \{x \in E \,:\, I \geq \delta(x)\},$$

for $X \subseteq E$ and $I \in \underline{D}$. Similarly, this Galois connection induces the closure operator $\phi_{\delta}^{\partial} := g^{\partial} \circ f^{\partial}$ on $\mathcal{P}(E)$. Galois closed entity sets relative to ϕ_{δ}^{∂} have been considered in the framework of symbolic data analysis [Bock and Diday, 2000].

Example 7 *Consider the meet-closed description context $\mathbb{K}_2$ given in Example 4. The pair $\{1,3\}$ is ϕ_δ-closed; but $\{1,2,3\}$ is not ϕ_δ-closed because* $\inf \delta_2(\{1,2,3\}) = (\emptyset, 30, \emptyset) \leq \delta_2(4)$. *On the other hand, the pair $\{4,5\}$ is ϕ_{δ}^{∂}-closed; but $\{1,2,3\}$ is not ϕ_{δ}^{∂}-closed because*

$$\delta_2(4) \leq \sup \delta_2(\{1,2,3\}) = ([0,3] \cup [21,24], 40, \{AH, BE, CI, NM, RS, SC\}).$$

The following proposition shows that the ϕ_δ-closure of any subset $X \subseteq E$ is contained in a ball centered in a subset of X and relative to some δ-meet compatible multiway dissimilarity.

Proposition 1 *Let d be a δ-meet compatible k-way dissimilarity measure on E and let $X \in E^*_{\leq k}$. Then for all $Y \in X^*_{\leq k-1}$ and all $y \in B^d(Y, d(X))$, $\phi_\delta(Y + y) \subseteq B^d(Y, d(X))$. Moreover, $\phi_\delta(X) \subseteq B^d(Y, d(X))$.*

The next proposition gives a necessary and sufficient condition for the ϕ_δ-closure of an entity subset X to be a ball (resp. k-ball) relative to some δ-meet compatible multiway dissimilarity.

Proposition 2 *Let d be a δ-meet compatible k-way dissimilarity measure on E. Then, for all $X \in E^*_{\leq k}$ and all $Y \in X^*_{\leq k-1}$:*

(i) $\phi_\delta(X) = B^d(Y, d(X))$ if and only if $\inf \delta(B^d(Y, d(X))) = \inf \delta(X)$.
(ii) $\phi_\delta(X) = B^d_X$ if and only if $\inf \delta(B^d_X) = \inf \delta(X)$.

We now go on stating the result showing the coincidence between nonempty Galois closed entity sets of a meet-closed description context and k-balls relative to some strictly description-meet compatible multiway dissimilarity.

Theorem 3 *For an integer $p \geq 2$, let d_p be a strictly δ-meet compatible p-way dissimilarity on E.*

(i) If $\delta(E)$ is of breadth one, the set ${\mathcal{F}_\delta}^*$ of nonempty ϕ_δ-closed subsets of E coincides with the set of $(d_2, 2)$-balls generated by singletons of E.

(ii) If $\delta(E)$ is of breadth $k \geq 2$, then ${\mathcal{F}_\delta}^*$ coincides with the set of (d_k, k)-balls.

Finally, as observed above, E being finite, there is an integer $k \geq 1$ such that k is the breadth of $\delta(E)$. Moreover, as any pairwise ultrametric dissimilarity is quasi-ultrametric, we derive the following from theorems 1 and 3.

Corollary 2 *There is an integer $k \geq 2$ such that nonempty Galois closed entity subsets of E coincide with k-balls relative to some k-way quasi-ultrametric dissimilarity on E.*

It may be noticed that, when $\underline{D}$ is a complete join-semilattice, similar results hold for Galois closed entity sets relative to ϕ_δ^∂, using δ-join compatible multiway dissimilarities.

Acknowledgements: The author is grateful to Boris Mirkin for helpful advise.

References

[Bandelt and Dress, 1994]H-J. Bandelt and A. W. M. Dress. An order theoretic framework for overlapping clustering. *Discrete Mathematics*, 136:21–37, 1994.

[Bennani and Heiser, 1997]M. Bennani and W. J. Heiser. Triadic distance models: axiomatization and least squares representation. *J. Math. Psychology*, 41:189–206, 1997.

[Birkhoff, 1967]G. Birkhoff. *Lattice theory.* 3rd edition, Coll. Publ., XXV. American Mathematical Society, Providence, RI, 1967.

[Bock and Diday, 2000]H.H. Bock and E. Diday, editors. *Analysis of Symbolic Data.* Springer-Verlag, 2000.

[Daws, 1996]J. T. Daws. The analysis of free-sorting data : beyond pairwise cooccurrences. *J. Classification*, 13:57–80, 1996.

[Diatta and Fichet, 1998]J. Diatta and B. Fichet. Quasi-ultrametrics and their 2-ball hypergraphs. *Discrete Mathematics*, 192:87–102, 1998.

[Diatta and Ralambondrainy, 2002]J. Diatta and H. Ralambondrainy. The conceptual weak hierarchy associated with a dissimilarity measure. *Mathematical Social Sciences*, 44:301–319, 2002.

[Diatta, 1997]J. Diatta. Dissimilarités multivoies et généralisations d'hypergraphes sans triangles. *Math. Inf. Sci. hum.*, 138:57–73, 1997.

[Domenach and Leclerc, 2002]F. Domenach and B. Leclerc. On the roles of Galois connections in classification. In O. Opitz M. Schwaiger, editor, *Explanatory Data Analysis in Empirical Research*, pages 31–40. Springer, 2002.

[Johnson, 1967]S. C. Johnson. Hierarchical clustering schemes. *Psychometrika*, 32:241–254, 1967.

[Joly and Le Calvé, 1995]S. Joly and G. Le Calvé. Three-way distances. *J. Classification*, 12:191–205, 1995.

One-mode Additive Clustering of Multiway Data

Dirk Depril and Iven Van Mechelen

KULeuven
Tiensestraat 103
3000 Leuven, Belgium
(e-mail: dirk.depril@psy.kuleuven.ac.be
iven.vanmechelen@psy.kuleuven.ac.be)

Abstract. Given a multiway data set, in several contexts it may be desirable to obtain an overlapping clustering of one of the modes implied by the data. For this purpose a one-mode additive clustering model has been proposed, which implies a decomposition of the data into a binary membership matrix and a real-valued centroid matrix. To fit this model to a data set, a least squares loss function can be minimized. This can be achieved by means of a sequential fitting algorithm (SEFIT) as developed by Mirkin. In this presentation we will propose a new algorithm for the same model, based on an alternating least squares approach.

Keywords: Additive Clustering, Approximation Clustering, ALS algorithm.

1 Introduction

N-way N-mode data often show up in statistical practice. The simplest instance is a two-way two-mode data set. In this paper we will focus on the latter type of data, but everything can easily be extended to the N-way case.

For two-mode two-way data sets a one-mode additive clustering model has been described by several authors, including [Mirkin, 1996]. The aim of the associated data analysis is to fit this model to a data set under study (either in a least squares or in a maximum likelihood classification sense). For this purpose [Mirkin, 1990] proposed a sequential fitting (SEFIT) algorithm. However, at this moment not much information is available about its performance; moreover, as will be discussed below, this algorithm implies some conceptual problems. As a possible way out, in this paper we will present a new algorithm to estimate the same model.

The remainder of this paper is then structured as follows. In Section 2 we will describe the one-mode additive clustering model and in Section 3 we will explain the aim of the associated data analysis. In Section 4 the SEFIT algorithm will be explained and in Section 5 we will present our new algorithm. In Section 6 we present a few concluding remarks.

2 Model

In one-mode additive clustering the data matrix X is approximated by a model matrix. This model matrix M with entries m_{ij} ($i = 1, \ldots, I$, $j = 1, \ldots, J$) can be decomposed as

$$m_{ij} = \sum_{r=1}^{R} a_{ir} g_{rj}, \tag{1}$$

with R denoting total number of clusters, with a_{ir} taking values 1 or 0 and with g_{rj} real-valued. A is called the *cluster membership matrix* with entries a_{ir} indicating whether entity i belongs to cluster r ($a_{ir} = 1$) or not ($a_{ir} = 0$). One may note that apart from the binary nature, no further restrictions are imposed on the values a_{ir}, implying that the resulting clustering may be an overlapping one. The vector $\mathbf{g}_r = (g_{rj})_{j=1}^{J}$ is called the *centroid* of cluster r and the entire matrix G with entries g_{rj} is called the *centroid matrix*. Equation (1) then means that the ith row of M is obtained by summing up the centroids of the clusters to which row i belongs. Note that (1) can also be written in matrix form as

$$M = AG. \tag{2}$$

In the past, this model has been described in [Mirkin, 1990] and [Mirkin, 1996].

To illustrate the conceptual meaningfulness of the one-mode additive clustering model we may refer to the following hypothetical medical example. Consider a patients by symptom data matrix, the entries of which indicate the extent to which each of a set of patients suffers from each of a set of symptoms. In such a context, symptom strength may be attributed to underlying diseases or syndromes, that correspond to clusters of patients. Given that patients might suffer from more than one syndrome (a phenomenon called syndrome co-morbidity), in such a case an overlapping patient clustering is justified. The measured values of symptom strength can be considered additive combinations of the underlying syndrome profiles formalized by the rows of the centroid matrix G of the additive clustering model (1).

3 Aim of the data analysis

A two-way two-mode data matrix X resulting from a real experiment can always be represented by the model in (1). However, in most cases, a large number of clusters R will be needed for this. Therefore one usually looks for a model with a small value for R that fits the data well in some way.

A first way to do this is a deterministic one. In that case one assumes that $X \approx M$ and the goal of the data analysis is then to find the model M with R clusters that optimally approximates the data X according to some loss function. In this paper, the quality of the approximation will be expressed in terms of a least squares loss function:

$$L^2 = \sum_{ij}(x_{ij} - \sum_{r=1}^{R} a_{ir} g_{rj})^2, \tag{3}$$

which needs to be minimized with respect to the unknown a_{ir} and g_{rj} ($i = 1, \ldots, I$, $r = 1, \ldots, R$, $j = 1, \ldots, J$). Note that, if the matrix A is given, then the optimal G according to (3) is the least squares multiple regression estimator $(A'A)^{-1}A'X$. Note that this implies, since we have only 2^{IR} possible binary matrices A, that the solution space of (3) is finite and that therefore in principle it is possible to find the global minimum enumeratively. However, as computation time is an exponential function of the size of the data matrix, an enumerative search will quickly become infeasible. Therefore, in practice suitable algorithms or heuristics need to be developed to find the global optimum of (3).

A second approach to the data analysis is of a stochastic nature. We now assume that:

$$x_{ij} = \sum_{r=1}^{R} a_{ir} g_{rj} + e_{ij}, \tag{4}$$

where e_{ij} is an error term with $e_{ij} \overset{iid}{\text{sim}} N(0, \sigma^2)$. The goal of the data analysis then is to estimate the a_{ir}, g_{rj} and σ that maximize the log-likelihood:

$$\begin{aligned} \log \ell &= \sum_{ij} \log f(x_{ij}|A, G, \sigma) \\ &= -IJ \log \sqrt{2\pi} - IJ \log \sigma - \frac{\sum_{ij}(x_{ij} - \sum_{r=1}^{R} a_{ir} g_{rj})^2}{2\sigma^2} \end{aligned} \tag{5}$$

This can be characterized as a classification likelihood problem. In the latter type of problem, the binary entries of A are considered fixed parameters that need to be estimated, rather than realisations of latent random variables as in mixture-like models. For the estimation of a_{ir} and g_{rj} we need to minimize the sum in the numerator of the right most term in (5). This means that the stochastic approach for estimating the memberships and centroids is fully equivalent to the deterministic approach as explained above. For the estimation of σ^2 we have

$$\hat{\sigma}^2 = \frac{\sum_{ij}(x_{ij} - \sum_{r=1}^{R} \hat{a}_{ir} \hat{g}_{rj})^2}{IJ}, \tag{6}$$

where $\hat{a}_{ir}$ and $\hat{g}_{rj}$ are the maximum likelihood estimators of a_{ir} and g_{rj} respectively.

4 SEFIT

As explained in the previous section, the minimization of the loss function (3) requires suitable algorithms. In this section we will explain a first such algorithm that has been developed by [Mirkin, 1990] and that is a sequentially fitting (SEFIT) algorithm. In this algorithm the membership matrix A is estimated column-by-column meaning that one sequentially looks for new clusters. Suppose $m-1$ clusters have already been found, the mth cluster is then estimated by making use of the residual data

$$x_{ij}^{m} = x_{ij} - \sum_{r<m} a_{ir} g_{rj} \tag{7}$$

and by minimizing the function

$$\sum_{ij} (x_{ij}^{m} - a_{im} g_{mj})^2. \tag{8}$$

Given the memberships a_{im} $(i = 1, \ldots, I)$ the least squares estimates for the centroid values g_{mj} $(j = 1, \ldots, R)$ are given by

$$g_{mj} = \frac{\sum_{i=1}^{I} a_{im} x_{ij}}{\sum_{i=1}^{I} a_{im}^2}, \tag{9}$$

which is the simple mean of the elements in the cluster.

The estimation of the memberships a_{im} $(i = 1, \ldots, I)$ proceeds as follows. We start with a zero memberships column (i.e., an empty cluster) and sequentially add elements of the first mode to the cluster in a greedy way, that is, add that row that yields the biggest decrease in the loss function (8), and continue until no further decrease is obtained.

A full loop of the algorithm then goes as follows. First estimate the memberships a_{im} $(i = 1, \ldots, I)$ using the residuals x_{ij}^{m} by means of the greedy procedure explained above and next estimate the centroid values g_{jm} $(j = 1, \ldots, R)$ by means of equation (9). Denote $\mathbf{a}_m = (a_{im})$ and $\mathbf{g}_m = (g_{mj})$, calculate the outer product $\mathbf{a}_m \mathbf{g}_m$ and subtract it from x_{ij}^{m} yielding the residuals $x_{ij}^{m+1} = x_{ij}^{m} - \mathbf{a}_m \mathbf{g}_m$. This loop is repeated on x_{ij}^{m+1} and the algorithm terminates if the prespecified number of clusters R is reached.

One may note that this algorithm may only yield a local minimum rather than the global optimum of the loss function. Moreover, SEFIT might also have problems in recovering a true underlying model. We now will illustrate this latter problem with a hypothetical example. Suppose the following true structure underlies the data X:

$$M = AG = \begin{pmatrix} 1\,0 \\ 1\,0 \\ 1\,1 \\ 1\,1 \\ 0\,1 \\ 0\,1 \end{pmatrix} \begin{pmatrix} g_{11}\ g_{12}\ g_{13} \\ g_{21}\ g_{22}\ g_{23} \end{pmatrix} = \begin{pmatrix} g_{11} & g_{12} & g_{13} \\ g_{11} & g_{12} & g_{13} \\ g_{11}+g_{21} & g_{12}+g_{22} & g_{13}+g_{23} \\ g_{11}+g_{21} & g_{12}+g_{22} & g_{13}+g_{23} \\ g_{21} & g_{22} & g_{23} \\ g_{21} & g_{22} & g_{23} \end{pmatrix}. \tag{10}$$

Suppose now that we estimate the first cluster and that the correct membership vector $\mathbf{a}_1 = (1,1,1,1,0,0)'$ has been recovered. Then according to (9) the estimate of the corresponding centroid $\mathbf{g}_1$ reads

$$\mathbf{g}_1 = (g_{11} + g_{21}/2, \quad g_{12} + g_{22}/2, \quad g_{13} + g_{23}/2), \tag{11}$$

which is not equal to the true (g_{11}, g_{12}, g_{13}). Clearly a bias has been introduced due to the overlapping nature of the clusters and all further estimates may be influenced by this wrong estimate since in the next step of the algorithm the centroid will be subtracted from the data.

5 ALS algorithm

Our new approach to find the optimum of the loss function (3) is of the alternating least squares (ALS) type: given a membership matrix A we will look for an optimal G conditional upon the given A; given this G we will subsequently look for a new and conditionally optimal A, and so on.

The easiest part is the search for G given the memberships A since this comes down to an ordinary multivariate least squares regression problem, with a closed form solution:

$$G = (A'A)^{-1}A'X. \tag{12}$$

For the estimation of A we can use a separability property of the loss function (3), see also [Chaturvedi and Carroll, 1994]. This loss function indeed can be rewritten as follows:

$$\begin{aligned} L^2 &= \sum_j (x_{1j} - \sum_{r=1}^{R} a_{1r} g_{rj})^2 \\ &\quad + \dots \\ &\quad + \sum_j (x_{Ij} - \sum_{r=1}^{R} a_{Ir} g_{rj})^2. \end{aligned} \tag{13}$$

The latter means that the contribution of the ith row of A has no influence on the contributions of the other rows. As a consequence, A can be estimated row-by-row, which reduces the work to evaluating $I\,2^R$ possible memberships (instead of the full 2^{IR}).

The alternating process is repeated until there is no more decrease in the loss function. Since in each step the optimal conditional solution is found, we create a decreasing row of positive loss function values. As a consequence, this row has a limit which moreover will be reached after a finite number of iterations since there are only a finite number of possible membership matrices A. The iterative process is to be started with some initial membership matrix A, which can for instance be user specified or randomly drawn. Since in the ALS approach entire matrices are estimated rather than single columns, a bias as implied by the SEFIT strategy is avoided. Nevertheless, the ALS algorithm could yield a local optimum of the loss function (3). The only solution for this inconvenience is to use a large number of starts and retain the best solution.

6 Concluding remark

In this paper we proposed a new algorithm for finding overlapping clusters of one mode of a multiway data set. It involves an alternating least squares approach and might overcome some limitations of Mirkin's original SEFIT algorithm. To justify the latter claim, however, an extensive simulation study in which the performance of both algorithms would be compared, would be needed.

References

[Chaturvedi and Carroll, 1994]A. Chaturvedi and J.D. Carroll. An alternating combinatorial optimization approach to fitting the indclus and generalized indclus models. *Journal of Classification*, pages 155–170, 1994.

[Mirkin, 1990]B. G. Mirkin. A sequential fitting procedure for linear data analysis models. *Journal of Classification*, pages 167–195, 1990.

[Mirkin, 1996]B. G. Mirkin. *Mathematical Classification and Clustering.* Kluwer Academic Publishers, Dordrecht The Netherlands, 1996.

Distances à trois voies : concepts théoriques et applications

Mohammed Bennani Dosse

Laboratoire de Statistiques
Université Rennes 2 Haute Bretagne
Place du Recteur Henri Le Moal CS 24307
35043 Rennes Cedex, France
(e-mail : mohamed.bennani@uhb.fr)

Abstract. Distance models for three-way proximity data, which consist of numerical values assigned to triples of objects that indicate their joint (lack of) resemblance, require a generalization of the usual distance concept defined on pairs of objets. An axiomatic framework is given for characterinzing three-way dissimilarity, three-way similarity and three-way distance. The Minkowski-p or $\mathcal{M}_p$ model, which includes the perimeter model, is studied and an Euclidean representation is introduced. Finally, two monotonically convergent algorithms are described that find weighted least squares representations under the Euclidean $\mathcal{M}_1$ and $\mathcal{M}_2$ models.

Keywords: Three-way dissimilarity, Three-way distances, Multidimensional scaling.

1 Introduction

De nombreux domaines scientifiques utilisent le concept de distance dans des contextes très différents. Le présent travail puise ses origines dans l'analyse des données où les distances sont principalement utilisées pour modéliser des jugements subjectifs de différence de façon à découvrir des structures latentes de représentation.

Ce concept intervient aussi lorsqu'on cherche, comme en classification, à transformer un tableau de données X en un tableau de distance D. Cette transformation peut entrainer une perte d'information comme le montrent les deux exemples suivants.

L'exemple 1, tiré de [Daws, 1996], concerne une expérience de libre classement. On demande à chacun des N sujets de produire une partition de n objets reflétant leurs ressemblances perçues. Classiquement, on détermine à partir de l'ensemble des partitions, un tableau de similarité de la manière suivante : pour deux objets i et j, la similarité s_{ij} est définie comme étant le nombre de sujets qui ont classé i et j ensemble. La distance entre i et j est égale à $\delta_{ij} = N - s_{ij}$. Le tableau 1 donne, pour deux groupes différents de sujets, les résultats d'un libre classement (la notation $12 - 3 - 4$ signifie que le sujet a produit trois classes : $\{1, 2\}$, $\{3\}$ et $\{4\}$).

partition	**groupe 1**	**groupe 2**
1234	0	0
123 – 4	5	1
124 – 3	0	0
134 – 2	0	0
1 – 234	1	2
12 – 34	0	1
13 – 24	1	2
14 – 23	0	0
12 – 3 – 4	1	4
13 – 2 – 4	0	3
1 – 23 – 4	1	4
14 – 2 – 3	0	0
1 – 24 – 3	2	0
1 – 2 – 34	2	0
1 – 2 – 3 – 4	5	1
Total	18	18

Table 1.

Pour les deux groupes on obtient :

$$s_{12} = 6,\ s_{13} = 6,\ s_{23} = 7,\ s_{14} = 0,\ s_{24} = 4,\ s_{34} = 3$$

On voit donc que la similarité à deux voies s ne permet pas de distinguer les deux groupes de sujets. Si, pour trois objets i, j et k, on définit la similarité à trois voies s_{ijk} comme étant le nombre de sujets qui ont classé i, j et k ensemble, alors on obtient :

- pour le groupe 1 :

$$s_{123} = 5,\ s_{124} = 0,\ s_{134} = 0,\ s_{234} = 1$$

- pour le groupe 2 :

$$s_{123} = 1,\ s_{124} = 0,\ s_{134} = 0,\ s_{234} = 2$$

La similarité à trois voies fait apparaitre clairement que les deux groupes n'ont pas classé de la même manière les quatre objets.

Dans le second exemple, emprunté à [Cox *et al.*, 1991], quatre individus sont décrits par sept variables binaires de la manière décrite dans le tableau 2.

Calculons, à l'aide de l'indice de Jaccard les dissimilarités entre les quatre individus :

$$\delta_{ij} = \frac{q_{ij}}{n_{ij} + q_{ij}}$$

	v_1	v_2	v_3	v_4	v_5	v_6	v_7
1	0	1	0	1	1	0	0
2	0	0	0	0	1	1	1
3	1	0	1	0	1	0	0
4	0	1	1	0	0	0	1

Table 2.

où n_{ij} (resp. q_{ij}) est le nombre de concordances positives (resp. nombre de discordances) entre les individus i et j. On a :

$$\delta_{12} = \delta_{13} = \delta_{14} = \delta_{23} = \delta_{24} = \delta_{34} = \frac{4}{5}$$

L'indice de Jaccard indique que ces quatre individus sont équidistants. Or, il suffit de réordonner le tableau 2 selon la forme suivante pour voir que si les individus 1,2 et 3 jouent des rôles symétriques il n'en est pas de même pour l'individu 4 :

	v_2	v_7	v_3	v_5	v_4	v_6	v_1
1	1	0	0	1	1	0	0
2	0	1	0	1	0	1	0
3	0	0	1	1	0	0	1
4	1	1	1	0	0	0	0

Table 3.

Cette conclusion est confirmée par le calcul de l'indice de Jaccard à trois voies (Bennani Dosse(1993)). En effet, cet indice donne :

$$\delta_{123} = \frac{6}{7},\ \delta_{124} = \delta_{134} = \delta_{234} = 1$$

Les deux exemples ci-dessus montrent l'intérêt de généraliser les concepts de similarité, dissimilarité et distances à trois (voire plusieurs) voies.

Dans la littérature, quelques auteurs se sont intéressés à ce problème. On peut citer les travaux de [Hayashi, 1972, Hayashi, 1989], [Gower, 1984], [Cox *et al.*, 1991], [Pan and Harris, 1991], [Daws, 1996]. Les premiers auteurs abordant les définitions axiomatiques et propriétés mathématiques sont [Joly and Le Calvé, 1989, Joly and Le Calvé, 1995], [Bennani Dosse, 1993] et [Heiser and Bennani Dosse, 1997]. Dans le paragraphe 2, nous allons faire quelques rappels de ces notions puis nous abordons quelques problèmes de représentations géométriques. Le dernier paragraphe traite un exemple réel à l'aide du modèle $\mathcal{M}_2$.

2 Définitions et propriétés

Soit **E** un ensemble fini non vide de cardinal n. On note ses éléments par $1, \ldots, i, j, k, \ldots, n$. Une dissimilarité à trois voies sur **E** est une mesure de dissemblance entre les éléments de **E** pris trois à trois. Plus la valeur de cette dissimilarité est grande, plus les éléments sont considérés comme différents.

Définir une dissimilarité à trois voies δ sur **E** consiste à associer à chaque triplet (i, j, k) de $\mathbf{E}^3$ un nombre réel positif ou nul, noté δ_{ijk}. Formellement :

Définition 1 *Une dissimilarité à 3 voies sur* **E** *est une application δ de $\mathbf{E}^3$ dans $\mathbb{R}^+$ telle que pour tout $i, j, k \in \mathbf{E}$ on a :*

$$\delta_{iii} = 0 \tag{1}$$

$$\delta_{ijk} = \delta_{ikj} = \delta_{jik} = \delta_{jki} = \delta_{kij} = \delta_{kji} \tag{2}$$

$$\delta_{iij} = \delta_{ijj} \tag{3}$$

Définition 2 *Soit δ une dissimilarité à 3 voies sur* **E**. *On appelle restriction de δ aux plans diagonaux l'application définie par :*

$$\rho_{ij} = \delta_{iij} (= \delta_{ijj})$$

Proposition 1 *L'application définie ci-dessus est une dissimilarité à 2 voies sur* **E**.

De façon duale, on définit le concept de similarité à 3 voies comme étant une mesure de ressemblance sur des triplets d'objets. Formellement :

Définition 3 *Une similarité à 3 voies sur* **E** *est une application s de $\mathbf{E}^3$ dans $\mathbb{R}^+$ telle que pour tout $i, j, k \in \mathbf{E}$ on a :*

$$s_{iii} = s_{jjj} = s_{kkk} \geqslant s_{ijk} \tag{4}$$

$$s_{ijk} = s_{ikj} = s_{jik} = s_{jki} = s_{kij} = s_{kji} \tag{5}$$

$$s_{iij} = s_{ijj} \tag{6}$$

Comme dans le cas "2 voies" les notions de dissimilarité et de similarités à 3 voies jouent des rôles opposés et on peut passer de l'une à l'autre par une fonction décroissante.

La généralisation de l'inégalité triangulaire qui a été proposé par Joly-Le Calvé(1989) est la suivante : pour tout $i, j, k, \ell \in \mathbf{E}^3$

$$\delta_{ijk} \leqslant \delta_{ik\ell} + \delta_{jk\ell} \tag{7}$$

Bennani Dosse(1993) propose l'inégalité suivante : pour tout $i, j, k, \ell \in \mathbf{E}^3$

$$2\,\delta_{ijk} \leqslant \delta_{ik\ell} + \delta_{jk\ell} + \delta_{ij\ell} \tag{8}$$

Proposition 2 *L'inégalité* (8) *implique l'inégalité* (7)*.*

On peut facilement vérifier que l'inégalité (8) n'est pas suffisante pour que ρ vérifie l'inégalité triangulaire. Par contre on montre (Heiser et Bennani Dosse(1997)) que l'on a :

$$\rho_{ij} \leqslant \frac{5}{4}(\rho_{ik} + \rho_{jk})$$

Pour que ρ vérifie l'inégalité triangulaire, Joly & Le Calvé(1989) introduisent la contrainte suivante :

$$\delta_{iij} \leqslant \delta_{ijk} \tag{9}$$

Définition 4 *Une application qui vérifie les axiomes* $(1), (2), (3), (7)$ *et* (9) *est appelée distance à 3 voies.*

Définition 5 *Une application qui vérifie les axiomes* $(1), (2), (3), (8)$ *est appelée distance triadique.*

Proposition 3 *Les indices de Daws et de Jaccard défnis dans le premier paragraphe sont des dissimilarités à 3 voies qui vérifient les inégalités* (8) *et* (9)*.*

Définition 6 *une application δ de $\mathbf{E}^3$ dans $\mathbb{R}^+$ est dite distance à centre à 3 voies s'il existe un vecteur $u \in \mathbb{R}^n_+$ tel que :*

$$\begin{aligned}\delta_{iii} &= 0\\ \delta_{iij} &= u_i + u_j\\ \delta_{ijk} &= u_i + u_j + u_k\end{aligned}$$

3 Modèles de Minkowski d'ordre p

Étant donnée une dissimilarité à 2 voies d sur $\mathbf{E}$, on peut construire de nombreux modèles de dissimilarités à 3 voies. Les modèles de Minkowski d'ordre p, $p \geqslant 1$, sont définis par :

$$\delta^p_{ijk} = d^p_{ij} + d^p_{ik} + d^p_{jk} \tag{10}$$

Proposition 4 *si d est une distance à 2 voies alors δ est une distance triadique. De plus, la restriction de δ aux plans diagonaux est une distance à 2 voies.*

Remarque 1 *Si $p = 1$ on obtient le modèle périmètre; si $p = 2$ on obtient le modèle $\mathcal{M}_2$ et si $p = \infty$ on obtient le modèle max.*

4 Représentations géométriques

Considérons le problème suivant : étant donnée une dissimilarité à trois voies δ sur $\mathbf{E}$, on cherche à représenter les éléments de $\mathbf{E}$ par des points dans un espace de dimension finie de manière que les distances à 3 voies dans cet espace approchent le plus possible les données initiales. Ce problème est une extension du multidimensional scaling (voir [Borg and Groenen, 1998]).

4.1 Approximation par une distance périmètre

Le problème posé est de minimiser la fonction :

$$\sigma_1 = \sum_i \sum_j \sum_k w_{ijk} \left(\delta_{ijk} - d_{ij} - d_{ik} - d_{jk}\right)^2$$

où d est une distance euclidienne dans un espace de dimension donnée $\mathbf{p}$ et les w_{ijk} sont des poids positifs ou nuls donnés.

4.2 Approximation par une distance $\mathcal{M}_2$

Le problème posé est de minimiser la fonction :

$$\sigma_2 = \sum_i \sum_j \sum_k w_{ijk} \left(\delta_{ijk} - \sqrt{d_{ij}^2 + d_{ik}^2 + d_{jk}^2}\right)^2$$

5 Application

Hayashi(1972) a collecté directement des données de dissimilarité à 3 voies portant sur l'improductivité d'équipes formées de trois individus. Vu la taille restreinte des données ($n = 6$), cet exemple présente surtout un caractère pédagogique. Les données sont présentées dans le tableau 4 :

Hayashi propose, pour faire une représentation euclidienne de la dissimilarité à 3 voies δ, d'utiliser le carré de la surface du triangle. La figure 1 présente le positionnement des 6 individus.

Nous avons analysé ces données à l'aide du modèle $\mathcal{M}_2$. La figure 2 montre que ces données mettent en evidence deux groupes d'individus $\{1, 2, 3\}$ et $\{4, 5, 6\}$.

$\delta_{123} = 1$	$\delta_{124} = 7$	$\delta_{125} = 6$	$\delta_{126} = 9$
$\delta_{134} = 7$	$\delta_{234} = 8$	$\delta_{135} = 6$	$\delta_{235} = 7$
$\delta_{136} = 9$	$\delta_{236} = 9$	$\delta_{145} = 4$	$\delta_{245} = 6$
$\delta_{345} = 3$	$\delta_{146} = 9$	$\delta_{246} = 8$	$\delta_{346} = 5$
$\delta_{156} = 6$	$\delta_{256} = 7$	$\delta_{356} = 3$	$\delta_{456} = 1$

Table 4.

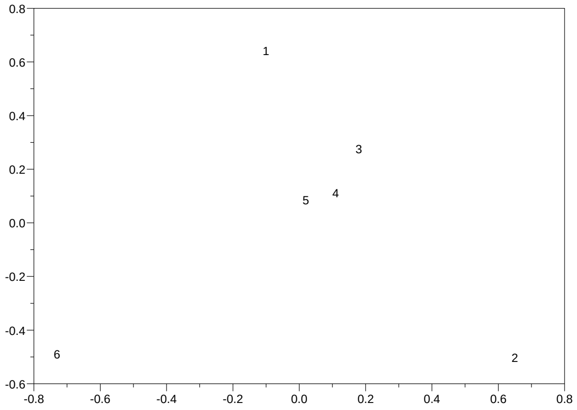


Fig. 1. données de Hayashi : modèle surface du triangle.

6 Conclusion

Ce travail montre qu'il est possible, grâce à quelques outils mathématiques élémentaires, d'étendre à plusieurs voies les notions de dissimilarité, similarité et distance. Nous avons choisi de mettre l'accent sur les représentations Euclidiennes mais d'autres sont possibles (comme les représentations hiérarchiques).

Un champ, particulièrement intéressant dans les applications, est celui où l'on dispose d'un tableau à trois voies où la donnée exprime une dissimilarité entre les éléments de trois ensembles disjoints. Cette approche est une généralisation du dépliage métrique (metric unfolding). Le lecteur intérréssé peut consulter Bennani Dosse(1995)[Bennani Dosse, 1995].

References

[Bennani Dosse, 1993]M. Bennani Dosse. *Analyses métriques à 3 voies.* Thèse de Doctorat-Université de Rennes 2 Haute Bretagne, Rennes, 1993.

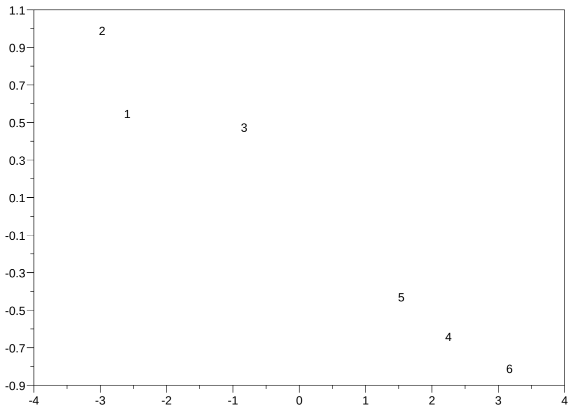


Fig. 2. données de Hayashi : modèle $\mathcal{M}_2$.

[Bennani Dosse, 1995]M. Bennani Dosse. Positionnement multidimensionnel d'un tableau à 3 voies. *Revue de Statistque Appliquée*, pages 63–75, 1995.

[Borg and Groenen, 1998]I. Borg and P. Groenen. *Modern Multidimensional Scaling : Theory and Applications.* Springer Series in Statistics, Rennes, 1998.

[Cox *et al.*, 1991]T.F. Cox, M.A.A. Cox, and J.A. Branco. Multidimensional scaling for n-tuples. *British Journal of Mathematical and Statistical Psychology*, pages 195–206, 1991.

[Daws, 1996]J.T. Daws. The analysis of free-sorting data : Beyond pairwise coocurrences. *Journal of classification*, pages 57–80, 1996.

[Gower, 1984]J.C. Gower. Multidimensional scaling displays. In H.G. Law, C.W. Snyder, J.A. Hattie, and R.P. MacDonald, editors, *Research methods for multimode data analysis*, 1984.

[Hayashi, 1972]C. Hayashi. Two dimensional quantification based on the measure of dissimilarity among three elements. *Annals of the Institute of Statistical Mathematics*, pages 251–257, 1972.

[Hayashi, 1989]C. Hayashi. Multiway data matrices and methods of quantification of qualitative data as strategy of data analysis. In R. Coppi and S. Bolasco, editors, *Multiway data analysis*, 1989.

[Heiser and Bennani Dosse, 1997]W. J. Heiser and M. Bennani Dosse. Triadic distance models : Axiomatization and least squares representation. *Journal of Mathematical Psychology*, pages 189–206, 1997.

[Joly and Le Calvé, 1989]S. Joly and G. Le Calvé. Three-way distances. *Rapport de recherche*, 1989.

[Joly and Le Calvé, 1995]S. Joly and G. Le Calvé. Three-way distances. *Journal of Classification*, pages 191–205, 1995.

[Pan and Harris, 1991]G. Pan and D.P. Harris. A new multidimensional scaling technique based upon association of triple objects p_{ijk} and its application to the analysis of geochemical data. *Mathematical Geology*, pages 861–886, 1991.

On the use of mutual information in data analysis : an overview

Ivan Kojadinovic

LINA CNRS FRE 2729, Site école polytechnique de l'université de Nantes
Rue Christian Pauc, 44306 Nantes, France
Email : ivan.kojadinovic@polytech.univ-nantes.fr

Abstract. An overview of the use of mutual information in data analysis is presented. Different normalized versions of this index of stochastic dependence are recalled, new approximations of it are proposed, its estimation in a discrete and in a continuous context is discussed, and some applications of it in data analysis are briefly reviewed.

Keywords: Mutual information, normalization, approximation, estimation, applications.

1 Introduction

Mutual information satisfies properties that make it an ideal measure of stochastic dependence [Cover and Thomas, 1991, Darbellay, 1999, Joe, 1989b] [Rényi, 1959]. Unlike Pearson's linear correlation coefficient which accounts only for linear relationships, or other well-known rank correlation coefficients that can detect monotonic dependencies, the mutual information takes into account all types of dependence.

In the first section, after introducing the notion of mutual information, we present its best-known normalized versions and we show how less computationally expensive approximations of it can be obtained by means of the concept of k-additive truncation. In the second section, its estimation is discussed both in a discrete and in a continuous context. The last section is devoted to a brief overview of some applications of mutual information in data analysis.

2 Mutual information

In the rest of the paper, random variables shall be denoted by uppercase letters, e.g. X, and random vectors by uppercase *black-board* letters, e.g. $\overrightarrow{\mathbb{X}}$. In order to unify the presentation of the mutual information in the discrete and in the continuous case, we shall classically further assume that the probability measures of the manipulated random vectors are absolutely continuous (a.c) with respect to (w.r.t) a σ-finite measure μ being either the counting measure in a discrete setting or the Lebesgue measure in a continuous framework.

2.1 Definition and properties

Let us consider a random vector $(\overrightarrow{\mathbb{X}}, \overrightarrow{\mathbb{Y}})$. The *mutual information* between $\overrightarrow{\mathbb{X}}$ and $\overrightarrow{\mathbb{Y}}$ is defined as the *distance from independence* between $\overrightarrow{\mathbb{X}}$ and $\overrightarrow{\mathbb{Y}}$ measured by the Kullback and Leibler divergence [Cover and Thomas, 1991] [Kullback and Leibler, 1951, Kus, 1999, Ullah, 1996].
For two densities p and q w.r.t μ with same support, the Kullback and Leibler divergence is defined by

$$KL(p,q) := \int p \log\left(\frac{p}{q}\right)\, d\mu \tag{1}$$

with the convention that $0 \log \frac{0}{0} := 0$.
Let us denote by $p_{(\overrightarrow{\mathbb{X}},\overrightarrow{\mathbb{Y}})}$, $p_{\overrightarrow{\mathbb{X}}}$ and $p_{\overrightarrow{\mathbb{Y}}}$ the joint and marginal densities of the random vectors. The mutual information between $\overrightarrow{\mathbb{X}}$ and $\overrightarrow{\mathbb{Y}}$ is then defined by

$$I(\overrightarrow{\mathbb{X}};\overrightarrow{\mathbb{Y}}) := KL(p_{(\overrightarrow{\mathbb{X}},\overrightarrow{\mathbb{Y}})}, p_{\overrightarrow{\mathbb{X}}} \otimes p_{\overrightarrow{\mathbb{Y}}}), \tag{2}$$

where $p_{\overrightarrow{\mathbb{X}}} \otimes p_{\overrightarrow{\mathbb{Y}}}$ denotes the tensor product of $p_{\overrightarrow{\mathbb{X}}}$ and $p_{\overrightarrow{\mathbb{Y}}}$. From the above definition, we see that the mutual information is symmetric and, by applying the Jensen inequality to the Kullback and Leibler divergence, we obtain that the mutual information is always non negative and zero if and only if $\overrightarrow{\mathbb{X}}$ and $\overrightarrow{\mathbb{Y}}$ are stochastically independent.
The mutual information can also be interpreted as the *H-information* obtained from the Shannon entropy [DeGroot, 1962, Morales *et al.*, 1996]. The Shannon entropy of a density p w.r.t μ, when it exists, is defined by

$$H(p) := -\int p \log(p)\, d\mu$$

with the convention that $0 \log 0 := 0$. In the discrete case, $H(p)$ always exists, is positive and can be interpreted as an *uncertainty* or an *information* measure [Rényi, 1965], whereas in the continuous case, when it exists, it can be negative and should be only interpreted as a measure of the *structure* contained in the density p.
With respect to the Shannon entropy, the mutual information between $\overrightarrow{\mathbb{X}}$ and $\overrightarrow{\mathbb{Y}}$ can be easily rewritten as

$$I(\overrightarrow{\mathbb{X}};\overrightarrow{\mathbb{Y}}) = H(p_{\overrightarrow{\mathbb{X}}}) - E_{p_{\overrightarrow{\mathbb{Y}}}}[H(p_{\overrightarrow{\mathbb{X}}|\overrightarrow{\mathbb{Y}}=y})] = H(p_{\overrightarrow{\mathbb{Y}}}) - E_{p_{\overrightarrow{\mathbb{X}}}}[H(p_{\overrightarrow{\mathbb{Y}}|\overrightarrow{\mathbb{X}}=x})]. \tag{3}$$

Hence, the mutual information can be interpreted as the reduction in the uncertainty of $\overrightarrow{\mathbb{X}}$ (resp. $\overrightarrow{\mathbb{Y}}$) due to the knowledge of $\overrightarrow{\mathbb{Y}}$ (resp. $\overrightarrow{\mathbb{X}}$) [Ullah, 1996]. Rewriting the expectation in Eq. (3), we obtain $E_{p_{\overrightarrow{\mathbb{Y}}}}[H(p_{\overrightarrow{\mathbb{X}}|\overrightarrow{\mathbb{Y}}=y})] = H(p_{(\overrightarrow{\mathbb{X}},\overrightarrow{\mathbb{Y}})}) - H(p_{\overrightarrow{\mathbb{Y}}})$, and therefore

$$I(\overrightarrow{\mathbb{X}};\overrightarrow{\mathbb{Y}}) = H(p_{\overrightarrow{\mathbb{X}}}) + H(p_{\overrightarrow{\mathbb{Y}}}) - H(p_{(\overrightarrow{\mathbb{X}},\overrightarrow{\mathbb{Y}})}). \tag{4}$$

2.2 Normalized versions of the mutual information in the discrete case

Consider two discrete random vectors $\overrightarrow{\mathbb{X}}$ and $\overrightarrow{\mathbb{Y}}$. Since the mutual information can be interpreted as the H-information obtained from the Shannon entropy, which is always non negative in the discrete case, a first normalized version of $I(\overrightarrow{\mathbb{X}};\overrightarrow{\mathbb{Y}})$ is given by

$$U(\overrightarrow{\mathbb{X}};\overrightarrow{\mathbb{Y}}) = \frac{H(p_{\overrightarrow{\mathbb{X}}}) - E_{p_{\overrightarrow{\mathbb{Y}}}}[H(p_{\overrightarrow{\mathbb{X}}|\overrightarrow{\mathbb{Y}}=y})]}{H(p_{\overrightarrow{\mathbb{X}}})} = \frac{I(\overrightarrow{\mathbb{X}};\overrightarrow{\mathbb{Y}})}{H(p_{\overrightarrow{\mathbb{X}}})}.$$

The quantity $U(\overrightarrow{\mathbb{X}};\overrightarrow{\mathbb{Y}})$, known as the *asymmetric uncertainty coefficient*, can be interpreted as the *relative* reduction of the uncertainty contained in $\overrightarrow{\mathbb{X}}$ given $\overrightarrow{\mathbb{Y}}$ [Särndal, 1974]. The above quantity is clearly not symmetric. A symmetric version of $U(\overrightarrow{\mathbb{X}};\overrightarrow{\mathbb{Y}})$, known as the *symmetric uncertainty coefficient* [Särndal, 1974], is defined by

$$S(\overrightarrow{\mathbb{X}};\overrightarrow{\mathbb{Y}}) := \frac{I(\overrightarrow{\mathbb{X}};\overrightarrow{\mathbb{Y}})}{\frac{1}{2}\left[H(p_{\overrightarrow{\mathbb{X}}}) + H(p_{\overrightarrow{\mathbb{Y}}})\right]}.$$

Although the values of the latter quantity are in $[0,1]$, it does not necessarily take the value 1 when there is a perfect functional dependence between $\overrightarrow{\mathbb{X}}$ and $\overrightarrow{\mathbb{Y}}$. This last observation led Joe [Joe, 1989b] to define a normalized version of the mutual information as

$$I_d^*(\overrightarrow{\mathbb{X}};\overrightarrow{\mathbb{Y}}) := \frac{I(\overrightarrow{\mathbb{X}};\overrightarrow{\mathbb{Y}})}{\min\left[H(p_{\overrightarrow{\mathbb{X}}}), H(p_{\overrightarrow{\mathbb{Y}}})\right]}. \tag{5}$$

The quantity $I_d^*(\overrightarrow{\mathbb{X}};\overrightarrow{\mathbb{Y}})$ clearly takes its values in $[0,1]$. Furthermore, $I_d^*(\overrightarrow{\mathbb{X}};\overrightarrow{\mathbb{Y}}) = 1$ if and only if $\overrightarrow{\mathbb{X}}$ and $\overrightarrow{\mathbb{Y}}$ are functionally dependent [Joe, 1989b, Theorem 2.3].

2.3 Normalized versions of the mutual information in the continuous case

Let (X,Y) be a normally distributed random vector with correlation coefficient ρ. The mutual information between X and Y is then given by $I(X;Y) = -1/2\log(1-\rho^2)$ [Cover and Thomas, 1991]. Starting from this observation and by analogy with the way Pearson's contingency coefficient was obtained, Joe [Joe, 1989b] defined a normalized version of the mutual information as

$$I_c^*(X;Y) := \sqrt{1-\exp[-2I(X;Y)]}. \tag{6}$$

The quantity $I_c^*(X,Y)$ clearly takes its values in $[0,1]$ and is equal to $|\rho|$ if (X,Y) is normally distributed with correlation coefficient ρ.

Let us now consider the case where X and Y are "approximately dependent". As in the case of the contingency coefficient, Joe [Joe, 1989b] conjectured that the "more X and Y are functionally dependent", the closer $I_c^*(X,Y)$ to 1 ; see also [Granger and Lin, 1994].

Note that the above quantity can immediately be generalized to random vectors.

2.4 Generalizations of the mutual information

Starting from Eq. (4), Abramson proposed a natural extension of the mutual information between more than two random vectors [Abramson, 1963]. The mutual information among three random vectors $\overrightarrow{\mathbb{X}}$, $\overrightarrow{\mathbb{Y}}$ and $\overrightarrow{\mathbb{Z}}$ having a joint density w.r.t μ is defined by

$$\begin{aligned} I(\overrightarrow{\mathbb{X}};\overrightarrow{\mathbb{Y}};\overrightarrow{\mathbb{Z}}) := {} & H(p_{\overrightarrow{\mathbb{X}}}) + H(p_{\overrightarrow{\mathbb{Y}}}) + H(p_{\overrightarrow{\mathbb{Z}}}) \\ & - H(p_{(\overrightarrow{\mathbb{X}},\overrightarrow{\mathbb{Y}})}) - H(p_{(\overrightarrow{\mathbb{X}},\overrightarrow{\mathbb{Z}})}) - H(p_{(\overrightarrow{\mathbb{Y}},\overrightarrow{\mathbb{Z}})}) + H(p_{(\overrightarrow{\mathbb{X}},\overrightarrow{\mathbb{Y}},\overrightarrow{\mathbb{Z}})}). \end{aligned}$$

More generally, for $r \geq 2$ random vectors $\overrightarrow{\mathbb{X}}_1,\ldots,\overrightarrow{\mathbb{X}}_r$ having a joint density w.r.t μ, the following definition was adopted by Abramson :

$$I(\overrightarrow{\mathbb{X}}_1;\ldots,\overrightarrow{\mathbb{X}}_r) := \sum_{k=1}^{r} \sum_{\{i_1,\ldots,i_k\}\subseteq\{1,\ldots,r\}} (-1)^{k+1} H(p_{(\overrightarrow{\mathbb{X}}_{i_1},\ldots,\overrightarrow{\mathbb{X}}_{i_k})}). \quad (7)$$

The mutual information among $r \geq 2$ random vectors $\overrightarrow{\mathbb{X}}_1,\ldots,\overrightarrow{\mathbb{X}}_r$ can be interpreted as a measure of their *simultaneous interaction* [Kojadinovic, 2004b] [Wienholt and Sendhoff, 1996]. It can equivalently be regarded as a sort of *multiway* similarity measure among variables. Should it be zero, the r random vectors do not simultaneously interact. Note that the mutual information between more than two random vectors is not necessarily non negative [Cover and Thomas, 1991].

Another straightforward generalization of the mutual information is frequently encountered in the literature under the name of *redundancy*. The redundancy [Wienholt and Sendhoff, 1996] among $r \geq 2$ random vectors $\overrightarrow{\mathbb{X}}_1,\ldots,\overrightarrow{\mathbb{X}}_r$ having a joint density w.r.t μ is defined by

$$R(\overrightarrow{\mathbb{X}}_1;\ldots;\overrightarrow{\mathbb{X}}_r) := KL(p_{(\overrightarrow{\mathbb{X}}_1,\ldots,\overrightarrow{\mathbb{X}}_r)}, p_{\overrightarrow{\mathbb{X}}_1} \otimes \cdots \otimes p_{\overrightarrow{\mathbb{X}}_r}),$$

which, in terms of the Shannon entropy, can be easily rewritten as

$$R(\overrightarrow{\mathbb{X}}_1;\ldots;\overrightarrow{\mathbb{X}}_r) = \sum_{i=1}^{r} H(p_{\overrightarrow{\mathbb{X}}_i}) - H(p_{(\overrightarrow{\mathbb{X}}_1,\ldots,\overrightarrow{\mathbb{X}}_r)}).$$

As previously, it is easy to verify that the redundancy is always positive and equal to zero if and only $\overrightarrow{\mathbb{X}}_1,\ldots,\overrightarrow{\mathbb{X}}_r$ are stochastically mutually independent. As for the mutual information, the higher the redundancy among the random vectors, the "stronger" their functional dependency [Joe, 1989b].

2.5 Approximations of the mutual information based on k-additive truncation

Consider a finite set $\aleph := \{X_1, \dots, X_m\}$ of random variables. The subsets of $\aleph$ will be denoted by uppercase *black-board* letters, e.g. $\mathbb{X}$. Given a subset $\mathbb{X} \subseteq \aleph$ composed of r variables, $\overrightarrow{\mathbb{X}}$ will denote an r-dimensional random vector whose coordinates are distinct elements from $\mathbb{X}$. We shall also assume that the variables in $\aleph$ have a joint density w.r.t μ.

Let $h : 2^\aleph \to \mathbb{R}$ and $i : 2^\aleph \to \mathbb{R}$ be set functions defined respectively by

$$h(\mathbb{X}) := \begin{cases} 0, & \text{if } \mathbb{X} = \emptyset, \\ H(p_{\overrightarrow{\mathbb{X}}}), & \text{if } \mathbb{X} \neq \emptyset, \end{cases}$$

and

$$i(\mathbb{X}) := \begin{cases} 0, & \text{if } \mathbb{X} = \emptyset, \\ I(X_{i_1}; \dots; X_{i_r}), & \text{if } \mathbb{X} = \{X_{i_1}, \dots, X_{i_r}\}. \end{cases}$$

Using concepts well-known in discrete mathematics such as the Möbius transform [Rota, 1964], it is easy to verify that i is an *equivalent representation* of h [Grabisch *et al.*, 2000, Kojadinovic, 2002]. Practically, this means that the numbers $\{h(\mathbb{X})\}_{\mathbb{X} \subseteq \aleph}$ can be recovered from the coefficients $\{i(\mathbb{X})\}_{\mathbb{X} \subseteq \aleph}$, and *vice versa*. More precisely, from Eq. (7) and using the *zeta transform* [Grabisch *et al.*, 2000], we have

$$i(\mathbb{X}) = \sum_{\mathbb{T} \subseteq \mathbb{X}} (-1)^{|\mathbb{T}|+1} h(\mathbb{T}) \qquad \text{and} \qquad h(\mathbb{X}) = \sum_{\mathbb{T} \subseteq \mathbb{X}} (-1)^{|\mathbb{T}|+1} i(\mathbb{T}), \qquad \forall \mathbb{X} \subseteq \aleph.$$

From the latter equation, it follows that the entropy of random vector $\overrightarrow{\mathbb{X}}$ whose coordinates are denoted $X_{i_1}, \dots, X_{i_r}$ can be rewritten as

$$\begin{aligned} H(p_{\overrightarrow{\mathbb{X}}}) = & \sum_{X_j \in \mathbb{X}} H(p_{X_j}) - \sum_{\{X_j, X_k\} \subseteq \mathbb{X}} I(X_j; X_k) \\ & + \sum_{\{X_j, X_k, X_l\} \subseteq \mathbb{X}} I(X_j; X_k; X_l) - \cdots + (-1)^{r+1} I(X_{i_1}; \dots; X_{i_r}). \end{aligned}$$

The entropy of $p_{\overrightarrow{\mathbb{X}}}$ is therefore calculated, first by summing the entropies of the singletons contained in $\mathbb{X}$, then by subtracting the sum of mutual informations among pairs of variables contained in $\mathbb{X}$, after by adding the sum of mutual informations among variables of 3-element subsets contained in $\mathbb{X}$, etc. The sums of mutual informations that are added or subtracted can be seen as *corrective terms* or *higher order terms.* In certain situations such as variable selection [Kojadinovic, 2004b], it may interesting, for computational reasons, to perform a *k-additive truncation* of H for a given $k \in \{1, \dots, m\}$, that is to neglect *corrective terms* of order greater than k in the expression of the entropy, which leads to an approximation of the mutual information between two random vectors. For instance, as shown in [Kojadinovic, 2002], for $k = 2$ and $k = 3$, we have respectively

$$I^{(2)}(\overrightarrow{\mathbb{X}};\overrightarrow{\mathbb{Y}}) = \sum_{X\in\mathbb{X}}\sum_{Y\in\mathbb{Y}} I(X;Y) \qquad \text{and}$$

$$\begin{aligned} I^{(3)}(\overrightarrow{\mathbb{X}};\overrightarrow{\mathbb{Y}}) = I^{(2)}(\overrightarrow{\mathbb{X}};\overrightarrow{\mathbb{Y}}) &- \sum_{X\in\mathbb{X}}\sum_{\{Y_1,Y_2\}\subseteq\mathbb{Y}} I(X;Y_1;Y_2) \\ &- \sum_{\{X_1,X_2\}\subseteq\mathbb{X}}\sum_{Y\in\mathbb{Y}} I(X_1;X_2;Y). \end{aligned}$$

Note that the lower the amount of interaction among random variables in a set $\mathbb{X}$, the closer the truncated entropy $H^{(k)}(p_{\overrightarrow{\mathbb{X}}})$ to $H(p_{\overrightarrow{\mathbb{X}}})$, with equality if there are no simultaneous interactions among more then k variables.

3 Estimation

3.1 In a discrete setting

Consider two discrete random vectors $\overrightarrow{\mathbb{X}}$ and $\overrightarrow{\mathbb{Y}}$ respectively taking their values in the sets $\{x_1, \dots, x_r\}$ and $\{y_1, \dots, y_s\}$. From Eq. (2), we see that their mutual information is clearly a function of their joint distribution $p_{(\overrightarrow{\mathbb{X}},\overrightarrow{\mathbb{Y}})}$, which is classically estimated by its maximum likelihood estimator (sample proportions). Using the well-know *delta* method [Agresti, 2002, Saporta, 1990], it can be shown that $KL(\hat{p}_{(\overrightarrow{\mathbb{X}},\overrightarrow{\mathbb{Y}})}, \hat{p}_{\overrightarrow{\mathbb{X}}} \otimes \hat{p}_{\overrightarrow{\mathbb{Y}}})$ is asymptotically normally distributed [Basharin, 1959, Menéndez *et al.*, 1995] with expectation $I(\overrightarrow{\mathbb{X}};\overrightarrow{\mathbb{Y}})$ and variance $\sigma^2_{KL}(p_{(\overrightarrow{\mathbb{X}},\overrightarrow{\mathbb{Y}})})/n$, where $\sigma^2_{KL}(p_{(\overrightarrow{\mathbb{X}},\overrightarrow{\mathbb{Y}})})$ is

$$\sum_{i=1}^{r}\sum_{j=1}^{s} p_{(\overrightarrow{\mathbb{X}},\overrightarrow{\mathbb{Y}})}(x_i,y_j)\left(\log\frac{p_{(\overrightarrow{\mathbb{X}},\overrightarrow{\mathbb{Y}})}(x_i,y_j)}{p_{\overrightarrow{\mathbb{X}}}(x_i)p_{\overrightarrow{\mathbb{Y}}}(y_j)}\right)^2 - KL(p_{(\overrightarrow{\mathbb{X}},\overrightarrow{\mathbb{Y}})}, p_{\overrightarrow{\mathbb{X}}} \otimes p_{\overrightarrow{\mathbb{Y}}})^2.$$

This result can be used to obtain approximate confidence intervals for the mutual information. When $\overrightarrow{\mathbb{X}}$ and $\overrightarrow{\mathbb{Y}}$ are stochastically independent, a classical calculation shows that the mutual information is asymptotically χ^2 distributed with $(r-1)(s-1)$ degrees of freedom [Menéndez *et al.*, 1995]. More details and further results can be found in [Fagen, 1978, Hutter and Zaffalon, 2005] [Roulston, 1999].

3.2 In a continuous setting

From Eqs. (4) and (7), we see that estimating mutual information amounts to estimating Shannon entropies.

Consider a random vector $\overrightarrow{\mathbb{X}}$ having a Lebesgue density. A point-wise estimation of the entropy of its density can be obtained in two steps : first,

by substituting the density of $\overrightarrow{\mathbb{X}}$ in the expression of the Shannon entropy by an estimate computed from available independent realizations; then, by computing the remaining integral by numerical quadrature [Granger and Lin, 1994] [Harvill and Ray, 2001, Joe, 1989b, Silverman, 1986].
The difficulties linked to numerical integration can however be avoided. Let $F_{\overrightarrow{\mathbb{X}}}$ be the cumulative distribution function of $\overrightarrow{\mathbb{X}}$ and let $\overrightarrow{\mathbb{X}}_1, \ldots, \overrightarrow{\mathbb{X}}_n$ be a random sample drawn from $p_{\overrightarrow{\mathbb{X}}}$. The Shannon entropy of $p_{\overrightarrow{\mathbb{X}}}$ can then be rewritten as

$$H(p_{\overrightarrow{\mathbb{X}}}) = -\int \log p_{\overrightarrow{\mathbb{X}}} dF_{\overrightarrow{\mathbb{X}}}.$$

Substituting $F_{\overrightarrow{\mathbb{X}}}$ by the empirical cumulative distribution function and $p_{\overrightarrow{\mathbb{X}}}$ by an estimate, we obtain a natural estimator of the Shannon entropy given by

$$\hat{H}(p_{\overrightarrow{\mathbb{X}}}) = -\frac{1}{n}\sum_{i=1}^{n} \log \hat{p}_{\overrightarrow{\mathbb{X}}}(\overrightarrow{\mathbb{X}}_i).$$

The above estimator was studied in [Hall and Morton, 1993, Joe, 1989a] in the case where $p_{\overrightarrow{\mathbb{X}}}(\overrightarrow{\mathbb{X}}_i)$ is estimated by *kernel density estimation* [Scott, 1992] [Silverman, 1986]. In that context, Hall and Morton showed that the estimator $\hat{H}(p_{\overrightarrow{\mathbb{X}}})$ is consistent if the dimension of $\overrightarrow{\mathbb{X}}$ is strictly inferior to 4 and if the density of $\overrightarrow{\mathbb{X}}$ satisfies certain regularity conditions. A synthesis on the estimation of the Shannon entropy in the continuous case can be found in [Beirlant *et al.*, 1997].
From a practical perspective, the use of two nonparametric density estimation technique is encountered in the literature : kernel density estimation [Granger and Lin, 1994, Harvill and Ray, 2001, Kojadinovic, 2004a] and *projection pursuit density estimation* [Friedman *et al.*, 1984, Kojadinovic, 2002].

Another approach to mutual information estimation is based on a prior discretization of the random vectors by means of recursive partitioning algorithms [Darbellay, 1999, Fraser, 1989]. The best studied and most promising approach is probably that proposed by Darbellay.

4 Some applications of mutual information in data analysis

In a discrete setting, unnormalized mutual information was used for discrete variable clustering [Benzécri, 1976, Chap. 5] (although the symmetric uncertainty coefficient or I_d^* seem more appropriate). Note that the approximation proposed in section 2.5 could be used to define new aggregation criteria. The asymptotic results presented in section 3.1 make it even possible to use the *analysis of the link likelihood* method [Lerman, 1981] in that context. The symmetric uncertainty coefficient was used for feature selection (see e.g. [Yei and Liu, 2003]), the use of the asymmetric version being even more natural in that context.

In a continuous setting, unnormalized mutual information was used for lag identification in nonlinear time series [Fraser, 1989, Granger and Lin, 1994] [Harvill and Ray, 2001, Kantz and Schreiber, 1997] and k-additive approximations of it for variable selection in regression problems [Kojadinovic, 2004a]. The coefficient I_c^* and redundancy was employed for continuous variable clustering [Kojadinovic, 2004b]. Redundancy minimization is at the root of some approaches to *independent component analysis* ; see e.g [Hyvärinen, 1999].

References

[Abramson, 1963]N. Abramson. *Information Theory and Coding.* McGraw Hill, New-York, 1963.

[Agresti, 2002]Alan Agresti. *Categorical Data Analysis.* Wiley, 2002. Second edition.

[Basharin, 1959]G.P. Basharin. On the statistical estimate of the entropy of a sequence of independent random variables. *Theory of Probability and its Applications*, 4:361–364, 1959.

[Beirlant *et al.*, 1997]J. Beirlant, E. Dudewicz, L. Györfi, and E.G. van der Meulen. Nonparametric entropy estimation: an overview. *Int. J. Math. Stat. Sci.*, 6:17–39, 1997.

[Benzécri, 1976]J.-P. Benzécri. *L'analyse de données: la taxonomie.* Dunod, 1976.

[Cover and Thomas, 1991]T. Cover and J. Thomas. *Elements of Information Theory.* John Wiley and Sons, 1991.

[Darbellay, 1999]G. A. Darbellay. An estimator for the mutual information based on a criterion for independence. *Computational Statistics and Data Analysis*, 32:1–17, 1999.

[DeGroot, 1962]M. H. DeGroot. Uncertainty, information and sequential experiments. *Ann. Math. Statist.*, 33:404–419, 1962.

[Fagen, 1978]R.M. Fagen. Information measures: statistical confidence limits and inference. *J. Theor. Biol.*, 73:61–79, 1978.

[Fraser, 1989]A. Fraser. Information and entropy in strange attractors. *IEEE Transactions on Information Theory*, 35(2):245–262, 1989.

[Friedman *et al.*, 1984]J. Friedman, W. Stuetzle, and A. Schroeder. Projection pursuit density estimation. *Journal of the American Statistical Association*, 79(387):599–608, 1984.

[Grabisch *et al.*, 2000]Michel Grabisch, Jean-Luc Marichal, and Marc Roubens. Equivalent representations of set functions. *Math. Oper. Res.*, 25(2):157–178, 2000.

[Granger and Lin, 1994]C. W. J. Granger and J. Lin. Using the mutual information coefficient to identify lags in nonlinear models. *J. Time Ser. Anal.*, 15:371–384, 1994.

[Hall and Morton, 1993]P. Hall and S.C. Morton. On the estimation of entropy. *Ann. Inst. Statist. Math.*, 45:69–88, 1993.

[Harvill and Ray, 2001]J. Harvill and B. Ray. Lag identification for vector nonlinear time series. *Communications Statistics: Theory and Methods*, 29:1672–1702, 2001.

[Hutter and Zaffalon, 2005]Marcus Hutter and Marco Zaffalon. Distribution of mutual information from complete and incomplete data. *Computational Statistics and Data Analysis*, 48:633–657, 2005.

[Hyvärinen, 1999]A. Hyvärinen. Survey on independent component analysis. *Neural Computing Surveys*, 2:94–128, 1999.

[Joe, 1989a]H. Joe. Estimation of entropy and other functionals of a multivariate density. *Ann. Inst. Statist. Math.*, 41:683–697, 1989.

[Joe, 1989b]H. Joe. Relative entropy measures of multivariate dependence. *J. Am. Statist. Assoc.*, 84:157–164, 1989.

[Kantz and Schreiber, 1997]H. Kantz and T. Schreiber. *Nonlinear time series analysis.* Cambridge University Press, 1997.

[Kojadinovic, 2002]I. Kojadinovic. *Modeling interaction phenomena using non additive measures : applications in data analysis.* PhD thesis, Université de La Réunion, France, 2002.

[Kojadinovic, 2004a]I. Kojadinovic. Agglomerative hierarchical clustering of continuous variables based on mutual information. *Computational Statistics and Data Analysis*, 46:269–294, 2004.

[Kojadinovic, 2004b]I. Kojadinovic. Relevance measures for subset variable selection in regression problems based on k-additive mutual information. *Computational Statistics and Data Analysis*, 2004. In Press.

[Kullback and Leibler, 1951]S. Kullback and R. A. Leibler. On information and sufficiency. *Ann. Math. Stat.*, 22:79–86, 1951.

[Kus, 1999]V. Kus. *Divergences and Generalized Score Functions in Statistical Inference.* PhD thesis, Czech Technical University, Prague, Czech Republic, 1999.

[Lerman, 1981]I.C. Lerman. *Classification et Analyse Ordinale de Données.* Dunod, Paris, 1981.

[Menéndez *et al.*, 1995]M.L. Menéndez, D. Morales, L. Pardo, and M. Salicrú. Asymptotic behaviour and statistical applications of divergence measures in multinomial populations: a unified study. *Statistical papers*, 36:1–29, 1995.

[Morales *et al.*, 1996]D. Morales, L. Pardo, and I. Vajda. Uncertainty of discrete stochastic systems: general theory and statistical theory. *IEEE Trans. on System, Man and Cybernetics*, 26(11):1–17, 1996.

[Rényi, 1959]A. Rényi. On measures of dependance. *Acta Mathematica Academiae Scientiarium Hungaricae*, 10:441–451, 1959.

[Rényi, 1965]A. Rényi. On the foundations of information theory. *Review of the International Statistical Institute*, 33(1):1–14, 1965.

[Rota, 1964]Gian-Carlo Rota. On the foundations of combinatorial theory. I. Theory of Möbius functions. *Z. Wahrscheinlichkeitstheorie und Verw. Gebiete*, 2:340–368 (1964), 1964.

[Roulston, 1999]M.S. Roulston. Estimating the errors on measured entropy and mutual information. *Phyisca D*, 125:285–294, 1999.

[Saporta, 1990]G. Saporta. *Probabilités, Analyse de Données et Statistique.* Editions Technip, Paris, 1990.

[Särndal, 1974]C.E. Särndal. A comparative study of association measures. *Psychometrika*, 39:165–187, 1974.

[Scott, 1992]D.W. Scott. *Multivariate Density Estimation: Theory, Practice, and Visualization.* Wiley Intersciences, 1992.

[Silverman, 1986]B. Silverman. *Density Estimation for Statistics and Data Analysis.* Chapman and Hall, New-York, 1986.

[Ullah, 1996]A. Ullah. Entropy, divergence and distance measures with econometric applications. *Journal of Statistical Planning and Inference*, 49:137–162, 1996.

[Wienholt and Sendhoff, 1996]W. Wienholt and B. Sendhoff. How to determine the redundancy of noisy chaotic time series. *International Journal of Bifurcation and Chaos*, 6(1):101–117, 1996.

[Yei and Liu, 2003]L. Yei and H. Liu. Feature selection for high-dimensional data : A fast correlation-based filter solution. In *Twentieth International Conference on Machine Learning (ICML 2003)*, pages 856–863, Washington, USA, 2003. AAAI Press.

Multivariate Statistical Analysis Approach in Modeling the Job Related Satisfaction

Marta Dziechciarz

Wrocław University of Economics
Komandorska 118/120
53-345 Wrocław, Poland
(e-mail: marta.dziechciarz@ae.wroc.pl)

Abstract. The main purpose of the study is to analyse and to model the perception of job related satisfaction. The survey of employee satisfaction was conducted to gain an understanding of employee's needs, opinions, concerns, skills, and perception of the organization. Paper shows the application of the cluster analysis framework for the employee classification in the organization. The classification into three clusters was chosen. The k-means method has been used for classification. According to the procedure proposed by Carmone, Kara, and Maxwell ([Carmone *et al.*, 1999]) – in the first step, 20 variables were used as a base for classification (selected out of complete list of 79 statements). The motivation for those analyses was to improve the quality of human resources management and to diversify the managerial approach towards distinctive groups of employees.

Keywords: Classification, Human Resources Management, Job Related Satisfaction.

1 Introduction

The main purpose of the study is to analyse and to model the perception of job related satisfaction. The level of employee satisfaction determines work quality, loyalty, engagement and identification with company objectives. The survey of employee satisfaction was conducted to gain an understanding of employee's needs, opinions, concerns, skills, and perception of the organization. The motivation for those analyses was to improve the quality of human resources management and to diversify the managerial approach towards distinctive groups of employees.

Analyses covered key areas of HR management – motivation and communication, evaluation of working conditions and working climate, measurement of attitudes and opinions. Data collected with the use of various measurements scales were analysed with the multivariate statistical techniques. Additionally, the classification of employees has been done.

Job related satisfaction is defined as positive attitude of employee towards the company, duties and co–workers (see [Levy Garboua and MontMarquette, 2001], and [Freeman, 1978]). The level of the perception of job related satisfaction is strongly related with the difference between subjective evaluation of the existing situation and the expectations. Employee's satisfaction level does not always reflect real situation, more often is determined by working conditions perception. It is assumed, that highly satisfied employee works harder, more efficiently, and with less slack and waste than unhappy person. He (she) is also more innovative and entrepreneurial, having the interest of the company at heart, given that they see their own satisfaction intimately and directly tied to the success of the company. Furthermore, it is often assumed that employees' satisfaction has a direct, positive impact on functioning of whole organization. It cannot be denied, that ineffective human resources management very often causes lack of job related satisfaction. It can be result of bad compensations system, unfair motivation system, ineffective internal communication or bad working climate [Armstrong, 1996]. Thus, those areas should be constantly monitored.

Companies, in order to measure satisfaction level, often conduct a questionnaire survey. Employees' opinion survey became an essential component of organizational culture and provides a picture of organizations' need for managerial improvements. These surveys can be used to solicit employee opinion on a variety of issues such as the company's success in communicating its mission to employees, or local issues such as quality of the working environment. A survey makes it possible to gather responses from individuals who would otherwise surely be hesitant to speak out their opinions and suggestions.

The findings and recommendations from the employee satisfaction survey enable human resources department to make essential improvements. Furthermore, identification of strengths and weaknesses is possible. Additionally, monitoring of job related satisfaction and identifying areas that need improvement could result in retaining highly performing staff members. Appropriate recommendations can help in eliminating of existing and potential problems and threats. Moreover, the results of feedback process provide an understanding how the employee perceives the organization along different dimensions. The role of the feedback can be summarized as follow:

- Is essential in facilitating development and organizational change.
- Allows the organization to focus on needs and leverage its strengths.
- Informs the organization management on which actions could create problems for the employees.
- Provides management with employees' feedback (both positive and negative) on the internal health of the organization.
- Measures the impact of current programs, policies and procedures.
- Can be used to motivate employees and improve job satisfaction.

Summing up, an employee satisfaction survey is essential source of information about organization and enables recognition of the perception of the job related satisfaction and employee happiness (in terms of every aspect in the organization). Moreover, it gives the opportunity to identify most motivating and de–motivating factors.

2 Surveyed Organization

The analysed company, a branch of multinational group is one of the biggest manufacturers of chassis systems on local market. Its products are sold to car manufactures, such as Fiat, Ford and General Motors. In view of long tradition and specific production profile, the company put special emphasis on the quality, thus plant has ISO/TS 16949 and ISO 14001 quality certificates.

The plant has long tradition, but in the recent period several times has been taken over by various multinational corporations. Hence, the organization culture has been destabilized. Consequently, whole organization suffers declining morale and motivation among employees. As a side effect, the perception of job related satisfaction also declined. To tackle the negative symptoms observed in the company, the least effective areas in the organisation had to be identified. Therefore an employee satisfaction survey was conducted.

3 Survey Questionnaire

The survey measures facets of the organization that employees feel satisfied with, those which are viewed less favourably, and opportunities for improvement. Hence, the survey questionnaire included items on the following areas:

- Motivation system.
- Internal communication and relation with superior.
- Job climate.
- Attitude to the company and duties.

To identify specific groups of employees, 568 people were surveyed. Over 216 responses were collected. This amounts to approximately 38% of employees. Relatively high percentage of persons working in supportive production took part in the survey, at the same time direct production employees participated the least. Furthermore, comparatively many young employees (26 to 32 years old) working between 2 and 5 years partake in survey. On the other hand, employees with job seniority above 16 years and older than 47 years old, did not respond in large numbers. Usually, this kind of survey is carried out with use of questionnaires, interviews and simply by observations. A questionnaire used for this employee opinion survey contained items that were rated on a five–point scale. Respondents were asked to choose category which best corresponds to their attitudes on a five point Likert scale: 1 – Strongly disagree, 2 – Disagree, 3 – No opinion, 4 – Agree, 5 – Strongly agree.

These items have been developed to measure different dimensions of the organization – communication, motivation, job climate and leadership.

Table 1. The survey's questions concern four major issues, which have strong effect on job satisfaction in organization. In order to find those factors, the survey covered major issues, which presumably have negative influence on the level of satisfaction perception.

	Current employment		**Surveyed employees**	
Employment category	Number of employees	Rate	Number of employees	Rate
1. Direct production	392	69%	123	57%
2. Indirect production	86	15%	64	30%
3. Administration	90	16%	29	13%
Job seniority				
1. Up to 1 year	132	23%	41	19%
2. Between 2 and 5 years	73	13%	62	29%
3. Between 6 and 10 years	12	2%	6	3%
4. Between11 and 15 years	26	5%	6	3%
5. Between 16 and 20 years	77	14%	21	10%
6. Longer than 21 years	248	44%	80	37%
Sex				
Female	103	18%	46	21%
Male	465	82%	170	79%
Age				
1. Up to 25 years old	90	16%	43	20%
2. Between 26 and 32 years old	108	19%	59	27%
3. Between 33 and 39 years old	60	11%	23	11%
4. Between 40 and 46 years old	130	23%	45	21%
5. Between 47 and 53 years old	147	26%	41	19%
6. Older than 54 years	33	6%	5	2%
Position				
1. Managerial	40	7%	17	8%
2. Blue–collar	528	93%	199	92%
Type of employment contract				
1. Contracted worker	478	84%	178	82%
2. Temporary worker	90	16%	38	18%
Total	**568**		**216**	

4 Classification of Employees

In the analysis it was highly important to identify, if there are groups of similar employees and to distinguish differences between those identified groups. As the criterion for clustering, the descriptive variables (questionnaire statements), that expressed employees' opinion about issues important for the company were used. In accordance with the character of the variables, the k-means technique was chosen for clustering. An extensive discussion on classification techniques may be found in several works dealing with k-means clustering algorithm (see [Gatnar and Walesiak, 2004], [Gordon, 1999], [Walesiak, 1996])

Although it limits the usefulness of the survey and the analysis and additionally limits the possibility of knowledge discovery – because of the explicitly expressed managerial needs formulated by the Human Resources Department, the classification into three clusters has been chosen. In the survey there were 79 variables. Not all were used for the analysis. The variables selection was based on the Heuristic Identification of Noisy Variables (HINoV) algorithm which uses Hubert and Arabie's ([Hubert and Arabie, 1985]) adjusted Rand index and k-means method of classification. Carmone, Cara, and Maxwell (1999, p. 507) demonstrate that using HINoV the results with simulated data helps identify noisy variables. Clustering objects based on only the non-noisy variables give better cluster recovery. The HINoV algorithm contains the following steps:

1. The starting point is data matrix for the analyzed objects $A = \{A_1, \ldots, A_n\}$ and for set of descriptive variables $M = \{M_1, \ldots, M_m\}$.

2. For each variable j the $k-$means clustering is performed. The arbitral (required) number of clusters is requested.

3. Hubert and Arabie's adjusted Rand indices R_{jl} $(j, l = 1, \ldots,\ m)$ for all $j \neq l)$ are calculated. Since R_{jl} are symmetric – $m(m-1)/2$ values are to be calculated.

4. The $(m \times m)$ matrix of adjusted Rand indices R_{jl} $(j, l = 1, \ldots,\ m)$ is constructed. Each row (or column) is summed up $R_{j\bullet} = \sum\limits_{l=1}^{m} R_{jl}$.

5. The measures (sums) $R_{1\bullet},\ R_{2\bullet}, \ldots,\ R_{m\bullet}$ are ranked. By analyzing scree diagram (looking for a kink point) – the subset of $h$variables with highest contribution to the cluster structure is selected. The value of the $R_{j\bullet}$ indicate the contribution of that variable to the cluster structure. Relatively low values of $R_{j\bullet}$ indicate noisy variables – those $(m-h)$ variables with insignificant contribution to the cluster structure are eliminated from the further analysis.

6. Rerun the $k-$means clustering of the set of analyzed objects $A = \{A_1, \ldots, A_n\}$ using only $h$selected variables.

As a result of the HINoV algorithm application, out of the complete list of 79 statements (variables), a sub–set of twenty variables was selected for cluster analysis.

Resulting classification of employees into three groups gave classes with 68, 85 and 63 respondents respectively. Third class, the smallest contains employees with somehow ambivalent opinions. One may notice that there is slight under–representation of men (of –5,1%). Employees working in the administration (white–collar) are overrepresented (14,9%). When job seniority is considered, there is substantial under–representation of employees working 11–15, and 16–20 years (–33,3% and –19,0% respectively). It receives reflection in the age of the class members.

Very interesting is the character of the class number one and two, which might be labeled *satisfied* and *dissatisfied.* In the class number one (*satisfied*) – the mean value in as many as 17 statements (out of twenty), reaches maximal value. In contrast – all 20 variables has minimal average value in the class number two (*dissatisfied*). The class, which is called *satisfied* has the structure very similar to the structure of the whole sample with respect to three characteristics – *employment category, position* and *sex.* There is substantial overrepresentation (46,7%) of the oldest employees in this class. When job seniority is considered, there is substantial overrepresentation of employees working less than one year, 11–15, and 16–20 years (22,8%, 16,7% and 9,5% respectively). Unexpectedly there is overrepresentation (of 8,8%) of temporary employees in this class. The class, which is called *dissatisfied* has the structure considerably different from the structure of the whole sample. There is substantial overrepresentation of *direct* and *supportive production* workers (8,1% and 8,9%) in this class. Thus the male employees are overrepresented by 7,3%.

5 Summary and Managerial Recommendations

The multivariate analysis and clustering of employees proved that there is clear distinction into at least three categories of employees. There is interesting impact resulting from the analysis for decision makers in the Enterprise. The most important is the information that there is need for HR policy differentiation. From the HR management point of view – the *satisfied* people do not require as much attention as those *dissatisfied.* Therefore it is clear that employees belonging to the class number two (*dissatisfied*) need special consideration. In order to formulate managerial recommendation – the analysis of the class number two (*dissatisfied*) is necessary. The lowest mean values may be observed for statements 18, 13, 20, 19 (mean values: 1,26; 1,45; 1,56; 1,64 respectively). The statements analysis shows that the most troublesome issues are connected with compensation and motivation system. This conclusion may be strengthen by the fact, that also in the class number one (*satisfied*) respective mean values are also low (although the highest in three classes). The corresponding values are: 2,96; 3,29; 3,19; and 3,51. Mean values in this class for other statements reach up to 4,71 (statement 17), 4,54 (statement 8).

References

[Armstrong, 1996]M. Armstrong. *A Handbook of Human Resources Management Practice.* WPSB, Kraków, 1996.

Table 2. Characteristics of the clusters

	Specification	Class		
		I	II	III
	Number of employees in each class	68	85	63
	Statements	Means		
1	Internal communication in company is functioning properly	3,69	2,60	2,68
2	I am always properly informed on results of my job evaluation	3,72	2,13	2,65
3	There is no exaggeration in assessing my behaviour	3,31	2,42	3,24
4	I could honestly recommend my company to my acquaintances as a good place to work	4,18	2,29	3,22
5	I am aware on my company condition and its future plans	3,65	2,15	3,78
6	My boss appreciates when I work well	3,84	2,08	2,48
7	In the company creative and energetic people are being promoted	3,94	1,85	3,24
8	I care for my company image	4,54	3,65	4,00
9	I am informed on objectives and tasks planed for my department for this year	4,01	2,52	3,29
10	I have never experienced not ethical behaviour	3,56	2,38	3,94
11	I consider my employment secure and stable	3,81	1,91	3,16
12	Our company cares for employees	3,90	1,76	2,40
13	I receive fair compensation for my efforts	3,29	1,45	1,97
14	In our department there is no unfair rivalry between co–workers	3,75	2,86	3,84
15	In case of difficulties I can surely count on my co–workers help	4,49	4,09	4,44
16	When I seek information for my work, I know where I can find it	4,43	3,24	3,90
17	I know what is expected from me on my workplace	4,71	4,15	4,32
18	In my company employees are compensated according to their achievements	2,96	1,26	2,11
19	I clearly understand compensation policy	3,51	1,64	2,10
20	The compensation system stimulates employees involvement and efficiency	3,19	1,56	1,87

[Carmone *et al.*, 1999]F. Carmone, A. Kara, and S. Maxwell. Hinov: A new model to improve market segments definition by identifying noisy variables. *Journal of Marketing Research*, 1999.

[Freeman, 1978]R. Freeman. Job satisfaction as an economic variable. *American Economic Review*, pages 135–141, 1978.

[Gatnar and Walesiak, 2004]E. Gatnar and M. Walesiak. *Metody statystycznej analizy wielowymiarowej w badaniach marketingowych [Multivariate statistical analysis in marketing research].* Wydawnictwo AE, Wrocław, 2004.

[Gordon, 1999]A. Gordon. *Classification.* Chapman & Hall, London, 1999.

[Hubert and Arabie, 1985]L. Hubert and P. Arabie. Comparing partitions. *Journal of Classification*, pages 193–218, 1985.

[Levy Garboua and MontMarquette, 2001]L. Levy Garboua and C. MontMarquette. *Satisfaction judgement and utility analysis.* Universite Paris, Paris, 2001.

[Walesiak, 1996]M. Walesiak. *Metody analizy danych marketingowych [Methods of market data analysis].* PWN, Warszawa, 1996.

Customer satisfaction and PLS structural equation modeling. An application to automobile market

Valentina Stan and Gilbert Saporta

Conservatoire National des Arts et Métiers
292 Rue Saint Martin
F 75141 Paris Cedex 03
(e-mail: valentina_titu@yahoo.fr, saporta@cnam.fr)

Abstract. We present the principal concepts of structural equation modeling and a comparison between the two main approaches: PLS (Partial Least Square) and LISREL (Linear Structural Relationship). A structural model uses 2 types of models: the measurement model (outer model) and the structural model (inner model). An application to real life data on customer satisfaction is given.

Keywords: Structural equation modeling, Partial least square, PLS approach.

1 An introduction to structural equation modeling

1.1 General considerations

Let p variables be observed upon n units. The p variables are partitioned in J subsets or blocks of k_j variables which are presumed to be pertinent for describing the phenomenon. Each of these blocks is designed to describe a theme of the general phenomenon. We shall designate these blocs by X_j and we shall consider them as matrices with dimension $(n \times k_j)$. In structural models the observed variables are called manifest variables. The latent variables are not observable: they exist by the relations they have with the manifest variables. In the following we shall always suppose that each block is associated with only one latent variable (unidimensionality). Therefore we can identify the blocks by the same name as their latent variable. The latent variable corresponding to the X_j block will be designated by ξ_j. A structural model needs 2 types of models: the measurement model (outer model) which connects the manifest variables to the latent variables and the structural model (inner model) which connects latent variables between them.

1.1.1 The measurement model (outer model) After having determined the blocks, we must specify the type of relationship between latent variables and manifest variables which correspond to block X_j. There are 3 ways: the reflective way, the formative way, the MIMIC way (Multiple effect Indicators for Multiple Causes).

The reflective way In this way, the manifest variables are considered like the "reflection" of their latent variables [Tenenhaus *et al.*, 2005]. This kind of situation exists for instance in models which analyse customer satisfaction of a particular kind of service: a set of questions about the image of the service which represents a latent variable in the model. Each manifest variable is related to its latent variable, as follows:

$$x_{jh} = \pi^0_{jh} + \pi_{jh}\xi_j + \epsilon_{jh} \quad \forall h = 1 \dots k_j$$

π^0_{jh} = constant term; π_{jh} = regression coefficient; ϵ_{jh} = residual term.

The formative way Here the latent variables represents the "reflection" of the manifest variables which belong to block X_j, and are thus a result of these [Tenenhaus and al., 2005]. In this type, the latent variable is a linear function of the manifest variables which generate it:

$$\xi_j = \Sigma^{k_j}_{h=1}\varpi_{jh}x_{jh+\delta_j}$$

;

$\varpi_{jh} \quad (h = 1 \dots h_j)$= multiple regression coefficients of ξ_j on ; δ_j = residual term.

1.1.2 The structural model (inner model)

Opposite to the measurement model, which deals with the relations between latent variables and their manifest, the structural model concerns the mode of estimation of latent variable between them. The relations between latent variables have the form:

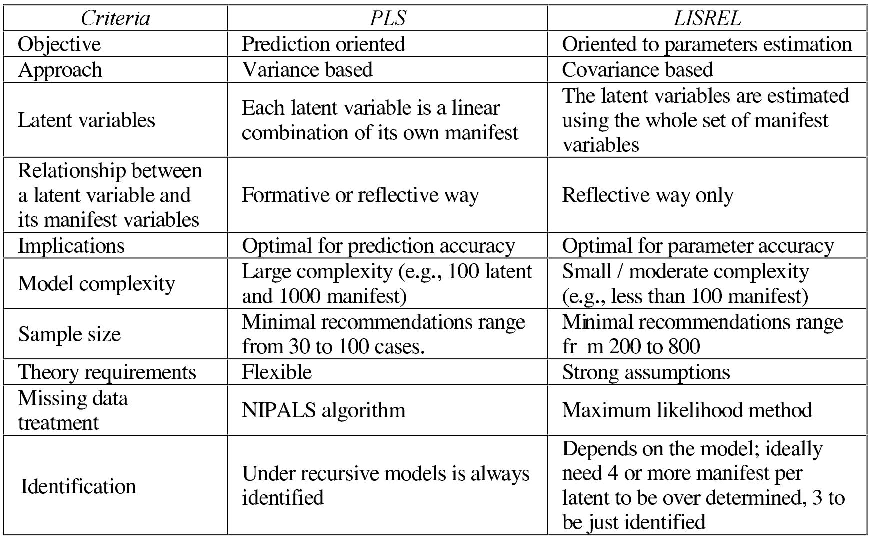
$$\xi_j = \beta^0_j + \Sigma^J_{i=1,i\neq j}\beta_{ji}\xi_i + \zeta_j \quad \forall j = 1 \dots J \tag{1}$$

β^0_j = constant term; β_{ji} = regression coefficient; ζ_j = residual term.

Wold [Wold, 1966] formalized the concept of partial least squares. His algorithm consists in estimating the latent variables (outer estimate and inner estimate) and the structural equations by OLS (Ordinal Least Squares) multiple regression with an iterative process. The initial value of the coefficients being equal to ±1, according to the sign of the correlation between latent variables or between latent and manifest variables.

1.2 A comparison between PLS and LISREL

We will follow here [Jöreskog and Wold, 1982], [Chin, 2000] and [Vinzi, 2003]. In PLS approach, there are less probabilistic hypotheses, data are modeled by a succession of simple or multiple regression and there is no identification problem. On the contrary in LISREL, the estimation is done by maximum likelihood, based on the hypothesis of multinormality and allows the modelisation of the variance-covariance matrix. However, identification problems and non-convergence of the algorithm are sometimes encountered. The differences between the estimations for a causal model using PLS and LISREL depends on the order in which the parameters of the model and latent variables are computed. For PLS the estimated latent variables are first computed by making them belong to the space spanned by their manifest variables. The model parameters are computed by using OLS multiple regression. With LISREL, one computes the model parameters by maximum likelihood and impose some constraints on latent variables. Consequently, the structural equations are more significant in LISREL than in PLS (the R^2 are larger) and the correlations between the manifest variables and their latent are larger in PLS. In LISREL approach, each latent variable is estimated by multiple regression, using all manifest variables. In PLS, latent variables are calculated as a linear combination of the associated manifest variables. PLS favours the outer model and LISREL the inner model. The table 1 summarizes criteria for choosing between PLS and LISREL.



Criteria	*PLS*	*LISREL*
Objective	Prediction oriented	Oriented to parameters estimation
Approach	Variance based	Covariance based
Latent variables	Each latent variable is a linear combination of its own manifest	The latent variables are estimated using the whole set of manifest variables
Relationship between a latent variable and its manifest variables	Formative or reflective way	Reflective way only
Implications	Optimal for prediction accuracy	Optimal for parameter accuracy
Model complexity	Large complexity (e.g., 100 latent and 1000 manifest)	Small / moderate complexity (e.g., less than 100 manifest)
Sample size	Minimal recommendations range from 30 to 100 cases.	Minimal recommendations range fr m 200 to 800
Theory requirements	Flexible	Strong assumptions
Missing data treatment	NIPALS algorithm	Maximum likelihood method
Identification	Under recursive models is always identified	Depends on the model; ideally need 4 or more manifest per latent to be over determined, 3 to be just identified

Table 1. Criteria for choosing between PLS and LISREL.

2 Practical application

2.1 Satisfaction in automobile market

Taking into account that the PLS approach is less used than LISREL in marketing research, even though it is more advantageous than the latter, our objective was to introduce how PLS works and to show its' capacities. To reach this goal, we used data provided by the PSA Company (Peugeot Citroën) on customers' satisfaction. We used the experimental PLSX module of the SPAD software, which has been developed within the framework of the ESIS project about the construction of a tool to analyze European customer satisfaction.

2.2 The questionnaire

The data obtained by questionnaire (which is confidential) represents satisfaction scores (with the scale of 1 to 10) on about thirty services. 2922 customers participated. Manifest variables are the followings (table 2):

Variable		Variable	
General satisfaction	S01	Radio - CD - rom	S17
General quality	S02	Heating - ventilation	S18
Quality -price ratio	S03	Boot capacity	S19
Absence of small, irritating defects	S04	Security	S20
Absence of noise and vibrations	S05	Braking	S21
General state of the paintwork	S06	Acceleration	S22
Robustness of commands, buttons	S33	Handling	S23
Solidity and robustness	S08	Suspension comfort	S24
Lock, door and window mechanisms	S09	Silence in rolling	S25
Inside space and seat modularity	S34	Maniability	S26
Inside habitability	S11	Direction	S27
Dashboard: quality of materials and finishing	S12	Gears	S28
Insider: quality of mat. and finishing	S13	Mechanic reliability	S29
Front seat comfort	S14	Oil consumption	S30
Driving position	S15	Mechanic's efficiency in solving problems	S31
Visibility from driver's seat	S16	Maintenance cost and repairs	S32

Table 2. Manifest variables.

Since we are interested in the relationships between variables, and not in their values, it was not necessary to rescale the answers, despite the fact that customers do not use the scale in the same way.

3 The analysis

3.1 Blocks building

We first had to partition the manifest variables (MV) into homogenous blocks, each one being explicitly associated with only one latent variable. After many trials and with the help of experts, we considered the following division of the 32 variables into 6 blocks (table 3):

3.2 The causality scheme

The measurement model has been established in the previous paragraph.

3.2.1 The structural model (inner model) Supposing that the themes reflect correctly the characteristics of the satisfaction, we must then propose relations between these themes, so as to explain the latent variable "general satisfaction". In the figure 1 we can visualize the structural model which shows the relations between the latent variables:

Block	Label	MV	Block	Label	MV	Block	Label	MV
General satisfaction	Gsat	S01	Internal comfort	Comf	S11	Driving quality	Drivq	S20
		S02			S12			S21
		S03			S13			S22
Construct quality	Conq	S04			S14			S23
		S05			S15			S24
		S06			S16			S25
		S28			S17			S26
		S29			S18			S27
Solidity	Soli	S08			S19	Costs	Costs	S30
		S09			S34			S31
		S33						S32

Table 3. The 6 blocks of manifest variables.

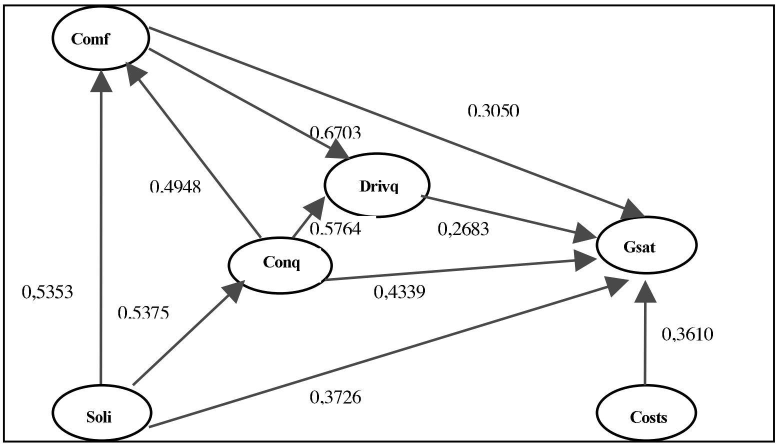


Fig. 1. The causality scheme with correlations values between latent variables.

3.3 Results and interpretations

We see that the variable "construction quality" is the most important variable for the "general satisfaction" (the correlation coefficient is 0,4339) and the less important is the "driving quality" (the correlation coefficient is 0,2683). Consequently, in order to increase the general satisfaction of the client, the producer should concentrate firstly on the "construction quality" and then on the "solidity", "costs", "internal comfort" and "driving quality". Let us now interpret the results in detail.

3.3.1 The measurement model After convergence of the PLS algorithm, one obtains the final weights which allow us to link the manifest variables with the latent variables:

$$\begin{aligned}
Gsat &= 0,2188 \times S01 + 0,5746 \times S02 + 0,4850 \times S03\\
Soli &= 0,4682 \times S08 + 0,4242 \times S09 + 0,4151 \times S33\\
Conq &= 0,2103 \times S04 + 0,2730 \times S05 + 0,3396 \times S06 + 0,3930 \times S28\\
&\quad +0,3787 \times S29\\
Drivq &= 0,1962 \times S20 + 0,1595 \times S21 + 0,1415 \times S22 + 0,1615 \times S23\\
&\quad +0,1775 \times S24 + 0,1658 \times S25 + 0,1728 \times S26 + 0,1805 \times S27\\
Comf &= 0,1492 \times S11 + 0,1795 \times S12 + 0,1756 \times S13 + 0,1542 \times S14\\
&\quad +0,1667 \times S15 + 0,1424 \times S16 + 0,1282 \times S17 + 0,1457 \times S18\\
&\quad +0,1092 \times S19 + 0,1513 \times S34\\
Costs &= 0,2396 \times S30 + 0,5707 \times S31 + 0,5042 \times S32
\end{aligned}$$

Table 4 presents only correlations larger than the mean of the absolute values (0,3723):

We observe that all latent variables are well correlated with their own manifest. So, the manifest variables "describe" their latent appropriately and the blocks are therefore validated. We see also that the largest correlation (0,8692) is between "general satisfaction" and their manifest "quality in general".

The R^2 coefficients between connected latent variables are:

$$\begin{aligned}
R^2(Conq; Soli) &= 0,2889\\
R^2(Comf; (Soli, Conq)) &= 0,3468\\
R^2(Drivq; (Conq, Comf)) &= 0,5286\\
R^2(Gsat; (Soli, Conq, Comf, Drivq, Costs)) &= 0,2516
\end{aligned}$$

In this table the most interesting relation concerns the "general satisfaction". For this variable, the R^2 coefficient generated by the other latent variables is 25%, and we consider that as satisfactory because there are 2922 individuals. The correlations between the latent variables are given below in table 5.

We can see that to improve "internal comfort", the producer should concentrate on "solidity" (correlation coefficient = 0,5353) and on the "construction quality" (0,4948). The producer?s efforts for improving "construction quality" also greatly affect the variable "leading quality" (0,5764). In order to obtain a good "construction quality" the producer could concentrate on "solidity" (0,5375).

We also observe an important correlation between "solidity" and "driving quality". We have chosen not to establish a relation between these two because this relation does not in any way influence the model. Given the causality scheme the determination of "general satisfaction" is a complex procedure in which almost all the latent variables are directly involved. The equation is as follows:

$$\begin{aligned}
Gsat = {} & 0,2721 \times Conq + 0,1678 \times Soli + 0,198 \times Costs\\
& +0,082 \times Comf + 0,095 \times Drivq
\end{aligned}$$

Variable	Solidity	Construct quality	Internal comfort	Driving quality	Costs	General satisfaction
S08	*0,7988*	0,4272	0,4568	0,4539		
S09	*0,7492*	0,3951	0,4040			
S33	*0,7425*	0,4093				
S04		*0,5456*				
S05		*0,5847*				
S06	0,4187	*0,5897*				
S28		*0,6666*	0,4336	0,5423		
S29		*0,6954*		0,4608		
S11			*0,6965*	0,4405		
S12	0,4173	0,3987	*0,7276*	0,4684		
S13	0,4089	0,3882	*0,7317*	0,4441		
S14			*0,7382*	0,4829		
S15			*0,7662*	0,5398		
S16			*0,6278*	0,4427		
S17			*0,5055*	0,3784		
S18			*0,5741*	0,4419		
S19			*0,5379*	0,3746		
S34	0,3956		*0,6535*	0,4048		
S20	0,4556	0,4276	0,6083	*0,7303*		
S21	0,3796	0,4079	0,4520	*0,6878*		
S22			0,4385	*0,6866*		
S23		0,3929	0,4903	*0,7914*		
S24		0,4267	0,5029	*0,7600*		
S25		0,4397	0,4051	*0,6424*		
S26		0,4575	0,5104	*0,7980*		
S27	0,3749	0,4826	0,5155	*0,7913*		
S30				0,3880	*0,4450*	
S31					*0,8374*	
S32					*0,8242*	
S01						*0,6399*
S02		0,4234				*0,8692*
S03						*0,7433*

Table 4. Correlations between manifest and latent variables.

Variable	Soli	Conq	Comf	Drivq	Costs	Gsat
Soli	1					
Conq	0,5375	1				
Comf	0,5353	0,4948	1			
Drivq	0,4949	0,5764	0,6703	1		
Costs	0,3116	0,4331	0,3144	0,3461	1	
Gsat	0,3726	0,4339	0,305	0,2683	0,361	1

Table 5. Correlations between latent variables.

The negative coefficient for "driving quality" can be explained by the fact that this variable increases with "construction quality" and the regression coefficient between "construction quality" and "general satisfaction" is 0,2721.

This multiplication coefficient is without doubt corrected by the negative coefficient on the "driving quality".

4 Conclusions

Firstly it must be underlined that this study did not follow the logical sequence of steps of the PLS approach: the construction of a model by experts, the construction of a questionnaire using this model, and the collection of customer data using this questionnaire. In our case, we have inverted the process: we have tried to build a model using data that had already been collected with a questionnaire. This fact has obviously effects on the final results which cannot be precisely measured. A hierarchy of the influence of the latent variables on general satisfaction can be established using the structural model: I. Construction quality; II. Solidity; III. Costs; IV. Internal comfort; V. Driving quality. The results obtained for general satisfaction are satisfactory: $R^2 = 25\%$ which is a good result for a large sample of almost 3000 respondents.

References

[Chin, 2000]W. W. Chin. *Partial Least Square for researchers: a overview and presentation of recent advances using the PLS approach*, http://discnt.cba.uh.edu/chin/indx.html, 2000

[Jöreskog and Wold, 1982]K. G. Jöreskog, H. Wold. *The ML and PLS techniques for modeling with latent variables: historical and competitive aspects.* In K. G. Jöreskog and H. Wold, editors, *Systems under indirect observation*, Part 1, North-Holland, Amsterdam, pages 263–270, 1982

[Tenenhaus *et al.*, 2005]M. Tenenhaus, V. E. Vinzi, Y-M. Chatelin, C. Lauro. *PLS path modeling.* Computational Statistics & Data Analysis, volume 48, issue 1, pages 159–205, January 2005

[Vinzi, 2003]V. E Vinzi. *The PLS Approach to Path Modeling.* IASC-IFCS Summer School, Lisbon, 2003

[Wold, 1966]H. Wold. *Estimation of principal components and related models by iterative least squares*, in Krishnaiah, P.R., editor, Multivariate Analysis, Academic Press, New York, pages 391–420, 1966

Combining Correspondence Analysis And Spatial Statistics For River Water Quality Assessment And Prediction

Henrique Garcia Pereira[1] and Jorge Ribeiro[1,2]

[1] CVRM - GeoSystems Centre of IST
Av. Rovisco Pais, 1
1049-001 Lisboa
(e-mail: hpereira@alfa.ist.utl.pt)

[2] Faculdade de Arquitectura da Universidade Técnica de Lisboa
Rua Prof. Cid dos Santos
Pólo Universitário do Alto da Ajuda
1349-055 Lisboa
(e-mail: jribeiro@fa.utl.pt or jribeiro@alfa.ist.utl.pt)

Abstract. In the context of a practical case study regarding an environment application, a methodology for river water quality assessment and prediction was developed. Such a methodology consists of calculating a quality index by correspondence analysis and predicting its value at non-sampled locations by spatial statistics.

Keywords: Environment, Water Quality Index, Correspondence Analysis, Cumulative Variogram, Kriging.

1 Introduction

When a river is submitted to anthropic environmental stress, *e.g.*, an industrial discharge, a variety of physical-chemical-biological variables are to be monitored in a series of downstream stations in order to guarantee the quality of its water.

For the sake of control by Environment concerned agencies (both official and NGOs), this disparate set of variables should be summarized in some kind of a global straightforward quality index, easy to be appreciated by public opinion and regulatory institutions. On the other hand, this summary measure should also account for all the available information related to the influence of the discharge onto the river water. Once calculated this index, an assessment can be made on the river water quality. But, obviously, if no modelling procedure is applied in order to provide some sort of prediction, this assessment refers only to the sampled points (the stations where the basic measurements are made).

Since no dispersion deterministic model is prone to be applied to the quality index (no mechanism can be assigned to the dynamics of such a hybrid combination of parameters), a stochastic forecasting methodology should be devised in order to predict the index at any non-sampled point (or domain), as required by the above mentioned Environment concerned agencies. Aiming at approaching this issue from the standpoint of spatial statistics, the standard estimation methodology should be adjusted in order to cope with the specific characteristics of such a problem, where geometry and dynamics play a determinant role. This entails the calculation of a non-Euclidean distance along the river and the development of a non-stationary estimation approach, adjusted to the river flow characteristics.

2 Methodology

The proposed methodology to address this two-fold problem consists of two steps:
In the first step, the barycentric affectation procedure put forward by Benzécri [Benzécri, 1980], and modified by Pereira [Pereira, 1988], was applied in order to produce a comprehensive quality index, ranging from -1 to +1, and accounting for the entire set of variables available at all monitoring stations. For this end, it is required that a panel of experts scrutinizes all measured parameters, split their range into p significant classes, and create two vectors in the variable classes space, designated by the 'GOOD' and 'BAD' poles. These poles represent, respectively, the 'ideal' water quality in its two extremes: pure and polluted water (according to the expert panel). These two 'ideal' vectors are arranged in a 2 x p matrix and submitted to Correspondence Analysis, providing an axis, onto the empirical samples (coded in complete disjunctive form) are projected as supplementary lines. The co-ordinate of each sample in this axis is the required index.
In the second step, the kriging technique, developed in [Matheron, 1965] for the case of space-stationary random functions, was adjusted to the specific features of river water flow according to the guidelines provided in Pereira *et al.* [Pereira *et al.*, 2000]. In particular, the lag for calculation of spatial auto-correlation function - Matheron's variogram - was not measured as an Euclidean distance, but as a 'meandric' one, which is the analogue, for the case of rivers, of the well known 'block distance', used in urban applications. Also, the variogram function and the resultant kriging system were modified to account for the fact that the index at a given point of space along the river depends only on the corresponding upstream values. Hence, a new auto-correlation tool - the cumulative variogram, as proposed by Sen [Sen, 1989] in a different context - was developed in order to avoid any stationarity assumption. This tool - which stands for the Probability Cumulative Function, as the "usual" variogram stands for the Probability Density Function, is defined by:

$$ {}_{w}^{a}\gamma\,[d(i)] \;=\; \sum_{i=1}^{m} (z_w - z_i)^2 \tag{1}$$

where $d(i)$ is the "meandric" distance between w and the station i $(i = 1, ..., m)$ and z_w is the index at the point w, to which the cumulative variogram refers. This tool allows to respect the practical order relationships between stations and points (or domains) to be predicted, as given by the river flow. Based on its auto-correlation with upstream values, the proposed index, viewed as a Regionalized Variable, can be estimated at any downstream non-sampled domain by the modified kriging system given below:

$$\begin{bmatrix} 0 & 0 & \ldots & 0 & 0 & 0 & 1 \\ \gamma_{1w} & 0 & & 0 & 0 & 0 & 1 \\ \gamma_{2w} & \gamma_{21} & & 0 & 0 & 0 & 1 \\ \vdots & \vdots & & 0 & 0 & 0 & 1 \\ \gamma_{iw} & \gamma_{i1} & \ldots & 0 & 0 & 0 & 1 \\ \vdots & \vdots & & \gamma_{(m-1)(m-2)} & 0 & 0 & 1 \\ \gamma_{mw} & \gamma_{m1} & & \gamma_{m(m-2)} & \gamma_{m(m-1)} & 0 & 1 \\ 1 & 1 & \ldots & 1 & 1 & 1 & 0 \end{bmatrix} \cdot \begin{bmatrix} \lambda_w \\ \lambda_1 \\ \lambda_2 \\ \vdots \\ \lambda_i \\ \vdots \\ \lambda_m \\ \mu \end{bmatrix} = \begin{bmatrix} \overline{\gamma}_w \\ \overline{\gamma}_1 \\ \overline{\gamma}_2 \\ \vdots \\ \overline{\gamma}_i \\ \vdots \\ \overline{\gamma}_m \\ 1 \end{bmatrix} \tag{2}$$

where w is the central point of the domain to be estimated on the grounds of i upstream stations $(i = 1, ..., m)$, λ are the kriging weights to be assigned to each sample value to predict the average index in the required domain, γ is the usual variogram deduced from the cumulative one by differentiation, and μ is the Lagrange parameter.

Details of the methodological framework where this step relies are given in Ribeiro [Ribeiro, 1999].

3 Case Study

In order to illustrate the above proposed methodology, a case study referring to the Oeiras River (south of Portugal, Fig. 1) is presented. The river is submitted to an industrial discharge and the non-sampled domain of concern on its water quality is located just before the junction with the main Guadiana River (domain W in Fig. 1). Along Oeiras River, water quality is monitored in a series of sampling stations (Fig. 1), for the variables given in the first column of Table 1. The second column of Table 1 contains the classes constructed by the panel of experts for each one of the variables, in order to define the 'Good' and 'Bad' poles, on which the index calculation relies. In the third and forth columns of Table 1, the weights assigned by experts to each variable modality are given.

Variables	Classes	Good Pole	Bad Pole
Biotic diversity based on macro-invertebrate *taxa*	1	0.00	0.80
	2	0.01	0.14
	3	0.05	0.05
	4	0.14	0.01
	5	0.80	0.00
Dissolved Oxygen (%)	[0 ; 50]	0.00	0.91
	] 50 ; 90]	0.09	0.09
	> 90	0.91	0.00
Temperature (^oC)	[0 ; 20]	0.50	0.02
	] 20 ; 30]	0.50	0.98
pH	[0 ; 6]	0.00	0.50
	] 6 ; 9]	1.00	0.00
	] 9 ; 14]	0.00	0.50
Conductivity (μS/cm)	[0 ; 400]	0.95	0.00
	] 400 ; 1500]	0.05	0.05
	> 1500	0.00	0.95
Chemical Oxygen Deficiency (mg/l)	[0 ; 10]	0.90	0.00
	] 10 ; 40]	0.10	0.10
	> 40	0.00	0.90
Sulphates (mg/l)	[0 ; 400]	0.99	0.01
	] 400 ; 3200]	0.01	0.99
Nitrates (mg/l)	[0 ; 25]	0.97	0.01
	] 25 ; 50]	0.02	0.02
	> 50	0.01	0.97
Phosphates (mg/l)	[0 ; 0.54]	0.96	0.01
	] 0.54 ; 0.94]	0.03	0.03
	> 0.94	0.01	0.96
Cu (mg/l)	[0 ; 0.005]	0.98	0.03
	> 0.005	0.02	0.97
Fe (mg/l)	[0 ; 0.3]	0.96	0.03
	> 0.3	0.04	0.97

Table 1. Weights defining the 'GOOD' and 'BAD' poles for river water quality.

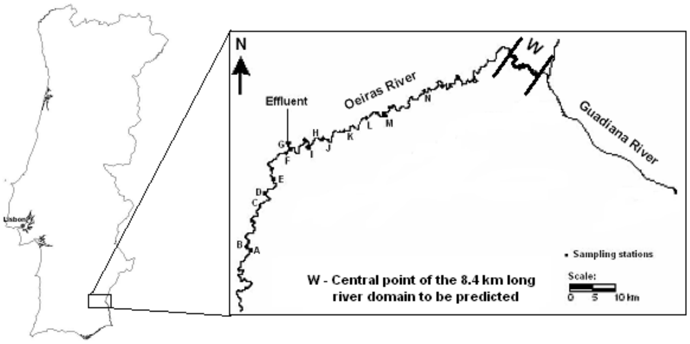


Fig. 1. Location of Oeiras River, sampling stations and domain of concern.

The results of the index calculation for each station according to the first step of the above described methodology are summarized in the histogram of Fig. 2.

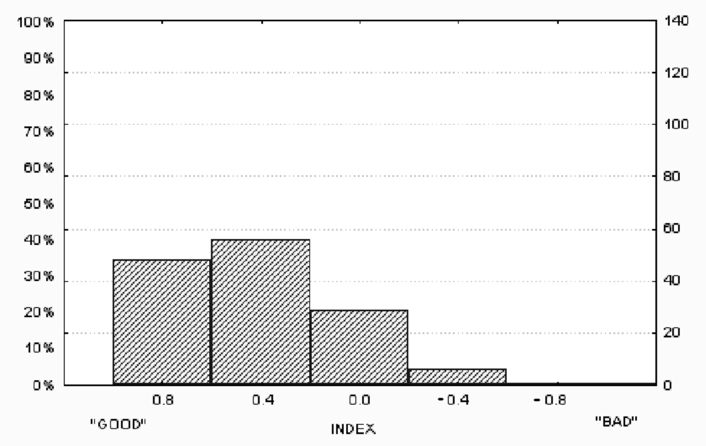


Fig. 2. Histogram summarizing the assignment of the index to the stations.

Since pluviometry can influence the dispersion of pollutants, the index was arranged in each station for the ”wet” and ”dry” months. The evolution of the average index along the river is shown in Fig. 3.

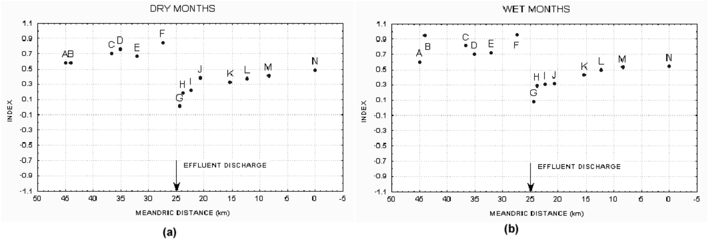


Fig. 3. Schematic representation of the average index for each station (a)-Dry, (b)-Wet months.

Regarding the second step of the methodology, the first point is to calculate the cumulative variogram for each sampling station according to equation 1, as given in Fig. 4.

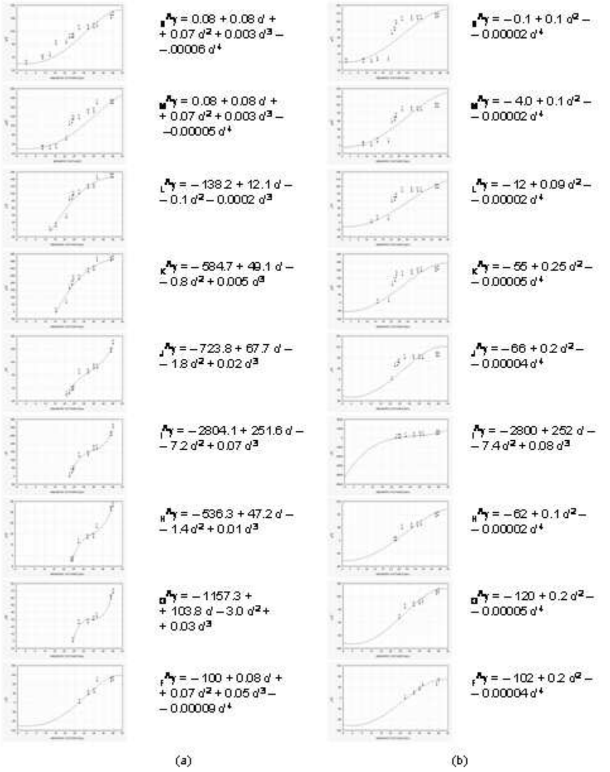


Fig. 4. Cumulative variogram for (a)-Dry and (b)-Wet months.

Differentiating the functions fitted to the curves of Fig. 4, the usual variogram is obtained per station and the system 2 is solved for obtaining the set of λ that permit to predict the value of the index in the non-sampled domain W of Fig. 1 by summing, for all stations, the product of for each λ by the corresponding average index (for dry and wet months).
The results of this calculation are given in Table 2, where the average value of the index in the domain W is compared with the corresponding values in the upstream A-B domain (before the effluent discharge, see Fig. 1).

	Average Index in the Domain W	Average Index in the Domain A-B
Wet Months	0.641	0.788
Dry Months	0.535	0.601

Table 2. Prediction of the Index after and before the effluent discharge.

Table 2 shows that, even though a small decrease in the quality index occurs from A-B to W, the contamination does not reach the Guadiana River, especially in the wet months.

4 Conclusions and Further Work

The proposed methodology allows the estimation of a river water quality index in a non-sampled spatial domain, using all upstream available information.
The point to be developed at this regard is the automatic selection of positive definite functions for the used variogram, obtained by differentiation of the empirical cumulative variogram.
In what concerns the forecasting of the index in time, the length of the available time series (7 years, two samples by year) does not allow any deeper approach than the split into "wet" and "dry" months. Nevertheless, when the time series will have some statistical significance, the parameters of their fitted models can be identified. Then, the same spatial estimation methodology can be applied to these parameters, allowing to predict their values in a non-sampled domain. Finally, the time series at this domain is simulated for future values and a space-time estimation is provided.

5 Aknowledgments

The authors are indebted to UE DG XII, that funded this research in the scope of the RIVERMOD project BE 97-4148 (BRITE - Industrial & Material Technologies).

References

[Benzécri, 1980]J.-P. Benzécri. *Pratique de l'analyse des données. Abrégé théorique - Cas modèle.* Dunod, Paris, 1980.

[Matheron, 1965]G. Matheron. *Les variables regionalisées et leur estimation.* Ed. Masson, Paris, 1965.

[Pereira *et al.*, 2000]H.G. Pereira, J. Ribeiro, A.J. Sousa, L. Ribeiro, A. Lopes, and J. Serôdio. Forecasting river water quality indices. In W.J. Kleingeld and D.G. Krige, editors, *Geostatistics 2000 - Cape Town, Proceedings of 6th International Geostatistics Congress*, pages 591–604. Geostatistical Association of Southern Africa, Cape Town, South Africa, 2000.

[Pereira, 1988]H.G. Pereira. Case study on application of qualitative data analysis to an uranium mineralization. *Quantitative Analysis of Mineral and Energy Resources*, pages 617–624, 1988.

[Ribeiro, 1999]J. Ribeiro. *Formulação de Índices Quantitativos com Base na Descriminação Baricêntrica. PhD Thesis.* IST, Lisbon, 1999.

[Sen, 1989]Z. Sen. Cumulative semivariogram models of regionalized variables. *Math. Geol.*, 21(8):891–903, 1989.

Part VIII

Mathematical statistics

Identification, estimation and control of uncertain dynamic systems: a nonparametric approach

Nadine Hilgert[1], Vivien Rossi[1,2], and Jean-Pierre Vila[1]

[1] UMR Analyse des Systèmes et Biométrie,
ENSA.M - INRA, 2 place Viala, Bât. 29
34060 Montpellier Cedex 1, France
(e-mail: hilgert@ensam.inra.fr, vila@ensam.inra.fr)

[2] I3M, UMR CNRS 5149, Université Montpellier 2,
Case Courrier 051, Place Eugène Bataillon
34095 Montpellier Cedex 5, France
(e-mail: rossiv@ensam.inra.fr)

Abstract. This paper is devoted to a short presentation of the use we did of nonparametric estimation theory for the estimation, filtering and control of uncertain dynamic systems. The fundamental advantage of this approach is its low dependence from any a priori modeling assumptions about uncertain dynamic components. It appears to be of great interest for the control of general discrete-time processes, and in particular biotechnological processes, which are emblematic of nonlinear uncertain and partially observed systems.

Keywords: Discrete-time stochastic systems, Markov controlled processes, Nonparametric identification, Predictive control, Nonlinear filtering, Fault detection.

1 Introduction

This paper is devoted to a survey of the use of nonparametric estimation theory for the estimation, filtering and control of uncertain dynamic systems. It relies on a set of works we have been developing for more than ten years and which emphasizes the efficiency of these nonparametric tools in functional estimation as well as in probability density estimation.

The frame of these developments is that of the control of general discrete-time processes, and in particular biotechnological processes, which are emblematic of nonlinear uncertain and partially observed systems. The field of bioprocess modeling and control offers typical examples of structural time-variations problems which cannot be handled by classic control methods: the dependence of the kinetic coefficients on biomass and substrate state variables is affected by functional fluctuations and not merely parametric ones. In that case, a more appropriate approach would be robust control, in which uncertainty is explicitly accounted for at the beginning of the control design through numerical or functional bounds. However, the performance of the related controllers can be sensitive to settings that are too much conservative or too much optimistic. The nonparametric approach is free from these prior assumptions: through a stochastic learning process, uncertain functional components are progressively and automatically estimated as deterministic or random functions of the measured quantities, in accordance with their actual but unknown and possibly time-varying structures. The use of this functional estimation procedure, compared with the usual and more or less arbitrary choice of these model components, contributes to the reduction of one source of model inadequacy. Moreover, the stochastic frame in which these nonparametric models are designed allows some uncontrolled disturbances such as measurement errors and parameter variations to be accounted for.

In the following we shall present successively application of this nonparametric approach to identification, filtering and control of dynamic systems.

2 Identification and estimation of nonlinear stochastic processes

The uncertain processes under consideration belong to the general class of controlled Markov chains.

They are represented by discrete-time autoregressive models of the following type:

$$X_{t+1} = F_t(X_t, U_t, \varepsilon_{t+1}), \tag{1}$$

where $X_t \in \mathbb{R}^s$, $U_t \in \mathbb{R}^m$ and ε_t are the output, input and noise of the system, respectively. Driving function F_t may be completely or partly unknown, according to the degree of uncertainty in the analytical knowledge of the process. This function may be deterministic or stochastic and is supposed to obey some regularity conditions (see §2.1). Moreover, when the state variable X_t is not observed, an observation model is supposed to be available, of the general form

$$Y_t = G_t(X_t, U_t, \eta_t) \tag{2}$$

where $Y_t \in \mathbb{R}^q$ and G_t is a known function and η_t an observation noise.

Estimating function F_t in model (1) may be intricate. The following particular case with an additive noise is more frequently met in practice:

$$X_{t+1} = f_t(X_t, U_t) + \varepsilon_{t+1}, \tag{3}$$

in which function f_t, from $\mathbb{R}^s \times \mathbb{R}^m$ to $\mathbb{R}^s$, may be completely or partly unknown. We are specifically interested in a type of non-linear models where the control variable U_n acts in a known part of function f_t. They are models of the field of bioprocess modeling and control, and are of form:

$$X_{t+1} = A_t(X_t)g_t(X_t) + B_t(X_t, U_t) + \varepsilon_{t+1}, \tag{4}$$

where A_t and B_t are known functions and g_t is unknown. Function g_t is for example the growth rate of some microorganism population whose concentration is a component of the state variable X_t. The control variable U_t is for example a dilution rate of a polluted water into a bioreactor.

Other examples of model (3) are for instance the evolution models of bacteria populations in food under the influence of environment covariates (U_t), or, in another field, models that describe the position of a space craft under control.

The following subsection is dedicated to the identification of model (3) when unknown (or partially unknown), with state X_t completely observed. The well-known convolution kernel method is applied to estimate function f_t (or only a subpart of it).

In subsection 2.2 state variables X_t are not supposed to be observed anymore and the issue considered is now that of their estimation, *i.e.* filtering, from knowledge of the observed variables Y_t and assuming knowledge of model F_t.

2.1 Identification of the model with convolution kernel estimators

Kernel smoothing methods are among the most reknown nonparametric estimation and prediction methods. They belong to the family of smoothing methods (orthogonal polynomials, splines,...) and are based on a local averaging procedure. They are widely used to estimate probability density functions and regression functions, see [Bosq, 1996].

When the whole function f_t is unknown in model (3), we can consider the following recursive kernel estimator, for all $x \in \mathbb{R}^s$ and $u \in \mathbb{R}^m$:

$$\widehat{f}_t(x,u) = \frac{\sum_{i=0}^{t-1} \delta_{1,i}^{-s} \delta_{2,i}^{-m} K_1(\frac{x-X_i}{\delta_{1,i}}) K_2(\frac{u-U_i}{\delta_{2,i}}) X_{i+1}}{\sum_{i=0}^{t-1} \delta_{1,i}^{-s} \delta_{2,i}^{-m} K_1(\frac{x-X_i}{\delta_{1,i}}) K_2(\frac{u-U_i}{\delta_{2,i}})}, \tag{5}$$

The functions K_1 and K_2 are two kernel functions. They are real positive symmetric functions integrating to one.
The sequences $(\delta_{1,i})$ and $(\delta_{2,i})$, called the bandwidths, have to be positive and decreasing. See [Georgiev, 1984] for the case of an *i.i.d.* sequence (U_t), and [Wagner and Vila, 2001] for a more general situation.

In the case of biotechnological processes, the partially known model (4) is the most frequently met. In that case, the kernel estimation of g_t is given by:

$$\widehat{g}_t(x) = \frac{\sum_{i=0}^{t-1} \delta_i^{-s} K(\frac{x-X_i}{\delta_i}) A_i^{-}(X_i)(X_{i+1} - B_i(X_i, U_i))}{\sum_{i=0}^{t-1} \delta_i^{-s} K(\frac{x-X_i}{\delta_i})}. \tag{6}$$

for all $x \in \mathbb{R}^s$. A_i^- is a general inverse of matrix A_i and K is the kernel function and (δ_i) the bandwidth.

The statistical convergence properties of kernel estimators (5) or (6) have been established under various assumptions about

- the probability distribution of the noise ε,
- the existence of admissible control strategies $(U_t)_{t\geq 1}$ able to stabilize the model (X_t)
- the behaviour of the unknown set of stochastic functions f_t (respectively g_t), which must be quite "stable", corresponding to a convergent sequence f_t (resp. g_t) or an *i.i.d.* functional sequence f_t (resp. g_t).

As regard the bandwidth parameters, the form $\delta_i = \gamma i^{-\alpha}$ is one for which convergence results have been established [Duflo, 1997], [Portier and Oulidi, 2000], [Hilgert *et al.*, 2000]. In some cases, an optimal choice of the bandwidth parameters can be determined by cross validation procedures, see [Vieu, 1991] for instance. From a theoretical point of view, we may distinguish between

- the *a.s.* uniform convergence on compact sets, which requires kernel functions with compact support, as the Epanechnikov kernel for example.
- the stronger *a.s.* convergence on dilated compact sets, which requires positive kernel functions, as the Gaussian kernel for example.

2.2 Estimation of state variables with convolution particle filters

Besides its efficiency in functional estimation of uncertain models as seen in the previous section, the nonparametric approach as proved to be useful as well in probability density estimation of unobserved state variables, *i.e.* in filtering problems.

The objective is now to estimate the unobserved state variable X_t from the analytical knowledge of state model F_t (1) and the observed variables $Y_{1:t} = (Y_1, \cdots, Y_t)$. When F_t and G_t correspond to linear functions of X_t and U_t with additive noises, the well-known Kalman filter provides an optimal estimate of the probability distribution of X_t conditionally to $Y_{1:t}$, $P(X_t|Y_{1:t})$. In the other cases, only the so-called Monte Carlo filters or particle filters (see [Doucet *et al.*, 2001] or [Del Moral, 2004]) provide consistent estimates of $P(X_t|Y_{1:t})$. The main principle of these filters is to build an estimate of $P(X_t|Y_{1:t})$ through the simulation of a large number N of random state particles $\{x_i\}$ which are then weighted according to their likelihoods with respect to the observed variables up to time t.

However the usual particle filters require, in practice, the function G_t to be additive in the observation noise η_t, and the analytic form of the density of η_t to be known.

This last assumption really reduces the applicative potential of these particle filters. The convolution particle filters we proposed in [Rossi, 2004] and [Rossi and Vila, 2004] drop this assumption thanks to the use of convolution kernels to estimate the conditional density $p(X_t|Y_{1:t})$ supposed to exist. The following algorithm shows the implementation of the Resampled-Convolution Filter, one of the filters we developed [Rossi, 2004]:

Starting from a given initial probability density $p_0(X_0)$ and N simulated state values $(\tilde{X}_0^1, \dots, \tilde{X}_0^N))\mathrm{sim}p_0(X_0)$,

At time t:

(i) Sampling Step:
$(\tilde{X}_t^1, \dots, \tilde{X}_t^N)\mathrm{sim}p_t^N$ where p_t^N is the last estimated state conditional density.

(ii) Evolving Step: for $i = 1..N$, $(\tilde{X}_t^i) \longrightarrow (\tilde{X}_{t+1}^i, \tilde{Y}_{t+1}^i)$ by simulation of model (1)-(2).

(iii) Approximation Step:
$$p_{t+1}^N(X_{t+1}|Y_{1:t+1}) = \frac{\sum_{i=1}^N K_{2,\delta_N}(Y_{t+1} - \tilde{Y}_{t+1}^i) K_{1,\delta_N}(X_{t+1} - \tilde{X}_{t+1}^i)}{\sum_{i=1}^N K_{2,\delta_N}(Y_{t+1} - \tilde{Y}_{t+1}^i)}$$
with $K_{1,\delta_N}(x) = \delta_N^{-s} K_1\left(\frac{x}{\delta_N}\right)$, $x \in \mathbb{R}^s$ and $K_{\delta_N}(y) = \delta_N^{-q} K_2\left(\frac{y}{\delta_N}\right)$, $y \in \mathbb{R}^q$.

This algorithm ensures to get an "on line" L$_1$-convergent estimate of the density $p_t(X_t|Y_{1:t})$ when the particles number N tends to infinity ([Rossi, 2004] or [Rossi and Vila, 2004]).

3 Nonparametric adaptive and predictive control

The objective considered in this section is to find a control sequence $(U_t)_{t\geq 1}$ which forces the state variables $(X_t)_{t\geq 1}$, to follow as best as possible a given bounded trajectory $(X_t^*)_{t\geq 1}$. The state variable X_t is now again supposed to be observed and to evolve according to model (3), with function f_t completely or partly unknown.

Two control strategies are considered in the following according to the immediate or anticipative trajectory fitness considered, the second one being furthermore a generalization of the first.

3.1 Adaptive tracking control

Consider the particular case of model (4) particularly convenient for the biotechnological systems, in which g_t is unknown. An adaptive control strategy has to be built from the nonparametric estimate (6), which ensures the stochastic closed-loop stability. This last property is indeed necessary to ensure the convergence properties of the kernel estimator $\widehat{g}_t$. When B_t is supposed to be invertible with respect to U_t, let us consider a solution U_t such that

$$B_t(X_t, U_t) = X_{t+1}^* - A_t(X_t)\widehat{g}_t(X_t)1\!\!1_{E_t}(X_t) - A_t(X_t)g^*(X_t)1\!\!1_{E_t^c}(X_t)$$

where E_t is a subset of the state space, depending on the kernel estimate $\widehat{g}_t$ and on g^*, an a priori knowledge of g_t.

It has been shown that this strategy is asymptotically optimal:

$$\tfrac{1}{t}\sum_{i=1}^{t} \|X_i - X_i^*\|^2 \xrightarrow{a.s.} trace(\Gamma) \quad \text{as } t \to \infty,$$

where Γ denotes the covariance matrix of the noise ε_t.

See [Portier and Oulidi, 2000] and [Hilgert, 1997] for more details.

3.2 Optimal predictive control

Let us consider again state model (3) with unknown function f_t and still the assumption of observed X_t.

The principle of the so-called predictive control is now well-known among control theorists (see for example [Camacho and Bordons, 1995]). The specificity of predictive control is to consider the future values to be followed by the state system in a near forward horizon of given length H. More precisely at each time step the future values of the state variables on the horizon are predicted conditionally to intermediary control values. These control values are then optimized in order to minimize some discrepancy function between the predicted state values and that of the trajectory on the same horizon. The first of these optimal values of the control variable is then applied to the system which enters then the following time step and the predictive horizon is translated. Such an anticipating strategy confers to predictive control a significant advantage among on-line tracking control strategies, and is particularly adapted to the control of processes with slow dynamic such as the biotechnological processes. The main question raised by the predictive control algorithms is that of the stability of the closed loop. For deterministic systems several constraint conditions have been designed to ensure this stability (see [Mayne *et al.*, 2000] for a recent survey). For stochastic system this issue is still open for the general case. We consider it in the nonparametric approach to follow and solve it in a simple case.

A nonparametric predictive control algorithm for uncertain system: At step t,

- let
$$J_t = \sum_{j=1}^{j=H} \| X^*_{t+j} - f^j_{t+j-1}\left(u^1,\ldots,u^j \,|\, X_{i,\, i\le t}\,;\, U_{i,\, i\le t-1}\right) \|^2$$
where
  - ○ H is the chosen length of the receding horizon
  - ○ $\widehat{X}_{t+j} = f^j_{t+j-1}\left(u^1,\ldots,u^j \mid X_{i,\, i\le t}\ ;\ U_{i,\, i\le t-1}\right)$ is a consistent estimate to be looked for $\mathrm{E}\left[X_{t+j} \,|\, X_{i,\ i\le t}\ ;\ U_{i,\ i\le t-1}\ ;\ U_t = u^1,\ldots,\right.$ $\left.U_{t+j-1} = u^j\right]$ which is itself the minimum variance predictor of the state value X_{t+j}.

- Find

$$\begin{aligned}\bar{U}_t &= (U_t^1, \ldots, U_t^H)\\ &= \mathrm{argmin}_{\|u^1\| \leq M, \ldots, \|u^H\| \leq M} \; J_t\end{aligned}$$

  with M: upper bound constraint in the control values.
- take $U_t \; = \; U_t^1$
- $t = t + 1$

A j-step-ahead nonparametric state predictor:
Let $Z_t^j \; = \; (X_t, U_t, \ldots, U_{t+j-1})^t$
Let us consider as estimate of $\mathrm{E}(X_{t+j} \mid Z_t^j = z)$

$$\widehat{X}_{t+j} = \widehat{\mathrm{E}}(X_{t+j} \mid Z_t^j = z) \; = \; \frac{\sum_{t=1}^{t-j} |\det(\delta_t^{-1})| K\left(\frac{z - Z_t^j}{h_t}\right) X_{t+j}}{\sum_{t=1}^{t-j} |\det(\delta_t^{-1})| K\left(\frac{z - Z_t^j}{h_t}\right)}$$

where K is a kernel of dimension $(s + jm)$ and the matrix δ_t, of same dimension, is the bandwidth parameter of K.
For uncontrolled process, the asymptotic behaviour of $\widehat{X}_{t+j}$ has been characterized under mixing conditions and stationarity assumptions [Bosq, 1996]. These results are not applicable for the controlled processes we consider in this paper since the applied control values are state dependent. However for the simplest case, $H = 1$, stability of the closed loop, almost sure uniform dilated convergence of the kernel predictor and suboptimality of the control strategy has been established under regular conditions ([Wagner, 2001], [Wagner and Vila, 2001]) in both cases of interest for the f_t sequence (see section 2.1).

Remark 1: the minimization of the criterion function J_t at step t with respect to the constrained control variables $(u^1, \cdots, u^H)$, can be done by standard descent algorithm. We developed also a more efficient neural network-based minimization procedure and applied it online on a real biotechnological depollution process [Vila and Wagner, 2003].

Remark 2: the choice of the length of the predictive horizon H must result from a case by case compromise between long term optimality of the predictive control (high values for H) and the quality of the kernel predictors (low values).

4 Conclusion and perspectives: towards the nonparametric supervision of uncertain systems

When dealing with process control, an unavoidable issue is that of supervision. Supervision consists in being able to detect any default in the process (*e.g.* pump clogging in a bioprocess), locating the default and remedying it (by an appropriate sequence of actions). From a statistical point of view, the problems of detection and isolation of a default are equivalent to detecting abrupt changes in a stochastic process, and testing multiple hypotheses to determine the faulty scenario among a number of possible scenarii of defaults [Dubuisson, 2001].

There exist many statistical procedures to answer such questions, see [Basseville and Nikiforov, 1993]. A well-known one is the CuSum test. It is based on a comparison, at each time instant, of the difference between the log-likelihood ratio value and its current minimal value, with respect to a fixed threshold. Most of these techniques require knowledge of both state and observation models.

When the state model is uncertain, the question is still open. However combining nonparametric estimates as (5) or (6) with classical test procedures gave us encouraging results on real experimental data issued from a depollution process.

Moreover, introducing filtering methods such as the one proposed above, will allow to generalize these nonparametric detection procedures to the most frequent situation of indirectly observed systems described by models (1) and (2).

References

[Basseville and Nikiforov, 1993]M. Basseville and I.V. Nikiforov. *Detection of abrupt changes - Theory and application.* Prentice Hall, 1993.

[Bosq, 1996]D. Bosq. *Nonparametric Statistics for Stochastic Processes, Estimation and Prediction.* Springer-Verlag, New York, 1996.

[Camacho and Bordons, 1995]E.F. Camacho and C. Bordons. *Model Predictive Control in the Process Industry.* Springer-Verlag, 1995.

[Del Moral, 2004]P. Del Moral. *Feynman-Kac Formulae. Genealogical and Interacting Particle Systems with Applications.* Springer, 2004.

[Doucet *et al.*, 2001]A. Doucet, N. De Freitas, and N. Gordon. *Sequential Monte Carlo Methods in Practice.* Springer, 2001.

[Dubuisson, 2001]B. Dubuisson. *Automatique et statistiques pour le diagnostic.* Hermes Science Europe Ltd, 2001.

[Duflo, 1997]M. Duflo. *Random Iterative Models.* Springer-Verlag, New York, 1997.

[Georgiev, 1984]A. Georgiev. Nonpararametric system identification by kernel methods. *IEEE Trans. Automat. Control*, pages 356–358, 1984.

[Hilgert *et al.*, 2000]N. Hilgert, R. Senoussi, and J.-P. Vila. Nonparametric identification of controlled nonlinear time varying processes. *SIAM J. Cont. Optim.*, pages 950–960, 2000.

[Hilgert, 1997]N. Hilgert. *Identification et contrôle de processus autorégressifs non linéaires incertains : application á des procédés biotechnologiques.* PhD Thesis, Paris XI, 1997.

[Mayne *et al.*, 2000]D.Q. Mayne, J.B. Rawlings, C.V. Rao, and P.O. Scokaert. Constrained model predictive control: stability and optimality. *Automatica*, pages 789–814, 2000.

[Portier and Oulidi, 2000]B. Portier and A. Oulidi. Nonparametric estimation and adaptive control of functional autoregressive models. *SIAM J. Cont. Optim.*, pages 411–432, 2000.

[Rossi and Vila, 2004]V. Rossi and J.-P. Vila. Nonlinear filtering in discrete time : A particle convolution approach. *Technical Report 04-03, Groupe biostatisque ENSAM/INRA/UM2 Montpellier*, 2004.

[Rossi, 2004]V. Rossi. *Filtrage Non Linéaire par Noyaux de Convolution. Application à un Procédé de Dépollution Biologique.* PhD Thesis, ENSA Montpellier, 2004.

[Vieu, 1991]P. Vieu. Nonparametric regression: optimal local bandwidth choice. *J. R. Statist. Soc. B*, pages 453–464, 1991.

[Vila and Wagner, 2003]J.-P. Vila and V. Wagner. Predictive neuro-control of uncertain systems: design and use of a neuro-optimizer. *Automatica*, pages 767–777, 2003.

[Wagner and Vila, 2001]V. Wagner and J.-P. Vila. Estimation non paramétrique et contrôle prédictif neuronal de processus autorégressifs non linéaires incertains. *Journées de Statistique de la SFdS*, pages 711–714, 2001.

[Wagner, 2001]V. Wagner. *Identification non paramétrique et contrôle prédictif neuronal de processus non linéaires incertains de type biotechnologique : application à un procédé de dépollution biologique.* PhD Thesis, ENSA Montpellier, 2001.

Nonparametric Frontier estimation: A Multivariate Conditional Quantile Approach

Abdelaati Daouia[1] and Léopold Simar[2]

[1] GREMAQ, Université de Toulouse I
21 allée de Brienne
31000 TOULOUSE, France
(e-mail: daouia@cict.fr)

[2] Institut de Statistique, Université de Louvain-la-Neuve
20 voie du roman pays
1348, Louvain-la-Neuve, Belgique
(e-mail: simar@stat.ucl.ca.be)

Abstract. This paper proposes a probabilistic framework for efficiency and productivity analysis in a complete multivariate setup (multiple inputs and multiple outputs). Properties of the Farrell's efficiency scores are derived in terms of the characteristics of the probability distribution of the data generating process. This allows to introduce a notion of α-quantile efficiency scores related to a non-standard conditional α-quantile frontier and nonparametric robust estimators are provided. The asymptotic behavior of the estimator is provided with numerical illustration.

Keywords: Frontier estimation, Robust nonparametric estimators, Conditional quantiles.

1 Introduction and Basic Concepts

Foundations of the economic theory on productivity and efficiency analysis date back to the works of [Koopmans, 1951] and [Debreu, 1951] on activity analysis. We consider a production technology where the activity of the production units is characterized by a set of inputs $x \in I\!\!R_+^p$ used to produce a set of outputs $y \in I\!\!R_+^q$. The production set is the set of technically feasible combinations of (x, y):

$$\Psi = \{(x, y) \in I\!\!R_+^{p+q} \mid x \text{ can produce } y\}. \tag{1}$$

Assumptions are usually done on this set, such as free disposability of inputs and outputs, meaning that if $(x, y) \in \Psi$, then $(x', y') \in \Psi$, as soon as $x' \geq x$ and $y' \leq y$.

The Farrell-Debreu efficiency scores for a given production scenario $(x, y) \in \Psi$, are defined as:

$$\text{Input oriented} \quad : \quad \theta(x, y) = \inf\{\theta \mid (\theta x, y) \in \Psi\} \tag{2}$$

$$\text{Output oriented} : \quad \lambda(x, y) = \sup\{\lambda \mid (x, \lambda y) \in \Psi\} \tag{3}$$

In practice Ψ is unknown and so has to be estimated from a random sample of production units $\mathcal{X} = \{(X_i, Y_i) \,|\, i = 1, \ldots, n\}$, where we assume that $\text{Prob}((X_i, Y_i) \in \Psi) = 1$ (called deterministic frontier models). So the problem is related to the problem of estimating the support of the random variable (X, Y) where Ψ is supposed to be compact. The most popular nonparametric estimators are based on the envelopment ideas (see *e.g.* [Simar and Wilson, 2000], for a recent survey).

The Free Disposal Hull (FDH) estimator ([Deprins *et al.*, 1984]) is provided by the free disposal hull of the sample points $\mathcal{X}$:

$$\widehat{\Psi}_{FDH} = \left\{(x, y) \in \mathbb{R}_+^{p+q} \,|\, y \leq Y_i,\; x \geq X_i, \quad i = 1, \ldots, n\right\}. \tag{4}$$

The FDH efficiency scores are obtained by plugging $\widehat{\Psi}_{FDH}$ in equations (2) and (3) in place of the unknown Ψ. The asymptotic properties of the resulting estimators are provided by [Park *et al.*, 2000]. In summary, the error of estimation converges at a rate $n^{1/(p+q)}$ to a limiting Weibull distribution.

The FDH estimators envelop all the data points and so are very sensitive to outliers and/or to extreme values. [Cazals *et al.*, 2002] have introduced the concept of partial frontiers (order-m frontiers) with a nonparametric estimator which does not envelop all the data points. The value of m may be considered as a trimming parameter and as $m \to \infty$ the partial order-m frontier converges to the full-frontier. It is shown that by selecting the value of m as an appropriate function of n, the non-parametric estimator of the order-m efficiency scores provides a robust estimator of the corresponding efficiency scores sharing the same asymptotic properties as the FDH estimators but being less sensitive to outliers and/or extreme values.

Recently [Aragon *et al.*, 2002] have proposed an alternative to order-m partial frontiers by introducing quantile based partial frontiers. The idea is to replace this concept of "discrete" order-m partial frontier by a "continuous" order-α partial frontier where $\alpha \in [0, 1]$ corresponds to the level of an appropriate non-standard conditional quantile frontier. Unlike the order-m partial frontiers, due to the absence of natural ordering of Euclidean spaces for dimension greater than one, the α-quantile approach is limited to one-dimensional input for the input oriented frontier and to one-dimensional output for the output oriented frontier.

In this paper, we overcome this difficulty and we propose an extension to the full multivariate case, introducing the concept of α-quantile efficiency scores and the corresponding α-quantile frontier set.

2 Probabilistic Formulation and Nonparametric Estimation

[Daraio and Simar, 2002] propose a probabilistic formulation of efficiency concepts. The Data Generating Process (DGP) of (X, Y) is completely characterized by

$$H_{XY}(x,y) = \text{Prob}(X \le x, Y \ge y). \tag{5}$$

The support of $H_{XY}(\cdot,\cdot)$ is Ψ and $H_{XY}(x,y)$ can be interpreted as the probability for a unit operating at the level (x,y) to be dominated. This joint probability can be decomposed as follows:

$$\begin{aligned} H_{XY}(x,y) &= \text{Prob}(X \le x \,|\, Y \ge y)\,\text{Prob}(Y \ge y) = F_{X|Y}(x|y)\,S_Y(y) & (6)\\ &= \text{Prob}(Y \ge y \,|\, X \le x)\,\text{Prob}(X \le x) = S_{Y|X}(y|x)\,F_X(x), & (7)\end{aligned}$$

where we suppose the conditional probabilities exit (*i.e.*, when needed, $F_X(x) > 0$ or $S_Y(y) > 0$).

An input oriented efficiency score $\tilde{\theta}(x,y)$ for $(x,y) \in \Psi$ is defined for all y with $S_Y(y) > 0$ as

$$\widetilde{\theta}(x,y) = \inf\{\theta \,|\, F_{X|Y}(\theta x|y) > 0\} = \inf\{\theta \,|\, H_{XY}(\theta x, y) > 0\}. \tag{8}$$

For the output oriented case, for all x such that $F_X(x) > 0$, we define the output efficiency score as

$$\widetilde{\lambda}(x,y) = \sup\{\lambda \,|\, S_{Y|X}(\lambda y|x) > 0\} = \sup\{\lambda \,|\, H_{XY}(x, \lambda y) > 0\}. \tag{9}$$

This input (resp. output) efficiency score can be interpreted as the proportionate reduction (resp. increase) of inputs (resp. outputs) a unit working at the level (x,y) should perform to be dominated with probability zero.

If Ψ is free disposal (a minimal assumption), it can be shown that: $\widetilde{\theta}(x,y) \equiv \theta(x,y)$ and $\widetilde{\lambda}(x,y) \equiv \lambda(x,y)$.

Natural nonparametric estimators of $\theta(x,y)$ and of $\lambda(x,y)$ are obtained by plugging the empirical distribution $\widehat{H}_{XY,n}$ in place of H_{XY} in the definition of the efficiency scores, where

$$\widehat{H}_{XY,n}(x,y) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}(X_i \le x, Y_i \ge y), \tag{10}$$

As pointed out in [Daraio and Simar, 2002], these estimators are the FDH estimators of the Farrell-Debreu efficiency scores.

3 Conditional Quantile Based Efficiency Scores

[Aragon *et al.*, 2002] have introduced the conditional quantile frontier function for a production (output) function when the output is unidimensional and for a cost (input) function when the input is one dimensional. We extend the ideas to a full multivariate setup. Since a natural ordering of Euclidean spaces of dimension greater than one does not exist, we overcome the difficulty by defining α-quantile efficiency scores as follows.

Definition 1 *For all y such that $S_Y(y) > 0$ and for $\alpha \in]0,1]$, the α-quantile input efficiency score for the unit $(x,y) \in \Psi$ is defined as*

$$\theta_\alpha(x,y) = \inf\{\theta \mid F_{X|Y}(\theta x|y) > 1-\alpha\} \tag{11}$$

For all x such that $F_X(x) > 0$ and for $\alpha \in]0,1]$, the α-quantile output efficiency score for the unit $(x,y) \in \Psi$ is defined as

$$\lambda_\alpha(x,y) = \sup\{\lambda \mid S_{Y|X}(\lambda y|x) > 1-\alpha\} \tag{12}$$

For instance, in the input case, $\theta_\alpha(x,y)$ is the proportionate reduction (if < 1) or increase (if > 1) of inputs, a unit working at the level (x,y) should perform to be dominated by firms producing more than the output level y with probability $1-\alpha$. If $\theta_\alpha(x,y) = 1$, we will say that the unit is input efficient at the level $\alpha \times 100\%$. Clearly when $\alpha = 1$, this is, under free disposability of Ψ, the Farrell-Debreu input efficiency score. In a certain sense, we can say that $\theta_\alpha(x,y)$ is the input efficiency of (x,y) at the level $\alpha \times 100\%$. The same is true in the output direction. We define Ψ^* as being the interior of Ψ.

Proposition 1 *Assume that $F_{X|Y}$ is continuous and monotone increasing in x and that $S_{Y|X}$ is continuous and monotone decreasing in y. Then, for all $(x,y) \in \Psi^*$, there exist α and β in $]0,1]$ such that*

$$\theta_\alpha(x,y) = 1, \qquad \textit{where } \alpha = 1 - F_{X|Y}(x|y) \tag{13}$$

$$\lambda_\beta(x,y) = 1, \qquad \textit{where } \beta = 1 - S_{Y|X}(y|x). \tag{14}$$

Proposition 1 shows that any point (x,y) in the interior of Ψ, belongs to an appropriate α-quantile efficient frontier in both directions (input and output). When $\alpha \to 1$, the α-quantile efficient scores converge monotonically to the Farrell-Debreu efficiency scores:

Proposition 2 *For all y such that $S_Y(y) > 0$, we have $lim_{\alpha\to 1} \searrow \theta_\alpha(x,y) = \theta(x,y)$ and for all x such that $F_X(x) > 0$, $lim_{\alpha\to 1} \nearrow \lambda_\alpha(x,y) = \lambda(x,y)$.*

The α-quantile input efficiency score $\theta_\alpha(x,y)$ is clearly monotone nonincreasing with x but it is in general not monotone in y, unless we add an assumption on $F_{X|Y}$:

Proposition 3 *Assume that $F_{X|Y}(\cdot|y)$ is continuous for any y. Then, the two following properties are equivalent.*

$$F_{X|Y}(x|y) \textit{ is monotone nonincreasing with } y \tag{15}$$

$$\theta_\alpha(x,y) \textit{ is monotone nondecreasing with } y \textit{ for all } \alpha. \tag{16}$$

Points (x,y) here are such that $F_{X|Y}(x|y) < 1$.

Proposition 4 *The two following properties are equivalent.*

$$S_{Y|X}(y|x) \textit{ is monotone nondecreasing with } x \tag{17}$$

$$\lambda_\alpha(x,y) \textit{ is monotone nondecreasing with } x \textit{ for all } \alpha. \tag{18}$$

Points (x,y) here are such that $S_{Y|X}(y|x) < 1$.

4 Nonparametric Estimator

A natural nonparametric estimator of the α-quantile efficiency scores is obtained by plugging the empirical $\widehat{H}_{XY,n}(x,y)$ in the above formulas

$$\widehat{\theta}_{\alpha,n}(x,y) = \inf\{\theta \mid \widehat{F}_{X|Y,n}(\theta x|y) > 1-\alpha\}, \tag{19}$$

$$\widehat{\lambda}_{\alpha,n}(x,y) = \sup\{\lambda \mid \widehat{S}_{Y|X,n}(\lambda y|x) > 1-\alpha\}, \tag{20}$$

These nonparametric estimators can be computed very easily. When $\alpha \to 1$, the estimators converge monotonically to the FDH efficiency scores $\widehat{\theta}_n(x,y)$ and $\widehat{\lambda}_n(x,y)$, respectively:

Proposition 5 *For all y such that $\widehat{H}_{XY,n}(\infty,y) > 0$, we have $lim_{\alpha\to 1} \searrow \widehat{\theta}_{\alpha,n}(x,y) = \widehat{\theta}_n(x,y)$ and for all x such that $\widehat{H}_{XY,n}(x,0) > 0$, $lim_{\alpha\to 1} \nearrow \widehat{\lambda}_{\alpha,n}(x,y) = \widehat{\lambda}_n(x,y)$.*

The asymptotic behavior of our estimator is given by the following theorems (only presented for the output direction: we have the same results for the input oriented case).

Theorem 1 *Let $(x,y) \in \Psi$ be such that $F_X(x) > 0$ and let $0 < \alpha < 1$. Assume that $\lambda \mapsto S_{Y|X}(\lambda y|x)$ is decreasing in a neighborhood of $\lambda_\alpha(x,y)$. Then, for every $\varepsilon > 0$,*

$$Prob(|\widehat{\lambda}_{\alpha,n}(x,y) - \lambda_\alpha(x,y)| > \varepsilon) \leq 2e^{-2n\delta^2_{\varepsilon,x,y}}, \quad \textit{for all } n \geq 1,$$

where

$$\delta_{\varepsilon,x,y} = \frac{F_X(x)}{(2-\alpha)} \min\big\{(1-\alpha) - S_{Y|X}((\lambda_\alpha(x,y)+\varepsilon)y|x) \\ ; S_{Y|X}((\lambda_\alpha(x,y)-\varepsilon)y|x) - (1-\alpha)\big\}.$$

Theorem 2 *Let $0 < \alpha < 1$ be a fixed order and let $(x,y) \in \Psi$ be a fixed unit such that $F_X(x) > 0$. Assume that $G(\lambda) = S_{Y|X}(\lambda y|x)$ is differentiable at $\lambda_\alpha(x,y)$ with negative derivative $G'(\lambda_\alpha(x,y)) =< \bigtriangledown S_{Y|X}(\lambda_\alpha(x,y)y|x), y >$. Then,*

$$\sqrt{n}\left(\widehat{\lambda}_{\alpha,n}(x,y) - \lambda_\alpha(x,y)\right) \xrightarrow{\mathcal{L}} N\left(0, \sigma^2_\alpha(x,y)\right) \quad \textit{as} \quad n \to \infty,$$

where

$$\sigma^2_\alpha(x,y) = \frac{\alpha(1-\alpha)}{[G'(\lambda_\alpha(x,y))]^2 F_X(x)}.$$

A more robust estimator of the Farrell-Debreu efficiency scores $\lambda(x,y)$ than the standard FDH estimator $\widehat{\lambda}_n(x,y)$, which however shares similar asymptotic properties with this latter one, can be derived as follows.

Lemma 1 *Assume that the support of Y is bounded. Then, for any $(x,y) \in \Psi$,*

$$n^{1/(p+q)}\left(\widehat{\lambda}_n(x,y) - \widehat{\lambda}_{\alpha(n),n}(x,y)\right) \xrightarrow{a.s.} 0 \quad as \quad n \to \infty,$$

where the order $\alpha(n) > 0$ is such that

$$n^{(p+q+1)/(p+q)}\left(1-\alpha(n)\right) \longrightarrow 0 \quad as \quad n \to \infty.$$

Making use of this lemma and the following decomposition

$$\begin{aligned} n^{1/(p+q)}(\lambda(x,y) - \widehat{\lambda}_{\alpha(n),n}(x,y)) = n^{1/(p+q)}(\lambda(x,y) - \widehat{\lambda}_n(x,y)) \\ +n^{1/(p+q)}(\widehat{\lambda}_n(x,y) - \widehat{\lambda}_{\alpha(n),n}(x,y)) \end{aligned}$$

we get immediately from Corollary 3.2 of [Park *et al.*, 2000] the following result:

Theorem 3 *Under Assumptions AI-AIII of [Park* et al.*, 2000], we have for any (x,y) interior to Ψ,*

$$n^{1/(p+q)}\left(\lambda(x,y) - \widehat{\lambda}_{\alpha(n),n}(x,y)\right) \xrightarrow{\mathcal{L}} Weibull(\mu_{NW,0}^{p+q}, p+q) \quad as \quad n \to \infty,$$

where $\mu_{NW,0}$ is a constant (see [Park et al.*, 2000]) .*

The latter results show that with an appropriate choice of α, we obtain a nonparametric estimator of the Farrell-Debreu efficiency score $\lambda(x,y)$ sharing the same properties than the FDH estimator, but since it does not envelop all the data points, it will be more robust to extreme and/or outlying observations.

5 Numerical Illustrations

We illustrate the α-quantile efficiency scores and their estimation by using some of simulated data set used in [Daraio and Simar, 2002] with multi-input ($p = 2$) and multi-output ($q = 2$) and Z is favorable to output production. The results are displayed in Figure 1. We see that all the ratios allow to detect the favorable effect of Z on the production process. The α-quantile measures being less sensitive to extreme values, give a better picture.

In order to appreciate the robustness to outliers, and compare the performance of the order-m and of the α-quantile measures, we introduce in the same data set 5 outliers by projecting, in the Y coordinates 5 points in a radial expansion by a factor 1/0.6. The results of this data set with $n = 105$ points are shown in Figure 2. It is clear that the full frontier approach is unable to detect the favorable effect of Z, at least for values larger than the mean of Z (2.5), the order-m does better but again fails for large values of Z. On the contrary, the order-α quantile frontier are much more robust to the 5 outliers and we obtain similar results as in Figure 1, where no outliers where introduced.

References

[Aragon *et al.*, 2002]Y. Aragon, A. Daouia, and C. Thomas-Agnan. Nonparametric frontier estimation: A conditional quantile-based approach. *to appear, Econometric Theory*, 2002.

[Cazals *et al.*, 2002]C. Cazals, J.P. Florens, and L. Simar. Nonparametric frontier estimation: A conditional quantile-based approach. *Journal of Econometrics*, 106:1–25, 2002.

[Daraio and Simar, 2002]C. Daraio and L. Simar. Introducing environmental variables in nonparametric frontier models: a probabilistic approach. *to appear, Journal of Productivity Analysis*, 2002.

[Debreu, 1951]G. Debreu. The coefficient of resource utilization. *Econometrica*, 19(3):273–292, 1951.

[Deprins *et al.*, 1984]D. Deprins, L. Simar, and H. Tulkens. Measuring labor inefficiency in post offices. In M. Marchand, P. Pestieau, and H. Tulkens, editors, *The Performance of Public Enterprises: Concepts and measurements*, pages 243–267. North-Holland, Amsterdam, 1984.

[Koopmans, 1951]T.C. Koopmans. An analysis of production as an efficient combination of activities. In T.C. Koopmans, editor, *Activity Analysis of Production and Allocation*. Cowles Commission for Research in Economics, Monograph 13, John-Wiley and Sons, Inc., New York, 1951.

[Park *et al.*, 2000]B. Park, L. Simar, and Ch. Weiner. The fdh estimator for productivity efficiency scores : Asymptotic properties. *Econometric Theory*, 16:855–877, 2000.

[Simar and Wilson, 2000]L. Simar and P.W. Wilson. The fdh estimator for productivity efficiency scores : Asymptotic properties. *Journal of Productivity Analysis*, 13:49–78, 2000.

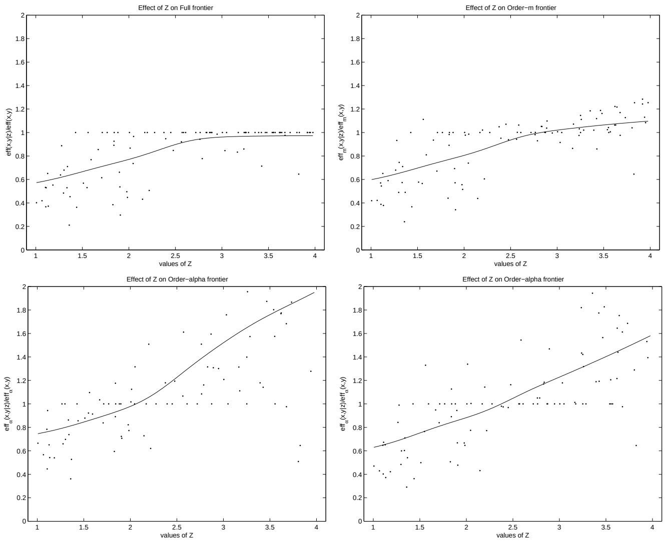


Fig. 1. *Simulated example, $n = 100$: "positive" effect of Z on production efficiency (output oriented framework). Scatterplot and smoothed regression of the ratios $\hat{\lambda}_n(x, y \mid z)/\hat{\lambda}_n(x, y)$ on Z (top left), of $\hat{\lambda}_{m,n}(x, y \mid z)/\hat{\lambda}_{m,n}(x, y)$ on Z (top right, with $m = 25$) and of $\hat{\lambda}_{\alpha,n}(x, y \mid z)/\hat{\lambda}_{\alpha,n}(x, y)$ on Z (bottom panel, left $\alpha = 0.80$ and right $\alpha = 0.90$). Here k-NN=17.*

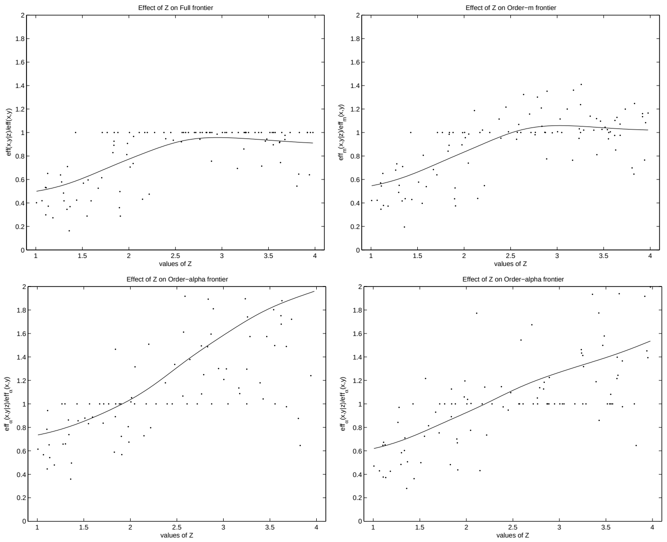


Fig. 2. *Simulated example, $n = 105$ including 5 outliers: "positive" effect of Z on production efficiency (output oriented framework). Scatterplot and smoothed regression of the ratios $\hat{\lambda}_n(x, y \mid z)/\hat{\lambda}_n(x, y)$ on Z (top left), of $\hat{\lambda}_{m,n}(x, y \mid z)/\hat{\lambda}_{m,n}(x, y)$ on Z (top right, with $m = 25$) and of $\hat{\lambda}_{\alpha,n}(x, y \mid z)/\hat{\lambda}_{\alpha,n}(x, y)$on Z (bottom panel, left $\alpha = 0.80$ and right $\alpha = 0.90$). Here k-NN=20.*

Empirical likelihood for non-degenerate U-statistics

Bing-Yi Jing[1], Junqing Yuan[1], and Wang Zhou[2]

[1] Department of Mathematics
Hong Kong Univ. of Sci. and Tech.
Clear Water Bay, Kowloon
Hong Kong
(e-mail: majing@ust.hk, yuanjq@ust.hk)

[2] Dept. of Stat. and Applied Prob.
National University of Singapore
Singapore 117543
(e-mail: stazw@nus.edu.sg)

Abstract. Standard empirical likelihood for U-statistics is too computationally expensive. To overcome this computational difficulty, we reformulate the empirical likelihood for non-degenerate U-statistics in terms of "pseudo" mean in this paper, and show that the empirical log-likelihood ratio has an asymptotic chi-squared distribution under second moment condition. The method is extremely simple to use, and yet provide better coverage accuracy in general than other alternative methods from our simulation studies.

Keywords: U-statistics, empirical likelihood, confidence interval.

1 Introduction

The empirical likelihood method was first introduced by [Owen, 1988] for constructing confidence intervals and [Owen, 1990] for confidence regions. [Hall and LaScala, 1990] has summarized its advantages over the bootstrap: the empirical likelihood regions are shaped "automatically" by the sample, Bartlett correctable, range preserving and transformation respecting. For these reasons, the empirical likelihood has found lots of applications such as in smooth functions of means [DiCiccio *et al.*, 1989], in nonparametric density [Chen, 1996], in regression function estimation [Owen, 1991] [Chen and Qin, 2000] and so on. For a more thorough review of the empirical likelihood method and its applications, the reader is referred to the recent monograph by [Owen, 2001].

In this paper, we are interested in applying the empirical likelihood method to U-statistics. Let $X, X_1, \cdots, X_n, n \geq 2$, be independent and identically distributed (i.i.d) random variables with common distribution function $F(x)$. A U-statistic of degree $m \geq 2$ with a symmetric kernel h is defined to be

$$U_n = \binom{n}{m}^{-1} \sum_{1 \leq i_1 < \ldots < i_m \leq n} h(X_{i_1}, ..., X_{i_m}) \tag{1}$$

where $\theta = Eh(X_1, ..., X_m)$ is a parameter of interest. Under very weak conditions, U_n is a Minimum Variance Unbiased Estimator of θ. On the other hand, U-statistics have many applications in hypothesis testing. For further details on U-statistic see [Lee, 1990]. Define

$$g(x) = Eh(x, X_1, ..., X_{m-1}) - \theta, \qquad \sigma_g^2 = var(g(X)). \tag{2}$$

Throughout this paper, we shall assume that $\sigma_g^2 > 0$.

The straightforward application of Owen's empirical likelihood in this context can be described as follows. Denote F_q to be the empirical distribution function which assigns probability q_i to observation X_i. Then, the empirical likelihood, evaluated at the true parameter value θ, can be defined by

$$\widetilde{L}(\theta) = \max_{\widetilde{\theta}(F_q)=\theta, \sum q_i=1} \prod_{i=1}^{n} q_i, \tag{3}$$

where

$$\widetilde{\theta}(F_q) = \binom{n}{m}^{-1} \sum_{1 \le i_1 < ..., < i_m \le n} n^m q_{i_1} ... q_{i_m} h(X_{i_1}, ..., X_{i_m}).$$

Note that $\prod_{i=1}^n q_i$, subject to $\sum_{i=1}^n q_i = 1$, attains its maximum n^{-n} at $q_i = n^{-1}$. Then, the empirical likelihood ratio at θ is given by

$$\widetilde{R}(\theta) = \widetilde{L}(\theta)/n^{-n} = \max_{\widetilde{\theta}(F_q)=\theta, \sum q_i=1} \prod_{i=1}^{n} (nq_i). \tag{4}$$

As mentioned in [Wood *et al.*, 1996], Wilks's theorem holds under mild conditions in this case, i.e., $-2\log \widetilde{R}(\theta) \xrightarrow{d} \chi_1^2$, where $\xrightarrow{d}$ means converges in distribution as $n \to \infty$, and χ_1^2 denotes the chi square distribution with one degree of freedom. This can be used to construct confidence intervals for the parameter θ. We shall refer to this procedure as Owen's *direct or "exact" empirical likelihood method* to U-statistics.

The major drawback of Owen's direct empirical likelihood method is its computational difficulty due to the presence of nonlinear constraints in the underlying optimization problem. [Wood *et al.*, 1996] proposed a so-called *sequential linerization method* for empirical likelihood methods with nonlinear constraints, and applied it to U-statistics. They found in their simulation studies that a single iteration of the linearization procedure may not be enough to achieve reliable coverage probabilities, and suggested to employ multiple (three to ten) iterations of the linearization procedure or bootstrap calibration in practice in order to improve coverage probabilities.

In this paper, we propose a new empirical likelihood method to U-statistics. The key idea of our method is to turn the U-statistic into a "sample mean" based on some "pseudo" observations, and then simply apply Owen's empirical likelihood to that "sample mean". As will be seen from the next section, those "pseudo" observations are in fact dependent random variables. Wilks's theorem will be established under mild conditions, which can then be used to construct confidence intervals for the parameter θ. The most attractive feature of our approach is its simplicity. Furthermore, our simulations results show that the coverage probabilities of our approach are in general better than alternative methods.

The paper is organized as follows. In Section 2, we introduce a new empirical likelihood method for U-statistics, and presents some theoretical results. Some simulation studies are conducted in Section 3 to compare the performances of the empirical likelihood and other methods. Proofs are deferred to Section 4.

2 Methodology and main results

First we rewrite U_n as

$$U_n = \frac{1}{n}\sum_{i=1}^{n} V_i,$$

where the "components" of U_n, defined by [Sen, 1960]

$$V_i = \binom{n-1}{m-1}^{-1} \sum_{\substack{1\le j_1<\cdots<j_{m-1}\le n \\ j_r\neq i, 1\le r\le m-1}} h(X_i, X_{j_1}, ..., X_{j_{m-1}}) \tag{5}$$

are treated as "pseudo" observations. Note that V_i's are dependent.

To employ empirical likelihood, let $p = (p_1, \cdots, p_n)$ be a probability vector, i.e., $\sum_{i=1}^n p_i = 1$ and $p_i \ge 0$ for $1 \le i \le n$. Let G_p be the distribution function which assigns probability p_i at the ith pseudo observation V_i, and hence $\theta(G_p) = \sum_{i=1}^n p_i V_i$. Then, the empirical likelihood ratio, evaluated at θ, is given by

$$L(\theta) = \max_{\theta(G_p)=\theta, \sum p_i=1} \prod_{i=1}^{n} p_i. \tag{6}$$

Note that $\prod_{i=1}^n p_i$, subject to $\sum_{i=1}^n p_i = 1$, attains its maximum n^{-n} at $p_i = n^{-1}$. So we define the empirical likelihood ratio at θ by

$$R(\theta) = L(\theta)/n^{-n} = \max_{\theta(G_p)=\theta, \sum p_i=1} \prod_{i=1}^{n}(np_i). \tag{7}$$

Using Lagrange multipliers, we have

$$p_i = \frac{1}{n}\frac{1}{1+\lambda(V_i-\theta)}, \tag{8}$$

where λ satisfies

$$g(\lambda) := \frac{1}{n}\sum_{i=1}^{n}\frac{V_i-\theta}{1+\lambda(V_i-\theta)} = 0. \tag{9}$$

After plugging the p_i's back into (7) and taking the logarithm of $R(\theta)$, we get the nonparametric log-likelihood ratio

$$\log R(\theta) = -\sum_{i=1}^{n}\log\{1+\lambda(V_i-\theta)\}.$$

The next theorem shows that Wilks's theorem holds here under a mild condition.

Theorem 1 *Assume that $Eh^2(X_1,...,X_m) < \infty$ and $\sigma_g^2 > 0$, then*

$$-\frac{2}{m^2}\log R(\theta) \xrightarrow{d} \chi_1^2.$$

The proof of Theorem 1 will be given in Section 4.

Remark 1 *Wilks's theorem, stated in Theorem 1, is slightly different from the ones we normally encounter. For example, for the Owen's direct empirical likelihood method, one has*

$$-2\log\widetilde{R}(\theta) \xrightarrow{d} \chi_1^2.$$

However, in our case here, we have

$$-\frac{2}{m^2}\log R(\theta) \xrightarrow{d} \chi_1^2.$$

Remark 2 *An approximate $1-\alpha$ level confidence interval for θ can be defined as*

$$\Re_c = \{\theta : -\frac{2}{m^2}\log R(\theta) \le c\},$$

where c is chosen to satisfy $P(\chi_1^2 \ge c) = \alpha$. From Theorem 1, we have

$$\lim_{n\to\infty} P\{\theta\in\Re_c\} = P(\chi_1^2 \le c) = 1-\alpha.$$

In other words, the interval $\Re_c$ gives asymptotic correct coverage probability.

3 Simulation results

In this section, we shall conduct some simulation studies to investigate the coverage accuracy of the empirical likelihood method proposed in this paper. Comparisons will be made with some alternative methods such as the normal approximation method, Owen's direct or "exact" empirical likelihood method, and the sequential linerization method proposed by [Wood *et al.*, 1996]. Three examples will be used for illustration: population variance, probability weighted moments, and Gini's mean difference as special cases of U-statistics. All the simulation results are based on 1,000 repetitions.

Example 1: population variance $\sigma^2 = var(X)$. In this case, the sample variance is a U-statistic with the kernel $h(x, y) = (x-y)^2/2$. For this example, it is also rather easy to apply Owen's empirical likelihood method directly by placing probability weight p_i on X_i and maximizing the empirical likelihood subject to

$$\sum_{i=1}^{n} p_i(X_i - \mu_X)^2 = \sigma^2 \qquad \text{with } \mu_X = \textstyle\sum_{i=1}^{n} p_i X_i.$$

Therefore, it would be interesting to compare this direct approach with the one proposed in this paper. For illustrative purposes, we shall include the normal approximation method as well for comparison. The underlying population is selected as standard Normal, then the actual value $\theta = 1$. The results are summarized in Table 1.

Table 1. Coverage accuracy for the variance

	nominal level	0.80	0.90	0.95
n=15	Normal Appr.	0.655	0.751	0.816
	Owen's EL	0.668	0.782	0.847
	Our EL	0.708	0.806	0.868
n=40	Normal Appr.	0.723	0.828	0.878
	Owen's EL	0.758	0.855	0.918
	Our EL	0.748	0.845	0.898
n=100	Normal Appr.	0.772	0.872	0.917
	Owen's EL	0.804	0.906	0.949
	Our EL	0.789	0.884	0.931

Example 2: probability weighted moment $E[XF(X)]$. In this case, the sample probability weighted moment is a U-statistic with the kernel $h(x, y) = \max\{x, y\}/2$. Coverage probabilities of the "exact" empirical likelihood method, described in the Introduction, were given in table 4 of [Wood *et al.*, 1996], which will be used for comparison with our own approach in this paper. Two underlying distributions are considered: the standard Normal and the exponential with mean 1. For these distributions, the population values are 0.282 and 0.75 respectively. Table 2 records the simulation results, with those in parentheses for the latter distribution.

Example 3: Gini's mean difference $E|X_1 - X_2|$. Gini's mean difference is an attractive measure for describing the population concentration. Its sample version is a U-statistics with the kernel $h(x, y) = |x - y|$. This

Table 2. Coverage accuracy for the probability weighted moment

	nominal level	0.80	0.90	0.95
n=15	"Exact" EL	0.745 (0.705)	0.844 (0.801)	0.896 (0.882)
	Our EL	0.746 (0.740)	0.845 (0.830)	0.912 (0.888)
n=40	"Exact" EL	0.742 (0.741)	0.849 (0.844)	0.922 (0.899)
	Our EL	0.768 (0.761)	0.866 (0.857)	0.923 (0.910)
n=100	"Exact" EL	0.787 (0.783)	0.895 (0.864)	0.944 (0.929)
	Our EL	0.821 (0.771)	0.904 (0.873)	0.941 (0.924)

example was also studied by [Wood *et al.*, 1996], who used their sequential linearization approach in this case. The comparisons with our method is presented in Table 3, where Wood *et al.*(r) denotes the sequential linearization approach with r iterations. For the underlying distribution, we use a standard Normal, so $\theta = 1.1284$.

Table 3. Coverage accuracy for Gini's mean difference

	nominal level	0.80	0.90	0.95
n=15	Wood *et al.*(1)	0.693	0.799	0.859
	Wood *et al.*(3)	0.737	0.864	0.932
	Our EL	0.741	0.846	0.889
n=40	Wood *et al.*(1)	0.756	0.862	0.919
	Wood *et al.*(3)	0.751	0.862	0.924
	Our EL	0.772	0.864	0.917
n=100	Wood *et al.*(1)	0.782	0.884	0.935
	Wood *et al.*(3)	0.780	0.887	0.939
	Our EL	0.787	0.889	0.936

The following observations can be made from our simulation studies:

(1) As expected, all methods improve as the sample size n increases.
(2) From Table 1, we see that, our method outperforms Normal Approximation method. Comparing with Owen's empirical likelihood method, our's looks better for small sample size.
(3) From Table 2, our method seems to perform slightly better than the "exact" empirical likelihood, mentioned in the Introduction. But our method is much simpler to use.
(4) From Table 3, we see that, overall, our method performs equally well as Wood *et al*'s sequential linearization approach with 3 iterations, and both are better than Wood *et al*'s approach with only 1 iteration. However, our method is the simplest amongst the three.

In summary, our empirical likelihood method for U-statistics in general performs better or as well as all other alternative methods such as normal approximation, exact empirical likelihood and sequential linearization procedure. Furthermore, our approach is the simplest one to use. For these reasons, our method should always be preferred.

4 Proof of main results

For notational simplicity, we shall prove our main results for U-statistics of order $m = 2$ only. The case for the general order $m \geq 2$ can be done similarly. But first we shall list several simple lemmas for easy reference later in the section.

Lemma 1 *[Hoeffding, 1948] Suppose* $Eh^2(X_1, X_2) < \infty$, *then*

$$\frac{\sqrt{n}(U_n - \theta)}{2\sigma_g} \xrightarrow{d} N(0, 1).$$

Corollary 1 *Assuming* $Eh^2(X_1, X_2) < \infty$, *then* $U_n - \theta = O_p(n^{-1/2})$.

Proof. This is a direct consequence of Lemma 1.

Lemma 2 *Let* $S = n^{-1} \sum_{i=1}^{n} (V_i - \theta)^2$, *if* $Eh^2(X_1, X_2) < \infty$, *then*

$$S = \sigma_g^2 + o(1) \qquad a.s.$$

Proof. Note that

$$S = \frac{1}{n}\sum_{i=1}^{n}(V_i - \theta)^2 = \frac{1}{n}\sum_{i=1}^{n}(V_i - U_n)^2 + (U_n - \theta)^2.$$

Let $\sigma^2 = var\{h(X_1, X_2)\} < \infty$, since $Eh^2(X_1, X_2) < \infty$, thus

$$var(U_n) = \frac{4(n-2)}{n(n-1)}\sigma_g^2 + \frac{2}{n(n-1)}\sigma^2.$$

Denote the jackknife estimate of $var(U_n)$ by $\widehat{var}(JACK)$, Lee (1990) identified that (page 223-4)

$$\frac{1}{n}\sum_{i=1}^{n}(V_i - U_n)^2 = \frac{(n-2)^2}{4(n-1)}\widehat{var}(JACK).$$

Since $\widehat{var}(JACK)$ is a consistent estimator of $var(U_n)$ in the sense that

$$n\{\widehat{var}(JACK) - var(U_n)\} \to 0, \qquad a.s.$$

then as $n \to \infty$, we have

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}(V_i - U_n)^2 &= \frac{(n-2)^2}{4(n-1)}\left(var(U_n) + o(n^{-1})\right)\\ &= \frac{(n-2)^2}{4(n-1)}\left(\frac{4(n-2)}{n(n-1)}\sigma_g^2 + \frac{2}{n(n-1)}\sigma^2 + o(n^{-1})\right)\\ &= \sigma_g^2 + o(1), \qquad a.s.\end{aligned}$$

In addition, the strong law of large number for U-statistics results in $U_n = \theta + o(1)$ *a.s.* Therefore, $S = \sigma_g^2 + o(1)$ *a.s.*, which ends the proof.

Lemma 3 *Let* $Y_n = \max_{1\le i\ne j\le n} |h(X_i, X_j)|$, *if* $Eh^2(X_1, X_2) < \infty$, *then*

$$Y_n = o(n^{1/2}) \qquad a.s.$$

Proof. Since $Eh^2(X_1, X_2) < \infty$, we have

$$\sum_{n=1}^{\infty} P\left(h^2(X_1, X_2) > n\right) < \infty,$$

which implies that

$$\sum_{n=1}^{\infty} P\left(h^2(X_i, X_j) > n\right) < \infty, \qquad \text{for any } 1 \le i \ne j \le n.$$

And hence by the Borel-Cantelli Lemma, with probability 1,

$$|h(X_i, X_j)| > n^{1/2}, \qquad \text{for any } 1 \le i \ne j \le n$$

finitely often. Thus with probability 1, $Y_n > n^{1/2}$ occurs finitely often. By the same argument $Y_n > An^{1/2}$ finitely often with probability 1 for any $A > 0$. Consequently,

$$\limsup_{n\to\infty} \frac{Y_n}{n^{1/2}} \le A \qquad a.s. \tag{10}$$

Inequality (10) holds simultaneously with probability 1 for any countable set of values for A. Therefore $Y_n = o(n^{1/2})$ *a.s.*

Corollary 2 *Let* $Z_n = \max_{1\le i\le n} |V_i - \theta|$, *if* $Eh^2(X_1, X_2) < \infty$, *then*

$$Z_n = o(n^{1/2}) \qquad a.s., \tag{11}$$

and

$$\frac{1}{n}\sum_{i=1}^{n} |V_i - \theta|^3 = o(n^{1/2}) \qquad a.s. \tag{12}$$

Proof. Note that

$$|V_i - \theta| \le \frac{1}{n-1}\sum_{\substack{j=1\\ j\neq i}}^{n} |h(X_i, X_j)| + |\theta| \le Y_n + |\theta|$$

for any $1 \le i \le n$. By Lemma 3, $Z_n = o(n^{1/2})$ *a.s.*

For the second assertion, by (11) and Lemma 2, with probability 1

$$\frac{1}{n}\sum_{i=1}^{n} |V_i - \theta|^3 \le Z_n \times \frac{1}{n}\sum_{i=1}^{n}(V_i - \theta)^2 = o(n^{1/2})$$

as has to be shown.

Proof of Theorem 1. We first show that the root of (9) satisfies $|\lambda| = O_p(n^{-1/2})$. Note that

$$\begin{aligned} 0 = |g(\lambda)| &= \frac{1}{n}\left|\sum_{i=1}^{n}(V_i - \theta) - \lambda\sum_{i=1}^{n}\frac{(V_i-\theta)^2}{1+\lambda(V_i-\theta)}\right| \\ &\ge \frac{|\lambda|}{n}\sum_{i=1}^{n}\frac{(V_i-\theta)^2}{1+\lambda(V_i-\theta)} - \frac{1}{n}\left|\sum_{i=1}^{n}(V_i-\theta)\right| \\ &\ge \frac{|\lambda| S}{1+|\lambda| Z_n} - \left|\frac{1}{n}\sum_{i=1}^{n}(V_i-\theta)\right|. \end{aligned}$$

By Corollary 1, the second term is $O_p(n^{-1/2})$. Recalling Lemma 2, $S = \sigma_g^2 + o(1)$ *a.s.*, it follows that $\frac{|\lambda|}{1+|\lambda| Z_n} = O_p(n^{-1/2})$, and hence by (11),

$$|\lambda| = O_p(n^{-1/2}). \tag{13}$$

For convenience, let $\gamma_i = \lambda(V_i - \theta)$ where λ is the root of (9). Then by (11) and (13),

$$\max_{1\le i\le n}|\gamma_i| = O_p(n^{-1/2})o(n^{1/2}) = o_p(1). \tag{14}$$

Expanding (9),

$$\begin{aligned} 0 = g(\lambda) &= \frac{1}{n}\sum_{i=1}^{n}(V_i-\theta)\left(1-\gamma_i+\gamma_i^2/(1+\gamma_i)\right) \\ &= \frac{1}{n}\sum_{i=1}^{n}V_i - \theta - S\lambda + \frac{1}{n}\sum_{i=1}^{n}(V_i-\theta)\gamma_i^2/(1+\gamma_i), \end{aligned} \tag{15}$$

The final term in (15) is bounded by

$$\frac{1}{n}\sum_{i=1}^{n}|V_i-\theta|^3\lambda^2|1+\gamma_i|^{-1} = o(n^{1/2})O_p(n^{-1})O_p(1) = o_p(n^{-1/2}) \tag{16}$$

using (12), (13) and (14). Therefore, we may write

$$\lambda = S^{-1}\left(\frac{1}{n}\sum_{i=1}^{n} V_i - \theta\right) + \beta = S^{-1}(U_n - \theta) + \beta, \tag{17}$$

where $|\beta| = o_p(n^{-1/2})$. By Taylor's expansion,

$$\begin{aligned} -\frac{1}{2}\log R(\theta) &= \frac{1}{2}\sum_{i=1}^{n}\gamma_i - \frac{1}{4}\sum_{i=1}^{n}\gamma_i^2 + \frac{1}{2}\sum_{i=1}^{n}\eta_i \\ &= \frac{1}{2}n\lambda(U_n - \theta) - \frac{1}{4}nS\lambda^2 + \frac{1}{2}\sum_{i=1}^{n}\eta_i \\ &= \frac{n(U_n - \theta)^2}{4S} - \frac{1}{4}nS\beta^2 + \frac{1}{2}\sum_{i=1}^{n}\eta_i, \end{aligned}$$

where $\eta_i = O(|\gamma_i|^3)$ *a.s.*. The first term has an asymptotic distribution χ_1^2 by Lemma 1 and 2. By Lemma 2 and (17), the second term is bounded by

$$\left|-\frac{1}{4}nS\beta^2\right| = n(\sigma_g^2 + o(1))o_p(n^{-1}) = o_p(1).$$

From (12) and (13), the final term is bounded by $o_p(1)$. Therefore applying Slutsky theorem completes the proof.

References

[Chen and Qin, 2000]S. X. Chen and Y. S. Qin. Empirical likelihood confidence intervals for local linear smoothers. *Biometrika*, pages 946–953, 2000.

[Chen, 1996]S. X. Chen. Empirical likelihood confidence intervals for nonparametric density estimation. *Biometrika*, pages 329–341, 1996.

[DiCiccio *et al.*, 1989]T. S. DiCiccio, P. Hall, and J. Romono. Comparison of parametric and empirical likelihood functions. *Biometrika*, pages 465–476, 1989.

[Hall and LaScala, 1990]P. Hall and B. LaScala. Methodology and algorithms of empirical likelihood. *International Statistical Review*, pages 109–127, 1990.

[Hoeffding, 1948]W. Hoeffding. A class of statistics with asymptotically normal distribution. *Ann. Math. Statist.*, pages 293 – 325, 1948.

[Lee, 1990]A. J. Lee. *U-statistics, Theory and Practice.* Marcel Dekker, Inc., New York, 1990.

[Owen, 1988]A. B. Owen. Empirical likelihood ratio confidence intervals for a single functional. *Biometrika*, pages 237–249, 1988.

[Owen, 1990]A. B. Owen. Empirical likelihood ratio confidence regions. *The Annals of Statistics*, pages 90–120, 1990.

[Owen, 1991]A. B. Owen. Empirical likelihood for linear models. *The Annals of Statistics*, pages 1725–1747, 1991.

[Owen, 2001]A. B. Owen. *Empirical Likelihood.* Chapman and Hall, London, 2001.

[Sen, 1960]P. K. Sen. On some convergence properties of u-statistics. *Calcutta Statistical Association Bulletin*, pages 1–18, 1960.

[Wood *et al.*, 1996]A. T. A. Wood, K. A. Do, and N. M. Broom. Sequential linearization of empirical likelihood constraints with application to u-statistics. *Journal of Computational and Graphical Statistics*, pages 365–385, 1996.

Two results in statistical decision theory for detecting signals with unknown distributions and priors in white Gaussian noise.

Dominique Pastor

GET - ENST Bretagne,
CNRS UMR 2872 TAMCIC,
Technopôle de Brest Iroise, CS 83818,
29238 BREST Cedex, France
(e-mail: dominique.pastor@enst-bretagne.fr)

Abstract. Two recent advances in statistical decision and estimation theory are presented. These results concern the detection of signals whose amplitudes are above or equal to some bound and that are less present than absent in a background of white Gaussian noise. The first result describes the non parametric detection of such signals when the noise standard deviation is known whereas the second result affords to perform the detection when this standard deviation is unknown. For both results, the role played by thresholding tests on the observation norms is crucial. The detection of radar targets is a typical field of application of these results.
Keywords: Estimation theory, Likelihood theory, Limit theorem, Non parametric decision, Thresholding test.

1 A sharp upper-bound for the probability of error of the MPE decision scheme and the MPE suboptimal test.

Albeit simple, a reasonable model for observations performed by sensors is that of signals randomly present or absent in additive and independent white Gaussian noise (WGN). In contrast with the simplicity of this model, the detection of such signals on the basis of a set of observations can be intricate. Actually, in many applications of most importance, very little is known about the observations or most of their parameters ([Kailath and Poor, 1998, section I]). In such situations, the detection of signals of interest cannot be achieved by standard likelihood theory based on the usual Bayes, minimax and Neyman-Pearson criteria for these ones require full knowledge of the signal distributions. Nonparametric and robust detection ([Poor, 1994, section III.E]), as well as Generalized Likelihood Ratio Tests ([Kay, 1998]), are then alternative formulations affording to deal with such cases. For instance, Constant False Alarm Rate (CFAR) systems standardly used in radar processing for detecting targets with a specified false alarm rate typically derive from such alternative approaches ([Minkler and Minkler, 1990]).

In [Pastor *et al.*, 2002], we investigate how far we can get if we assume only two hypotheses on the signal to detect. First, the signal is supposed to be less present than absent in the sense that its prior is less than or equal to one half; second, the norm of this same signal is assumed to be larger than or equal to some positive real number A. The purpose of such assumptions is to bound our lack of prior knowledge. The following theorem is then established in [Pastor *et al.*, 2002]. In the statement of this theorem, $\mathbf{I}_n$ stands for the identity matrix with size $n \times n$; by thresholding test with threshold height T, we mean the binary hypothesis test whose decision is that some signal is present if the observation norm exceeds T and whose decision is that noise only is present otherwise; finally, we remind the reader that the so-called MPE decision scheme is basically the likelihood ratio test that yields the least probability of error amongst all possible binary hypothesis tests ([Poor, 1994]).

Theorem 1 *Let* $U, \Lambda, X : \Omega \to \mathbf{R}^n$ *be three random vectors and let* $\epsilon : \Omega \to \{0,1\}$ *be a random variable defined on the same probability space* $(\Omega, \mathcal{B}, P)$ *such that* Λ*,* X *and* ϵ *are independent,* $X \in \mathcal{N}(0, \sigma_0^2 I_n)$ *and* $U = \epsilon\Lambda + X$*.*

Let $V(\rho)$ *be the function of the positive real variable* ρ

$$\begin{aligned} V(\rho) = {} & \frac{e^{-\rho^2/2}}{2^{n/2}\Gamma(n/2)} \int_0^{\xi(\rho)} e^{-t^2/2} t^{n-1} {}_0F_1(n/2\,;\rho^2 t^2/4)dt \\ & + \frac{1}{2}\left[1 - \frac{2^{1-n/2}}{\Gamma(n/2)} \int_0^{\xi(\rho)} e^{-t^2/2} t^{n-1} dt\right]. \end{aligned} \tag{1}$$

where $\xi(\rho)$ *is the unique positive solution for* x *in the equation*

$${}_0F_1(n/2; \rho^2 x^2/4) = e^{\rho^2/2}. \tag{2}$$

Then, given any positive real number $A > 0$*, for any* Λ *less present than absent with norm almost surely larger than or equal to* A*,* $V(A/\sigma_0)$ *is an upper-bound for the probability of error of both the MPE decision scheme and the threshold test with threshold height* $\sigma_0\xi(A/\sigma_0)$*. This bound is reached by both tests when the prior* $P(\{\varepsilon = 1\})$ *equals* $1/2$ *and* Λ *is uniformly distributed on the sphere with radius* A *centred at the origin.*

The thresholding test with threshold height $\sigma_0\xi(A/\sigma_0)$ is hereafter called the MPE suboptimal test. It is basically nonparametric in the sense given by [Poor, 1994] since $V(A/\sigma_0)$ is the constant performance measurement this test guarantees over the whole class of those signals less present than absent with norms larger than or equal to A.

2 Detection of relatively big signals in WGN with unknown level: the Essential Supremum Test.

The thresholding test introduced by theorem 1 is workable in practice only if the noise standard deviation is known. On the basis of theorem 2 stated in subsection 2.2 below, subsection 2.3 then introduces an algorithm named the Essential Supremum Test (EST) and aimed at detecting signals of interest even if the noise standard deviation is unknown. Beforehand, we need some appropriate notations and pieces of terminology.

2.1 Some notations.

The characteristic function of a set K will be denoted by $\mathcal{X}_K$: $\mathcal{X}_K(x) = 1$ if $x \in K$ and $\mathcal{X}_K(x) = 0$ otherwise. A real number x (resp. an integer k) is said to be positive if $x > 0$ (resp. $k > 0$). The real number x (resp. the integer k) is said to be non negative if $x \geq 0$ (resp. $k \geq 0$).

Only one probability space $(\Omega, \mathcal{M}, P)$ is considered in what follows.

Given any positive integer n, $\| \cdot \| : \mathbf{R}^n \to [0, \infty)$ will stand for the usual euclidean norm on $\mathbf{R}^n$. Given any n-dimensional random vector $Y : \Omega \to \mathbf{R}^n$, $\|Y\|$ will stand for the random variable $\|Y\| : \Omega \to [0, \infty)$ that assigns the non negative real number $\|Y(\omega)\|$ to every given $\omega \in \Omega$.

Let $\mathbf{S}$ henceforth stands for the set of all the sequences of n-dimensional real random vectors defined on Ω. Given some positive real number σ_0 and some natural number n, an element $X = (X_k)_{k \in \mathbf{N}}$ of $\mathbf{S}$ will be called an *n-dimensional WGN with standard deviation* σ_0 if the random vectors X_k, $k \in \mathbf{N}$, are mutually independent and identically Gaussian distributed with null mean vector and covariance matrix $\sigma_0^2 \mathbf{I}_n$ ($X_k \mathrm{sim} \mathcal{N}(0, \sigma_0^2 \mathbf{I}_n)$).

As usual, we denote by $L^2(\Omega, \mathbf{R}^n)$ the Hilbert space of those n-dimensional real random vectors $Y : \Omega \to \mathbf{R}^n$ such that $E[\|Y\|^2] < \infty$. We will hereafter deal with the set of those elements $\Lambda = (\Lambda_k)_{k \in \mathbf{N}}$ of $\mathbf{S}$ such that $\Lambda_k \in L^2(\Omega, \mathbf{R}^n)$ for every $k \in \mathbf{N}$ and $\sup_{k \in \mathbf{N}} E[\|\Lambda_k\|^2]$ is finite. According to standard notations, we denote this subset of $\mathbf{S}$ by $\ell^\infty(\mathbf{N}, L^2(\Omega, \mathbf{R}^n))$.

2.2 A limit theorem

The subsequent theorem derives from a more general result established in [Pastor, 2004] and suffices for achieving our purpose, that is introducing the EST.

Theorem 2 *Let $U = (U_k)_{k \in \mathbf{N}}$ be some element of $\mathbf{S}$ such that $U = \varepsilon \Lambda + X$ where $\Lambda = (\Lambda_k)_{k \in \mathbf{N}}$, $X = (X_k)_{k \in \mathbf{N}}$ and $\varepsilon = (\varepsilon_k)_{k \in \mathbf{N}}$ are respectively an element of $\mathbf{S}$, some n-dimensional WGN with standard deviation σ_0 and a sequence of random variables valued in $\{0, 1\}$.*

Assume that

(H1) *for every $k \in \mathbf{N}$, Λ_k, X_k and ε_k are mutually independent;*

(H2) *the random vectors U_k, $k \in \mathbf{N}$, are mutually independent;*

(H3) *the set of priors $\{\{P(\{\varepsilon_k = 1\}) : k \in \mathbf{N}\}$ has a maximum p in $[0, 1)$ and the random variables ε_k, $k \in \mathbf{N}$, are mutually independent;*

(H4) *$\Lambda \in \ell^\infty(\mathbf{N}, L^2(\Omega, \mathbf{R}^n))$ and there exists $A \in (0, \infty)$ such that, for every $k \in \mathbf{N}$, $\|\Lambda_k\| \geq A$ almost surely.*

Then, σ_0 is the only strictly positive real number σ, such that, for every $\beta_0 \in (0, 1]$,

$$\lim_{A\to\infty} \left\| \overline{\lim_m} \left| \frac{\displaystyle\sum_{k=1}^m \|U_k\| \mathcal{X}_{[0,\beta\sigma\xi(A/\sigma)]}(\|U_k\|)}{\displaystyle\sum_{k=1}^m \mathcal{X}_{[0,\beta\sigma\xi(A/\sigma)]}(\|U_k\|)} - \sigma G_n\left(\beta\xi(A/\sigma)\right) \right| \right\|_\infty = 0 \qquad (3)$$

uniformly in $\beta \in [\beta_0, 1]$ where, for every non negative real value x,

$$G_n(x) = \frac{\displaystyle\int_0^x t^n e^{-t^2/2} dt}{\displaystyle\int_0^x t^{n-1} e^{-t^2/2} dt}.$$

2.3 The Essential Supremum Test

Let L be some natural number and set $\beta_\ell = \ell/L$ for every $\ell \in \{1, \ldots, L\}$. On the basis of theorem 2, given some elementary event $\omega \in \Omega$ and m vectors $U_1(\omega), \ldots, U_m(\omega)$, the idea is then to estimate σ_0 by an eventually local minimum $\hat{\sigma}_0(m, \omega)$ of

$$\sup_{\ell\in\{1,\ldots,L\}} \left\{ \left| \frac{\displaystyle\sum_{k=1}^m \|U_k(\omega)\| \mathcal{X}_{[0,\beta_\ell\sigma\xi(A/\sigma)]}(\|U_k(\omega)\|)}{\displaystyle\sum_{k=1}^m \mathcal{X}_{[0,\beta_\ell\sigma\xi(A/\sigma)]}(\|U_k(\omega)\|)} - \sigma G_n\left(\beta_\ell\xi(A/\sigma)\right) \right| \right\}, \qquad (4)$$

when σ runs through the search interval $(0, \sigma_{max}(m, \omega)]$ where

$$\sigma_{\max}(m, \omega) = \sup_{k\in\{1,\ldots,m\}} \{\|U_k(\omega)\|\}/\sqrt{n}.$$

When σ runs through the search interval proposed above, the discrete cost (4) is a scalar bounded nonlinear function of σ. We thus seek an eventual local minimum of the discrete cost (4) by means of a standard minimization routine such as the golden section search and parabolic interpolation ([Forsythe *et al.*, 1976] and [Press *et al.*, 1992]). Given $k \in \mathbf{N}$, the decision on the value of ε_k is then achieved by replacing, in the expression of the MPE suboptimal test, the exact value of σ_0 by its estimate. The resulting binary hypothesis test is then the map of Ω into $\{0, 1\}$ defined by $\hat{\mathcal{T}}_k = \mathcal{X}_{[0,\infty)}\left(\|U_k\| - \hat{\sigma}_0(m, \omega)\xi(A/\hat{\sigma}_0(m, \omega))\right)$.

Our choice for the search interval upper bound is then justified as follows. If $\hat{\sigma}_0(m,\omega)$ might be larger than $\sigma_{\max}(m,\omega)$, we would take the risk to get an estimate larger than every ratio $\|U_k(\omega)\|/\xi(A/\sigma_{\max}(m,\omega))$, when $k \in \{1,\ldots,m\}$. Indeed, $\xi(\rho) \geq \sqrt{n}$ ([Pastor *et al.*, 2002]) for all non negative real value ρ. Thereby, the outcome of the test $\widehat{\mathcal{T}}_k$ could be that no signal is present whereas the full absence of signals of interest amongst m observations is hardly probable when m is large.

2.4 Some experimental results

The performance of the EST should be less than that of the MPE suboptimal test. However, when m and A increase, we also can expect that the performance measurements of these two tests become close to each other. If so, above which values for m and A can the essential supremum test be considered as workable in practice? Till now, we have no theoretical answer to this question and it seems hardly feasible to get an experimental answer to it because we simply do not known which priors and distributions to choose for such experiments? Therefore, in this section, we will be satisfied with some experimental results concerning the following basic case.

With the same notations as those used so far, we suppose that for every given $k \in \mathbf{N}$, U_k, Λ_k and X_k are two-dimensional random vectors ($n = 2$) where Λ_k is uniformly distributed on the circle centred at the origin with radius A. We further assume that $P(\varepsilon_k = 1\}) = 1/2$. Given $k \in \mathbf{N}$, the two components of Λ_k can be regarded as the in-phase and quadrature components of a sinusoidal carrier with amplitude A and phase uniformly distributed in $[0, 2\pi]$. Thereby, deciding whether ε_k equals 0 or 1 is the standard "Non coherent Detection of a Modulated Sinusoidal Carrier" problem ([Poor, 1994, Example III.B.5, p. 65]). The MPE decision scheme for making a decision on the value of ε_k is the thresholding test whose threshold height is the unique solution in x to the equation $I_0\left(A/\sigma_0 x\right) = e^{A^2/2\sigma_0^2}$, where I_0 is the zeroth order modified Bessel function of the first kind ([Poor, 1994, Example II.E.1, p. 34]). Since $I_0(x) = {}_0F_1(1; x^2/4)$, the reader will easily verify that the result is also a straighforward consequence of theorem 1.

Suppose now that the noise standard deviation is unknown. If we dispose of m observations U_k, $k = 1,\ldots,m$, we can estimate this standard deviation by minimizing the discrete cost (4) on the basis of those m references. This estimate can then be used for tuning the EST and, on the basis of the $(m+1)$th observation U_{m+1}, make a decision on the value of ε_{k+1}. This decision making has a certain probability of error $\hat{V}_m(A/\sigma_0)$. If m and A are large enough, $\hat{V}_m(A/\sigma_0)$ and $V(A/\sigma_0)$ are expected to draw near to each other. In other words, when m and $\rho \in (0,\infty)$ are large enough, $\hat{V}_m(\rho)$ and $V(\rho)$ should be close to each other. We thus carry out simulations so as to experimentally verify this intuitive claim.

In these simulations, $\sigma_0 = 1$ for this choice induces no loss of generality; given $\rho \in (0,\infty)$, $\hat{V}_m(\rho)$ is computed by choosing signals uniformly distributed on the sphere centred at the origin with radius ρ. We minimize the discrete cost (4) with $L = m$ as a trade-off between accuracy of the estimate and computational cost. Given $m \in \mathbf{N}$ and $\rho \in (0, \infty)$, we approximate $V_m(\rho)$ by the EST Binary Error Rate (BER), computed as follows. Given $j \in \mathbf{N}$, the EST estimates σ_0 on the basis of the m observations $U_{(j-1)(m+1)+k}$, $k = 1, 2, \ldots, m$ and makes a decision on the value of $\varepsilon_{j(m+1)}$. If I_j stands for the indicator variable defined by $I_j = 1$ if the EST makes the wrong decision on the value of $\varepsilon_{j(m+1)}$ and by $I_j = 0$ otherwise, the random variables I_j, $j \in \mathbf{N}$, are mutually independent because of the mutual independence of the trials. It turns out that estimating $\hat{V}_m(\rho)$ by the sample proportion S_k/k, where $S_k = \sum_{j=1}^{k} I_j$ and k is some specified number of trials, is not suitable with respect to our purpose. Indeed, $\hat{V}_m(\rho)$ is expected to approximate reasonably well $V(\rho)$ for large values of m and ρ; now, $V(\rho)$ rapidly decreases with ρ; hence, the accuracy of the sample proportion S_k/k may significantly depend on the value of $\hat{V}_m(\rho)$. Thence, we resort to inverse binomial sampling as practitioners in telecommunication systems usually do since error probabilities also decrease rapidly with input signal to noise ratios. The BER is thus defined as the ratio i/K where $K = \inf\{k \in \mathbf{N} : S_k = i\}$ is the minimum number of trials experimentally required for achieving a predefined number of errors equal to i.

Figures 1 to 3 present experimental results obtained for different values for m. Each figure displays $V(\rho)$ and the BERs of the EST for $\rho = 0.5, 1, 1.5, \ldots, 5$ and a pre-specified number of errors i equal to 400, which is a reasonable choice according to practitioners in telecommunication systems. As expected, the larger m and ρ, the closer $V(\rho)$ and $\hat{V}_m(\rho)$.

Consider now the case of sinusoidal carriers with amplitudes all equal to $C\rho$ with $C > 1$. According to theorem 2, the least we can expect is that the larger C, the better the performance of the EST. For instance, the results displayed in figure 4 were obtained for $m = 300$ and signals of interest with amplitude A one dB larger than the value ρ, that is $A = 1.2589\rho$. These results strongly suggest that the asymptotic conditions of theorem 2 are not so constraining in practice and can probably be relaxed.

3 Perspectives and extensions

Forthcoming work should address the respective influence of the EST various parameters, analyse how the asymptotic conditions of theorem 2 can actually be relaxed and assess the quality of EST estimate of the noise standard deviation.

A natural application of the approach presented in this paper is the design of Constant False Alarm Rate (CFAR) systems used in radar processing for detecting targets. Our intention is then to study to what extent theorems 1 and 2 are complementary to standard results and algorithms such as those described in [Minkler and Minkler, 1990].

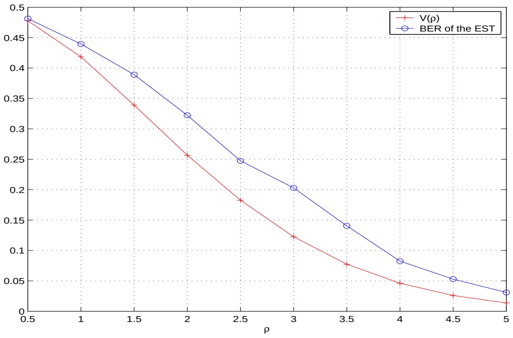


Fig. 1. Performance of the EST with $L = 100$ and $m = 100$ references for the non coherent detection of modulated sinusoidal carriers .

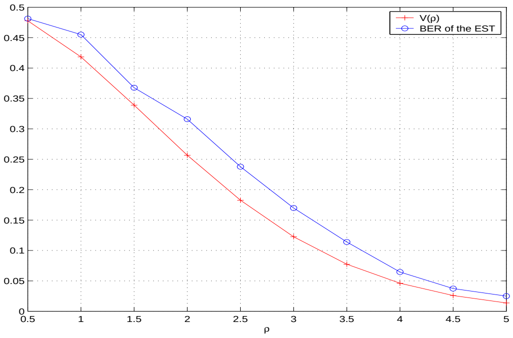


Fig. 2. Performance of the EST with $L = 200$ and $m = 200$ references for the non coherent detection of modulated sinusoidal carriers.

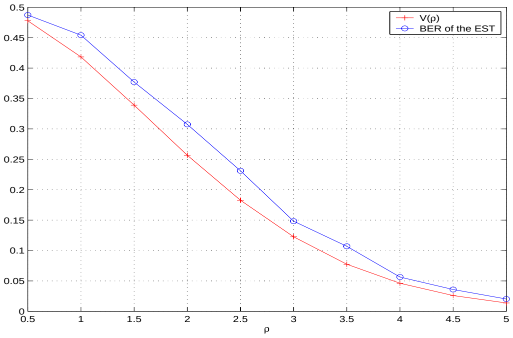


Fig. 3. Performance of the EST with $L = 300$ and $m = 300$ references for the non coherent detection of modulated sinusoidal carriers.

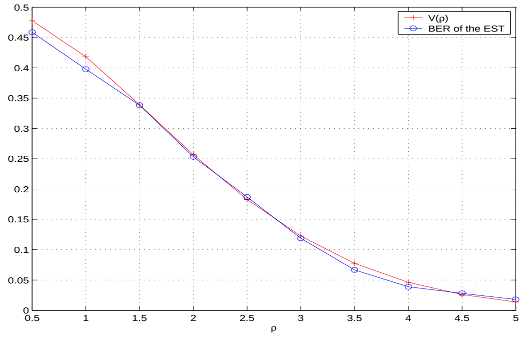


Fig. 4. Performance of the EST with $L = 300$ and $m = 300$ references for the non coherent detection of modulated sinusoidal carriers with amplitudes $A[dB]$ equal to $\rho[dB] + 1$.

References

[Forsythe *et al.*, 1976]G. E. Forsythe, M. A. Malcolm, and C. B. Moler. *Computer Methods for Mathematical Computations*. Prentice-Hall, 1976.

[Kailath and Poor, 1998]T. Kailath and H. V. Poor. Detection of stochastic processes. *IEEE Transactions On Information Theory*, pages 2230–2259, 1998.

[Kay, 1998]S. M. Kay. *Fundamentals of Statistical Signal Processing, Volume II: Detection Theory*. Prentice Hall, Upper Saddle River, 1998.

[Minkler and Minkler, 1990]G. Minkler and J. Minkler. *The Principles of Automatic Radar Detection In Clutter, CFAR*. Magellan Book Company, Baltimore, 1990.

[Pastor *et al.*, 2002]D. Pastor, R. Gay, and A. Groenenboom. A sharp upper-bound for the probability of error of the likelihood ratio test for detecting signals in white gaussian noise. *IEEE Transactions On Information Theory*, pages 228–238, 2002.

[Pastor, 2004]D. Pastor. *A limit theorem for sequences of independent random vectors with unknown distributions and its application to nonparametric detection.* GET/ENST-Bretagne Internal Report, Paris, 2004.

[Poor, 1994]H. V. Poor. *An Introduction to Signal Detection and Estimation.* Springer-Verlag, 1994.

[Press *et al.*, 1992]W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery. *Numerical recipes in C, The art of Scientific Computing*. Cambridge University Press, 1992.

Asymptotic Efficiency in Censored Alternating Renewal Processes

Enrique E. Alvarez

University of Connecticut
215 Glenbrook Road CLAS 328
Storrs, CT 06269 - 4120, USA
(e-mail: ealvarez@stat.uconn.edu)

Abstract. Consider a process that jumps back and forth between two states, with random times spent in between. Suppose the durations of subsequent on and off states are i.i.d. and that the process has started far in the past, so it has achieved stationarity. We estimate the sojourn distributions through maximum likelihood when data consist of several realizations observed over windows of fixed length. For discrete and continuous time Markov chains, we also examine if there is any loss of efficiency when ignoring the stationarity structure in the estimation.

Keywords: Alternating renewal process, Asymptotic efficiency, Window censoring.

1 Introduction

Consider a machine which periodically fails, undergoes technical service, and is put to work again, so that the working and out-of-service times form an alternating renewal process (ARP). Suppose further that the machine was placed in service in the indefinite past, so that the process may be regarded as stationary. Our interest here is to estimate the distribution of the on and off times when several such processes are observed over a time interval, or when the same process is observed over several "well separated" windows.

Such alternating renewal processes have been taken as models for diverse phenomena such as system availability and reliability in engineering [Pham-Gia and Turkkan, 1999], or the behavior of healthy-sick cycles in actuarial and insurance mathematics [Ramsay, 1984]. They have also been of interest as building blocks for other processes where the cumulative count from many alternating renewal processes whose inter-arrival times have high or infinite variance can produce aggregate network traffic that exhibits long range dependence [Murad S. Taqqu and Sherman, 1997].

The present study is concerned with estimating the distribution of the time spent in each of the states with maximum likelihood methods, when the data consist of "windows" from several stationary ARPs.

2 Construction and stationarity of ARPs

Consider a set of pairs of positive random variables $\{(Z_*, Y_*), (Z_1, Y_1), \ldots\}$ with the property that the first pair (Z_*, Y_*)simQ_0 and it is independent from the remaining $(Z_i; Y_i) \overset{\text{iid}}{\text{sim}}\, Q$. That kind of arrangement constitutes an *alternating renewal sequence* with *inter-arrival times* $X_* = Z_* + Y_*$, $X_i := Z_i + Y_i$, and *renewal times* $S_0 := X_*$ and $S_n := S_0 + \sum_1^n X_i$ for $n > 0$.

Consider the counting process $N(t) := \sum_0^\infty I\{S_n \in [0; t]\}$ and in order to record the state of the process at each time, introduce $W(t) := I\{S_{N(t)-1} + Z_{N(t)} > t\}$, which is the *alternating renewal process* associated with the renewal sequence. Thus the distribution of $W := \{W(t), t \geq 0\}$ is determined by Q_0 and Q; call the process *pure* if $X_* \equiv 0$ or *delayed* otherwise. Think of the Z's and Y's denoting durations of *on* and *off* times respectively; and for identifiability assume throughout that $P(Z_i = 0) = P(Y_i = 0) = 0$ for all $i \in \mathbb{Z}^+$.

Note also that the initial random vector (Z_*, Y_*) can be thought of as resulting from an ordinary pair (Z_0, Y_0)sim$Q$ through truncation, as

$$Z_* = (X_* - Y_0)^+ \text{ and } Y_* = X_* \wedge Y_0. \tag{1}$$

In particular, situations with $Z_* = 0$ correspond to paths beginning in the off-state.

In this study we are concerned not with pure but with delayed alternative renewal processes, the importance of which is that with an appropriate choice of Q_0 the process W is stationary, in a sense to be defined shortly. Figure 1 shows a typical sample path observed over the "window" of time $[0, T]$.

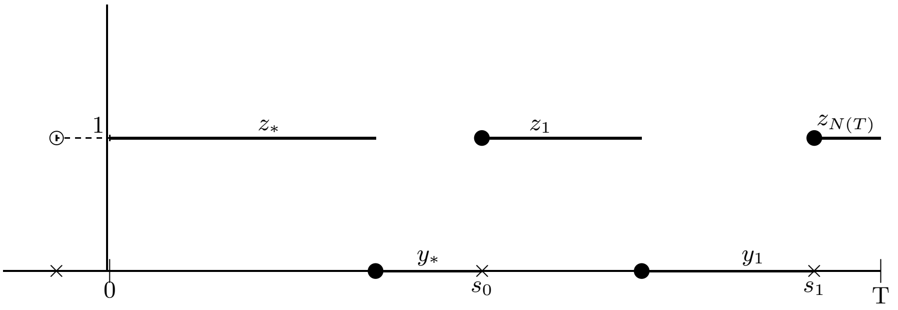


Fig. 1. A Sample Path from a Delayed ARP over $[0, T]$

2.0.0.1 Stationarity Choose any $t \in \mathbb{R}^+$ (deterministically or randomly but independent of the process) and construct a new alternating renewal sequence $\{(Z_i^t, Y_i^t), i \geq 0\}$ by censoring everything to the left of t. This is, the new sequence has an initial pair

$$\begin{aligned}Z_*^t &= (S_{N(t)-1} + Z_{N(t)} - t)^+,\\ Y_*^t &= Y_{N(t)} - (t - S_{N(t)-1} - Z_{N(t)})^+;\end{aligned}$$

and subsequently $Z_i^t = Z_{N(t)+i}$ and $Y_i^t = Y_{N(t)+i}$, for $i \geq 1$. Notice that because this construction implies that $Z_*^t = 0$ on the event $C := \{S_{N(t)-1} + Z_{N(t)} \leq t\}$, the distribution of the random variable Z_*^t has a point mass at zero whenever C has positive probability.

Definition 1 *Call the ARP stationary if and only if the two sequences* $\{(Z_*, Y_*), (Z_i, Y_i), i \geq 1\}$ *and* $\{(Z_*^t, Y_*^t), (Z_i^t, Y_i^t), i \geq 1\}$ *are equal in distribution for every* $t \in [0, \infty)$.

Assume that $X := X_1$ has finite expectation μ_X and denote $Z := Z_1$, $Y := Y_1$.

Theorem 21 *If the distribution of the initial pair* (Z_*, Y_*) *is given by*

$$Q_0(z, y) = \frac{1}{\mu_X} E_Q \ \{(z \wedge Z) \ 1 \,[Y \leq y] + (y \wedge Y)\}, \qquad (2)$$

then process $\{W(t), t \geq 0\}$ *is stationary in the sense of definition 1.*

See [4]. In the special case when the on-time $Z\mathrm{sim}H$ is independent of the off-time $Y\mathrm{sim}G$ this gives

$$Q_0\,(z, y) = \frac{\mu_Y}{\mu_X} \int_0^y \frac{1 - G(u)}{\mu_Y} du + \frac{\mu_Z}{\mu_X} G\,(y) \int_0^z \frac{1 - H(u)}{\mu_Z} du. \qquad (3)$$

3 A two-states Markov chain

The simplest example of a window censored alternating renewal process is a pair of consecutive observations from a Markov chain on {0,1}. When the transition probabilities are $\pi_0 := P(W_{t+1} = 1 | W_t = 0)$ and $\pi_1 := P(W_{t+1} = 1 | W_t = 1)$, the stationary distribution is given by

$$q := P\{W_t = 0\} = \tfrac{1-\pi_1}{1-\pi_1+\pi_0},\ p := P\{W_t = 1\} = \tfrac{\pi_0}{1-\pi_1+\pi_0}\,.$$

The joint density of a pair of consecutive observations is

$$P(W_t = x_i; W_{t+1} = y_i) = \frac{\pi_0(1-\pi_1)}{1-\pi_1+\pi_0} \left(\frac{\pi_1}{1-\pi_1}\right)^{x_i y_i} \left(\frac{1-\pi_0}{\pi_0}\right)^{(1-x_i)(1-y_i)}. \qquad (4)$$

This is of exponential family form with complete sufficient statistic T, and canonical parameter η given respectively by

$$T = \begin{pmatrix} X_i Y_i \\ (1 - X_i)(1 - Y_i) \end{pmatrix} \text{ and } \eta = \begin{pmatrix} \ln \pi_1 - \ln(1 - \pi_1) \\ \ln(1 - \pi_0) - \ln \pi_0 \end{pmatrix}.$$

By standard results in exponential families theory [11], the maximum likelihood estimators are

$$\widehat{\pi_0} = \tfrac{\sum_{i=1}^n (X_i-Y_i)^2}{2n-\sum_{i=1}^n X_i-\sum_{i=1}^n Y_i} \text{ and } \widehat{\pi_1} = \tfrac{2\sum_{i=1}^n X_i Y_i}{\sum_{i=1}^n X_i+\sum_{i=1}^n Y_i};$$

and $\sqrt{n}(\hat{\pi}-\pi) \Rightarrow N(0;\hat{\Sigma})$, where

$$\hat{\Sigma} = \frac{1}{2}\left(1-\pi_1+\pi_0\right)\begin{pmatrix} \pi_0\left(1-\pi_0\right)\frac{1+\pi_0}{1-\pi_1} & -\pi_1\left(1-\pi_0\right) \\ -\pi_1\left(1-\pi_0\right) & \pi_1\frac{2-\pi_1}{\pi_0}\left(1-\pi_1\right) \end{pmatrix}.$$

Alternatively, we could ignore stationarity in order to estimate π_0 and π_1 by the sample proportion of transitions into each state, i.e.

$$\widetilde{\pi_0} = \tfrac{\sum_{i=1}^n (1-X_i)Y_i}{\sum_{i=1}^n (1-X_i)} \text{ and } \widetilde{\pi_1} = \tfrac{\sum_{i=1}^n X_i Y_i}{\sum_{i=1}^n X_i}.$$

By the multivariate central limit theorem and the delta method, $\sqrt{n}(\tilde{\pi}-\pi) \Rightarrow N(0;\tilde{\Sigma})$, with

$$\tilde{\Sigma} = \left(1-\pi_1+\pi_0\right)\begin{pmatrix} \frac{\pi_0(1-\pi_0)}{1-\pi_1} & 0 \\ 0 & \frac{\pi_1(1-\pi_1)}{\pi_0} \end{pmatrix}.$$

At this point, it is natural to ask what is lost in terms of efficiency by ignoring stationarity in the estimation. To address this question, consider the difference matrix $\hat{\Sigma} - \tilde{\Sigma} =: (1-\pi_1+\pi_0)\ \Delta$. It is easy to check that the diagonal entries of Δ are strictly negative and that the cross-products are equal. Therefore, the matrix difference $(\hat{\Sigma}-\tilde{\Sigma})$ has one eigenvalue which is negative and the other is zero. This result is surprising, because it implies that there exist functions of the transition probabilities for which ignoring stationarity is of no consequence asymptotically. Essentially, any function of (π_0, π_1) with gradient proportional to the eigenvector corresponding to the null eigenvalue of Δ will have that property. This will be explored further for continuous time Markov chains in section 4.

4 A continuous time Markov chain

When the on and off times follow independent exponential distributions Z_i sim Q_z = exp (λ_1) and $Y_i\text{sim}Q_y = \exp(\lambda_2)$, the process $\{W(t), t \geq 0\}$ is a continuous time Markov chain. At any given time, the excess life is independent of the history of the process.

The stationary distribution is, according to equation (3):

$$Q_0(z,y) = \frac{\lambda_2}{\lambda_1+\lambda_2}(1-e^{-z\lambda_1})\left(1-e^{-\lambda_2 y}\right) + \frac{\lambda_1}{\lambda_1+\lambda_2}(1-e^{-y\lambda_2}), \tag{5}$$

with marginal distributions

$$Q_0(\infty,y) = (1-e^{-y\lambda_2}) \text{ and } Q_0(z,\infty) = \tfrac{\lambda_2}{\lambda_1+\lambda_2}(1-e^{-z\lambda_1}) \ + \ \tfrac{\lambda_1}{\lambda_1+\lambda_2}.$$

Notice that Z_0 is independent of Y_0, since $Q_0(z, y) = Q_0(\infty, y)Q_0(z, \infty)$.

Reference [Alvarez, 2003] investigates how to obtain a likelihood for a sample path of an ARP observed on a window $[0, T]$, as a Radon-Nykodym derivative with respect to an appropriately chosen dominating measure and restricted to a filtration that corresponds to the censoring mechanism. The main result is that the window-censored likelihood ratio is a product of three types of factors:

i) In a typical sample path where at least one transition in observed, we multiply
  (a) the value of initial density
  (b) the values of the densities at all non-censored on and off times
  (c) the survival function for the duration of the last state in the window

ii) Secondly, if the window $[0, T]$ contains no jumps, the likelihood equals the survival function of the excess life in either state.

Using the above recipe, after some algebra we obtain the likelihood over a window $[0, T]$ as

$$l(T) = \frac{{\lambda_1}^{\tau+1\{W(T)=0\}}{\lambda_2}^{\tau+1\{W(0)=1\}}}{\lambda_1 + \lambda_2} \exp\left[-\lambda_1 \mathrm{on}(T) - \lambda_2 \mathrm{off}(T)\right], \qquad (6)$$

where $\mathrm{on}(t) := \int_0^t W(t)dt =: t - \mathrm{off}(t)$. This additive property is characteristic to the Markov chain and it is fairly intuitive. Because of the memoryless property of the exponential distribution, the break up of the total on or off times into subperiods does not provide any additional information on their distribution. When we observe m windows independently up to a same time T, the log-likelihood over the sample is the sum of the corresponding path likelihoods.

4.1 Asymptotic normality

Following standard theorems in asymptotic statistics it is established that the likelihood equation has a unique root with probability tending to 1 as $m \to \infty$ and that $\sqrt{n}\left(\widehat{\lambda}_n - \lambda_0\right) \Rightarrow N(0, \hat{\Sigma})$ with

$$\hat{\Sigma} = \frac{(\lambda_1 + \lambda_2)}{(\lambda_1 T + \lambda_2 T + 2)} \begin{pmatrix} \lambda_1 \frac{\lambda_1 T + \lambda_2 T + 1}{\lambda_2 T} & 1/T \\ 1/T & \lambda_2 \frac{\lambda_1 T + \lambda_2 T + 1}{\lambda_1 T} \end{pmatrix}.$$

Notice that while the main diagonal entries are $O(1/T)$, the off-diagonal entries are $O(1/T^2)$ as $T \to \infty$. This is intuitive, since the only reason why the estimators of λ_1 and λ_2 are dependent is the presence in the data of the initial (left censored) observations. As the observation window enlarges, the information provided by the first two observations becomes negligible and the estimators closer to being independent.

4.2 Comparison with classic estimators

As in the discrete Markov chain example of Section 3, it is natural to ask if there is any loss in efficiency by ignoring stationarity in the estimation.

Suppose that we "condition away" the initial states. That is, we seek a log-likelihood function conditioned on $\sigma\left\{Z_0 1(Z_0 > 0), Y_0 1(Z_0 = 0)\right\}$. This is given over a single window by

$$\ln l^c(T) = [\tau + r_1 + d_0 - 1]\,(\ln \lambda_1) + \tau\,(\ln \lambda_2) - \lambda_1 \mathrm{on}(T) - \lambda_2 \mathrm{off}(T),$$

and its gradient is

$$\nabla \ln l^c(T) = \begin{pmatrix} (\tau + r_1 + d_0 - 1)/\lambda_1 - \mathrm{on}(T) \\ \tau/\lambda_2 - \mathrm{off}(T) \end{pmatrix}.$$

The conditional maximum likelihood estimators can be easily found over m windows to be

$$\widetilde{\lambda_1} = \tfrac{\tau + r_1 + d_0 - m}{\mathrm{on}(T)} \text{ and } \widetilde{\lambda_2} = \tfrac{\tau}{\mathrm{off}(T)}.$$

It is easy to check that

$$E\left[-\nabla^2 \ln l^c(T)\right]^{-1} = \frac{\lambda_1 + \lambda_2}{T} \begin{pmatrix} \frac{\lambda_1}{\lambda_2} & 0 \\ 0 & \frac{\lambda_2}{\lambda_1} \end{pmatrix}.$$

Therefore, $\sqrt{m}(\tilde{\lambda} - \lambda) \Rightarrow N(0; \tilde{\Sigma})$ with

$$\tilde{\Sigma} = \frac{\lambda_1 + \lambda_2}{T} \begin{pmatrix} \frac{\lambda_1}{\lambda_2} & 0 \\ 0 & \frac{\lambda_2}{\lambda_1} \end{pmatrix},$$

which coincides with the approximation for the unconditional m.l.e's for large T's. To compare the two methods asymptotically let

$$\hat{\Sigma} - \tilde{\Sigma} =: \tfrac{\lambda_1 + \lambda_2}{T} \tfrac{1}{\lambda_1 T + \lambda_2 T + 2} \Delta \quad \text{with} \quad \Delta = \begin{pmatrix} -\frac{\lambda_1}{\lambda_2} & 1 \\ 1 & -\frac{\lambda_2}{\lambda_1} \end{pmatrix}.$$

As in the discrete chain, Δ is negative semidefinite since $\mathrm{tr}(\Delta) < 0$ and $|\Delta| = 0$. The m.l.e. is then better than its conditional version, with a gain in efficiency that depends inversely on the truncation time and which is also affected by the relative means of the on and off times.

On the other hand, Δ has eigenpairs

$$\left[0, (\lambda_2, \lambda_1)'\right] \text{ and } \left[\left(-\tfrac{\lambda_1}{\lambda_2} - \tfrac{\lambda_2}{\lambda_1}\right), (-\lambda_1, \lambda_2)'\right],$$

which can be used to decompose $\Delta = PDP'$, with

$$P = \frac{1}{\sqrt{\lambda_1^2 + \lambda_2^2}} \begin{pmatrix} \lambda_2 & -\lambda_1 \\ \lambda_1 & \lambda_2 \end{pmatrix} \text{ and } D = \begin{pmatrix} 0 & 0 \\ 0 & -\frac{\lambda_1}{\lambda_2} - \frac{\lambda_2}{\lambda_1} \end{pmatrix}.$$

This suggests the definition of a new parameter $\eta = \eta(\lambda)$ by

$$\begin{pmatrix} \eta_1(\lambda_1, \lambda_2) \\ \eta_2(\lambda_1, \lambda_2) \end{pmatrix} := \begin{pmatrix} \lambda_1 \lambda_2 \\ \frac{1}{2}\lambda_2^2 - \frac{1}{2}\lambda_1^2 \end{pmatrix}.$$

This map is continuous and has the Jacobian matrix

$$D_\eta = \begin{pmatrix} \frac{\partial}{\partial \lambda_1}\eta_1(\lambda_1, \lambda_2) & \frac{\partial}{\partial \lambda_2}\eta_1(\lambda_1, \lambda_2) \\ \frac{\partial}{\partial \lambda_1}\eta_2(\lambda_1, \lambda_2) & \frac{\partial}{\partial \lambda_2}\eta_2(\lambda_1, \lambda_2) \end{pmatrix} = \begin{pmatrix} \lambda_2 & \lambda_1 \\ -\lambda_1 & \lambda_2 \end{pmatrix}.$$

By the delta method, the estimators $\widehat{\eta} = \eta\left(\widehat{\lambda}\right)$ and $\widetilde{\eta} = \eta\left(\widetilde{\lambda}\right)$ are asymptotically normal and the difference in covariance matrices is

$$D_\eta\left(\hat{\Sigma} - \tilde{\Sigma}\right) D'_\eta = \frac{1}{T}\frac{\lambda_1 + \lambda_2}{\lambda_2 \lambda_1}\frac{1}{\lambda_1 T + \lambda_2 T + 2}\begin{pmatrix} 0 & 0 \\ 0 & -\left(\lambda_1^2 + \lambda_2^2\right)^2 \end{pmatrix}.$$

The product of the hazard rates is estimated equally efficiently by the two methods, asymptotically, but for estimation of the difference in the square of the hazard rates the unconditional m.l.e. is better. As before, the gain in efficiency depends inversely on the truncation time.

For the parameter $\eta_2(\lambda_1, \lambda_2) = \frac{1}{2}\lambda_2^2 - \frac{1}{2}\lambda_1^2$ the asymptotic relative efficiency (ARE) of $\tilde{\eta_2}$ w.r.t. $\hat{\eta_2}$ is given by

$$\text{A.R.E.}(\tilde{\eta_2}, \hat{\eta_2}) = 1 - \frac{(\lambda_1^2 + \lambda_2^2)^2}{2(\lambda_1^4 + \lambda_2^4)} \Big/ \left[1 + \frac{1}{2}(\lambda_1 + \lambda_2)T\right].$$

The fraction in the numerator varies between 0 when $\lambda_1 \to 0$ and 1 when $\lambda_1 = \lambda_2$. When T is small the gains in efficiency could be substantial. As an example, Table 1 quantifies these gains for a few combination of parameters values.

Case:	**i**	**ii**	**iii**	**iv**	**v**
λ_1	0.5	0.5	0.5	0.5	0.5
λ_2	1	1	0.5	0.5	0.5
T	4	20	2	1	0.5
A.R.E.$(\tilde{\eta_2}, \hat{\eta_2})$	**0.82**	**0.95**	**0.50**	**0.33**	**0.20**

.

Table 1. A.R.E. of $\tilde{\eta_2}$ w.r.t. $\hat{\eta_2}$

References

[Alvarez, 2003]Enrique E. Alvarez. *Likelihood Based Estimation of Stationary Semi-Markov Processes Under Window Censoring.* PhD thesis, Universtity of Michigan, 2003.

[Murad S. Taqqu and Sherman, 1997]Walter Willinger Murad S. Taqqu and Robert Sherman. Proof of a fundamental result in self-similar traffic modeling. *Computer Communication Review*, 2:5–23, 1997.

[Pham-Gia and Turkkan, 1999]T. Pham-Gia and N. Turkkan. System availability in a gamma alternating renewal process. *Naval Res. Logist.*, 46:822–844, 1999.

[Ramsay, 1984]Colin M. Ramsay. The asymptotic ruin problem when the healthy and sick periods form an alternating renewal process. *Insurance Mathematics and Economics*, 3:139–143, 1984.

Asymptotic results for the MPL estimators of the Contact Process

Xavier Guyon[1] and Besnik Pumo[2]

[1] SAMOS, Université Paris 1,
90 rue de Tolbiac, 75634 Paris 13e, France
(e-mail: Xavier.Guyon@univ-paris1.fr)
[2] Institut National d'Horticulture - UMR SAGAH,
42 rue G. Morel, BP 60057, 49071 Beaucouzé, France
(e-mail: Besnik.Pumo@inh.fr)

Abstract. Let X be a discrete time contact process (CP) on $\mathbf{Z}^2$ as defined by Durrett and Levin (1994). We study the estimation of the model based on space-time evolution of X, that is, $T+1$ successive observations of X on a finite subset S of sites. We consider the maximum marginal pseudo-likelihood (MPL) estimator and show that, when $T \to \infty$, this estimator is consistent and asymptotically normal for a non vanishing supercritical CP. Numerical studies confirm the theoretical results and compare the MPL estimators with coding method estimators. Finally we present some results on CP of order d.

Keywords: Contact process, supercritical process, marginal pseudo-likelihood, identifiability of a model, consistency, asymptotic normality.

1 Introduction and description of the model

Consider a simple model of spread of a single species population evolving in $\mathbf{Z}^2$. Depending on some biological parameters, the dynamics is determined by specifying, for each site $s \in \mathbf{Z}^2$, the conditional probability that site s will be in state $X_{t+1}(s) = y \in \{0,1\}$ at time $t+1$ given X_t, the configuration at time t. State 1 (respectively 0) means that there is a (respectively no) plant in s. In this paper we propose an estimator for the parameters of the model, based on observations of X at instants $t = 0, \ldots, T$ on a finite and fixed subset S of $\mathbf{Z}^2$ and study the asymptotic properties of the estimator when the process is non vanishing on S. Fiocco and Zwet considered the estimation problem based on one observation at time t, when t is sufficiently large ([Fiocco and Zwet, 2003]).

We consider the *discrete time version* of the *Contact Process* (CP) as defined by Durrett & Levin [Durrett and Levin, 1994]. Suppose that the transition probability at a site s and at time t is stationary in space and time and depends locally on $x_{t-1}(\mathcal{N}_1(s))$, the first order neighbourhood of the site s at time $t-1$, where $\mathcal{N}_d(s) = \{u \in \mathbf{Z}^2 : \|s-u\|_1 \leq d\}$.

The system evolves as follows:

a. Each plant alive at time t dies with a probability γ at time $t+1$,

b. If the plant in s survives, then it produces an offspring that is dispersed to $u \in \partial s$, where $\partial s = \mathcal{N}_1(s) \setminus \{s\}$, with probability λ; the reproduction events for different values of s and different $u \in \partial s$ are independent,

c. If one or more plants are dispersed to s, or if there is a plant at s that survives between t and $t+1$, then $X_{t+1}(s) = 1$; otherwise $X_{t+1}(s) = 0$.

Furthermore, events defined on (a) and (b) are independent in time.

This model depends on the parameter $\theta = (\gamma, \lambda)$ and we suppose that $\theta \in (0,1)^2$. Other models are possible by defining different rules of evolution (cf. [Mollison, 1977] for example). Finally, some of the methods developed in our paper can be generalized for non stationary processes in space and/or in time.

The 'all 0' in $\mathbf{Z}^2$ state is an absorbing state. So, to make sense, for the asymptotic study, we need a condition **(I)**, verified with probability 1 conditionally to the non-extinction of X on S, the fixed domain of observation. Note that a CP survives with positive probability for a supercritical process that is CP such that $P(\tau = +\infty) > 0$ where τ gives the extinction time of the process ([Durrett and Levin, 1994]).

The paper is organized as follows. In section 2 we define the *marginal pseudo-likelihood* (MPL) estimator of θ. The identifiability of MPL is presented in section 3 and asymptotic results of MPL estimator in section 4. In section 5 we consider some simulations studies and compare numerically MLP estimators with coding method estimators proposed by Besag ([Besag, 1972]). A brief discussion on CP of order d is given in section 6.

Proofs of results are to be found in [Guyon and Pumo, 2004].

2 Marginal pseudo-likelihood (MPL)

Let $\mathrm{x(T)} = (\mathrm{x_0}, \mathrm{x_1}, \cdots, \mathrm{x_T})$ be $(T+1)$ successive configurations of X, S a finite subset of $\mathbf{Z}^2$ and $S_1 = \{u \in \mathbf{Z}^2 : \exists v \in S \text{ such that } \|u - v\|_1 \leq 1\}$. The estimator of θ we choose is a value which maximize a MPL of $\mathrm{x}(T)$ observed on S_1. The idea of pseudo-likelihood is classic in statistic: gaussian pseudo-likelihood for stationary field on $\mathbf{Z}^d$ ([Whittle, 1963]), conditional pseudo-likelihood for a Markov field on a lattice ([Besag, 1974]).

For a subset $A \subset S$, let denote $P_A(x_t, x_{t+1}; \theta)$ the transition-probability $P(X_{t+1}(A) = x_{t+1}(A) \mid X_t(S_1) = x_t(S_1))$. As the transition-probability for $A = S$ is analytically intractable, as $\#(S)$, the number of sites of S, is important, we will use the following *marginal pseudo-transition probability* $M_S(x_t, x_{t+1}; \theta)$ on S, in order to estimate θ. $M_S(x_t, x_{t+1}; \theta)$ is the product of $P_{\{s\}}(x_t, x_{t+1}; \theta)$ for $s \in I(x_t)$, where:

$$I(x_t, S) = \{s \in S : \; \exists x_{t+1} \; \text{ s.t. } P_{\{s\}}(x_t, x_{t+1}; \theta) > 0\}$$

The product of these marginal pseudo-transitions at consecutive instants define the MPL. For $s \in S$ and A a finite subset of $\mathbf{Z}^2$, denote $m(x_t, A) =$

$\sum_{s \in A} x_t(s)$, the number of sites of A occupied by x_t. As the model is isotropic in space, the law of $X_{t+1}(s)$ given x_t depends only on $c(x_t, s)$:

$$c(x_t, s) = (x_t(s), m(x_t, \partial s)) \in \mathcal{C}_1 = \{0, 1\} \times \{0, 1, 2, 3, 4\}. \tag{1}$$

More precisely, $X_{t+1}(s)$ conditionally on x_t is a Bernoulli random variable:

$$P_{\{s\}}(x_t, x_{t+1}; \theta) = p(x_t, s; \theta)^{1-x_{t+1}(s)}(1 - p(x_t, s; \theta))^{x_{t+1}(s)},$$

where $p(x_t, s; \theta) = \gamma^{x_t(s)} \delta^{m(x_t, \partial s)}$ and $\delta = \gamma + (1 - \gamma)(1 - \lambda)$ controls non-proliferation at time $(t + 1)$ in a site $s' \in \partial s$ of a plant present in s at time t. Since $X_{t+1}(s) = 0$ if $c(x_t, s) = (0, 0)$, only sites $s \in I(x_t)$ are informative in the transition $t \mapsto t + 1$. So:

$$M_S(x_t, x_{t+1}; \theta) = \prod_{s \in I(x_t)} p(x_t, s; \theta)^{1-x_{t+1}(s)}(1 - p(x_t, s; \theta))^{x_{t+1}(s)} \tag{2}$$

with convention $M(0, 0; \theta) = 1$ if $I(x_t) = \emptyset$. Denote $\eta = \gamma + (1 - \gamma)(1 - \lambda)^2$: η controls non-proliferation at time $(t+1)$ in the set $\{s, s'\}$ of a plant present in $u \in \partial s \cap \partial s'$ at time t.

By a direct calculation it follows that:

$$Cov(X_{t+1}(s), X_{t+1}(s') \mid x_t) = p(x_t, s; \theta)\ p(x_t, s'; \theta)\ [b(x_t, s, s'; \theta) - 1]$$

where

$$b(x_t, s, s'; \theta) = \begin{cases} \delta^{-m(x_t, \{s, s'\})} & if\ s' \in \mathcal{N}_1(s) \setminus \{s\} \\ \delta^{-2m(x_t, \partial s \cap \partial s')} \eta^{m(x_t, \partial s \cap \partial s')} & if\ s' \in \mathcal{N}_2(s) \setminus \mathcal{N}_1(s) \\ 1 & if\ s' \notin \mathcal{N}_2(s). \end{cases} \tag{3}$$

In particular if $s' \notin \mathcal{N}_2(s)$, $(X_{t+1}(s) \mid x_t)$ and $(X_{t+1}(s') \mid x_t)$ are independent.

Using (2) for $t = 0, \cdots, T - 1$, let us give the explicit expression of MPL based on observation of x(T) on S_1. Denote $n(x_t)$ (respectively $n(x_t, c)$) the number of informative sites of the configuration x_t on S (respectively with configuration $c \in \mathcal{C}_1$) and:

$$n(T) = \sum\nolimits_{t=0}^{T-1} n(x_t), \quad n(T, c) = \sum\nolimits_{t=0}^{T-1} n(x_t, c).$$

Clearly $n(T) = \sum_{c \neq (0,0)} n(T, c)$. The normalized log-marginal pseudo-likelihood of x(T) observed on S_1 is:

$$l_T(\theta) = \frac{1}{n(T)} \sum_{t=0}^{T-1} \sum_{s \in I(x_t)} \{\log[p(x_t, s; \theta)]^{\bar{x}_{t+1}(s)} + \log[\bar{p}(x_t, s; \theta)]^{x_{t+1}(s)}\} \tag{4}$$

where $\bar{x}_{t+1}(s) = 1 - x_{t+1}(s), \bar{p}(x_t, s; \theta) = 1 - p(x_t, s; \theta)$. The maximum MPL estimator of θ (or MPLE) is a value which maximize the MPL,

$$\hat{\theta}_T = \arg_\theta \max l_T(\theta).$$

3 MPL allows identification of $\boldsymbol{\theta}$

In order to prove that MPL allows identification of θ, we need to show that π_c is strictly positive for two linearly independent configurations, where:

$$\pi_c = \underline{\lim}_{T\to\infty}\frac{n(T,c)}{n(T)}.$$

The positivity of $\pi_c > 0$ for $c \in \mathcal{C}_1^*$, the set of configurations on $\mathcal{N}_1(0)$ such that $x(0) = 1$, is obtained by the following Lemma under the condition **(I)** of non-extinction of X on S :

$$(\mathbf{I}) : I_\infty = \{\mathrm{x} = (x_t, t \geq 0) \ \textit{such} \ n(\mathrm{x(T)}) \to \infty \text{ as } \mathrm{T} \to \infty\}.$$

Lemma 1 *Let $\mathcal{C}_1^*$ be the set of configurations on $\mathcal{N}_1(0)$ such that $x(0) = 1$. Then there exists $\alpha > 0$ such that, $\forall c \in C_1^*$, and $\forall x \in I_\infty$, we have $\pi_c \geq \alpha$.*

From the positivity of π_c, it follows that under **(I)** and for large T, $\theta \to l_T(\theta)$ allows identification of θ. Indeed:

- if $\mathrm{x}(T)$ realizes two linearly independent configurations $c_a = (u_a, v_a)$ and $c_b = (u_b, v_b)$, then $\theta \mapsto l_T(\theta)$ is an injective function;
- under **(I)**, the probability that each configuration c of $\mathcal{C}_1^*$ appears on S converges to 1 when $T \to \infty$.

In conclusion let as make two important remarks:

i) As X_∞ is spatially translation-invariant and ergodic, [Durrett, 1995], it follows that $\lim_{T\to\infty}\frac{n(T,c)}{n(T)}$ exists and is strictly positive for $c \in \mathcal{C}_1$.

ii) Space and/or time invariance of the model is not crucial on the proof of the subergodicity result: a similar result can be proved for non translation invariant models under the supplementary condition that transition probabilities are uniformly positive.

4 Consistency and normality of the MPL estimator

Let $f : U \to R$ be a real function twice continuously differentiable on an open subset U of $\mathbf{R}^d$ and $f^{(1)}(\theta)$ the vector of first derivatives. The following result sets up the consistency and asymptotic normality of the maximum MPLE $\hat{\theta}_T$ associated to (4). The proofs are based on Theorem 3.4.3 and 3.4.5 of Guyon ([Guyon, 1995]). In order to prove the positivity of $J_T(\theta_o)$ we used an idea of Jensen and Künsch ([Jensen and Künsch, 1994]) and a subergodicity result which generalize Lemma 1.

Let I_2 be the 2×2 identity matrix, $A_T(\theta_o)$, $B_T(\theta_o)$ the 2×2 matrices:

$$A_T(\theta_o) = \frac{1}{n(T)} \sum_{t=0}^{T-1} \sum_{s \in I(x_t)} \frac{p^{(1)t}[p^{(1)}]}{p(1-p)}(x_t, s; \theta_o) \tag{5}$$

$$B_T(\theta_o) = \frac{1}{n(T)} \sum_{t=0}^{T-1} \sum_{s,s' \in I(x_t)} [b(x_t, s, s'; \theta_o) - 1] \; \frac{p^{(1)}(x_t, s; \theta_o)\, {}^t[p^{(1)}(x_t, s'; \theta_o)]}{[\bar{p}(x_t, s; \theta_o)]\, [\bar{p}(x_t, s'; \theta_o)]} \tag{6}$$

with $b(x_t, s, s'; \theta_o)$ given by (3).

Theorem 1 *Let us suppose that $\theta_o = (\gamma_o, \lambda_o)$, the true unknown value of the parameter, is an interior point of a compact $\Theta \subset]0,1[^2$. Then, under condition **(I)** the maximum MPL estimator is consistent:*

$$\lim_{T \to \infty} \hat{\theta}_T \overset{a.s.}{=} \theta_o.$$

and asymptotically normal:

$$\sqrt{n(T)}\, [A_T(\theta_o) + B_T(\theta_o)]^{-1/2} A_T(\theta_o)(\hat{\theta}_T - \theta_o) \xrightarrow{d} \mathbf{G}_2(0, I_2).$$

5 Numerical studies

In this section we give some empirical results with S the 64×64 square lattice and initial configuration 'all sites occupied'. To avoid boundary effects we have used periodic boundary conditions. In Fig. 1 we present the evolution of the bias and the standard deviation of $\hat{\gamma}_T$ and $\hat{\lambda}_T$ for $T = 1, \ldots, 99$ for the supercritical CP with parameters $\gamma_o = 0.35, \lambda_o = 0.25$.

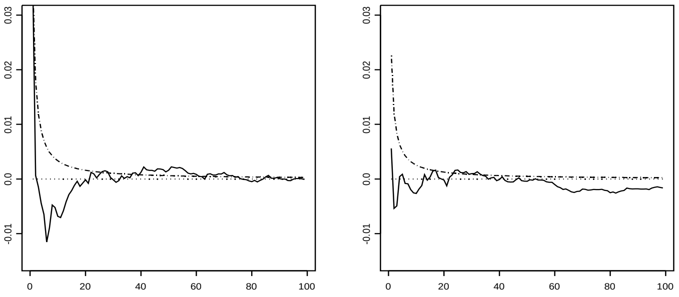


Fig. 1. Evolution of the bias (solid lines) and standard deviation (multiplied by 100, dotted lines) for the estimators of γ_o (left) and λ_o (right) for the supercritical CP with parameters $\gamma_o = 0.35, \lambda_o = 0.25$.

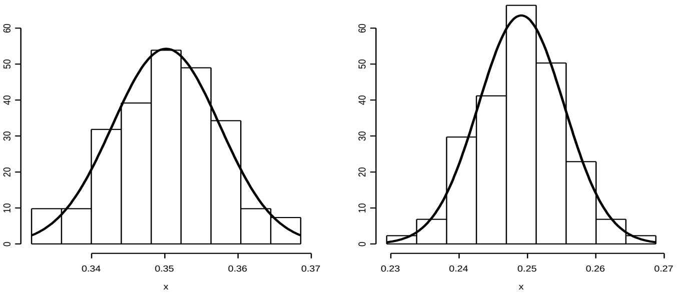


Fig. 2. Histograms of 100 estimations of γ_o (left) and λ_o (right) for the supercritical CP with parameters $\gamma_o = 0.35, \lambda_o = 0.25$.

Empirical study of asymptotic normality of estimators for a supercritical CP is based in 100 simulations with $T = 99$. Histograms are presented in Fig. 2. Asymptotic normality is checked by using a chi-squared test at level 5% and defining 9 equiprobable classes. Normality is accepted for $\hat{\gamma}$ (respectively $\hat{\lambda}$) since $\chi^2 = 1.7$ (respectively $\chi^2 = 4.4$) and $\chi^2_{0.95}(6) = 12.59$.

We also compared the estimated standard errors $\hat{\sigma}_{\hat{\gamma}}, \hat{\sigma}_{\hat{\lambda}}$ and empirical standard errors $s_{\hat{\gamma}}, s_{\hat{\lambda}}$ for the supercritical CP with parameter $\gamma_o = 0.35$, $\lambda_o = 0.25$. The values $\hat{\sigma}_{\hat{\gamma}_4}, \hat{\sigma}_{\hat{\lambda}_4}$ are obtained from a single simulation with $T = 4$ by applying Theorem 1 where $A_4(\theta_o)$ (respectively $B_4(\theta_o)$) are approximated by $A_4(\hat{\theta}_4)$ (respectively $B_4(\hat{\theta}_4)$). The empirical standard errors $s_{\hat{\gamma}_4}, s_{\hat{\lambda}_4}$ are obtained from 100 estimations for the 100 simulations. The results are presented in Table 1. As expected, there are few differences between estimated standard errors and empirical standard errors. Finally, Table 2 gives the estimations of γ_o and λ_o for six CP with parameters (γ_o, λ_o) $\in$ (0.2, 0.4, 0.6) × (0.1, 0.2). In these simulations, $T = 4$ and 40% of sites, randomly chosen, were occupied at time $t = 0$. We compare MPL estimators with coding method of estimation introduced by Besag ([Besag, 1972]). Let $K = 3 \times \mathbf{Z}^2 \cap S$, a *strong-coding subset* that is $\partial s \cap \partial s' = \emptyset$ for $s \neq s'$ of K. As variables $\{(X_{t+1}(s) \mid X_t = x), s \in K\}$ are independent, the normalized log-conditional likelihood of the CP restricted on sites s of K is given by:

	$\hat{\sigma}_{\hat{\gamma}_4}$	$s_{\hat{\gamma}_4}$	$\hat{\sigma}_{\hat{\lambda}_4}$	$s_{\hat{\lambda}_4}$
MPL estimations	0.0074	0.0074	0.0063	0.0058

Table 1. Comparison of estimated and empirical standard deviation

$$l_{T,K}(\theta) = \frac{1}{n_K(T)} \sum_{t=0}^{T-1} \sum_{s \in I_K(x_t)} \{\log[p(x_t, s; \theta)]^{\bar{x}_{t+1}(s)} + \log[\bar{p}(x_t, s; \theta)]^{x_{t+1}(s)}\}$$

	$\lambda = 0.1$					$\lambda = 0.2$				
γ	$\hat{\gamma}_4$	$\hat{\sigma}_{\hat{\gamma}_4}$	$\hat{\lambda}_4$	$\hat{\sigma}_{\hat{\lambda}_4}$	$n(4)$	$\hat{\gamma}_4$	$\hat{\sigma}_{\hat{\gamma}_4}$	$\hat{\lambda}_4$	$\hat{\sigma}_{\hat{\lambda}_4}$	$n(4)$
0.2	0.210	0.006	0.104	0.003	14600	0.189	0.006	0.193	0.004	15285
0.4	0.391	0.008	0.106	0.004	12406	0.399	0.008	0.188	0.005	13573
0.6	0.597	0.009	0.100	0.005	9223	0.607	0.009	0.206	0.008	10457

Table 2. Estimation of the parameters and their standard deviation.

where $I_K(x_t)$ gives the set of informative sites of K, $n_K(x_t) = \sharp(I_K(x_t))$ and $n_K(T) = \sum_{t=0}^{T-1} n_K(x_t)$. The *$K$-coding estimator* of θ is a value which maximize $l_{T,K}(\theta)$. By applying this method of estimation for six CP we obtained the results presented in Table 3.

	$\lambda = 0.1$					$\lambda = 0.2$				
γ	$\hat{\gamma}_4$	$\hat{\sigma}_{\hat{\gamma}_4}$	$\hat{\lambda}_4$	$\hat{\sigma}_{\hat{\lambda}_4}$	$n(4)$	$\hat{\gamma}_4$	$\hat{\sigma}_{\hat{\gamma}_4}$	$\hat{\lambda}_4$	$\hat{\sigma}_{\hat{\lambda}_4}$	$n(4)$
0.2	0.217	0.014	0.100	0.008	2442	0.183	0.013	0.170	0.010	2568
0.4	0.400	0.019	0.113	0.011	2106	0.400	0.019	0.184	0.014	2256
0.6	0.590	0.022	0.084	0.013	1542	0.602	0.022	0.198	0.023	1766

Table 3. Estimation of the parameters and their standard deviation obtained by K-coding method

6 Estimation of parameters of CP of order *d*

In this section we briefly present results for the CP of order d presented also in [Pumo and Le Corff, 2001] and which generalize the standard CP defined in the introduction. Denote ∂s a general neighbourhood of s. In order to define the CP of order d we only substitute b in the definition of the standard CP with b':

b'. If the plant in s survives, then it produces an offspring that is dispersed to $u = z + s \in \partial s$ with probability $g(z)$; the reproduction events for different values of s and different $u \in \partial s$ are independent,

Denote $\lambda = (\lambda_1, \ldots, \lambda_d)'$ the vector of different values of $g(z), z \in \partial 0 \backslash \{0\}$. Then we call d the order of the CP. The unknown parameter θ is defined now by $\theta = (\gamma, \lambda')$. It can be shown that similar results remains valid for the CP of order d. Furthermore, by applying Theorem 3.4.6 in [Guyon, 1995] we can do tests on parameters λ in order to determine the optimal neighbourhood for the definition of the model. In Table 4 we give estimations of six CP of order 2 with parameters $\theta_o = (\gamma_o, \lambda_{1o}, \lambda_{2o})$ where $(\gamma_o, \lambda_{1o}) \in (0.2, 0.4, 0.6) \times (0.1, 0.2)$ and $\lambda_{2o} = \lambda_{1o}/\sqrt{2}$. In these simulations we considered a 100×100 lattice and at time $t = 0$ all sites were occupied.

	$\lambda_1 = 0.1, \lambda_2 = 0.0707$			$\lambda_1 = 0.2, \lambda_2 = 0.1414$		
γ	$\hat{\gamma}$	$\hat{\lambda}_1$	$\hat{\lambda}_2$	$\hat{\gamma}$	$\hat{\lambda}_1$	$\hat{\lambda}_2$
0.2	0.199	0.098	0.072	0.200	0.197	0.144
0.4	0.399	0.101	0.072	0.399	0.200	0.141
0.6	0.597	0.102	0.065	0.598	0.199	0.135

Table 4. Estimation of parameters of CP of order 2.

References

[Besag, 1972]J. Besag. On the statistical analysis of nearest-neighbour systems. In *Proc. 9th Europian Meetying of Statist., Budapest*, pages 101–105, 1972.

[Besag, 1974]J. Besag. Spatial interaction and the statistical analysis of lattice systems. *JRSS B*, pages 192–225, 1974.

[Durrett and Levin, 1994]R. Durrett and S.A. Levin. Stochastic spatial models: a user's guide to ecological applications. *Phil. Trans. R. Soc. Lond.*, pages 329–350, 1994.

[Durrett, 1995]R. Durrett. *Ten lectures on particle systems. Cours de Saint Flour (1993).* Springer, USA, 1995.

[Fiocco and Zwet, 2003]M. Fiocco and W.R. Zwet. Parameter estimation for the supercritical contact process. *Bernoulli*, pages 1071–1092, 2003.

[Guyon and Pumo, 2004]X. Guyon and B. Pumo. Estimation spatio-temporelle d'un système de particules. *Préprint SAMOS*, pages 1–21, 2004.

[Guyon, 1995]X. Guyon. *Random fields on a network: modelling, statistics and applications.* Springer, Berlin, 1995.

[Jensen and Künsch, 1994]J.L. Jensen and H.R. Künsch. On asymptotic normality of pseudo-likelihood estimate for pairwise interaction processes. *Ann. Inst. Statist. Math.*, pages 475–486, 1994.

[Mollison, 1977]D. Mollison. Spatial contact models for ecological and epidemic spread, with discussion. *JRSS B.*, pages 283–326, 1977.

[Pumo and Le Corff, 2001]B. Pumo and J. Le Corff. Parameter estimation of the contact process on a lattice. In G. Govaert, J. Janssen, and N. Limnios, editors, *ASMDA 2001 Proceddings*, pages 866–871, 2001.

[Whittle, 1954]P. Whittle. On stationary process in the plane. *Biometrika*, pages 434–449, 1954.

A Practical Implementation of the Gibbs Sampler for Mixture of Distributions: Application to the Determination of Specifications in Food Industry

Julien Cornebise[1], Myriam Maumy[2], and Philippe Girard[3]

[1] E.S.I-E-A
72 avenue Maurice Thorez,
94200 IVRY SUR SEINE, France
(e-mail: cornebis@et.esiea.fr)
LSTA
Université Pierre et Marie Curie - Paris VI
175 rue du Chevaleret,
75013 PARIS, France

[2] IRMA
Université Louis Pasteur
7 rue René Descartes,
67084 STRASBOURG Cedex, France
(e-mail: mmaumy@math.u-strasbg.fr)

[3] Nestlé, Quality Management Department
Av. Nestlé, 55. CH-1800 VEVEY, Switzerland
(e-mail: philippe.girard@nestle.com)

Abstract. This article, mainly targeted to practitioners, illustrates practical issues that may arise when applying MCMC technics to a mixture of distributions model on real data. This data is provided by coffee manufacturer to determine specifications for soluble coffee. Assuming a known number of components, parameters of each component are estimated using the Gibbs sampler and specifications are derived as the 99% quantile of the first distribution. Convergence and label-switching are discussed. Determination of the number of components is also considered, via model selection using the Bayes Factors.

Keywords: MCMC, Mixture, Gibbs Sampler, Label switching, Bayes factors.

1 The Problem and its Modelling

Following an international agreement, a commercial product sold as pure soluble coffee must have been manufactured using green coffee only. However, in a minority of cases, economic adulteration of soluble coffee has been observed in some countries. As a matter of fact, few commercial soluble coffees have been shown to be adulterated with coffee husks/parchments, cereals, and some other plant extracts. In such cases, glucose and xylose contents have proven to be the most discriminant indicators to detect the adulteration. For pure soluble coffee, their concentration are low whereas they become high in case of adulteration. Provided a set of 1002 soluble coffee samples, on which both glucose and xylose concentrations have been measured, we are interested in determining:

- the number K of kinds of production, and their parameters (mean, standard deviation) : $(K-1)$ different frauds, plus one for pure coffee;
- the proportion of each population;
- from the first population and its corresponding characteristics, the specifications within which a soluble coffee can be considered as pure coffee ?

In this article, we only consider the univariate case. Therefore, glucose and xylose concentrations are considered as separate quantities. The approach could, in a further work, be generalised to the bivariate case.

In the univariate case, the observed distribution of the glucose (resp. the xylose) measured on the 1002 coffee samples is modeled as a mixture of normal distributions. We consider that the $T = 1002$ observations from the sample come from K distinct populations (1 pure and $(K-1)$ adulterated), each population $k \in \{1, \dots, K\}$ following a normal distribution of density f_k and of parameters $\theta_k = (\mu_k, \sigma_k^2)$.

Therefore, the likelihood of an observation x_i, $1 \leq i \leq T$ is:

$$[x_i|\boldsymbol{\theta}, \boldsymbol{\pi}] = \sum_{k=1}^{K} \pi_k f_k(x_i|\theta_k),$$

where $f_k(\bullet|\theta_k)$ is the probability density function (*pdf*) of a normal distribution with parameters θ_k and π_k is the probability of belonging to population k, such that $\sum_{k=1}^{K} \pi_k = 1$. The parameters of interest are $\boldsymbol{\theta} = (\theta_1, \dots, \theta_K)$ and $\boldsymbol{\pi} = (\pi_1, \dots, \pi_K)$. The choice of the value of K will be mentioned in section 6.

Other parametrisations for mixtures of normal distributions have been published: interested readers can refer to Robert in [Droesbeke *et al.*, 2002], or in [Marin *et al.*, to appear]. However, this parametrization lacks the natural physical interpretation of the parameters achieved with the actual one.

In the bayesian paradigm, parameters θ_k of each distribution are considered as random variables, having their own distribution. Starting from an initial knowledge about a phenomena described in the *prior distribution* of the parameters $(\boldsymbol{\theta}, \boldsymbol{\pi})$, the Bayes formula enables to update this information by adding the information brought by the data provided the model definition. The *prior* distribution, and its parameters, called *hyperparameters*, are a way to take mathematically into account prior knowledge of the experts of the field, if available (f.i., the potential informations held by the chemists).

To ease the reading, we use throughout the article the notation [.] introduced by [Gelfand *et al.*, 1990] to denote any pdf. In this notation, $[\boldsymbol{\theta}, \boldsymbol{\pi}]$ denotes the *prior* distribution for $(\boldsymbol{\theta}, \boldsymbol{\pi})$, $[y|\boldsymbol{\theta}, \boldsymbol{\pi}]$ the likelihood and $[\boldsymbol{\theta}|\boldsymbol{\pi}, x]$ is the conditional pdf of $\boldsymbol{\theta}$.

Finally, the eventual goal of this application is to estimate a function of the parameters $F(\boldsymbol{\theta}, \boldsymbol{\pi})$ where F can be either the identity function for each parameter or a quantile function. This is generally assessed by

$\mathbb{E}[F(\boldsymbol{\theta},\boldsymbol{\pi})|\boldsymbol{x}] = \int F(\boldsymbol{\theta},\boldsymbol{\pi})[\boldsymbol{\theta},\boldsymbol{\pi}|\boldsymbol{x}]d(\boldsymbol{\theta},\boldsymbol{\pi})$, where $[\boldsymbol{\theta},\boldsymbol{\pi}|\boldsymbol{x}]$ is the pdf of the *posterior* distribution, i.e. the distribution of the parameters conditionnaly to the observations $\boldsymbol{x} = (x_1,\ldots,x_T)$. This pdf is computed via the Bayes formula : $[\boldsymbol{\theta},\boldsymbol{\pi}|\boldsymbol{x}] = [\boldsymbol{x}|\boldsymbol{\theta},\boldsymbol{\pi}][\boldsymbol{\theta},\boldsymbol{\pi}]/[\boldsymbol{x}]$, where $[\boldsymbol{x}]$ is the *prior* distribution of the observations, which can be taken as constant and thus ignored.

The first question is therefore to choose the *prior* distribution of the parameters, $[\boldsymbol{\theta},\boldsymbol{\pi}]$.

Before going any further, we have to mention that a hidden variable $z_i,\ i \in \{1,\ldots,T\}$ has been introduced in the model mentioned above to ease its Bayesian analysis (see below). This variable is not observed and thus named *latent variable.* $z_i \in \{1,\ldots,K\}$ indicates the original population of the observation x_i, and $\boldsymbol{z} = (z_1,\ldots,z_T)$. The z_i are i.i.d, with $[z_i = k|\boldsymbol{\pi}] = \pi_k$, $[x_i|\boldsymbol{\theta},\boldsymbol{\pi},z_i = k] = \mathcal{N}(x|\mu_k,\sigma_k^2)$, where $\mathcal{N}(\bullet|\mu,\sigma^2)$ denotes the univariate normal pdf. Analysis of mixture distributions by MCMC methods have been the subject of many publications, for example [Diebolt and Robert, 1990], [Richardson and Green, 1997], [Stephens, 1997], [Marin *et al.*, to appear].

2 Choosing the Prior and its Hyperparameters

As part of the Bayesian analysis, *prior* definition is the first step to go through. Several cases may arise:

- either the experts of the field have valuable information about the distribution of the parameters that should be taken into account : for example, they approximately know what the mean of each component should be.
- or they do not have any information at all - or this information should be ignored, in order to check their results by an objective analysis. Two approaches can be chosen by the statistician :
  - using empirical *prior*, i.e. hyperparameters built upon the data.
  - using non-informative *prior*, i.e. *prior* carrying no information at all. This is somewhat hard to achieve, because purely non-informative *prior* can be improper (for example, uniform distribution on the whole space) and cause troubles. We can also use poorly informative *prior*, for example very dispersed normal distributions.

Moreover, the *prior* is often chosen in a closed-by-sampling or *conjugate prior* family, *i.e.* such that conditionning by the sample (passing from the *prior* to the posterior distribution) only result in a change of the hyperparameters, not in a change of family. This simplifies implementation.

Here we have chosen the following model, that often arises, because each distribution belongs to a closed-by-sampling family :

$$\begin{aligned}
\boldsymbol{\pi} = (\pi_1,\ldots,\pi_K) &\text{ sim } \mathrm{Di}(a_1,\ldots,a_K)\\
\mu_k|\sigma_k^2 &\text{ sim } \mathcal{N}(m_k,\sigma_k^2/c_k)\\
\sigma_k^{-2} &\text{ sim } \mathcal{IG}(\alpha_k,\beta_k)
\end{aligned}$$

where Di is a Dirichlet distribution, and $\mathcal{IG}$ an Inverse Gamma. Thus, the hyperparameters are a_k, m_k, c_k, α_k, and β_k, $k \in \{1, \ldots, K\}$. For non-informative *prior*, we should have a Uniform distribution for $\boldsymbol{\pi}$, which can be obtained with $a_1 = \ldots = a_K = 1$, a Dirac pdf for σ_k^2, which can be obtained with limit values for α_k and β_k, and a unary density on $\mathbb{R}$ for each μ_k, which is more difficult to implement with limit values on c_k.

In practice, non-informative *prior* is not so easy to deal with : for instance, when programming the algorithm with Matlab, it is not always possible to deal with infinite values of the parameters, of with such particular densities. Therefore, poorly informative *prior*, or even empirical *prior*, should be used. This is what we have done here.

We have chosen : $\forall\, k \in \{1, \ldots, K\}$, $\pi_k = 1$, $c_k = 1$, $m_k = \bar{x}$ (empirical mean), $\alpha_k = K$, and $\beta_k = (T(K-1))^{-1} \sum_{i=1}^{T} (x_i - \bar{x})^2$. Thus, the proportions of each component are non-informative, the means of the μ_k are equal to the empirical mean of the sample, and the means of σ_k^2 is equal to the empirical variance ($\mathbb{E}[\sigma_k^2] = (\alpha_k - 1)\beta_k$ for Inverse Gamma). For reasons that should become clear further (related to label-switching problems and Bayes factors), we have chosen the same hyperparameters for each components: this maintains the symmetry of the density (and therefore of the likelihood).

This part of the Bayesian analysis is certainly the most subjective. The choice of *prior* is clearly the weakest point of the analysis, and the more arguable and argued. Many discussions exist about it, and each approach has its pros and cons. The approach chosen here is neither the most rigorous one, nor the purest, but allows easy implementation. In order to overcome these discussions, a sensitivity analysis needs to be done after the study. Further discussions about the choice of the *prior* can be found in almost any reference: see for example [Droesbeke *et al.*, 2002], [Gelman *et al.*, 2003].

3 Gibbs Sampling, Complete Conditionnal distributions

The evaluation of the expectation is hard to achieve, either analitically or numerically (due either to its complex expression or to its highly multidimensional nature). MCMC methods are a good way to solve this problem. We recall that the principle of Monte-Carlo methods is to generate N independent realizations $(\boldsymbol{\theta}^{(i)}, \boldsymbol{\pi}^{(i)})$ of random variables following the posterior distribution $[\boldsymbol{\theta}, \boldsymbol{\pi}|\boldsymbol{x}]$, and to approximate : $\int F(\boldsymbol{\theta}, \boldsymbol{\pi})[\boldsymbol{\theta}, \boldsymbol{\pi}|\boldsymbol{x}] d(\boldsymbol{\theta}, \boldsymbol{\pi}) \approx \sum_{i=1}^{N} F(\boldsymbol{\theta}^{(i)}, \boldsymbol{\pi}^{(i)})/N$. Here the function F is either the identity function to estimate the parameters or the 99%-quantile function of the first component of the mixture (i.e. the component corresponding to pure coffee powder, without any kind of fraud). In this last case, in order to be the most conservative possible, the empirical 95%-quantile of the sampled values $F(\boldsymbol{\theta}^{(i)}, \boldsymbol{\pi}^{(i)})$, instead of their empirical mean, has been used.

Given the conditional structure of the model of interest, the Gibbs sampler has been used to generate a N-sample from the *posterior* distribution. Quite straightforward to implement, it relies on the availability of all *complete conditional distributions*. Let $\theta = (\theta_1, \ldots, \theta_n)$ the vector of the parameters of the model. Here, we have $\theta = (\boldsymbol{\theta}, \boldsymbol{\pi}) = (\mu_1, \sigma_1^2, \ldots, \mu_K, \sigma_K^2, \pi_1, \ldots, \pi_K)$. Let $\theta_{(i)} = (\theta_1, \ldots, \theta_{i-1}, \theta_{i+1}, \ldots, \theta_n)$ the vector θ without its i^{th} component. The density of the complete conditional distribution of θ_i is $[\theta_i | \theta_{(i)}, \boldsymbol{x}]$. Let us assume that we know its closed form for all $i \in \{1, \ldots, n\}$, wich is often the case. Let us also take arbitrary initial values $\theta^{(0)} = (\theta_1^{(0)}, \ldots, \theta_n^{(0)})$. The Gibbs sampler's algorithm consists in successively sampling:

- $\theta_1^{(i)}$ from $\left[\theta_1 | \theta_2^{(i-1)}, \theta_3^{(i-1)}, \theta_4^{(i-1)}, \ldots, \theta_n^{(i-1)}\right]$
- $\theta_2^{(i)}$ from $\left[\theta_2 | \theta_1^{(i)} \quad, \theta_3^{(i-1)}, \theta_4^{(i-1)}, \ldots, \theta_n^{(i-1)}\right]$
- $\cdots$
- $\theta_n^{(i)}$ from $\left[\theta_n | \theta_1^{(i)} \quad, \theta_2^{(i)} \quad, \theta_3^{(i)} \quad, \ldots, \theta_{n-1}^{(i)}\right]$

for $i \in \{1, \ldots, N + M\}$, where M is a number of iterations that will be discarded. They are called "burn-in" iterations, and correspond to the time before convergence. It can be shown that $\boldsymbol{\theta}^{(M)} = (\theta_1^{(M)}, \ldots, \theta_n^{(M)})$ converges in distribution to the posterior joint distribution $[\theta_1, \ldots, \theta_n | \boldsymbol{x}]$. The following N values are then considered as a sample from this distribution, and can be used to approximate $F(\theta)$ by empirical mean, as mentioned above.

In our model with the above assumptions, we have the following complete conditional distributions (in order to simplify the notations, the list of all parameters but θ are figured by (θ)):

$$\begin{aligned}
z_i | (\boldsymbol{z_i}), \boldsymbol{x} &\text{ sim } \mathrm{Mu}(\pi_1, \ldots \pi_K) \\
\boldsymbol{\pi} | (\boldsymbol{\pi}), \boldsymbol{x} &\text{ sim } \mathrm{Di}(a_1 + n_1, \ldots, a_K + n_K), \text{ where } n_k = \textstyle\sum_{i:z_i=k} 1 \\
\mu_k | (\mu_k), \boldsymbol{x} &\text{ sim } \mathcal{N}(\tfrac{m_k c_k + s_k}{c_k + n_k}, \tfrac{\sigma_k^2}{n_k + c_k}) \\
\sigma_k^{-2} | (\sigma_k^2), \boldsymbol{x} &\text{ sim } \mathcal{IG}\left(\alpha_k + \tfrac{n_k+1}{2}, \beta_k + \tfrac{1}{2}\left(c_k(\mu_k - m_k)^2 + \textstyle\sum_{i:z_i=k}(x_k - \mu_k)^2\right)\right)
\end{aligned}$$

We actually run the sampler with $M = 5000$ and $N = 5000$. Convergence issues and choice of M are discussed in the next section. Metropolis Hastings algorithm details, as well as variants of the Gibbs Sampler can be found in [Droesbeke *et al.*, 2002], or in [Richardson and Green, 1997] or also in [Marin *et al.*, to appear] for an approach more directly linked to mixture of distributions.

4 Convergence Issues

Since the first historically convergence diagnosis (known as the "thick pen" one, [Gelfand *et al.*, 1990]), there exist two main kinds of methods to determine whether the sampler has reached convergence or not, i.e. whether the values from the current generation can be considered as a sample of the target distribution. Two kinds of diagnoses can be considered depending on the number of chains run for carrying out the diagnosis: we consider either the single or multi-chain, where the single ones are rejected by [Gelman and Rubin, 1992b]. We then consider only the multi-chain procedure.

The principle of muti-chains diagnoses is to run multiple independent chains from very different starting points, and to test whether the last values of each chain come from the same distribution or not. If that is the case, we can assume that convergence has been reached. [Gelman and Rubin, 1992a] and [Gelman and Rubin, 1992b], have proposed a method which is often used, based on the comparison of within and between-chains variances. At the beginning of the sampling, the chains are much influenced by the starting point, and the between-chains variance is high above the within-chain one. When each chain has reached the target distribution, the ratio between within and between-chain variance should be around 1. This method has been much improved since then, and some more subtel tests are avalaible, though not implemented here.

If we note $x_{i,j}$ the i^{th} value of chain j, $i \in \{1, \ldots, M+N\}$, $j \in \{1, \ldots, J\}$, we compute the empirical within-chain and between-chain (respectively W and B) as follows

$$\begin{array}{r} W = \frac{1}{m}\sum_{j=1}^{m} \frac{1}{n-1}\sum_{i=1}^{n} \left(x_{i,j} - x_{\cdot,j}\right)^2 \\ B = \frac{n}{m-1}\sum_{j=1}^{m} \left(x_{\cdot,j} - x_{\cdot,\cdot}\right)^2 \end{array} \text{ with } \begin{array}{r} x_{\cdot,j} = \frac{1}{n}\sum_{i=1}^{n} x_{i,j} \\ x_{\cdot,\cdot} = \frac{1}{m}\sum_{j=1}^{m} x_{\cdot,j}. \end{array} .$$

The quantity $\hat{\sigma}_+^2 = \frac{n-1}{n}W + \frac{1}{n}B$ can be interpreted as an estimate of the variance of the target distribution. Gelman and Rubin show that, when the initial values of the J chains are chosen "sufficiently different", $\hat{\sigma}_+^2$ systematically overestimates the variance while chains have not reached convergence. Convergence diagnosis is thus based on the statistic $\sqrt{\hat{R}_{GR}} = \sqrt{\hat{\sigma}_+^2/W}$ which tends to 1 when $n \to +\infty$. Practically, convergence is considered as achieved when $\sqrt{\hat{R}_{GR}} < 1,2$. In a multiple parameters case, this diagnosis must be carried out for each parameter separately, the overall convergence being attained when all parameters have converged to its target distribution.

This method, quite straightforward to implement, has proven to be efficient in many cases. However, due to the label-switching issue (see below), this method appeared to be inefficient in our case. Future developments of this study will see this point worked through.

5 The Label Switching Problem

A particularity of the mixture of distributions is that the likelihood and the joint *posterior* pdf (which is the target distribution of the Gibbs Sampler) is symmetrical, i.e. invariant by permutation of the components. Therefore, this last pdf has up to $K!$ duplications of each mode, and the sampler can move from one mode to another freely, thus permuting the components.

The absence of label-switching means that the sampler is stuck in a local mode, maybe because modes are well separated (e.g. when K is very low). The space of parameters is thus not completely explored, which is dangerous.

When estimating a function $F(\boldsymbol{\theta}, \boldsymbol{\pi})$ which is also invariant by permutation of the components (e.g. estimating the pdf at a given point), label-switching is not a problem. But when trying to estimate the parameters of each component, this label-switching has to be undone ([Stephens, 2000b]). Two approaches can be foreseen: imposing an identification constraint during the sampling, or post-processing the generated N-sample to undo the permutation.

The first solution consists of constraining the exploration of the space of parameters by the sampler, which alters the results, see [Celeux *et al.*, 2000], [Marin *et al.*, to appear], [Stephens, 2000b]. Forcing the *prior* to be highly separable (using much different hyperparameters for components) is not a good idea neither : label-switching arises anyway, and the resulting lack of correspondance between the *prior* and the component would corrupt any further use of the *prior* (such as Bayes factor).

We applied here the second solution, i.e. the post-processing. Ordering on μ_k, or on σ_k^2, or on π_k is not a good idea. Some components may be close in mean but not in variance, and vice-versa. Some methods use the Kullback-Leibler "distance" and clustering algorithms (e.g. K-means with $K!$ classes) to determine to which mode (i.e. permutation) belongs each of the sampled vectors of parameters. The reader is invited to read the three references above for more details.

It has to be noted that label-switching is incompatible with convergence diagnoses mentionned in section 4 : comparing the variance between chains is meaningless when components can swap ! Moreover, clustering assumes that convergence has been reached, it would thus be non-sense to use variance-based diagnoses on post-processed samples.

6 Choosing a model : Bayes Factors

Until now, we have worked with a given number K of components. The question is now to choose between different models. Let $\mathcal{M} = \{M_1, \ldots, M_K\}$ be a finite ensemble of possible models (each one with k components, up to K, $K = 7$ in our application) to explain the observations x_i, parametrized by $\boldsymbol{\theta}$. The best model within the K possible is the one with the highest *posterior* probability.

The *posterior* probability of model M_k is calculated via Bayes formula as follows: $[M_k|\boldsymbol{x}] = ([M_k][\boldsymbol{x}|M_k]) \,/\, \left(\sum_{M_k \in \mathcal{M}} [M_k][\boldsymbol{x}|M_k]\right)$, where $[M_k]$ is an *prior* probability of M_k, with $\sum_{M_k \in \mathcal{M}} [M_k] = 1$ (e.g. $\forall\, k$, $[M_k] =$

$1/K$), and $[\boldsymbol{x}|M_k]$, defined by $[\boldsymbol{x}|M_k] = \int [\boldsymbol{x}|\theta_k][\theta_k]\, d\theta_k$ is the *prior* predictive distribution of $\boldsymbol{x}$ under model M_k.

Let us consider M_1 and M_2. The ratio between *posterior* probabilities $[M_2|\boldsymbol{x}]/[M_1|\boldsymbol{x}] = ([M_2][\boldsymbol{x}|M_2]) \,/\, ([M_1][\boldsymbol{x}|M_1])$ is a *posterior* bet in favor of M_2 compared to M_1. The ratio $B_{21} = [\boldsymbol{x}|M_2]/[\boldsymbol{x}|M_1]$, modifying the *prior* bet $[M_2]/[M_1]$ is called *Bayes factor* of model M_2 relatively to model M_1.

Bayes factors are the basis of bayesian model selection. A ratio close to 1 means that both models equivalently explain the observations, whereas much higher than 1 indicate that the model in the numerator is preferable. [Kass and Raftery, 1995] suggest a scale based on $2\ln(B_{21})$, which can be valid for a first indication, but is far from being general.

The evaluation of Bayes factors, and specially of $[\boldsymbol{x}] = [\boldsymbol{x}|M_k] = \int [\boldsymbol{x}|\theta_k] \times [\theta_k]\, d\theta_k$, relies on the Gibbs outputs. $[\boldsymbol{x}]$ could be estimated by Monte-Carlo sum on the likelihood with values sampled from the *prior* distribution of θ. Nevertheless, *prior* distributions are often very flat, much more than the *posterior* ones, and such method would not be much significant. Newton and Raftery advises to use another method, based on samples from the *posterior* distribution $[\theta|\boldsymbol{x}]$. Bayes formula gives: $[\theta]/[\boldsymbol{x}] = [\theta|\boldsymbol{x}]/[\boldsymbol{x}|\theta]$. By integration on θ, we have: $[\boldsymbol{x}]^{-1} = \int ([\theta|\boldsymbol{x}]/[\boldsymbol{x}|\theta])\, d\theta$, and thus can estimate $[\boldsymbol{x}]^{-1}$ by Monte-Carlo methods, sampling from $[\theta|\boldsymbol{x}]$, or more precisely continuing the Gibbs Sampler, setting each parameter one after the other to its estimate, as described in [Chib, 1995] and [Carlin and Chib, 1995].

Again, label-switching is a problem: the estimation by Monte-Carlo method involves here a function which is not invariant by permutation of the component, so permutations have to be undone. That's why, in such cases, other methods such as reversible jump or birth-death processes are preferred (see [Stephens, 2000a]).

7 Possible Improvements, Future Work

This study is an illustration of practical issues encountered when applying MCMC technics for the Bayesian Analysis of the mixture of distribution model. Some issues have already addressed in the literature (label-switching) but without clear and straightforward solutions and some others are pending (*prior* definition). Even though, the methods presented here are quite straightforward to implement, and thus can be easily used in a first approach of the problem.

The following conclusions were attained: due to the recurrent problem of label-switching (caused by the intrisic structure of the dataset), immediate interpretation and efficient model selection (we can not actually choose between 3, 4, or 5 components) were not carried out. Further work in terms of model selection must be definitely done. EM algorithm (using Mixmod software, developed by INRIA's IS2 team) has been used but gave completely different results, incoherent with chemists' interpretations: the algorithm seems trapped in local optimum. MCMC methods are therefore concluded more satisfactory: they at least give meaningfull results.

The following points have to be worked further: the hypothesis of normality of the components (log-normality would seem more coherent), the choice of the *prior* (less informative *prior*, maybe with a learning sub-sample in order to create more informative *prior* that can be used in Bayes factor). A sensitivity analysis is therefore needed to further validate the approach and assess the influence of each assumptions. The question of the convergence is a tricky point to address, provided the label-swistching issue. No rigorous diagnosis has been envisaged: birth-death processes or reversible jump methods need to be considered.

This is an illustration of the difficulties that praticians may face before benefiting from the powerful tools tha are MCMC Bayesian methods.

8 Acknowledgments

The authors would like to thank Eric Parent, from ENGREF (Paris, France), and Luc Perrault, from Hydro-Québec (Varennes, Québec, Canada) for their help and support.

References

[Carlin and Chib, 1995]B.P. Carlin and S. Chib. Bayesian model choice via markov chain monte carlo methods. *J. R. Statist. Soc. B*, 57:473–484, 1995.

[Celeux *et al.*, 2000]G. Celeux, M. Hurn, and C. Robert. Computational and inferential difficulties with mixture posterior distributions. *J. Am. Stat. Assoc.*, 95:957–970, 2000.

[Chib, 1995]S. Chib. Marginal likelihood from the gibbs output. *J. Am. Stat. Assoc.*, 90:1313–1321, 1995.

[Diebolt and Robert, 1990]J. Diebolt and C.P. Robert. Bayesian estimation of finite mixture distributions, part i : Theoretical aspects. Technical Report 110, LSTA, Université Paris VI, 1990.

[Droesbeke *et al.*, 2002]J.J. Droesbeke, J. Fine, and G. Saporta, editors. *Méthodes Bayésiennes en statistique.* Technip, 2002.

[Gelfand *et al.*, 1990]A.E. Gelfand, S.E. Hills, A. Rancine-Poon, and A.F.M. Smith. Illustration of bayesian inference in normal data models using gibbs sampling. *J. Amer. Stat. Assoc*, 85:972–985, 1990.

[Gelman and Rubin, 1992a]A. Gelman and D.B. Rubin. Inference from iterative simulation using multiple sequence (with discussion). *Statistical Science*, 7:457–511, 1992.

[Gelman and Rubin, 1992b]A. Gelman and D.B. Rubin. A single series from the gibbs sampler provides a false sense of security (with discussion). In J.M. Bernardo, J.O. Berger, A.P. Dawid, and A.F.M. Smith, editors, *Bayesian Statistics 4*, pages 625–631. Oxford University Press, 1992.

[Gelman *et al.*, 2003]A. Gelman, J.B. Carlin, H.S. Stern, and D.B. Rubin. *Bayesian Data Analysis.* Chapman and Hall, 2003.

[Kass and Raftery, 1995]R.E. Kass and A.E. Raftery. Bayes factor. *J. Am. Stat. Assoc.*, 90:773–795, 1995.

[Marin *et al.*, to appear]J.M. Marin, K. Mengersen, and C.P. Robert. *Bayesian modelling and inference on mixtures of distributions.* Elsevier-Sciences, (to appear).

[Richardson and Green, 1997]S. Richardson and P.J. Green. On bayesian analysis of mixtures with an unknown number of components. *J. R. Statist. Soc. B*, 59(4):731–792, 1997.

[Stephens, 1997]M. Stephens. *Bayesian Methods for Mixtures of Normal Distributions.* PhD thesis, Magdalen College, Oxford, 1997.

[Stephens, 2000a]M. Stephens. Bayesian analysis of mixtures with an unknown number of components — an alternative to reversible jump methods. *Annals of Statistics*, 2000.

[Stephens, 2000b]M. Stephens. Dealing with label-switching in mixture models. *J. R. Stat. Soc. B*, 62:795–809, 2000.

Repeated Significance Tests for Distributions with Heavy Tails

Joseph Glaz and Vladimir Pozdnyakov

Department of Statistics
University of Connecticut
Storrs CT, 06269-4120, USA
(e-mail: glaz@uconnvm.uconn.edu, vladimir.pozdnyakov@uconn.edu)

Abstract. In this article we survey recent results on the development of nonparametric repeated significance tests for distributions with heavy tails. The implementation of these tests depends on the invariance theorem for partial sums of truncated independent and identically distributed random variables. We also discuss a method for evaluating the power function and the expected stopping times associated with these testing procedures.

Keywords: Repeated Significance Test, Heavy Tails, Invariance Principle.

1 Introduction

Repeated significance tests were introduced in [Armitage, 1958]. Major developments in the theory and applications of repeated significance tests have been reported among others in: [Dias and Garsia, 1999], [Hu, 1988], [Jennison and Turnbull, 2000], [Lai and Siegmund, 1977, 1979], [Lalley, 1983], [Lerche, 1986], [Sellke and Siegmund, 1983], [Sen, 1981, 1985, 2002], [Siegmund, 1982, 1985], [Takahashi, 1990], [Whitehead, 1997], [Woodroofe, 1979, 1982], and [Woodroofe and Takahashi, 1982]. Sequential tests and repeated significance tests have been developed mostly for normal or a t-distribution models ([Siegmund, 1985] and [Takahashi, 1990]. For independent and identically distributed (iid) observations from a distribution with a finite mean nonparametric repeated significance tests have been discussed in [Sen, 1981, 1985, 2002].

In this article we review recent results in [Glaz and Pozdnyakov, 2005] and [Pozdnyakov and Glaz, 2005] for repeated significance tests for iid observations from a continuous symmetric distribution with heavy tails and infinite variance and possibly no mean, as in the case of the Cauchy distribution. The repeated significance test developed in [Glaz and Pozdnyakov, 2005] is nonparametric in nature and is applicable to a class of stable distributions with a specified tail behavior. The method used in deriving this repeated significance extends the approach in [Sen, 1981] and [Sen, 1985]. It is based on the invariance principle for partial sums of truncated random variables ([Pozdnyakov, 2003]). Pozdnyakov and Glaz ([Pozdnyakov and Glaz, 2005]) introduce a sequential test that is a repeated significance test with a random target sample size. It depends on the gross rate of the sample variance of the tests statistics used in the repeated significance test. This test is fully nonparametric and its implementation does not depend on the asymptotic tail behavior of the underlying model for the observed data.

The article is organized as follows. In Section 2, we describe the repeated significance test for the median of a continuous symmetric distribution with heavy tails in a class of stable distributions with a specified tail behavior. We discuss how one selects a continuation region associated with this repeated significance test for specified significance level, initial sample size and target sample size. Power calculations and evaluation of expected stopping times are discussed in [Glaz and Pozdnyakov, 2005].

In Section 3 we describe a repeated significance test with a random target sample size. An application of this repeated significance test for a shift model is presented along with evaluation its performance. Concluding results are presented in Section 4.

2 A Repeated Significance Test for a Median

Let $\{X, X_i; i \geq 1\}$ be iid observations from a continuous distribution F symmetric about $-\infty < \theta < \infty$. Assume that $E(X^2) = \infty$. We present below a repeated significance test with initial sample size n_0 and target sample size N for testing

$$H_0 : \theta = 0 \ vs \ H_a : \theta \neq 0. \tag{1}$$

Assume that $H_0 : \theta = 0$ and F belongs to the domain of attraction of a stable distribution with exponent $0 < \gamma < 2$, i.e.

$$E\left(X^2 I_{(|X \leq t|)}\right) \sim t^{2-\gamma} L(t),$$

where L is a slowly varying function. Here and in what follows we denote by $u(t) \sim v(t)$ the asymptotic equivalence of two functions $u(t)$ and $v(t)$, in the sense that $\lim_{t\to\infty} [u(t)/v(t)] = 1$. The classical repeated significance test is based on a sequence of partial sums ([Siegmund, 1985]). The problem we encounter here is that the sequence of partial sums from a distribution with an infinite second moment does not converge to a Brownian motion. To overcome this difficulty we employ partial sums of truncated random variables. Let $\{d_n; n \geq 1\}$ be an increasing sequence of positive numbers such that

$$nP\left(|X| > d_n\right) \sim \gamma_n,$$

where $\gamma_n \nearrow \infty$, as $n \to \infty$. Define the truncated partial sums:

$$S_n = \sum_{i=1}^{n} X_i I_{(|X_i| \leq d_n)}. \tag{2}$$

Denote by

$$B_n = Var(S_n).$$

The classical invariance principle (Donsker Theorem, [Billingsley, 1995, p. 520]) is not applicable here as the sequence of truncated partial sums, $\{S_n; n \geq 1\}$, does not have independent increments . Let $\{W(t); 0 \leq t < \infty\}$ be the standard Brownian motion. The following invariance principle will be used to construct the continuation region for the repeated significance test:

Theorem 1 *(Pozdnyakov 2003) If the random variable X belongs to the Feller class, i.e.*

$$\limsup_{t\to\infty} \frac{t^2\mathbf{P}(|X|>t)}{\mathbf{E}\big(X^2 \mathrm{I}_{|X|\leq t}\big)} < \infty,$$

the average number of the excluded variables

$$n\mathbf{P}\big(|X| > d_n\big)\mathrm{sim}\gamma_n \nearrow \infty$$

and $\lim_{n\to\infty} B_n/B_{n+1} = 1$, *then* $S_n(t) \xrightarrow{d} W(t)$ *in the sense* $(\mathcal{C}[0,1], \rho)$, *where* $S_n(t)$ *is the linear interpolation between points*

$$\Big(0,0\Big), \Big(\frac{B_1}{B_n}, \frac{S_1}{\sqrt{B_n}}\Big), ..., \Big(1, \frac{S_n}{\sqrt{B_n}}\Big).$$

Since B_n is unknown, following Sen ([Sen, 1981, p. 249]), we replace it with an almost sure equivalent sequence of sample variances:

$$A_n = \sum_{i=1}^{n} X_i^2 \mathrm{I}_{(|X_i|\leq d_n)} - \frac{S_n^2}{\sum_{i=1}^{n} \mathrm{I}_{(|X_i|\leq d_n)}}. \tag{3}$$

Let

$$\tau = \min\Big\{n_0 \leq n \leq N; |S_n| \geq b\sqrt{A_n}\Big\}$$

be the stopping time associated with the repeated significance test, where n_0 is the initial sample size and N is the target sample size. This test stops and rejects H_0, given in Equation (1), if and only if $\tau \leq N$. Its power function is given by:

$$\pi(\theta) = P_\theta\,(\tau \leq N) = 1 - \beta(\theta),$$

where

$$\beta(\theta) = P_\theta\,\Big(|S_n| < b\sqrt{A_n}; n_0 \leq n \leq N\Big),$$

is the probability of type II error function and $\big\{b_n = b\sqrt{A_n}; n_0 \leq n \leq N\big\}$ is a sequence of constants that determine its continuation region. The significance level of this repeated significance test is given by:

$$\begin{aligned}\alpha &= \pi(0) = P_0\,(\tau \leq N) = 1 - \beta(0)\\ &= P_0\left\{\max_{n_0\leq n\leq N}\left(\frac{|S_n|}{\sqrt{A_n}}\right) \geq b\right\}.\end{aligned}$$

To implement this test an accurate approximation for $b = b(\alpha, n_0, N)$ is needed. The following result is central for achieving this goal.

Proposition 1 *(Glaz and Pozdnyakov 2005) Assume that F belongs to the domain of attraction of a continuous symmetric stable distribution with exponent $0 < \gamma < 2$. Then the following results are true:*
1) F belongs to the Feller class.
2) The average number of the excluded terms $n\mathbf{P}(|X| > dn^{\delta}) \nearrow \infty$ whenever $1 - \gamma\delta > 0$. In particular, any $0 < \delta < 1/2$ guarantees it for all $0 < \gamma < 2$.
3) If $1 - \gamma\delta > 0$ and $\lim_{n_0,N\to\infty}(n_0/N) = c < 1$, then

$$\max_{n_0 \le n \le N} \frac{|S_n|}{\sqrt{A_n}} \xrightarrow{d} \sup_{[t_0 \le t \le 1]} \frac{|W(t)|}{\sqrt{t}},$$

where

$$t_0 = c^{1+(2-\gamma)\delta}. \tag{4}$$

In view of this result, let the truncating levels $d_n = dn^{\delta}$, where $d > 0$ and $0 < \delta < 1/2$. Let $c = n_0/N$, be the ratio of the initial and target sample sizes of the repeated significance test. Then, $b = b(\alpha, n_0, N)$ can be approximated by $b_{t_0}(\alpha)$ by solving

$$\mathbf{P}\left(\sup_{[t_0,1]} \frac{|W(t)|}{\sqrt{t}} > b_{t_0}(\alpha)\right) = \alpha,$$

where t_0 is given in Equation (4). The algorithm in [De Long, 1981] is used to evaluate $b_{t_0}(\alpha)$.

Example 1 *Domain of attraction of a Cauchy distribution with location parameter θ and scale parameter 1.*

Let $\{X_i; i \ge 1\}$ be a sequence of iid observations from a distribution F in the class of distributions with a domain of attraction of a Cauchy distribution with location parameter θ and scale parameter 1. Assume that $H_0 : \theta = 0$ is true. We consider here truncation levels $d_n = n^{1/4}, n_0 \le n \le N$. In Table 1, the performance of the proposed repeated significance test is evaluated in terms of accuracy of achieving an assigned significance level α, for given values of n_0 and N. The theoretical critical values $b_{t_0}(\alpha)$ and the corresponding targeted significance levels have been obtained from [De Long, 1981]. The achieved significance levels were evaluated from a simulation with $10,000$ trials.

Table 1. Simulated Significance Levels

n_0	N	t_0	$b_{t_0}(\alpha)$	*targeted* α	*simulated* α
100	303	1/4	2.7	.0503	.0541
100	303	1/4	3.3	.0098	.0094
100	754	1/12.5	2.6	.0989	.1012
30	91	1/4	2.7	.0503	.0638
30	91	1/4	3.3	.0098	.0167
30	226	1/12.5	2.6	.0989	.1119

For small values, the achieved significance levels are close to targeted significance levels even for a moderate value of $n_0 = 30$. For larger significance levels one has to use higher initial values to get accurate approximations. A value of $n_0 = 100$ produced accurate results even for $\alpha = .10$.

3 A Repeated Significance Test with Adaptive Target Sample Size

The implementation of the repeated significance test in the previous section requires specification of the asymptotic tail behavior of the distribution under the null hypothesis. In some applications this might not be known. To address this issue, in ([Pozdnyakov and Glaz, 2005]) we introduced a nonparametric repeated significance test with adaptive target sample size.

Let T_n be a sequence of test statistics associated with a repeated significance test. Let A_n be a sample variances of T_n. Define a stopping time $\mathcal{N}$ by

$$\mathcal{N} = \inf\{k \geq n_0 : \frac{A_k}{A_{n_0}} \geq \frac{1}{t_0}\}, \tag{5}$$

where $0 < t_0 < 1$ is a design parameter. A repeated significance test with adaptive target sample size is defined as follows. At time $k \geq n_0$ observe T_k. Stop and reject H_0, if k is the smallest integer such that $A_k/A_{n_0} < 1/t_0$ and $|T_k| \geq b\sqrt{A_k}$. Otherwise, we stop monitoring at time $\mathcal{N}$ and accept H_a. The following result is central to the implementation of the repeated significance test with adaptive target sample size.

Theorem 2 *(Pozdnyakov and Glaz 2005) Assume that the functional central limit theorem for the sequence $\{T_n\}$ holds, and there exists a sequence of numbers $B_n \nearrow \infty$ and $B_n/Var(T_n) \to 1$ as $n \to \infty$. If the sample variance A_n satisfies*

$$\frac{A_n}{B_n} \to 1 \quad a.s.,$$

then

$$P\left(\max_{n_0 \leq k \leq \mathcal{N}} \left|\frac{T_k}{\sqrt{A_k}}\right| > b\right) \longrightarrow \alpha(t_0, b) \text{ as } n_0 \to \infty.$$

Theorem 2 is applied to a functional central limit theorem for a sequence of truncated partial sums to develop a repeated significance test with random sample size for the shift model. In what follows we describe this test and present a simulation study to evaluate its performance.

Let $\{X, X_i; i \geq 1\}$ be iid observations from a continuous distribution F symmetric about $-\infty < \theta < \infty$. We are interested in testing sequentially

$$H_0 : \theta = 0 \text{ vs } H_a : \theta \neq 0.$$

Define the sequence of truncated partial sums S_n as in Section 2, Equations (2). Note that the variances of the truncated partial sums satisfy the monotonicity condition needed in Theorem 2 and that one can employ the version of the sample variances given in Equation (3). It was shown in [Glaz and Pozdnyakov, 2005], that the conditions of Theorem **??** along with

$$\lim_{n\to\infty} \frac{n\mathbf{P}\big(|X| > d_n\big)}{\ln\ln(n)} = \infty \tag{6}$$

and

$$\ln\ln(B_n) = o(n), \tag{7}$$

imply that A_n is almost sure equivalent to the population variance B_n. Note that conditions (6) and (7) are not restrictive from the practical point of view.

Based on these results, the following repeated significance test with adaptive target sample size is developed. Let

$$\tau = \inf\Big\{k \geq n_0 : |S_k| \geq b\sqrt{A_k}\Big\}$$

be a stopping time, where n_0 is the initial sample size, and $\mathcal{N}$ is the adaptive target sample size defined by (5). The repeated significance test stops and rejects H_0 if and only if $\tau \leq \mathcal{N}$. Hence, $\tau \wedge \mathcal{N}$ is the stopping time associated with this repeated significance test.

The following class of heavy tail distributions will be used in evaluating the performance of this repeated significance test. We say that a random variable X has a *Cauchy^p* distribution iff

$$X \overset{d}{=} \mathrm{sign}(Y)|Y|^p,$$

where $p > 0$ and Y has a standard Cauchy distribution. If X has a *Cauchy^p* distribution, then it is symmetric and belongs to the Feller class for any $p > 0$. Moreover, $E(|X|^q) < \infty$, if $q < 1/p$.

To evaluate the performance of the proposed repeated significance test, we consider the following four distributions: Normal, Cauchy$^{1/2}$, Cauchy, and Cauchy^2. These distributions have different tail behaviors and it is impossible to specify a deterministic target sample size in the repeated significance test based on the truncated sums S_n, discussed in [Glaz and Pozdnyakov, 2005], that guarantees a correct significance level α for all four distributions. Numerical results presented in Table 1 show that the introduction of an adaptive target sample size successfully addresses this problem. The truncation level $d_n = n^{1/4}$ was used. The design parameters corresponding to targeted values of $\alpha = .01$ and $.05$ were evaluated from the tables in [De Long, 1981]. The simulated significance levels are presented as top values in the table. The simulated values of $E(\tau \wedge \mathcal{N})$ are rounded to whole numbers and are presented as the bottom values in the table. These values are based on a simulation with 10000 trials.

Table 2. Simulated Significance Levels and Expected Stopping Times, $n_0 = 100$, $d_n = n^{1/4}$

t_0^{-1}	b	*Normal*	$Cauchy^{1/2}$	*Cauchy*	$Cauchy^2$
4	3.3	.0097	.0104	.0093	.0076
		391	319	276	260
	2.7	.0508	.0465	.0473	.0463
		382	313	272	256
7.5	3.4	.0099	.0117	.0099	.0085
		729	544	439	397
	2.8	.0511	.0484	.0480	.0447
		711	533	429	391

4 Concluding Remarks

In this article we reviewed two recently developed repeated significance tests for distributions with heavy tails. For additional results and discussion the readers are referred to the articles [Glaz and Pozdnyakov, 2005] and [Pozdnyakov and Glaz, 2005]. We would like to note that currently the applications of these results are restricted to symmetric distributions. To extend these results to non symmetric distributions presents new challenges. The first step in this direction is to extend the invariance theorem that has been established in [Pozdnyakov, 2003]. The development of inference procedures following these repeated significance tests are also of great interest in applications.

References

[Armitage, 1958]P. Armitage. Numerical studies in the sequential estimation of a binomial parameter. *Biometrika*, pages 1-15, 1958.

[Billingsley, 1995]P. Billingsley. *Probability and Measure, 3nd ed.* Wiley-Interscience, New York, 1995.

[De Long, 1981]D. De Long. Crossing probabilities for a square root boundary by a Bessel process, *Communications in Statistics: Theory & Methods, Series A*, pages 2197-2213, 1981.

[Dias and Garsia, 1999]R. Dias and N.L. Garcia. Approximations for non-symmetric truncated sequential and repeated significance tests. *Sequential Analysis*, pages 107-117, 1999.

[Glaz and Pozdnyakov, 2005]J. Glaz and V. Pozdnyakov. A repeated significance test for distributions with heavy tails, *Sequential Analysis*, pages 77-98, 2005.

[Hu, 1988]I. Hu. Repeated significance tests for exponential families. *Annals of Statistics*, pages 1643-1666, 1988.

[Jennison and Turnbull, 2000]C. Jennison and B.W. Turnbull. *Group Sequential Methods with Applications to Clinical Trials.* Chapman & Hall/CRS, Boca Raton, 2000.

[Lai and Siegmund, 1977]T.L. Lai and D. Siegmund. A nonlinear renewal theory with application to sequential analysis I. *Annals of Statistiscs*, pages 946-954, 1977.

[Lai and Siegmund, 1979]T.L. Lai and D. Siegmund. A nonlinear renewal theory with application to sequential analysis II. *Annals of Statistiscs*, pages 60-76, 1979.

[Lalley, 1983]S. Lalley. Repeated likelihood ratio tests for curved exponential families. *Z. Wahrsch. Verw. Gebiete*, pages 293-321, 1983.

[Lerche, 1986]H.R. Lerche. An optimality property of the repeated significance test. *Proceedings of the National Academy of Science USA*, pages 1546-1548, 1986.

[Pozdnyakov, 2003]V. Pozdnyakov. A Note on functional CLT for truncated sums, *Statistics & Probability Letters*, pages 277-286, 2003.

[Pozdnyakov and Glaz, 2005]V. Pozdnyakov and J. Glaz. A nonparametric repeated significance test with adaptive target sample size. *Journal of Statistical Planning and Inference*, (to appear).

[Sellke and Siegmund, 1983]T. Sellke and D. Siegmund. Sequential analysis of the proportional hazards model. *Biometrika*, pages 315-326, 1983.

[Sen, 1981]P.K. Sen. *Sequential nonparametrics: invariance principles and statistical inference.* Jonh Wiley & Sons, New York, 1981.

[Sen, 1985]P.K. Sen. *Theory and Applications of Sequential Nonparametrics.* CBMS-NSF Regional Conference Series in Applied Mathematics 49, SIAM, Philadelphia, 1985.

[Sen, 2002]P.Q. Sen. Repeated significance tests in frequency and time domains. *Sequential Analysis*, pages 249-283, 2002.

[Shiryaev, 1995]A.N. Shiryaev. *Probability.* Springer, New York, 1995.

[Siegmund, 1982]D. Siegmund. Large deviations for boundary crossing probabilities. *Annals of Probability*, pages 581-588, 1982.

[Siegmund, 1985]D. Siegmund. it Sequential Analysis. Springer-Verlag, New York, 1985.

[Takahashi, 1990]H. Takahashi. Asymptotic expansions for repeated sequential tests for the normal means. *Journal of the Japanese Statistical Association*, pages 51-60, 1990.

[Whitehead, 1997]J. Whitehead. *The Design and Analysis of Sequential Clinical Trials, 2nd ed.*, John Wiley & Sons Ltd., Chichester, West Sussex, 1997

[Woodroofe, 1979]M. Woodroofe. Repeated likelihood ratio tests. *Biometrika*, pages 453-463, 1979.

[Woodroofe, 1982]M. Woodroofe. *Nonlinear Renewal Theory in Sequential Analysis.* SIAM, Philadelphia, 1982.

[Woodroofe and Takahashi, 1982]M. Woodroofe and H. Takahashi. Asymptotic expansions for the error probabilities of some repeated significance tests. *Annals of Statistics*, pages 895-908, 1982.

Sequential Sampling Inspection Could Save Money: A Case in Connecticut in Point

Nitis Mukhopadhyay

Department of Statistics
University of Connecticut
CLAS Building, , U-4120
215 Glenbrook Road
Storrs, CT 06269-4120, U.S.A.
(e-mail: mukhop@uconnvm.uconn.edu)

Abstract. The State of Connecticut bought 10, 000 computers/servers from a contracted supplier. These were supposed to include some special internal hardware. The technology department inspected 4, 000 pieces from the delivered batch and found only 58 "good" ones! It is shown that the inspection protocol that allowed checking 4, 000 computers was at best outrageously wasteful. An appropriately designed strategy with fewer than 10% inspections could conclude with near certainty that the batch was far below expectation.

Keywords: Inspection protocol, Inspection sampling, Percentage saving, Sampling strategy.

1 Introduction

On Tuesday, June 8, 2004, the Hartford Courant's Connecticut section's headline read "Woman Accused of Bilking State" which drew my immediate attention. I became intrigued as I read the story, "... In March 2001, the Computer Plus Center won a $17.2 million state contract, making it the exclusive vendor of Dell computers and servers for all state agencies, the arrest affidavit states. In January 2003, the state Department of Information Technology filed a complaint about the company, ... announcing the arrest. The servers would not work, according to the affidavit." The article quoted Chief State's Attorney Christopher Morano saying, "The servers that were delivered did not have the amount of memory, or the quality memory, in them, that was required." The article then went on to report, "... The state's technology department took apart 4, 000 of the 10, 000 (computers/servers) delivered by the said company. Of those, Morano said, only 58 contained the required network interface cards, ... "

This note is not about allegations or legal posturing. I was struck by the fact that Connecticut's technology department took apart 4, 000 from 10, 000 computers/servers delivered by the said company whereas only 58 good items were found! Given this batch, a moot question is this: Was it possible to come to the conclusion of alleged fraud by inspecting a small fraction of $10,000$ items? The answer is, 'of course, yes'. I will substantiate this with appropriately designed random sampling strategies to gather just the right amount of information much sooner.

Now, suppose that the State's technology department could come to the same conclusion of alleged fraud by inspecting n items (computers) where n was decisively "small" compared with $4,000$. Also, suppose that the inspection per computer takes x minutes and a skilled technologist charges $\$a$ per x minutes of inspection. When an item is checked out, it is out of commission so that the State loses $\$b$ per piece per inspection. A technologist will probably be paid $\$c$ for the mileage and per diem on an average per inspection assuming that these 10,000 computers are scattered in different locations. Then, we have:

$$\begin{aligned} \text{Savings: } SV &= \$(a+b+c)(4000-n), \\ \text{Percentage Savings: } PSV &= \left(1-\tfrac{n}{4000}\right)\times 100. \end{aligned} \%tag1 \tag{1}$$

Let us throw in some realistic numbers. For example, suppose that $a = 150, b = 50, c = 10$ and the savings would amount to $\$735,000$ or $\$420,000$ if one could arrive at the conclusion of alleged fraud by inspecting only 500 or $2,000$ computers respectively. These savings could be in cash or kind, for example, in the form of savings from cost-share or overtime payments.

There are other expenses too when a computer is inspected. For example, there is cost for electricity and for storage of non-functioning computers. Also, the supplier was already paid and the State "lost" interest income from that fund! Then, waiting for a year or more to bring lawsuits against supplier(s) drains the State's resources even further. The term SV in (1) may not take into account all kinds of costs borne by the State. Yet, one cannot deny that the term PSV from (1) portrays a realistic quantification of percentage savings regardless of the magnitudes of a, b, c and other costs involved.

2 A statistical formulation

We face a large population of $10,000(= R)$ items where each item is either 'good' or 'bad'. When an item is randomly selected, suppose that the probability that it is good (or bad) is p (or $q = 1-p$), $0 < p < 1$. The percentage of good items ($= 100p\%$) is assumed unknown.

Clearly, I can set the following lower and upper bounds for p:

$$0.0058 \approx \tfrac{58}{10000} \le p \le \tfrac{6058}{10000} \approx 0.6058 \tag{2}$$

The lower (upper) bound for p in (2) is obtained by assuming that there were no (all) good items among $6,000$ remaining uninspected items. On the other hand, it appears that p should be closer to $\frac{58}{4000} \approx 0.0145$ rather than the most pessimistic value 0.0058 or the overly optimistic value 0.6058.

To estimate p, one would inspect n items selected randomly from the batch to check how many items $(= X)$ out of n items are indeed good. Having a large population on hand, I treat X as an approximately binomial random variable with n and p. An estimator of p is

$$\widehat{p}_n = \frac{\#\text{ good items in the random sample}}{n}. \tag{3}$$

This $\widehat{p}_n$ has the following variance and estimated variance:

$$Var\left(\widehat{p}_n\right) = p(1-p)/n,\ \widehat{Var\left(\widehat{p}_n\right)} = \widehat{p}_n\left(1-\widehat{p}_n\right)/n, \tag{4}$$

by disregarding the finite population correction factor $1 - nR^{-1}$. See Sukhatme *et al.* (1984, p. 43).

Now, how many items (that is, n) should one inspect so that the standard confidence interval $\widehat{p}_n \pm E$ for p would have $100(1-\alpha)\%$ confidence? By appealing to the central limit theorem, the required sample size n is then approximately given by

$$n \equiv n(p) = \left(z_{\alpha/2}/E\right)^2 p(1-p). \tag{5}$$

Since p is unknown, one may opt for the maximum possible value of $n(p)$ that would work for all possible values of $p, 0 < p < 1$. This maximum occurs when $p = \frac{1}{2}$ which motivates the following expression for n:

$$n \equiv n_{\max} = \tfrac{1}{4}\left(z_{\alpha/2}/E\right)^2. \tag{6}$$

α $z_{\alpha/2}$	E						
	0.10	0.08	0.05	0.02	0.016	0.012	0.01
	$\mathbf{n}_{\max}$ **values:**						
0.10 1.645	68	108	271	1692	3007	4699	6766
	n(p) values:						
$\mathbf{p = 0.2}$	43.3	67.7	173.2	1082.4	1691.3	3006.7	4329.6
$\mathbf{p = 0.1}$	24.4	38.1	97.4	608.9	951.3	1691.3	2435.4
$\mathbf{p = 0.0145}$	3. 87	6.04	15.5	96.7	151.1	268.5	386.7
	$\mathbf{n}_{\max}$ **values:**						
0.05 1.96	97	151	385	2401	4269	6670	9604
	n(p) values:						
$\mathbf{p = 0.2}$	61.5	96.0	245.9	1536.6	24010	4268.4	6146.6
$\mathbf{p = 0.1}$	34.6	54.0	138.3	864.4	1350.6	2401.0	3457.4
$\mathbf{p = 0.0145}$	5.49	8.58	22.0	137.2	214.4	381.2	549.0

Table 1. Sample size $n(p)$ from (5) and $n_{\max}$ from (6).

The recommended expression $n_{\max}$ is used in many practical problems. For example, one may refer to Chase and Bown (2000, p. 330). But, $n_{\max}$ is rather conservative because it works for all p values across the board. In fact, $n_{\max}$ may be viewed as a minimax choice for the required sample size.

In table 1, the first (second) block corresponds to 90% (95%) confidence intervals with a particular E value. First, $n_{\max}$ values are provided and then for each fixed E, I provide $n(p)$ values for $p = 0.2, 0.1, 0.0145$. Note that $p = 0.0145$ corresponds to $\frac{58}{4000}$. From this table, I immediately summarize some features, namely:

(1) $n(p)$ goes up for fixed p if $E \downarrow$;

(2) $n(p)$ goes down for fixed E if $p \downarrow$;

(3) $n(p)$ goes down significantly compared with $n_{\max}$ for fixed E if $p \downarrow$.

3 Two-stage sampling to determine sample size

Recall that $\widehat{p}_n \pm E$ would have approximately $100(1-\alpha)\%$ confidence when the required sample size $n \equiv n(p)$ is approximated by the expression from (5). But, this expression involves unknown p to begin with! Hence, one must inspect items at least in two steps. This is called *two-stage* or *double sampling* strategy. See Stein (1945,1949), Ghosh *et al. (*1997, Chapter 6), and Mukhopadhyay and Solanky (1994, Chapter 2). Robbins and Siegmund (1974) and Mukhopadhyay and Cicconetti (2004) respectively proposed purely sequential and two-stage estimation strategies for p under various kinds of loss functions. One may also take a look at Corneliussen and Ladd (1970), Ghosh (1970), Wald (1947), and other sources.

I propose to inspect $m(\geq 2)$ pilot items and obtain only a preliminary estimate $\widehat{p}_m$ in the first stage. Here, $< u >$ stands for the largest integer $< u$. Now, I let

$$N = \max\left\{m, \left\{\left(\frac{t_{m-1,\alpha/2}}{E}\right)^2 \left(\widehat{p}_m + m^{-1}\right)\left(1 - \widehat{p}_m - m^{-1}\right)\right\} + 1\right\} \qquad (7)$$

where $t_{m-1,\alpha/2}$ is the upper $50\alpha\%$ point of the Student's t distribution with $m-1$ degrees of freedom. If one believes that p is rather small, then $\widehat{p}_m$ may be zero and hence $\widehat{p}_m$ is replaced by $\widehat{p}_m + m^{-1}$ in (7).

If $N = m$, there will be no need for more inspections beyond the pilot stage. But, if $N > m$, then I propose to inspect $N - m$ additional items and record the number of good items in the second stage. The final confidence interval estimator for p is going to be $\widehat{p}_N \pm E$ where

$$\widehat{p}_N = \frac{\text{\# good items in the combined random sample from both stages}}{N}. \qquad (8)$$

p $\frac{58r}{4000}$	Observed sample sizes in ten replications	% Savings PSV $100\left(1-\frac{\text{Ave }N}{4000}\right)\%$
$r=1$	$378, 620, 620, 254, 378,$ $128, 500, 254, 254, 128$ $\overline{N}=351.4$, $\mathrm{S_N}=181.4$, $\widetilde{N}=316$	91.22%
$r=2$	$620, 966, 853, 620, 254,$ $620, 378, 378, 254, 620$ $\overline{N}=556.3$, $\mathrm{S_N}=240.0$, $\widetilde{N}=620$	86.1%
$r=3$	$737, 966, 500, 500, 128,$ $853, 378, 853, 853, 620$ $\overline{N}=638.8$, $\mathrm{S_N}=262.9$,$\widetilde{N}=678.5$	84. 0%
$r=10$	$2165, 1986, 2251, 2251, 1893,$ $1893, 2498, 2251, 2417, 2165$ $\overline{N}=2177.0$, $\mathrm{S_N}=204.2$, $\widetilde{N}=2208.0$	45. 6%
$r=41.779$	$3896, 3896, 3850, 3974, 3541,$ $3663, 3824, 3796, 3734, 3963$ $\overline{N}=3813.7$, $\mathrm{S_N}=136.1$, $\widetilde{N}=3837$	4. 7%

Table 2. Values of N using (7) from ten replications with 2% over-sampling on an average compared with n(p) from (5) and $\alpha = 0.05$, $m = 124$.

By the way, Mukhopadhyay (2004) gave a practical way to determine the pilot sample size m as follows:

$$m = \text{smallest positive integer such that } t^2_{m-1,\alpha/2}/z^2_{\alpha/2} \leq 1+\varepsilon, \tag{9}$$

assuming that one can comfortably entertain $100\varepsilon\%$ over-sampling on an average compared with $n(p)$ from (5). Then, one arrives at the following choice for the pilot sample size m depending upon ε:

$$m = \left\langle \frac{1}{2\varepsilon}\left(\frac{1}{2}(z^2_{\alpha/2}+1) + \left\{2\left[\frac{1}{3}z^4_{\alpha/2} + \frac{23}{12}z^2_{\alpha/2} + \frac{5}{4}\right]\varepsilon\right\}^{1/2}\right)\right\rangle + 1. \tag{10}$$

Now, in order to have a feel for what one may face in practice, I decided to generate a Bernoulli population where $p = \frac{58r}{4000}$ with $r = 1, 2, 3, 10$, and 41.779. The case "$r = 1$" simulates the situation on hand where we are told that 58 good items have been observed among 4000 inspected items and no more. The cases "$r = 2, 3, 10$" respectively simulate situations where we may expect to see good items at the rate (p) of two, three or ten times the rate of what we have been told to have happened. The case "$r = 41.779$" simulated a situation where one may expect to see good items at the most optimistic rate given what has happened in the situation on hand, that is with $p = 0.6058(\equiv \frac{6058}{10000})$. We fixed $\alpha = 0.05$, $\varepsilon = 0.02$ and hence (10) suggested a pilot sample size $m = 124$. I determined N ten separate times from independent replications in each situation. Table 2 provides all ten observed N values along with their average $\overline{N}$, standard deviation S $_N$, and the median $\widetilde{N}$, in each case. The last column provides the estimated average percentage savings compared with $n = 4000$. One sees unbelievable percentage savings in sample size on an average when $r = 1, 2, 3$ and 10. Even in the most optimistic situation ($r = 41.779$) described in the last block of table 2, we note 4.7% average savings in sample size compared with $n = 4000$. This saving may appear insignificant, but then one should consider this: After observing only 58 good items among 4000 inspected items, what is the likelihood that all remaining 6000 uninspected items would be judged good if inspected? The point is that even under such rarest of rare occurrence, the present sampling strategy could have saved us by inspecting nearly 3800 items on an average instead of 4000 items!

4 Sequential testing to determine a sample size

I continue with random sampling from a Bernoulli(p) population where p is the fraction of good items in a very large population having $R(= 10,000)$ items. The inspection team must have certain high value $p_0, 0 < p_0 < 1$, in mind that it expects the vendor to comply with in good faith. The State may hope that $p_0 \approx 1.0$. Obviously, a small percentage of items may turn out bad, but those bad items would be expected to be properly 'corrected' by the supplier. So, the inspection team could set up a sampling strategy for the following testing problem:

$$H_0\colon\ p \geq p_0 \text{ versus } H_1\colon\ p < p_0 \tag{11}$$

Suppose that one fixes $p_0 = 0.95$ or 0.99 and it means that the State considers 9500 or 9900 good items found among 10,000 items is within reason. But, if $p < p_0$ where p_0 is a set number, then the inspection team will 'raise a flag' in favor of possible suspicion of receiving lesser than expected quality. I would like to clear one important point. The number p_0 ought to be specified by the State. Such specification may take into consideration the inspection team's mindset that is consistent with the State's budgetary constraints plus other protocols as required.

One may feel tempted to use customary normal approximation to a binomial distribution and hence having $n(\geq 30)$ observations, one would reject the null hypothesis H_0 if and only if

$$\widehat{p}_n < p_0 - z_\alpha \sqrt{\tfrac{p_0(1-p_0)}{n}}, \tag{12}$$

with the level of significance, $\alpha = P\{\text{ Rejecting} H_0 \mid H_0 \text{ is true}\}$.

But, what should be the appropriate sample size, n? Now, suppose that one asks that the power of the test (12) when $p = p_1(< p_0)$ be at least $1 - \beta$

where $0 < \beta < 1$ is a small and fixed number. In many investigations, one fixes power 80% (that is, $\beta = 0.20$ to detect a certain "effect size" (that is, $p_1 - p_0$). Let me denote $\sigma_i = \sqrt{p_i(1-p_i)}, i = 0, 1$. For large n, the power of the test (12) when $p = p_1 (< p_0)$ can be expressed as

$$P\left\{\text{Rejecting } H_0 \mid p = p_1\right\} \approx P\left\{Z < \tfrac{\sqrt{n}(p_0 - p_1)}{\sigma_1} - z_\alpha \tfrac{\sigma_0}{\sigma_1}\right\}. \qquad (13)$$

Now, it ought to be clear that the power given in (13) would be at least $1-\beta$ provided that the sample size n is chosen as follows:

$$n \geq \frac{(z_\alpha \sigma_0 + z_\beta \sigma_1)^2}{(p_1 - p_0)^2} = n^*(p_0, p_1).$$

I define

$$n \equiv n(p_0, p_1) = \max\left\{30, \langle n^*(p_0, p_1)\rangle + 1\right\}, \qquad (14)$$

so that customary normal approximation to a binomial distribution will be expected to work (since $n \geq 30$).

In table 3, I provide values of $\langle n^*(p_0, p_1)\rangle + 1$ for $p_0 = 0.95, 0.99$ with $\alpha = 0.05$ and $\beta = 0.05, 0.10, 0.20$. When the null hypothesis tests a large p_0 value, naturally the required sample size becomes rather too small in order to detect p_1 far away from p_0 whether the power is set at 80%, 90% or 95%. It is clear that one needs to inspect nearly 30 or so items while testing a null hypothesis with large $p_0 (= 0.95, 0.99)$ when the true fraction of good items is indeed only $0.0145 (\approx \frac{58}{4000})$ or close to the most optimistic value $0.6058 (\approx \frac{6058}{10000})$.

p_0	β					p_1	
		0.90	0.80	0.60	0.50	0.25	0.0145
0.99	0.20	22	7	3	2	1	1
	0.10	38	13	5	3	1	1
	0.05	54	19	7	5	2	1
						p_1	
		0.90	0.80	0.60	0.50	0.25	0.0145
0.95	0.20	150	22	5	3	2	1
	0.10	221	34	8	5	2	1
	0.05	291	46	12	7	3	1

Table 3. Values of $\langle n^*(p_0, p_1)\rangle + 1$ from (14) with 5% level and power $1 - \beta$ with $\beta = 0.05, 0.10$, and 0.20. Sample size n is max{table entry,30}.

It is obvious that the best fixed-sample-size test (12) could arrive at a decision with very few inspections if p was indeed as small as it was in the supplied batch. It is also well known, however, that the test (12) is not really 'optimal' in a larger class of sequential tests. Wald's (1947) *sequential* *probability ratio test* (SPRT) which is optimal [Wald, 1947; Wald and Wolfowitz, 1948] would have required the least number of inspections N on an average with comparable error rates α and β for testing H_0: $p = p_0$ versus H_1: $p = p_1 (< p_0)$.

In table 4, I summarize some findings obtained from 1000 independently run simulations in each case. I generated Bernoulli populations with $p = 0.1045$ and 0.6058, the most pessimistic and optimistic values of p respectively, that are possible for the supplied batch of computers. In no situation, I ended up accepting H_0 which postulated a higher p than that under H_1 as indicated by the entry '$\#H_0 = 0$'. The entries $\overline{N}, S_{\mathrm{N}}, N_{\min}, N_{\max}$ respectively stand for the average, standard deviation, the minimum and the maximum obtained from 1000 iterations. Even $N_{\max}$ ranged from merely 9 to 385! The rest of the numbers speak for themselves.

	$p_0 = 0.90, p_1 = 0.80$	$p_0 = 0.75, p_1 = 0.70$
$\mathbf{p = 0.0145}$	$\#H_0 = 0, \overline{N} = 7.11, S_{\mathrm{N}} = 0.34$ $N_{\min} = 7, N_{\max} = 9$	$\#H_0 = 0, \overline{N} = 26.37, S_{\mathrm{N}} = 0.64$ $N_{\min} = 26, N_{\max} = 30$
$\mathbf{p = 0.6058}$	$\#H_0 = 0, \overline{N} = 23.39, S_{\mathrm{N}} = 9.49$ $N_{\min} = 7, N_{\max} = 66$	$\#H_0 = 0, \overline{N} = 152.95, S_{\mathrm{N}} = 49.09$ $N_{\min} = 57, N_{\max} = 385$

Table 4. Summary from 1000 simulations in each case
for Wald's SPRT with $\alpha = \beta = 0.01$.

5 Concluding thoughts

It is clear that the protocol that allowed inspecting $4,000$ computers to detect *only* 58 good ones was at best outrageously wasteful. This is a stunning example of the fleecing of taxpayer's money! An appropriately designed sampling strategy could conclude with near certainty (that is, $\alpha = \beta = 0.01$) that the supplied batch was far below any expected standard with fewer than 10% inspections. Hiring a qualified statistical consultant at the right time would have saved the State of Connecticut much wasted resources amounting to hundreds of thousands of dollars in this one project alone.

References

[Chase and Bown, 2000]W. Chase and F. Bown. *General Statistics*, 4^{th} edition. Wiley, New York, 2000.

[Corneliussen and Ladd, 1970]A.H. Corneliussen and D.W. Ladd. On sequential tests of the binomial distribution. *Technometrics*, 12, pages 635-646, 1970.

[Ghosh, 1970]B.K. Ghosh. *Sequential Tests of Statistical Hypotheses.* Addison-Wesley, Reading, 1970.

[Ghosh *et al.*, 1997]M. Ghosh, N. Mukhopadhyay, and P.K. Sen. *Sequential Estimation.* Wiley, New York, 1997.

[Hartford Courant, 2004]Hartford Courant. The lead article in "Connecticut" section B of the newspaper. Copyright 2004, The Hartford Courant Co. A Tribune Publishing Company, Tuesday, June 8, 2004.

[Mukhopadhyay, 2004]N. Mukhopadhyay. A new approach to determine the pilot sample size in two-stage sampling. *Communication in Statistics-Theory & Methods* (A special issue dedicated to Milton Sobel's memory), P. Chen, L. Hsu, and S. Panchapakesan, editors, 33, 2004, in press.

[Mukhopadhyay and Cicconetti, 2004]N. Mukhopadhyay and G. Cicconetti. How many simulations to run? In N. Mukhopadhyay, S. Datta, and S. Chattopadhyay, editors, *Applied Sequential Methodologies*, pages 261-292; Marcel Dekker: New York, 2004.

[Mukhopadhyay and Solanky, 1994]N. Mukhopadhyay and T.K.S. Solanky. *Multistage Selection and Ranking Procedures*. Marcel Dekker: New York, 1994.

[Robbins and Siegmund, 1974]H. Robbins and D. Siegmund. Sequential estimation of p in Bernoulli trials. In E.J. Williams, editor, *Studies in Probability and Statistics*, E.J.G. Pitman Volume, pages 103-107; Jerusalem Academic Press: Jerusalem, 1974.

[Stein, 1945]C. Stein. A two sample test for a linear hypothesis whose power is independent of the variance. *Annals of Mathematical Statistics*, 16, pages 243-258, 1945.

[Stein, 1949]C. Stein. Some problems in sequential estimation (abstract). *Econometrica*, 17, pages 77-78, 1949.

[Sukhatme *et al.*, 1984]P.V. Sukhatme, B.V. Sukhatme, S. Sukhatme, and C. Asok. *Sampling Theory of Surveys Applications*, 3^{rd} edition. Iowa State University Press, Ames, 1984.

[Wald, 1947]A. Wald. *Sequential Analysis*. Wiley, New York, 1947.

[Wald and Wolfowitz, 1948]A. Wald and J. Wolfowitz. Optimum character of the sequential probability ratio test. *Annals of Mathematical Statistics*, 19, pages 326-339, 1948.

The Distributions of Stopping Times For Ordinary And Compound Poisson Processes With Non-Linear Boundaries: Applications to Sequential Estimation.

S. Zacks

Binghamton University
Department of Mathematical Sciences
Binghamton, NY 13902-6000
(e-mail: shelly@math.binghamton.edu)

Abstract. Distributions of the first-exit times from a region with non-linear upper boundary are discussed for ordinary and compound Poisson processes. Explicit formulae are developed for the case of ordinary Poisson processes. Recursive formulae are given for the compound Poisson case, where the jumps are positive, having continuous distributions with finite means. Applications to sequential point estimation are illustrated.

Keywords: Stopping times, sequential estimation, non-linear boundaries, compound Poisson processes.

1 Introduction

The distributions of stopping times for ordinary or compound Poisson processes when the boundaries are linear were studied in a series of papers by [Perry *et al.*, 1999a] [Perry *et al.*, 1999b] [Perry *et al.*, 2002a], [Perry *et al.*, 2002b], [Stadje and Zacks, 2003], [Zacks, 1991], [Zacks, 1997] and [Zacks *et al.*, 1999]. In particular, see the survey paper of [Zacks, 2005]. In the present paper we discuss the problem when the boundaries are non-linear. [Picard and Lefevre, 1996] studied crossing times of counting processes with non-linear lower boundaries, using pseudo-polynomials. We are developing a different approach for ordinary or compound Poisson processes with upper non-linear boundaries. In a recent paper by [Zacks and Mukhopadhyay, 2005], the theory presented here was applied to find the exact risk of sequential point estimators of the mean of an exponential distribution. Five different stopping rules with corresponding estimators were considered. By converting the problems to stopping times of an ordinary Poisson process, the boundaries were of two types: concave $B(t) = \gamma t^{\alpha}$, $0 < \alpha < 1$, and convex $B(t) = \gamma t^{\alpha}$, $\alpha > 1$. Explicit solutions were given there for the distributions of the estimators and their moments. When the distributions of the observed random variables are not exponential the situation is much more difficult. We assume that the observed random variables $X_1, X_2, \ldots$ are i.i.d. positive and that, for each $n \geq 1$, the sequence (n, S_n), where $S_n = \sum_{i=1}^{n} X_i$, is transitively sufficient. We apply the Poissonization method (see [Cesaroli, 1983], [Zacks, 1994]) to approximate the distribution of the stopping variable M in the sequential estimation by the distribution of a stopping time T. Here T is the first time that the compound Poisson Process $X(t) = S_{N(t)}$ crosses above an increasing boundary $B(t)$, where $\lim_{t\to\infty} B(t) = \infty$ and $\lim_{t\to\infty} B(t)/t = 0$.

While the derivation of the distributions of stopping times in the ordinary Poisson case is immediate, that for the compound Poisson process is complicated. We outline a solution by solving a sequence of related problems. In Section 2 we derive the distribution of a stopping time T, where an ordinary Poisson process $\{N(t), t \geq 0\}$ crosses an upper boundary $B(t)$. In Section 3 we discuss the problem when a compound Poisson process $X(t)$ crosses $B(t)$. In Section 4 we present an application to sequential estimation and some numerical results from [Zacks and Mukhopadhyay, 2005]. All lemmas and theorems are presented without formal proofs.

2 The Distribution of The First Crossing Time Of An Ordinary Poisson Process

Consider an ordinary Poisson process (OPP) $\{N(t), t \geq 0\}$ with $N(0) = 0$. This is a homogeneous process with intensity λ, $0 < \lambda < \infty$. For the properties of an OPP see [Kao, 1977].

Let $B(t)$ be strictly increasing, non-linear function of t, with $B(0) = 0$, $B(t) \nearrow \infty$ and $B(t)/t \to 0$ as $t \to \infty$. We are interested in the distribution of the stopping time

$$T = \inf\{t \geq t_k : N(t) \geq B(t)\}, \tag{1}$$

where for $l \geq k$ $t_l = B^{-1}(l)$. Since $B(t)$ is strictly increasing, $t_k < t_{k+1} < t_{k+2} < ....$ Notice that the distribution of $N(t_k)$ is Poisson(λt_k). Accordingly,

$$P(T = t_k) = 1 - P(k-1; \lambda t_k), \tag{2}$$

where $P(\cdot;\mu)$ is the cdf of Poisson(μ). We denote by $p(\cdot;\mu)$ the pdf of Poisson(μ).

Lemma 1 *For each λ, $0 < \lambda < \infty$,*

$$P_\lambda\{T < \infty\} = 1.$$

□

Define the defective probability function

$$g_\lambda(j;t) = P\{N(t) = j, T > t\}, \quad j = 0, 1, ... \tag{3}$$

for $t \geq t_k$. Since $T \geq t_k$ with probability one,

$$\begin{aligned} g_\lambda(j;t_k) &= p(j;\lambda t_k), \quad && j = 0, ..., k-1 \\ &= 0, && j \geq k. \end{aligned} \tag{4}$$

Furthermore we have, for $t_{l-1} < t \leq t_l$, $l \geq k$, and $j = 0, ..., l-1$

$$g_\lambda(j;t) = \sum_{i=0}^{j \wedge (l-2)} g_\lambda(i;t_{l-1}) p(j-i;\lambda(t-t_{l-1})), \tag{5}$$

where $j \wedge (l-2) = \min(j, l-2)$. Thus, according to (5),

$$\begin{aligned} P_\lambda\{T > t\} &= \sum_{l=k+1}^{\infty} I(t_{l-1} < t \leq t_l) \sum_{j=0}^{l-1} g_\lambda(j;t) \\ &= \sum_{l=k+1}^{\infty} I(t_{l-1} < t \leq t_l) \sum_{j=0}^{l-2} g_\lambda(j;t_{l-1}) P(l-1-j;\lambda(t-t_{l-1})). \end{aligned} \tag{6}$$

Theorem 1 *The distribution function of T is absolutely continuous on (t_k, ∞) with density*

$$\Psi_T(t;\lambda) = \lambda \sum_{l=k+1}^{\infty} I(t_{l-1} < t < t_l) \cdot \sum_{j=0}^{l-2} g_\lambda(j;t_{l-1}) \cdot p(l-1-j;\lambda(t-t_{l-1})). \tag{7}$$

□

Theorem 2 *The r-th moment of T, $(r \geq 1)$, is*

$$\begin{aligned} E_\lambda\{T^r\} &= t_k^r(1 - P(k-1;\lambda t_k)) \\ &+ r! \sum_{l=k+1}^{\infty} t_{l-1}^r \sum_{j=0}^{l-2} g_\lambda(j;t_{l-1}) \sum_{i=0}^{r} \frac{1}{(r-i)!} \binom{l-1-j+i}{i} \cdot \\ &\cdot \frac{1}{(\lambda t_{l-1})^i} (1 - P(l-1-j+i;\lambda \Delta_l)), \end{aligned} \tag{8}$$

where $\Delta_l = t_l - t_{l-1}$. □

3 The Distribution of The First Crossing Time of a Compound Poisson Process

We consider here a compound Poisson process (CPP) with positive jumps. Accordingly, let $X_0 \equiv 0, X_1, X_2, ...$ be i.i.d. positive random variables having a common absolutely continuous distribution F, with density f. We assume that $F(0) = 0$.

Let $\{N(t), t \geq 0\}$ be an OPP, with intensity λ. We assume that $\{N(t), t \geq 0\}$ and $\{X_n, n \geq 1\}$ are independent. The CPP is $\{X(t), t \geq 0\}$, where

$$X(t) = \sum_{n=0}^{N(t)} X_n. \tag{9}$$

The distribution function of $X(t)$, at time t, is

$$H(x;t) = \sum_{n=0}^{\infty} p(n;\lambda t) F^{(n)}(x), \tag{10}$$

where $F^{(0)}(x) \equiv 1$ and $F^{(n)}(x)$ for $n \geq 1$ is the n-fold convolution

$$F^{(n)}(x) = \int_0^x f(y) F^{(n-1)}(x-y) dy. \tag{11}$$

The density of $H(x;t)$ for $x > 0$ is

$$h(x;t) = \sum_{n=1}^{\infty} p(n;\lambda t) f^{(n)}(x), \tag{12}$$

where $f^{(n)}$ is the n-fold convolution of f. For some $t > 0$ we are interested in the distribution of the stopping time

$$T_c = \inf\{t \geq t^* : X(t) \geq B(t)\}, \tag{13}$$

where $B(t)$ is the non-linear increasing boundary, as in Section 2. The distribution of T_c has an atom at $t = t_k$, given by

$$P\{T_c = t^*\} = 1 - H(B(t^*); t^*), \tag{14}$$

and

$$P\{T_c > t^*\} = H(B(t^*); t^*). \tag{15}$$

Moreover, see [Gut, 1988],

$$\lim_{t \to \infty} \frac{X(t)}{t} = \mu t,$$

where $\mu = E\{X_1\}$. Hence since $\dfrac{B(t)}{t} \to 0$ as $t \to \infty$ we obtain

Lemma 2 *For a CPP* $\{X(t), t \geq 0\}$

$$P\{T_c < \infty\} = 1. \tag{16}$$

□

Define the defective distribution

$$G(x;t) = P\{X(t) \leq x, T_c > t\}. \tag{17}$$

Clearly,

$$P\{T_c > t\} = G(B(t);t). \tag{18}$$

Let $g(x;t)$ denote the defective density of $G(x;t)$. An explicit formula of $g(x;t)$ was derived by [Stadje and Zacks, 2003] for the case of a linear boundary $B(T) = \beta + t$. In the case of a non-linear boundary we follow the following steps.

First, define a sequence $\{B^{(m)}(t), m \geq 1\}$ of step-functions, such that $B^{(m)}(t) \leq B^{(m+1)}(t)$ for all $m \geq 1$, all $0 \leq t < \infty$, and such that $\lim_{m\to\infty} B^{(m)}(t) = B(t)$.

Second, define the stopping time

$$T_c^{(m)} = \inf\{t \geq t^* : X(t) \geq B^{(m)}(t)\}, \tag{19}$$

and correspondingly

$$G^{(m)}(x;t) = P\{X(t) \leq x, T_c^{(m)} > t\}. \tag{20}$$

Notice that $\{T_c^{(m)} > t\} \subset \{T_c^{(m+1)} > t\}$, for all $m \geq 1$. Hence, by monotone convergence

$$\lim_{m\to\infty} G^{(m)}(x;t) = G(x;t). \tag{21}$$

Thus, we approximate $G(B(t);t)$ by $G^{(m)}(B^{(m)}(t);t)$ for m sufficiently large. For $m \geq 1$, let $\{t_l^{(m)}, l \geq 0\}$ be the end points of partition intervals of $[t^*, \infty)$, where

$$t_l^{(m)} = B^{-1}\left(B(t^*) + \frac{l}{m}\right), \quad l \geq 0. \tag{22}$$

The corresponding boundary $B^{(m)}(t)$ is given by the step-function

$$B^{(m)}(t) = \sum_{l=1}^{\infty} I\{t_{l-1}^{(m)} \leq t < t_l^{(m)}\}\left(B(t^*) + \frac{l-1}{m}\right). \tag{23}$$

We develop now recursive formula for $G^{(m)}(x;t)$, $t \geq t^*$. Notice that $t_0^{(m)} = t^*$ for all $m \geq 1$.

Since $T_c^{(m)} \geq t^*$ for all $m \geq 1$,

$$G^{(m)}(x;t_0^{(m)}) = I\{x < B(t^*))H(x;t^*) + I(x \geq B(t^*))H(B(t^*),t^*). \tag{24}$$

Furthermore, since $B^{(m)}(t) \geq B(t^*)$ for all $t \geq t^*$ and $m \geq 1$,

$$G^{(m)}(x;t) = H(x;t), \quad x \leq B(t^*), \tag{25}$$

for all $t > t^*$. In addition, for all $l \geq 1$,

$$G^{(m)}(x; t_l^{(m)}) = G^{(m)}(B^{(m)}(t_{l-1}^{(m)}); t_l^{(m)}), \quad \text{all} \quad x \geq B^{(m)}(t_{l-1}^{(m-1)}). \tag{26}$$

Finally, for every $l \geq 1$, $t_{l-1}^{(m)} < t \leq t_l^{(m)}$ and $x \leq B^{(m)}(t_{l-1}^{(m)})$,

$$G^{(m)}(x; t) = \int_0^x G^{(m)}(y; t_{l-1}^{(m)}) h(x - y; t - t_{l-1}^{(m)}) dy. \tag{27}$$

Thus, for each $m \geq 1$,

$$P\{T_c^{(m)} > t\} = \sum_{l=1}^{\infty} I\{t_{l-1}^{(m)} \leq t < t_l^{(m)}\} \cdot G^{(m)}(B^{(m)}(t_{l-1}^{(m)}); t). \tag{28}$$

Functionals of the distribution of $T_c^{(m)}$ can be derived from (28).

4 Application In Sequential Estimation: The Exponential Case.

Let $X_1, X_2, \ldots$ be i.i.d. random variables having an exponential distribution with mean β, $0 < \beta < \infty$. For estimating β consider the sequential stopping variable

$$M = \min\{m \geq k : m \geq (A/c)^{1/2} \bar{X}_m\}, \tag{29}$$

where A, c are positive constants and $\bar{X}_m = \dfrac{1}{m}\sum_{i=1}^{m} X_i$. For background information on this stopping rule and references see [Zacks and Mukhopadhyay, 2005]. At stopping we estimate β with the estimator $\bar{X}_M$. The corresponding risk is

$$R(\bar{X}_M, \beta) = A \cdot E\{(\bar{X}_M - \beta)^2\} + cE\{M\}. \tag{30}$$

[Zacks and Mukhopadhyay, 2005] applied the theory of Section 2 to evaluate exactly the functionals $E\{\bar{X}_M\}$ and $R(\bar{X}_M, \beta)$. This was done by considering the OPP $\{N(t), t \geq 0\}$ with intensity $\lambda = 1/\beta$. If we replace m and $m\bar{X}_m$, respectively, with $N(t)$ and t we obtain from (29) the related stopping time

$$T = \inf\{t \geq t_k : N(t) \geq \gamma t^{1/2}\}, \tag{31}$$

where $\gamma = (A/c)^{1/4}$ and $t_k = (k/\gamma)^2$. Here we have $M = N(T)$ and $\bar{X}_M = T/N(T)$. By slight modification of equation (8) we get the moments of $\bar{X}_M$. In Table 1 we present some exact values of $E\{M\}$, $E\{\bar{X}_M\}$ and $R(\bar{X}_M, \beta)$.

c	$\beta = 1$			$\beta = 1.25$		
	$E\{M\}$	$E\{\bar{X}_M\}$	$R(\bar{X}_M, \beta)$	$E\{M\}$	$E\{\bar{X}_M\}$	$R(\bar{X}_M, \beta)$
0.5	4.712	0.8663	4.1318	5.584	1.0739	5.4454
0.1	9.482	0.8757	2.2889	11.867	1.1145	2.9793
0.05	13.472	0.0915	1.7079	16.987	1.1500	2.1487
0.01	31.892	0.9597	0.7207	39.076	1.2124	0.8687
0.005	44.305	0.9742	0.4834	55.515	1.2249	0.5931

Table 1. Exact Values of $E\{M\}$, $E\{\bar{X}_M\}$ and $R(\bar{X}_m, \beta)$ for $A = 10$, $k = 3$.

References

[Cesaroli, 1983]M. Cesaroli. Poisson randomization in occupancy problems. *J. Math. Anal. Appl., 94*, pages 150–165, 1983.

[Gut, 1988]A. Gut. *Stopped Random Walks: Limit Theorems and Applications.* Springer-Verlag, New York, 1988.

[Kao, 1977]E.P.C. Kao. *An Introduction to Stochastic Processes.* Duxbury, New York, 1977.

[Perry *et al.*, 1999a]D. Perry, W. Stadje, and S. Zacks. Contributions to the theory of first-exit times of some compound poisson processes in queueing theory. *Queueing Systems*, pages 369–379, 1999a.

[Perry *et al.*, 1999b]D. Perry, W. Stadje, and S. Zacks. First-exit times for increasing compound processes. *Comm. Statist-Stochastic Models*, pages 977–992, 1999b.

[Perry *et al.*, 2002a]D. Perry, W. Stadje, and S. Zacks. Boundary crossing for the difference of two ordinary or compound poisson processes. *Ann. Oper. Res.*, pages 119–132, 2002a.

[Perry *et al.*, 2002b]D. Perry, W. Stadje, and S. Zacks. First-exit times of compound poisson processes for some types of positive and negative jumps. *Stochastic Models*, pages 139–157, 2002b.

[Picard and Lefevre, 1996]P. Picard and C. Lefevre. First crossing of basic counting processes with lower non-linear boundaries: a unified approach through pseudopolynomials (i). *Adv. Appl. Prob.*, pages 853–876, 1996.

[Stadje and Zacks, 2003]W. Stadje and S. Zacks. Upper first-exit times of compound poisson processes revisited. *Prob. Eng. Inf. Sci.*, pages 459–465, 2003.

[Zacks and Mukhopadhyay, 2005]S. Zacks and N. Mukhopadhyay. Exact risks of sequential point estimators of the exponential parameter. *Sequential Analysis*, 2005.

[Zacks *et al.*, 1999]S. Zacks, D. Perry, D. Bshouty, and S. Bar-Lev. Distribution of stopping times for compound poisson processes with positive jumps and linear boundaries. *Stochastic Models*, pages 89–101, 1999.

[Zacks, 1991]S. Zacks. Distributions of stopping times for poisson processes with linear boundaries. *Comm. Statist.-Stochastic Models*, pages 233–242, 1991.

[Zacks, 1994]S. Zacks. The time until the first two order statistics of independent poisson processes differ by a certain amount. *Comm. Statist.-Stochastic Models*, pages 853–866, 1994.

[Zacks, 1997]S. Zacks. Distributions of first-exit times for poisson processes with lower and upper linear boundaries. In N.L. Johnson and N. Balakrishnan, editors, *A Volume in Honor of Samuel Kotz*, pages 339–350, 1997.

[Zacks, 2005]S. Zacks. Some recent results on the distributions of stopping times of compound poisson processes with linear boundaries. *J. Statist. Planning and Inf.*, pages 95–109, 2005.

Last exit times for a class of asymptotically linear estimators*

M. Atlagh[1], M. Broniatowski[2], and G. Celant[3]

[1] Département de Mathématiques,
Université Louis Pasteur, 4 Rue René Descartes, 67000 Strasbourg, France
[2] LSTA,
Université Paris 6, 4 Place Jussieu, 75005 Paris, France
[3] Dipartimento di Scienze Statistiche,
Via C.Battisti 241, Padova, Italy

Abstract. We study the last exit time for the Glivenko-Cantelli statistics indexed by some class of functions. Also we provide upper bounds for its tail distribution. Our first example is the Glivenko-Cantelli statistics indexed by a subclass of a Sobolev space; we next consider last exit times for adaptive semiparametric estimates in the spirit of Klaassen, for which we provide the distribution and tail bounds uniformly upon the nuisance parameter.

AMS 2000 Classification:Primary 62F35
Keywords: Last exit time, Functional Glivenko-Cantelli statistics, Adaptive estimation, Semiparametric estimation.

1 Introduction

Rates of convergence for statistical estimators usually focus of asymptotic equivalents for the distance between the estimate and the parameter it intends to approximate. When the estimate is strongly consistent, which is to say that it converges almost surely, then the time which is necessary in order that it stays in some neighborhood of the true value of the parameter from then on is a well defined random variable, which bears a very intuitive sense and which, sometimes, can be evaluated, at least for small neighborhoods of the parameter. In this context, the situation is quite similar to the case when we consider a deterministic sequence x_n converging to x in a metric space: given some (small) ε, which is the order of magnitude of the integer $N(\varepsilon)$ such that, for all n larger than $N(\varepsilon)$, the distance between x_n and x remains forever smaller than ε? This class of problems is usually referred to as "last exit times" problems, considering that the terms of the sequence of estimates may stay outside the ε-neighborhood of its limit only when when n is smaller than $N(\varepsilon)$. This notion has been presented for sequences of M-estimates by [Stute, 1983], and has been extended to the last exit time for the Glivenko-Cantelli statistics by [Hjort and Fenstad, 1992]; extensions for the case when the sample is drawn from a Markov chain or from a strongly mixing

* Work partly supported by Ateneo di Padova, n. CPDAO23415, Italy

sequence have been studied by [Barbe *et al.*, 1999], and some extensions for U-statistics have recently been proposed by [Bose and Chatterjee, 2001]. The present paper follows this chain of works and provides some insight in the range of adaptive semi parametric estimates; it also provides some information on the tail of the distribution of last exit times for those types of estimates. The main tool to be imported for the obtention of the law of $N(\varepsilon)$ for such estimates is uniformity with respect to the nuisance parameter. This is achieved through Gaussian approximations for the so-called sequential empirical process, a device which has been proposed in the form which is to be used here by [Sheehy and Wellner, 1992].

The structure of the paper is as follows: The first section is devoted to the obtention of a general result on last exit times for the Glivenko-Cantelli statistics indexed by a class of functions, following the line defined by [Stute, 1983]. The second section specializes this result for various situations, with an emphasis towards semi parametric adaptive estimators in the spirit of [Klaassen, 1987].

2 Last exit time for the functional Glivenko-Cantelli Statistics

A sample $(X_1, ..., X_n)$ is given , with i.i.d. components following a common distribution P on some space $\mathcal{X}$. For f a real valued measurable function on $\mathcal{X}$ we denote Pf the expectation of f with respect to P, i.e. $Pf := \int f dP$. Denote P_n the empirical measure pertaining to the sample, $P_n := \frac{1}{n}\sum_{i=1}^{n} \delta_{X_i}$ where δ_x is the Dirac mass at point x. In the sequel $\mathcal{F}$ denotes a subclass of functions in $\mathcal{L}^2(P)$ for which we assume that for all x in $\mathcal{X}$, the condition

$$\sup_{f\in\mathcal{F}} |f(x) - Pf| \quad \text{is finite} \tag{1}$$

holds. Define

$$N_\varepsilon := \sup\left\{ n \geq 1 : \sup_{f\in\mathcal{F}} |(P_n - P)(f)| \geq \varepsilon \right\},$$

which denotes the last exit time for the Glivenko–Cantelli statistics indexed by $\mathcal{F}$. Since $\mathcal{F}$ is a Donsker class, it satisfies the Glivenko–Cantelli property, namely

$$\lim_{n\to\infty} \sup_{f\in\mathcal{F}} |(P_n - P)f| = 0 \quad \text{a.s.},$$

which implies that N_ε is a.s. finite. We consider the limiting distribution of $\varepsilon^2 N_\varepsilon$ when ε tends to 0.

Following [Stute, 1983], with $N_\varepsilon(f) := \sup\{n \geq 1 : |P_n f - Pf| > \varepsilon\}$, we have, for fixed f in $\mathcal{F}$,

Proposition 1 *Let f belongs to $\mathcal{L}^2(P)$. Then*

(i) $\lim_{\varepsilon\to 0}\varepsilon^2 N_\varepsilon(f)\overset{\mathrm{d}}{=} W^2_{\max}(f) := \sigma^2(f)\sup_{0\le s\le 1} W^2(s)$,
where $\sigma^2(f) = Pf^2-(PF)^2$*, and* $W(s)$ *is the standard Wiener process.*

(ii) $\lim_{\lambda\to\infty}\lim_{\varepsilon\to 0}\dfrac{P\{\varepsilon^2 N_\varepsilon(f)>\lambda\}}{\psi\left(\frac{\sqrt{\lambda}}{\sigma(f)}\right)} = 1$*, where* ψ *denotes the upper tail of the standard normal distribution* $\psi(\lambda) := P[N(0,1)>\lambda]$.

We obtain an information pertaining to the moments of the r.v. $\varepsilon^2 N_\varepsilon(f)$ for small ε.

A sequence of r.v.'s Y_n is r-quick convergent to 0 whenever, for all $\varepsilon>0$, $E(N^r_\varepsilon) := E(\sup\{n\ge 1 : |X_n|\ge\varepsilon\})^r$ is finite.

This property has been used by [Lai, 1981] in order to assess optimality properties of probability ratio tests in sequential analysis.

As a consequence of Proposition 1 we have

Corollary *Let* f *belong to* $\mathcal{L}^2(P)$*. The sequence* $(P_n-P)(f)$ *is* r*-quick convergent to 0 for all* $r>0$.

Proof of Proposition 1

(i) for all $f\in\mathcal{L}^2(P)$, it holds $N_\varepsilon(f) = \sup\left\{n\ge 1 : \frac{1}{n}\left|\sum_{i=1}^n (f(X_i)-Pf)\right|\ge\varepsilon\right\}$
$= \sup\left\{n\ge 1 : \frac{\nu_n(f)}{\sqrt{n}}\ge\varepsilon\right\}$.
Let y be a positive number. Then $P\left\{\varepsilon^2 N_\varepsilon(f)\ge y\right\} = P\left\{N_\varepsilon(f)\ge\frac{y}{\varepsilon^2}\right\}$.
Define $m :=< y/\varepsilon^2 >$, the smallest integer larger or equal y/ε^2. Then

$$\begin{aligned}P\{\varepsilon^2 N_\varepsilon(f)\ge y\} &= P\{N_\varepsilon(f)\ge m\}\\ &= P\left\{\sup\left\{n\ge 1 : \frac{1}{n}\left|\sum_{i=1}^n (f(X_i)-Pf)\right|\ge\varepsilon\right\}\ge m\right\}\\ &= P\left\{\sup_{n\ge m}\frac{1}{n}\left|\sum_{i=1}^n (f(X_i)-Pf)\right|\ge\sqrt{\frac{y_0}{m}}\right\},\end{aligned}$$

where $y_0 := m\varepsilon^2$. We thus have $0\le y_0-y\le\varepsilon^2$, which entails that y_0 tends to y as ε tends to 0.
For all f in $L^2(P)$, we therefore have

$$\begin{aligned}&P\{\varepsilon^2 N_\varepsilon(f)\ge y\}\le P\left\{\sqrt{m}\sup_{n\ge m}\frac{1}{n}\left|\sum_{i=1}^n (f(X_i)-Pf)\right|\ge\sqrt{y_0}\right\}\\ &= P\left\{\sqrt{m}\sup_{t\ge 1}\frac{1}{tm}\left|\sum_{i=1}^{mt}(f(X_i)-Pf)\right|\ge\sqrt{y_0}\right\}\\ &= P\left\{\sigma(f)\sup_{t\ge 1}\frac{1}{t}\left|\frac{1}{\sigma(f)\sqrt{m}}\sum_{i=1}^{mt}(f(X_i)-Pf)\right|\ge\sqrt{y_0}\right\}\\ &=: P(E_m).\end{aligned}$$

For all $r > 1$, set

$$A_m(r) := \left\{ \sigma(f) \sup_{1\le t\le r} \frac{1}{t} \left| \frac{1}{\sigma(f)\sqrt{m}} \sum_{i=1}^{mt} f(X_i) - Pf) \right| \ge \sqrt{y_0} \right\}$$

and

$$B_m(r) := \left\{ \sigma(f) \sup_{t>r} \frac{1}{t} \left| \frac{1}{\sigma(f)\sqrt{m}} \sum_{i=1}^{mt} (f(X_i) - Pf) \right| \ge \sqrt{y_0} \right\}.$$

For all $r > 1$,

$$P(E_m) = P(A_m(r) \cup B_m(r)),$$

and therefore

$$\begin{aligned}\lim_{\varepsilon\to 0} P\left(\varepsilon^2 N_\varepsilon(f) \ge y\right) &= \lim_{m\to\infty} P(E_m) = \lim_{r\to\infty}\lim_{m\to\infty} P(A_m(r)\cup B_m(r)) \\ &= \lim_{r\to\infty}\lim_{m\to\infty} P(A_m(r)) + P(B_m(r)) - P(A_m(r)\cap B_m(r)).\end{aligned}$$

If

$$\lim_{r\to\infty}\lim_{m\to\infty} P(B_m(r)) = 0, \tag{2}$$

then

$$\lim_{m\to\infty} P(E_m) = \lim_{r\to\infty}\lim_{m\to\infty} P(A_m(r)).$$

The proof of (2) is easy.
Following Donsker Invariance Principle, the processes

$$\left(\frac{1}{\sigma(f)\sqrt{m}} \sum_{i=1}^{mt} (f(X_i) - Pf); \quad t \in [1, r] \right)_{m\ge 1}$$

converge in distribution in $D[1, r]$ to the Brownian process $W(t)$. By continuity of the supremum it therefore holds

$$\lim_{m\to\infty} \sup_{1\le t\le r} \frac{1}{t\sigma(f)\sqrt{m}} \left| \sum_{i=1}^{mt} (f(X_i) - Pf) \right| \overset{\mathrm{d}}{=} \sup_{1\le t\le r} \left| \frac{W(t)}{t} \right|.$$

For all $t \in [1, r]$, the process $W^*(\frac{1}{t}) = \frac{W(t)}{t}$ is also a Brownian process. Therefore

$$\sup_{1\le t\le r} \left| \frac{W(t)}{t} \right| \overset{\mathrm{d}}{=} \sup_{\frac{1}{r}\le s\le 1} |W^*(s)|.$$

Hence whenever (2) holds, we have

$$\begin{aligned}\lim_{\varepsilon\to 0} P(\varepsilon^2 N_\varepsilon(f) \ge y) &= P\left(\sigma(f) \lim_{r\to\infty} \sup_{\frac{1}{r}\le s\le 1} |W(s)| \ge \sqrt{y} \right) \\ &= P\left\{ \sigma(f) \sup_{0\le s\le 1} |W(s)| \ge \sqrt{y} \right\} \\ &= P\left\{ W^2_{\max}(f) \ge y \right\},\end{aligned}$$

where $W^2_{\max}(f) = \sigma^2(f) \sup_{0\le s\le 1} W^2(s)$.

(ii) The maximal variance of $\sigma(f)W(s)$ equals $\sigma^2(f)$ and is obtained when $s = 1$.

Let

$$I_h := \{s \in [0,1] : s\sigma^2(f) \geq \sigma^2(f) - h^2\} = \left[1 - \frac{h^2}{\sigma^2(f)}, 1\right],$$

and note that $E(\sigma^2(f)W(1)) = \sigma^2(f)$.

The uniform a.s. continuity of $\sigma(f)W(s)$ on $[0,1]$ entails that

$$\lim_{h \to 0} \frac{1}{h} E\left\{\sup_{s \in I_h} \sigma(f)|W(s) - W(1)|\right\} = 0,$$

which proves that the conditions in [Adler, 1990], Theorem 5.5 are fulfilled, proving the claim. ■

Let us turn to the uniform case, that is, consider the limiting distribution of $\varepsilon^2 N_\varepsilon$, $\varepsilon \to 0$. We will assume

H1: $\mathcal{F}$ is a Donsker class

H2: For all $x \in \mathcal{X}$, $\sup_{f \in \mathcal{F}} |f(x) - P(f)|$ is finite

H3: $\mathcal{F}$ is a permissible class of function, implying that $\sup_{f \in \mathcal{F}} |P_n f - Pf|$ is measurable.

Define a Gaussian process Z_P defined on $[0,1] \times \mathcal{F}$, a version of which has uniformly bounded sample paths which are uniformly continuous on $[0,1] \times \mathcal{F}$ when equipped with the $\tilde{\rho}_P$ pseudo-metric defined on $[0,1] \times \mathcal{F}$ defined through $\tilde{\rho}_P\left((s,f),(t,g)\right) := |s - t| + P(f - g)^2$. The existence of such process is a consequence of hypothesis (H1) above (see [Sheehy and Wellner, 1992]). The *Kiefer-Müller* process Z_P is centered for any s and f, and its covariance operator is given by

$$cov\left[Z_P(s,f), Z_P(t,g)\right] = (s \wedge t)(Pfg - PfPg).$$

We then have

Proposition 2 *When $\mathcal{F}$ satisfies (H1), (H2) and (H3),*

$$\lim_{\varepsilon \to 0} \varepsilon^2 N_\varepsilon \overset{\mathrm{d}}{=} \sup_{(s,f) \in \mathcal{F}'} |Z_P(s,f)|^2,$$

where $\mathcal{F}' := [0,1] \times \mathcal{F}$.

Proof

Let y be some positive number, and $m :=< y/\varepsilon^2 >$. It holds, setting $y_0 = m\varepsilon^2$,

$$\begin{aligned}
&P(\varepsilon^2 N_\varepsilon \geq y) = P(N_\varepsilon \geq m)\\
&= P\left(\sqrt{m}\sup_{n\geq m}\sup_{f\in\mathcal{F}}\frac{1}{n}\left|\sum_{i=1}^{n}(f(X_i)-P(f)\right| \geq \sqrt{y_0}\right)\\
&= P\left(\sup_{f\in\mathcal{F}}\sup_{t\geq 1}\frac{1}{t}\left|\frac{1}{\sqrt{m}}\sum_{i=1}^{mt}(f(X_i)-Pf)\right| \geq \sqrt{y_0}\right)\\
&= P\left(\sup_{f\in\mathcal{F}}\sup_{t\geq 1}\frac{1}{t}\left|\mathbb{Z}_m(t,f)\right| \geq \sqrt{y_0}\right),
\end{aligned}$$

where $\mathbb{Z}_m$ is the sequential empirical process, an element of $\ell^\infty([0,\infty)\times\mathcal{F})$ defined through $\mathbb{Z}_m(t,f) := \frac{1}{\sqrt{m}}\sum_{i=1}^{mt}(f(X_i)-Pf)$. Note that for all fixed t, by (1), $\mathbb{Z}_m(t,.)$ belongs to $l^\infty(\mathcal{F})$, the set of all bounded sequences defined from $\mathcal{F}$ onto $\mathbb{R}$.

Following [Sheehy and Wellner, 1992], Theorem 11, for any $r > 0$, $\mathbb{Z}_m$ converges in distribution in $\ell^\infty([0,r]\times\mathcal{F})$ to Z_P. For all $r \geq 1$, it thus holds

$$\lim_{m\to\infty}\sup_{f\in\mathcal{F}}\sup_{1\leq t\leq r}\left|\frac{\mathbb{Z}_m(t,f)}{t}\right| \overset{\mathrm{d}}{=} \sup_{f\in\mathcal{F}}\sup_{1\leq t\leq r}\left|\frac{Z_P(t,f)}{t}\right|.$$

Define $Z^*(\frac{1}{t},f) := \frac{Z_P(t,f)}{t}$, for $t\in[1,r]$ and $f\in\mathcal{F}$. The centered Gaussian process Z^* is a Kiefer-Müller process indexed by $[1,r]\times\mathcal{L}^2(\mathcal{X})$; its covariance operator is defined, for f,g in $\mathcal{L}^2(\mathcal{X})$ and s,t in $[1,r]$, by

$$E\left(Z^*\left(\frac{1}{s},f\right)Z^*\left(\frac{1}{t},g\right)\right) = \left(\frac{1}{s}\wedge\frac{1}{t}\right)(Pfg-(Pf)(Pg)),$$

whence

$$\sup_{1\leq t\leq r}\left|\frac{Z_P(t,f)}{t}\right| \overset{\mathrm{d}}{=} \sup_{1\leq t\leq r}\left|Z^*\left(\frac{1}{t},f\right)\right| \overset{\mathrm{d}}{=} \sup_{\frac{1}{r}\leq s\leq 1}|Z^*(s,f)|.$$

It follows by continuity that for any $r\geq 1$,

$$\lim_{m\to\infty}P\left(\sup_{f\in F}\sup_{1\leq t\leq r}\frac{1}{t}\mathbb{Z}_m(t,f)\geq\sqrt{y_0}\right) = P\left(\sup_{\frac{1}{r}\leq s\leq 1}|Z^*(s,f)|\geq\sqrt{y_0}\right),$$

which, since y tends to 0 as $m\to\infty$ equals $P\left(\sup_{\frac{1}{r}\leq s\leq 1}|Z^*(s,f)|^2\geq y\right)$.

In order to prove Proposition 2, it remains to prove that

$$\lim_{r\to\infty}\lim_{m\to\infty}P\left(\sup_{f\in\mathcal{F}}\sup_{t\geq r}\frac{1}{tm}\left|\sum_{i=1}^{mt}(f(X_i)-Pf)\right|\geq\sqrt{y_0}\right) = 0.$$

This follows , as (2), selecting some f in $\mathcal{F}$ and noting that

$\sup_{f\in\mathcal{F}}\sup_{t\geq r}\frac{1}{tm}\left|\sum_{i=1}^{mt}(f(X_i)-Pf)\right| \geq \frac{1}{[rm]}\left|\sum_{i=1}^{[rm]}(f(X_i)-Pf)\right|$. ■

3 Last exit times for adaptive estimates

Let $X_1^n := (X_1, \ldots, X_n)$ be an i.i.d. sample with X_1 distributed by $P_{\theta,g}$ on $\mathbb{R}^k$. The parameter of interest is θ, with $\theta \in \Theta$, an open set, an $g \in \mathcal{G}$, the set of nuisance parameters. A locally asymptotically linear estimator T_n of θ satisfies

$$\sqrt{n}\left(T_n - \theta_n - \frac{1}{n}\sum_{i=1}^{n} J(X_i, \theta_n, g)\right) = o_{\theta_n,g}(1). \tag{3}$$

In (3) (θ_n) is any sequence such that

$$\sqrt{n}(\theta_n - \theta) = O(1) \tag{4}$$

and J satisfies

$$\int J(x, \theta, g) dP_{\theta,g}(x) = 0 \tag{5}$$

together with

$$\int |J(x, \theta, g)|^2 dP_{\theta,g}(x) < \infty, \tag{6}$$

for all $\theta \in \Theta$ and $g \in \mathcal{G}$.

All the o and O notation are meant "in probability" where random variables are involved.

The function J in (3) is the influence function for T_n. An estimate S_n of θ is adaptive and efficient whenever there exists a function $J(x, \theta, g)$ such that, for all sequence (θ_n) satisfying (4), it holds

$$\lim_{n\to\infty} \sqrt{n}(S_n - \theta_n) \overset{\mathrm{d}}{=} N\left(0, \Sigma_{\theta,g}^{-1}\right) \tag{7}$$

where $\Sigma_{\theta,g}$ is the covariance matrix of $J(X, \theta, g)$ for fixed regular θ and g. When the J function coincides with the usual score function $\frac{\dot{f}(x,\theta,g)}{f(x,\theta,g)}$ with f the density of $P_{\theta,g}$, (7) is the classical normal limit behavior for ML estimates under contiguity of the sequence of measures $P_{\theta_n,g}$ to $P_{\theta,g}$ for all (θ_n) satisfying (7) and $g \in \mathcal{G}$, which we will assume from now on. Regularity of θ and g is defined in [Bickel, 1982].

Adaptive estimates are efficient for all $g \in \mathcal{G}$, even though the knowledge of g may not be used in the construction of the estimates. Under the above setting, constructions of efficient adaptive estimates have been proposed in [Beran, 1978], [Schick, 1986] and [Klaassen, 1987] among others. We will follow Klaassen's approach based on influence functions and provide some insight on last exit times for his estimates, strenghtening his assumptions when needed. We just need some kind of uniformity with respect to g as stated now.

Assume that there exists a sequence of functions $J_n(x, \theta_n, X_1^n)$ defined on $(\mathbb{R}^k, \Theta, \mathbb{R}^{kn})$ and a function $J(x, \theta, g)$ defined on $(\mathbb{R}^k, \Theta, \mathcal{G})$, such that

$$\sup_{g\in\mathcal{G}} \int J_n(x, \theta_n, X_1^n) dP_{\theta_n,g}(x) = o_{\theta_n}(1) \tag{8}$$

and

$$\sup_{g\in\mathcal{G}}\sqrt{n}\int[J_n(x,\theta_n,X_1^n)-J(x,\theta_n,g)]^2df_{\theta_n g}(x)=o_{\theta_n}(1). \tag{9}$$

Assume further that we can construct a sequence of preliminary estimates S_n of θ such that

$$\sqrt{n}(S_n-\theta)=O_\theta(1). \tag{10}$$

Then we can construct a uniformly locally asymptotically linear adaptive estimate T_n of θ, satisfying therefore

$$\sup_{g\in\mathcal{G}}\sqrt{n}\left(T_n-\theta_n-\frac{1}{n}\sum_{i=1}^{n}J(X_i,\theta_n,g)\right)=o_{\theta_n}(1) \tag{11}$$

for all sequence (θ_n) satisfying (4). Display (11) proves that the estimate J_n of the Influence function J, together with an initial estimate of θ, provides explicit estimates of θ enjoying asymptotic normality and second order efficiency in the sense of Rao.

We now state additional conditions which entail some knowledge n the last exit time for the above adaptive estimate T_n. We assume that $J(x,\theta,g)$ is regular on a bounded open subset S of the image of X_1, say I_mX_1, uniformly upon θ and g. Also J is assumed to be constant outside S. Such conditions entail robustness for T_n. Precisely, assume

(H1) There exists $q>k/2$ such that $\sup_{\theta,g}\|J(\cdot,\theta,g)\|_{W_2^q}<\infty$, where $\|\cdot\|_{W_2^q}$ is the L_2–Sobolev norm of order q on S.

(H2) There exists $K>0$ such that for all a in $I_m\ X_1\setminus S$, for all (θ,g), $J(a,\theta,g)=K$.

(H3) $I_m\ X_1$ is convex or is a countable union of convex sets with non intersecting closures.

Under (H1), (H2) and (H3) the class $\mathcal{J}$ of functions $J(\cdot,\theta,g)$ is Donsker. When $I_mX_1=[0,1]^k$, Theorem 7.7.1 in [Dudley, 1982], entails that for some $K_1>0$, for any $\varepsilon>0$, the entropy number of $\mathcal{J}$ satisfies

$$\log N(\varepsilon,W_2^q,\mathcal{J})\leq K_1e^{-k/q}.$$

Denote $N_\varepsilon:=\sup\{n\geq 1:|T_n-\theta|>\varepsilon\}$ where $|\cdot|$ is the Euclidian norm. Applying Proposition 3 yields

Corollary *Under all the above assumptions, plus (H1), (H2) and (H3),*

(i) $\lim_{\varepsilon\to 0}\varepsilon^2N_\varepsilon\overset{\mathrm{d}}{=}\sup_{\theta,g}\sup_{0\leq s\leq 1}|Z(s,J(\cdot,\theta,g))|^2$ *where* $Z(s,f)$ *is the functional Kiefer-Müller Process as defined in Section 1.*

(ii) *When* $I_mX_1=[0,1]^k$*, then there exists some positive constant* K *such that for all* $\lambda>0$*,*

$$\lim_{\varepsilon\to 0}P(\varepsilon^2N_\varepsilon>\lambda)\leq 2K(\sqrt{\lambda})^{1+k/q}\psi(\sqrt{\lambda}/\sigma_y)$$

where $\sigma_y^2:=\sup_{\theta,g}\int J^2(x,\theta,g)dP_{\theta,g}(x)$.

References

[Adler, 1990]R.J. Adler. *An Introduction to Continuity, Extrema and Related Topics for General Gaussian Processes.* Lecture Notes Monograph Series IMS, 1990.

[Barbe *et al.*, 1999]P. Barbe, M. Doisy, and B. Garel. Last passage time for the empirical mean of some mixing processes. *Stat. Prob. Letters*, 40(3):237–245, 1999.

[Beran, 1978]R. Beran. An efficient and robust adaptive estimator of location. *Annals of Statistics*, 6(2):293–313, 1978.

[Bickel, 1982]P.J. Bickel. On adaptive estimation. *Annals of Statistics*, 10(3):647–671, 1982.

[Bose and Chatterjee, 2001]A. Bose and S. Chatterjee. Last passage times of minimum contrast estimators. *Jour. Austral. Math. Soc.*, 71:1–10, 2001.

[Dudley, 1982]R.M. Dudley. *A Course on Empirical Processes.* Lecture Notes in Mathematics, Springer, 1982.

[Hjort and Fenstad, 1992]N.L. Hjort and G. Fenstad. On the last time and the number of times an estimator is more than ϵ from its large value. *Annals of Statistics*, 820(1):469–489, 1992.

[Klaassen, 1987]A.J. Klaassen. Consistent estimation of the influence function of locally asymtotically linear estimators. *Annals of Statistics*, 15(4):1548–1562, 1987.

[Lai, 1981]T.L. Lai. Asymptotic optimality of invariant sequential probability ratio tests. *Annals of Statistics*, 9(1):318–333, 1981.

[Schick, 1986]A. Schick. On asymptotically efficient estimation in semiparametric models. *Annals of Statistics*, 14(3):1139–1151, 1986.

[Sheehy and Wellner, 1992]A. Sheehy and J.A. Wellner. Uniform Donsker classes of functions. *Annals of Statistics*, 20(4):1983–2030, 1992.

[Stute, 1983]W. Stute. Last passage times of M-estimators. *Scand. Jour. Statistics*, 10:301–305, 1983.

Adaptive M-Estimators For Robust Covariance Estimation

Christopher L. Brown, Ramon F. Brcich, and Abdelhak M. Zoubir

Signal Processing Group, Institute of Telecommunications
Darmstadt University of Technology
Merckstrasse 25, D-64283 Darmstadt, Germany.
(e-mail: chris.brown@ieee.org, r.brcich@ieee.org, zoubir@ieee.org)

Abstract. Robust covariance estimates are required in many applications. Here, a promising adaptive robust scale estimator is extended to this problem and compared to other robust estimators. Often the performance analysis of covariance estimators is performed from the perspective of the final application. However, different applications have different requirements, hence we make a comparison based on some general metrics. Results show that the adaptive scheme shows good performance under the nominal case and graceful degradation in performance with increasing levels of contamination.

Keywords: robust estimation, covariance, M-estimators.

1 Introduction

Numerous problems in signal processing require estimates of covariance. This occurs, e.g., in array processing where the objective is to either detect the number of sources impinging on an array or their directions of arrival (DOA). Unfortunately, the sample covariance estimator has poor performance when there are model deviations or outliers in the observations [Williams and Johnson, 1993].

Robust estimators protect against this, usually for only a small decrease in performance at the nominal model. Robustness is recognised as a favourable property since, in practice, it is more the norm than the exception that such disturbances exist.

Here we concentrate on robust covariance estimation for multidimensional observations. In the context of robust estimation, the covariance matrix is also referred to as the association or scatter matrix to allow for nonexistence of the second order moments. Several approaches have been suggested including:

i) FLOM (Fractional Lower Order Moment) estimators based on covariation [Shao and Nikias, 1993, Tsakalides and Nikias, 1996, Liu and Mendel, 2001].

ii) Nonparametric estimators using signs or ranks [Visuri et al., 2001, Kendall and Gibbons, 1990].

iii) Expectation maximisation (EM) applied to Gaussian mixture models [Kozick and Sadler, 2000].
iv) Ellipsoidal trimming [Cook *et al.*, 1993]
v) Huber's robust M-estimators [Huber, 1981, Williams and Johnson, 1993].

The first two methods are computationally inexpensive, however they may sacrifice too much performance degradation under nominal conditions in order to be robust. The ellipsoidal trimming procedure and the iterative nature of the EM algorithm make their computational complexity an issue when considering implementation. The last is arguably the method of choice but is difficult to use in practice due to the multi-dimensional optimisation it requires.

To avoid these issues, the simplest class of estimators applies a one-dimensional scale estimator to robustly estimate each element of the covariance matrix. The problem then reduces to one of finding robust estimators of scale. To this end, we will investigate a number of robust scale estimators, including an adaptive M-estimator[1] which was shown to improve upon the existing robust estimators of scale [Brcich *et al.*, 2004].

This paper is organised as follows: in Section 2 we define the signal model and describe the scale estimators to be used for element-wise covariance matrix estimation. These methods will be compared with the covariance estimators to be described in Section 3. The final application, e.g. DOA, for these covariance estimates will determine the best metric to be used. However, since we do not wish to restrict our study to one application, we must consider general metrics. Hence, for the simulation results shown in Section 4, comparisons are made using a number of metrics. Finally, conclusions are drawn and directions for future work described.

2 Robust covariance estimation using scale estimators

One approach to the estimation of covariance matrices is to estimate individual matrix elements using robust estimators of scale. Consider the following signal model

$$\boldsymbol{x}(n) = A\boldsymbol{u}(n)\ , \quad n = 1, \ldots, N \tag{1}$$

where $\boldsymbol{x}(n) = [x_1(n), x_2(n), \ldots, x_M(n)]^T$ is the observation vector, $\boldsymbol{u}(n) = [u_1(n), u_2(n), \ldots, u_P(n)]^T$ is a vector of independent and identically distributed (iid) standard (zero mean and unit variance) random variables and A is the $M \times P$ mixing matrix. The true covariance matrix is $C = \mathsf{E}\left[\boldsymbol{x}\boldsymbol{x}^H\right] = AA^H$ and each matrix element is

$$C(i,k) = \mathsf{E}\left[x_i x_k^*\right]\ . \tag{2}$$

[1] In this paper, the term "adaptive" will be used to refer to techniques that are data-dependent, i.e. parameters used in the procedure are set based on the values of the observations

However, rather than simply replacing the expectation operation in (2) with the sample average to estimate matrix elements, a more robust operation is to use [Huber, 1981]

$$\hat{C}_{\sigma}^{*}(i,k) = \frac{\hat{\sigma}^2(x_i + x_k) - \hat{\sigma}^2(x_i - x_k)}{\hat{\sigma}^2(x_i + x_k) + \hat{\sigma}^2(x_i - x_k)}\hat{\sigma}(x_i)\hat{\sigma}(x_k) \tag{3}$$

where $\hat{\sigma}(\cdot)$ is a robust scale estimator. We will now investigate a number of possible robust scale estimators for this procedure.

2.1 Sample estimator

The sample estimator of standard deviation is known [Huber, 1981] to have poor resistance to outliers. Despite this, we will include it in this study to provide a frame of reference.

2.2 Median absolute deviation

The median absolute deviation (MAD)

$$MAD(\boldsymbol{x}) = \text{median}(|\boldsymbol{x} - \text{median}(\boldsymbol{x})|) \tag{4}$$

has been described as a 'candidate for being the "most robust estimate of scale" '[Huber, 1981]. For symmetric distributions, this is approximately half the interquartile range. Hence, to convert this measure to a true scale estimate, it must be normalised. For nominally Gaussian distributions, a MAD-scale estimator is given by $\hat{\sigma}_{\text{MAD}}(\boldsymbol{x}) = \frac{MAD(\boldsymbol{x})}{\Phi^{-1}(0.75)}$ where $\Phi^{-1}(\cdot)$ is the inverse Gaussian cdf.

2.3 M-estimators of scale

The ML estimate of scale may be found by solving the log-likelihood equation,

$$\sum_{n=1}^{N} \psi\left(\frac{x(n)}{\sigma}\right) = 0 \tag{5}$$

for σ where

$$\psi(x) = -1 - x\frac{\dot{f}_X(x)}{f_X(x)} \tag{6}$$

is the scale score function associated with the density $f_X(x)$ and $\dot{f}_X(x)$ denotes the derivative of $f_X(x)$. By contrast, an M-estimator [Huber, 1981] replaces the nominal score function $\psi(x)$ with a similarly behaved function $\varphi(x)$ chosen to confer robustness on the estimator under deviations from the assumed density. With this in mind, Huber proposed that a clipped quadratic score function

$$\varphi_H(x;k) = \min(x^2, k^2) - \delta = \begin{cases} x^2 - \delta, \; |x| \leq k \\ k^2 - \delta, \; |x| > k, \end{cases} \tag{7}$$

be used in the M-estimator for scale as it minimises the maximum relative asymptotic variance of the scale estimate in the case of a contaminated Gaussian distribution. δ is determined such that the estimator is unbiased for the nominal Gaussian distribution. The parameter k controls the sensitivity of the estimator to the contamination and should decrease as the proportion of outliers increases.

2.4 Adaptive M-estimators of scale

One of the drawbacks of the M-estimators described above is that the best value of the cut-off parameter k is dependent on the degree of contamination [Brcich and Zoubir, 2002, Brown *et al.*, 2003]. In [Brcich *et al.*, 2004], an adaptive scheme was presented that sought to relieve this restriction. There, the score function is composed of a family of basis functions, the weights of which are chosen adaptively from the data. By using bases that were appropriate for a range of levels of contamination, the adaptive scheme was able to maintain high performance for a wider range of scenarios than the "static" M-estimators. Of course, as well as finite sample performance, the asymptotic performance of the adaptive scheme will also be dependent on selecting appropriate bases that can adequately represent the optimum score function. For a full description of the adaptive algorithm, see [Brcich et al., 2004].

3 Alternative Robust Covariance Estimators

Together with the element-wise scale estimation based approaches described in the previous section, we will also consider FLOM and sign covariance matrix (SCM) methods.

3.1 FLOM based estimator

The use of FLOMs has been shown to have strong motivation and impressive performance when impulsive noise exists [Shao and Nikias, 1993]. They estimate the covariation of α-stable random processes – analogous to the covariance of Gaussian random variables. Recognising this, FLOM based measures of association were proposed in [Tsakalides and Nikias, 1996] for the purpose of determining DOA. The "covariation" matrices are found by

$$\hat{C}_{\text{FLOM}}(i,k;p) = \frac{\sum_{n=1}^{N} x_i(n) |x_k(n)|^{p-2} x_k^*(n)}{N} \quad . \tag{8}$$

The parameter p is the order of the moments. Setting $p = 2$ reduces (8) to a sample covariance – appropriate under the condition of Gaussianity. However, as contamination occurs, to prevent estimator breakdown, p should be set to a lower value. The lower the value, the greater the degree of robustness, at the cost of less accuracy under the nominal case.

The form of (8) is very similar to that used in ROC-MUSIC [Liu and Mendel, 2001] differing only by the normalisation factor of each column. Further, for identically distributed observations, this normalisation factor will be approximately equal for all columns. Here, only the FLOM-based method will be considered. To ensure Hermitian matrices, the estimated matrix is averaged with its Hermitian, as in [Tsakalides and Nikias, 1996].

3.2 Sign covariance matrix

The SCM was suggested as a robust estimate of covariance in [Visuri et al., 2001]. The concept is to take the sample covariance of some function, $\tilde{\boldsymbol{x}} = \boldsymbol{S}(\boldsymbol{x})$, of the multi-variate observations. In [Visuri *et al.*, 2001] $\boldsymbol{S}(\cdot)$ was the spatial sign function which normalises each observation to a unit vector. Hence, the spatial sign function can be viewed as the multi-dimensional version of the sign function and from this interpretation the robust behaviour of the SCM is clear. Due to the normalisation of the observations, scale information is lost. However, it was also shown that the subspace estimates from the sample SCM converge to the true subspace.

When more than just a good subspace estimate is required, results in [Visuri *et al.*, 2001] showed that for small samples it is better to whiten the observations using the eigenvectors of the sample SCM and then estimate the eigenvalues as the marginal variances of the transformed observations. To estimate the marginal variances the MAD was used.

4 Results

Herein, and without loss of generality, we only consider real random variables. In the results shown here, iid samples of $\boldsymbol{u}(t)$ for $P = 4$ and $N = 100$ were drawn from the selected distribution. The first six distributions were Gaussian mixtures where the nominal distribution was $\mathcal{N}(0, 1)$ distribution and the contaminating distribution was $\mathcal{N}(0, 100)$. The probability of contamination took values $\varepsilon = 0, 0.01, 0.02, 0.05, 0.1, 0.2$. The last two distributions were the t_3 and t_4 distributions respectively.

A number of mixing matrices were considered, however, due to space limitations, results are only shown for two representative cases

$$\tilde{A}_1(i,k) = 0.4^{|i-k|} \text{ and } \tilde{A}_2(i,k) = \begin{cases} 1 & i = k \\ \frac{1}{4} & i \neq k \end{cases} .$$

The mixing matrices are then standardised so that the true covariance matrices have unit diagonals,

$$A(i,k) = \frac{\tilde{A}(i,k)}{\sqrt{\sum_{j=1}^{P} \tilde{A}^2(i,j)}} \quad . \tag{9}$$

Our objective here is to compare the static and adaptive M-estimators of scale with existing methods for the purposes of covariance matrix estimation. The comparison is not straightforward as it can either depend on the final application, i.e., mean squared error (MSE) of DOA estimates, or on more general metrics, such as the Frobenius norm. The former approach is popular in signal processing, however, good performance in one application does not necessarily imply similar performance in others. Hence, we now study the performance of the estimators using three different metrics: the Frobenius norm, relative MSE of the eigenvalues and the sphericity statistic. Average values for the metrics over 500 Monte Carlo runs were calculated.

Frobenius norm: The element-wise sum of squared differences between $\hat{C}$ and C

$$L_F(\hat{C}, C) = \text{trace}\{(\hat{C} - C)(\hat{C} - C)^H\}. \tag{10}$$

This measures the MSE. Results using the Frobenius norm are shown in Table 1 and Table 2 for A_1 and A_2 mixing matrices respectively for the following methods: sample scale estimator, adaptive scale M-estimator with basis functions $\varphi_H(x;k), k = 1.5, 2, 2.5$, static scale M-estimator (Huber) with basis functions $\varphi_H(x;k), k = 1, 1.5, 2, 2.5$, MAD, FLOM based estimator with $p = 1, 1.5, 1.8$ and the SCM in its original (SCM1) and whitened (SCM2) forms.

	Noise distribution							
Estimator	1	2	3	4	5	6	7	8
Sample	0.45	3.4	5.8	15	29	55	6.2	3
Adaptive	0.47	0.6	0.82	1.6	3.8	15	2.5	1.9
Huber 1.0	0.64	0.71	0.8	1.3	2.7	8.8	1.9	1.5
Huber 1.5	0.53	0.64	0.74	1.5	3.9	15	2.3	1.5
Huber 2.0	0.48	0.59	0.79	2.1	6.2	24	2.8	1.8
Huber 2.5	0.46	0.66	1.1	3.4	10	35	3.2	2
MAD	0.72	0.78	0.85	1.2	2.7	7.9	2	1.4
FLOM 1.0	0.61	0.47	0.57	1.5	3.2	6.2	0.82	0.46
FLOM 1.5	0.5	1	1.8	4.5	9	18	2.2	1.2
FLOM 1.8	0.43	1.9	3.8	9.3	18	35	3.9	2
SCM1	2.2	2.2	2.2	2.2	2.2	2.1	2.2	2.2
SCM2	0.73	0.74	0.87	1.5	3.6	14	2.6	1.9

Table 1. Estimator performance using the Frobenius norm and the A_1 mixing matrix.

	Noise distribution							
Estimator	1	2	3	4	5	6	7	8
Sample	0.46	3.2	6.2	14	28	53	5.5	2.9
Adaptive	0.47	0.59	0.84	1.8	4.7	16	2.5	1.7
Huber 1.0	0.64	0.69	0.83	1.5	3.3	11	2	1.4
Huber 1.5	0.52	0.61	0.79	1.7	4.5	16	2.2	1.5
Huber 2.0	0.48	0.6	0.91	2.3	7.1	23	2.7	1.8
Huber 2.5	0.44	0.68	1.2	3.6	10	31	3.1	2
MAD	0.72	0.79	0.88	1.4	3.1	9.9	1.9	1.3
FLOM 1.0	0.61	0.48	0.61	1.6	3.4	6.6	0.85	0.49
FLOM 1.5	0.51	1	1.8	4.7	9.2	18	2.2	1.2
FLOM 1.8	0.43	2	3.7	9.4	18	34	3.9	2
SCM1	2.2	2.2	2.1	2.1	2.1	2.1	2.1	2.1
SCM2	0.67	0.75	0.85	1.5	3.6	14	2.6	1.8

Table 2. Estimator performance using the Frobenius norm and the A_2 mixing matrix.

Though not shown here, when considering the robust scale estimator based techniques as described in Section 2, investigations showed that the use of (3) did indeed improve robustness considerably. No change was observed in the case of the sample estimator. Therefore, all results shown here for these scale estimator based techniques utilised this procedure.

Inspecting the two tables, similar observations can be made,

- The performance when using the sample scale estimator quickly breaks down with contamination.
- Comparing the adaptive and static M-estimators shows that the adaptive scheme tends to follow the best performance of the static schemes – i.e. for low contamination levels, the adaptive scheme shows similar performance to the static case with high k and for high contamination levels it is similar to the static estimator with low k.
- MAD is indeed showing very robust performance, with little deterioration in performance, however, the SCM techniques show themselves to be insensitive to contamination, especially SCM1. In both cases, however, poor performance relative to some of the other techniques is observed in the nominal case (Gaussianity) – as expected of nonparametric techniques.
- Both static M-estimator and FLOM techniques can be "tuned" through parameters k and p respectively. For low contamination, high parameter values are best, while for high contamination, low parameter settings are best.

Note that incorporation of additional bases with smaller k can increase the robustness of the adaptive scheme. This comes at the expense of a slightly higher computational burden and reduction in performance for the nominal and lightly contaminated cases.

Relative MSE of eigenvalues: Let $\lambda_{i,\hat{C}}$, $\lambda_{i,C}$, $i = 1, \ldots, M$ be the ordered eigenvalues of $\hat{C}$ and C. This metric measures the relative squared difference between the eigenvalues $\lambda_{i,\hat{C}}$ and $\lambda_{i,C}$

$$L_E(\hat{C}, C) = \sum_{i=1}^{M} \left(\frac{\lambda_{i,\hat{C}} - \lambda_{i,C}}{\lambda_{i,C}} \right)^2 \quad . \tag{11}$$

Results are shown in Table 3 and similar qualitative conclusions could be drawn as those from Tables 1 and 2. This confirms that an estimator with good performance in a Frobenius norm sense will also produce good eigenvalue estimates. In particular, relevant to this investigation, it confirms that the adaptive M-estimator scheme exhibits good performance compared to static schemes. However, for high contamination levels, again, the MAD and SCM1 methods gain strong justification for their use.

	Noise distribution							
Estimator	1	2	3	4	5	6	7	8
Sample	0.078	5.5	16	1e+02	4.1e+02	1.5e+03	33	5.2
Adaptive	0.089	0.15	0.33	1.3	7.4	1e+02	3.2	1.7
Huber 1.0	0.2	0.24	0.28	0.81	3.6	36	1.9	1.1
Huber 1.5	0.12	0.17	0.25	1.2	7.5	1e+02	2.6	1.2
Huber 2.0	0.096	0.17	0.34	2.4	20	2.8e+02	3.9	1.6
Huber 2.5	0.086	0.22	0.63	6.2	51	5.9e+02	5.4	2.2
MAD	0.31	0.34	0.37	0.83	3.3	27	1.8	0.99
FLOM 1.0	0.27	0.27	0.32	0.73	2.2	7.2	0.35	0.22
FLOM 1.5	0.16	0.31	0.85	5.1	21	92	1.5	0.47
FLOM 1.8	0.093	1.4	5.1	30	1.2e+02	4.9e+02	6.4	1.7
SCM1	1.8	1.8	1.8	1.8	1.9	1.9	1.8	1.8
SCM2	0.22	0.25	0.38	1.3	6.3	89	3.5	1.7

Table 3. Estimator performance using the relative MSE of the eigenvalues and the A_1 mixing matrix.

Sphericity statistic: The ratio of the geometric mean to the arithmetic mean of the eigenvalues,

$$L_{SS}(\hat{C}) = \frac{\left(\prod_i \lambda_{i,\hat{C}} \right)^{1/M}}{\frac{1}{M} \sum_i \lambda_{i,\hat{C}}}. \tag{12}$$

A normalised sphericity metric is then obtained as $L_S(\hat{C}, C) = L_{SS}(\hat{C})/L_{SS}(C)$. The sphericity statistic indicates the shape of the distribution. If C has equal eigenvalues the scale is equal in all directions. It also appears in the likelihood function of model selection criteria, such as the MDL, for source detection with Gaussian observations. Hence an L_S nears 1 would indicate good performance of model selection criteria when using robust eigenvalue estimates. Results are shown in Table 4 for A_2.

- The M-estimator based techniques, both static and adaptive, show steady degeneration with increasing contamination.
- Encouraging results are found for SCM2 and for the higher order FLOMs.
- Results for the sample estimator are seen to be excellent.

	Noise distribution							
Estimator	1	2	3	4	5	6	7	8
Sample	0.98	0.89	0.87	0.89	0.93	0.96	0.95	0.97
Adaptive	0.99	0.96	0.95	0.9	0.79	0.67	0.94	0.96
Huber 1.0	0.97	0.95	0.93	0.89	0.82	0.68	0.92	0.93
Huber 1.5	0.97	0.96	0.94	0.89	0.79	0.68	0.94	0.95
Huber 2.0	0.98	0.96	0.94	0.87	0.76	0.81	0.95	0.96
Huber 2.5	0.99	0.96	0.92	0.83	0.79	0.9	0.95	0.97
MAD	0.94	0.93	0.91	0.88	0.81	0.67	0.9	0.92
FLOM 1.0	0.98	0.89	0.83	0.69	0.56	0.49	0.84	0.88
FLOM 1.5	0.99	0.89	0.84	0.78	0.76	0.79	0.9	0.93
FLOM 1.8	0.98	0.89	0.85	0.82	0.85	0.89	0.92	0.95
SCM1	1.2	1.2	1.2	1.2	1.2	1.1	1.2	1.2
SCM2	0.97	0.96	0.97	0.95	0.91	0.82	0.92	0.94

Table 4. Estimator performance using the sphericity ratio and the A_2 mixing matrix

This can be explained as follows. Both the nominal and contaminating components have the same correlation matrix, differing only in their relative powers. Hence large amounts of contamination do not significantly affect this statistic. However, with only small amounts of contamination the sample subspace can be perturbed. If the nominal and contaminating components possessed different correlation structures, we would expect a steady deterioration in performance with respect to the amount of contamination.

5 Conclusions

The proposed adaptive scheme shows significant advantages over the static M-estimator. In particular, when considering the possibility of an unknown degree of contamination, performance follows the properties of the appropriate static estimator. When compared to other estimators, the adaptive scheme shows good performance in the nominal case, while also showing graceful degradation as contamination increases. Other schemes were shown to have either poor nominal performance (e.g. M-estimator with small k, MAD, SCM, FLOM with small p) or more rapid breakdown (e.g. M-estimator with large k and FLOM with large p).

It is observed that SCM2, i.e. the whitened SCM, shows considerable improvement across *all* metrics for light contamination when compared to the unmodified SCM1. This motivates future investigation into the application of a similar transformation for other estimators.

References

[Brcich and Zoubir, 2002]R. F. Brcich and A.M. Zoubir. Robust estimation with parametric score function estimation. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2002*, Orlando, FL, USA, May 2002.

[Brcich *et al.*, 2004]R. F. Brcich, C. L. Brown, and A. M. Zoubir. An adaptive robust estimator for scale in contaminated distributions. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2004*, Montreal, Canada, May 2004.

[Brown *et al.*, 2003]C. L. Brown, R. F. Brcich, and A. Taleb. Suboptimal robust estimation using rank score functions. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2003*, Hong Kong, April 2003.

[Cook *et al.*, 1993]R. D. Cook, D. M. Hawkins, and S. Weisberg. Exact iterative computation of the robust multivariate minimum volume ellipsoid estimator. *Statistics and Probability Letters*, 16:213–218, 1993.

[Huber, 1981]P. Huber. *Robust Statistics*. John Wiley, 1981.

[Kendall and Gibbons, 1990]M. Kendall and J. D. Gibbons. *Rank Correlation Methods*. Oxford Univ. Press, 5th edition, 1990.

[Kozick and Sadler, 2000]R. Kozick and B. Sadler. Maximum likelihood array processing in non-Gaussian noise with Gaussian mixtures. *IEEE Trans. on Signal Processing*, 48(12):3520–35, December 2000.

[Liu and Mendel, 2001]T. Liu and J. Mendel. A subspace-based direction finding algorithm using fractional lower order statistics. *IEEE Trans. on Signal Processing*, 49(8):1605–13, August 2001.

[Shao and Nikias, 1993]M. Shao and C. L. Nikias. Signal processing with fractional lower order moments: Stable processes and their applications. *Proc. IEEE*, 81(7):986–1010, July 1993.

[Tsakalides and Nikias, 1996]P. Tsakalides and C. Nikias. The robust covariation-based MUSIC (ROC-MUSIC) algorithm for bearing estimation in impulsive environments. *IEEE Trans. on Signal Processing*, 44(7):1623–33, July 1996.

[Visuri *et al.*, 2001]S. Visuri, H. Oja, and V. Koivunen. Subspace-based direction-of-arrival estimation using nonparametric statistics. *IEEE Trans. on Signal Processing*, 49(9):2060–73, September 2001.

[Williams and Johnson, 1993]D. Williams and D. Johnson. Robust estimation of structured covariance matrices. *IEEE Trans. on Signal Processing*, 41(9):2891–906, September 1993.

Minimum Entropy Estimators in Semiparametric Regression Problems

Eric Wolsztynski, Eric Thierry, and Luc Pronzato

Laboratoire I3S, UNSA-CNRS
2000, route des lucioles, Les Algorithmes - bât. Euclide B, BP.121
06903 Sophia Antipolis Cedex, France
(e-mail: {pronzato,et,wolsztyn}@i3s.unice.fr)

Abstract. We consider semiparametric regression problems for which the response function is known up to some vector of parameters θ and the errors have an unknown density f, treated as an infinite-dimensional nuisance parameter for the estimation of θ. The maximum likelihood (ML) estimator is clearly unapplicable in this context, and classical approaches like least squares or M-estimation may perform poorly. Since the results of Stein in 1956, a large amount of work was dedicated to the construction of adaptive estimators that have the same asymptotic behavior as the ML estimator (*asymptotic efficiency*). The focus has been mainly set on the asymptotic theory and the practical results seem to be restricted to the case of scalar observations.

We presented in [Pronzato *et al.*, 2004] an estimator that minimizes the entropy of the symmetrized sample of the residuals. In [Wolsztynski *et al.*, 2005] we show the link between this Minimum Entropy (ME) estimator, the ML estimator, and the two-stage adaptive estimator of [Bickel, 1982]. Also, we show that the shift-invariance property of entropy confers some robustness to ME estimation.

Adaptive estimation has important applications in Signal and Image Processing. The present paper summarizes the theoretical aspects of the ME approach and focuses on such applications. Although asymptotic properties are commonly the main concern, we illustrate the performances of estimators for finite samples through simulations, including multidimensional situations. The examples we consider also illustrate the robustness of ME estimation.

Keywords: Adaptivity, efficiency, entropy estimation, multivariate regression, semiparametric estimation.

1 Introduction

We consider nonlinear regression models that we assume to be known up to some vector of parameters $\theta \in \Theta \subset \mathbb{R}^p$. We denote $\eta(\bar{\theta}, x)$ the model response, where $\bar{\theta} \in \text{int}(\Theta)$ is the true unknown value of θ and $x \in \mathcal{X} \subset \mathbb{R}^q$ are the design variables. In what follows the design can be randomized or fixed. The observations $Y_i \in \mathbb{R}^d$, $d \geq 1$, are given by

$$Y_i = \eta(\bar{\theta}, X_i) + \varepsilon_i \, , \quad i = 1, \ldots, n \, , \tag{1}$$

with (ε_i) a sequence of independent and identically distributed (i.i.d.) random variables with probability density function (p.d.f.) f with respect to the

Lebesgue measure. For a given measure μ on the design variable x we suppose that the identifiability condition $\left[\int_{\mathcal{X}} [\eta(\theta, x) - \eta(\bar{\theta}, x)]^2 \mu(dx) = 0 \;\Rightarrow\; \theta = \bar{\theta}\right]$ is satisfied. The only assumptions that we make on f, along with some usual regularity conditions, are that it is (centrally) symmetric about 0 and has unbounded support. The density of the noise then corresponds to an infinite-dimensional nuisance parameter for the estimation of θ, and an estimator that remains *asymptotically efficient* in this context is termed *adaptive* (whenever it exists). [Bickel, 1982] and then [Manski, 1984] established that adaptivity was possible for nonlinear regression models.

Consider the residuals $e_i(\theta)$ obtained from the observations (1),

$$e_i(\theta) = Y_i - \eta(\theta, X_i) = \varepsilon_i + \eta(\bar{\theta}, X_i) - \eta(\theta, X_i),\; i = 1, \ldots, n. \tag{2}$$

We suggest in [Pronzato *et al.*, 2004] an estimator of θ that minimizes an estimate of the entropy of the residuals in the univariate case. Since entropy is shift-invariant, we use the $2n$ symmetrized residuals $\pm e_i(\theta)$ with density given X_i

$$f^s_{e,X_i}(u) = \frac{1}{2}\left[f(u - \eta(\bar{\theta}, X_i) + \eta(\theta, X_i)) + f(u + \eta(\bar{\theta}, X_i) - \eta(\theta, X_i))\right]. \tag{3}$$

Using classical results of Information Theory, we show in [Wolsztynski *et al.*, 2005] that the (Shannon) entropy $H(f^s_e) = -\int f^s_e(e) \log f^s_e(e) \mu(de)$ of the marginal distribution of the symmetrized residuals, $f^s_e(u) = \int_{\mathcal{X}} f^s_{e,x}(u)\mu(dx)$, is minimum for $\theta = \bar{\theta}$ when the identifiability condition given above is satisfied. When f is unknown, an estimator of $H(f^s_e)$ thus provides a criterion for the estimation of θ. Moreover, we shall see that the shift-invariance property of the entropy makes minimum entropy (ME) estimation robust with respect to the presence of outlying data.

In [Wolsztynski *et al.*, 2005] we show the link between a two-step ME estimation procedure and the adaptive Stone-Bickel approach for univariate observations. The construction involves data splitting, which allows for the estimate of the density to be independent of that of the entropy, and the application of a single Newton step onto a preliminary locally sufficient estimator then provides an asymptotically efficient estimator of θ.

In the next section we consider two direct ME estimation procedures (without data splitting) for multidimensional data samples. Two examples illustrate the performance of our technique in Section 3.

2 Direct Minimum Entropy estimation procedures

The direct ME estimator that we proposed for univariate data is constructed by plugging a kernel density estimate $\hat{f}^\theta_n$ of f^s_e based on the $2n$ symmetrized residuals $\pm e_i(\theta)$ in an empirical expression of the entropy. The density estimate we use is given by

$$\hat{f}_n^{\theta}(u) = \frac{1}{2nh_n} \sum_{i=1}^{n} \left[K\left(\frac{u - e_i(\theta)}{h_n}\right) + K\left(\frac{u + e_i(\theta)}{h_n}\right) \right] ,$$

with h_n a smoothing parameter, and is used to construct the Ahmad-type plug-in entropy estimator

$$\hat{H}_n(\theta) = -\frac{1}{n} \sum_{i=1}^{n} \log \hat{f}_n^{\theta}(e_i(\theta)) \,. \tag{4}$$

This provides a fully non parametric estimate that can be used for the estimation of θ without data splitting. An alternative entropy estimator can also be constructed by using a truncated integral of $\hat{f} \log \hat{f}$ instead of the sum in (4), but in practice the two estimators turn out to be quite close in performance (although the results might vary in function of the selected bandwidth and of the nature of the problem). For the simple case of the location model we give in [Pronzato *et al.*, 2004] a justification for this method and in [Wolsztynski *et al.*, 2004] we show that $\hat{H}_n(\theta) \xrightarrow{p} H(f_e^s) \geq H(f)$ uniformly in θ, $n \to \infty$, with $H(f_e^s) = H(f)$ for $\theta = \bar{\theta}$, provided that the kernel bandwidth h_n decreases slowly enough and f and K satisfy some regularity conditions. Under slightly stronger conditions, we also prove that $\nabla^2 \hat{H}_n(\theta) \xrightarrow{p} \nabla^2 H(f_e^s)$ uniformly in θ, $n \to \infty$, with $\nabla^2 H(f_e^s) = \nabla^2 H(f) = \mathcal{I}(f)$ for $\theta = \bar{\theta}$. However, proving adaptivity of this direct approach remains an open challenge.

Consider now multidimensional observations. In the case of independent components, the entropy of the residuals is the sum of the entropies of each component (i.e. $H(f_e^s) = \sum_{j=1}^{d} H(f_{e^j}^s)$). The construction used in (4) is therefore suitable to obtain the entropy of the residuals as the sum of the entropies of each marginal distribution.

In the general situation where independence of components does not necessarily hold, we can extend the procedure above by simply using techniques of multivariate density estimation. For small dimensions (2 or 3), techniques based on products of univariate kernels are computationally efficient, see for instance [Scott, 1992]. Given $K(.)$ a univariate density that is symmetric about zero, we thus propose to use

$$\hat{f}_n^{\theta}(u) = \frac{1}{2\,n} \left[\sum_{i=1}^{n} \prod_{j=1}^{d} \frac{1}{h_j} K\left(\frac{u^j - e_i^j(\theta)}{h_j}\right) + \sum_{i=1}^{n} \prod_{j=1}^{d} \frac{1}{h_j} K\left(\frac{u^j + e_i^j(\theta)}{h_j}\right) \right] , \tag{5}$$

where $h_j = h_j(\pm e_1^j(\theta), \ldots, \pm e_n^j(\theta), K)$ is the bandwidth of the univariate kernel K based on the j-th component of the symmetrized sample of residuals, with K satisfying common regularity conditions. One can choose, e.g., K as the standard normal, in which case the optimal bandwidth (in the sense of the asymptotic mean integrated squared error) is given by $h^\star = (4/(d+2))^{1/(d+4)} \sigma_i n^{-1/(d+4)}$, see [Scott, 1992]. Plugging (5) into an Ahmad-type estimate of the entropy similar to (4), we obtain a criterion for estimating θ from multidimensional data. In practice, one can define a data-driven selection of h by substituting the estimated standard deviation of the residuals on each component for its exact value into the expression of $h^\star$. Notice that $\hat{H}_n(\theta)$ is two times continuously differentiable w.r.t. $\theta \in \mathrm{int}(\Theta)$ when $\eta(\theta, x)$ is smooth enough.

In higher dimensions, however, kernel estimation techniques rapidly become inefficient. The major limitation comes from the choice of the bandwidth $h(\pm e_1(\theta), \ldots, \pm e_n(\theta), K)$: due to the curse of dimensionality, the bandwidth for each kernel must be large enough to take a sufficient number of data points into account, which causes oversmoothing. The main alternatives involve kernels that are not positive everywhere [Härdle and Linton, 1994], which is not suitable for computing entropy, non-differentiable density estimates, see for instance [Türlach, 1994], and kernel methods with variable bandwidth [Scott, 1992, Devroye and Lugosi, 2000]. We consider now a special case of the latter.

We suggest here a simple alternative that uses the k-nearest neighbor (kNN) entropy estimator as introduced by [Kozachenko and Leonenko, 1987] for $k = 1$ and extended to $k > 1$ in [Goria *et al.*, 2005], where its consistency is proved for general dimension d under very weak conditions on f.

Consider the open ball $v(x, r)$ centered on $x \in \mathbb{R}^d$ with radius $r > 0$; its volume is given by $|v(x, r)| = r^d c_1(d)$, where $c_1(d) = 2\pi^{d/2}/(d\Gamma(d/2))$. Denote the Euclidean distance from $e_i(\theta)$ to its k-th nearest neighbor by $\rho_{i,k}(\theta)$. For the symmetrized residuals $\pm e_j(\theta)$, the kNN-ME estimator of θ then minimizes

$$H_{k,n}(\theta) = d \log \bar{\rho}_k(\theta) + T(n, k)\,, \tag{6}$$

where $\bar{\rho}_k(\theta) = \left(\Pi_{i=1}^{2n} \rho_{i,k}(\theta)\right)^{1/2n}$ is the geometric mean of the kNN distances and $T(n, k) = \log(n-1) - \psi(k) + \log c_1(d)$ does not depend on θ, with $\psi(k) = \Gamma'(k)/\Gamma(k)$, the digamma function.

The parameter k can be chosen so that $k/n \to 0$, $k \to \infty$ when $n \to \infty$; a typical choice is $k = \sqrt{n}$. We shall take $k > p$ where $p = \dim(\theta)$ to avoid singularities. Notice that (6) is not differentiable in θ.

Although asymptotic results are not yet available for this procedure, we present it here as a simple computational alternative. In the next section we present two examples in image processing for 3-dimensional data.

Note that one could consider the estimate (6) of the entropy as another plug-in entropy estimate, that is, as a generalization of the method of kernels (but avoiding the tricky problem of bandwidth selection). Indeed, consider the ball $v(x, \rho_k)$ mentioned above, where $\rho_k = \rho_k(x)$ is the distance from $x \in \mathbb{R}^d$ to its k-th closest point; [Devroye and Wagner, 1977] proved the strong consistency of the kNN p.d.f. estimate $\hat{f}_n(x) = k \left[n \left(\rho_k(x)\right)^d c_1(d)\right]^{-1}$. The ME estimator based on (6) can thus be written as a multivariate plug-in estimator (with a different bias correction term).

3 Examples

We present some simulation results obtained on images, where the estimator of θ is obtained through an exhaustive search on a finite grid. In this context, entropy is a very natural criterion given its key role in coding theory for the definition of maximum compression rates (or equivalently of minimum description lengthes). Minimizing the entropy of the errors between two signals or two images is equivalent to choosing the parameters for which we achieve the maximum compression rate.

We take a 176×144 png picture for the first example (scalar residuals), and a 352×288 jpg one for the second example, which gives 3-dimensional residuals. Here the observations correspond to a bloc A of an image that is contaminated with additive noise. The problem is to locate the corresponding bloc in a copy X of the original image, also contaminated with noise. We suppose that this copy has not suffered from any nonlinear transformation. The coordinates $\bar{\theta}$ of A are measured from the top-left corner of the original image, and θ is therefore a two-dimensional vector. The dimension of the observations corresponds to the number of channels that make each pixel: 1 channel describes the gray level in the black and white png file, whereas 3 channels (RGB) contain the levels of coloring in the color jpg file. Figure 1 shows, clockwise from top left, (a) the 15×15 bloc A, within the small square, to be identified in (b), the working image, that contains 2×6 outliers (white patch); (c) the 30×30 bloc A to be located in the color image (d). (a) and (b) are black and white pictures contaminated by gaussian noise of variance 6; (c) and (d) are in color and are contaminated by salt and pepper noise, where 6% of pixel values are replaced by the maximum or minimum possible values and contaminated pixels are randomly distributed on the image.

In the first example, images (a) and (b), we compare the LS estimator, the Least Absolute Values (LAV) estimator (which minimizes the sum of the absolute values of the residuals), the plug-in ME (piME) estimator given by (4) and the kNN-ME estimator given by (6). The bandwidth h for piME is set to $.2345\ \sigma\,(2n)^{-1/5}$ (which is optimal in the sense of minimal mean integrated squared error for gaussian kernels, see e.g. [Berlinet and

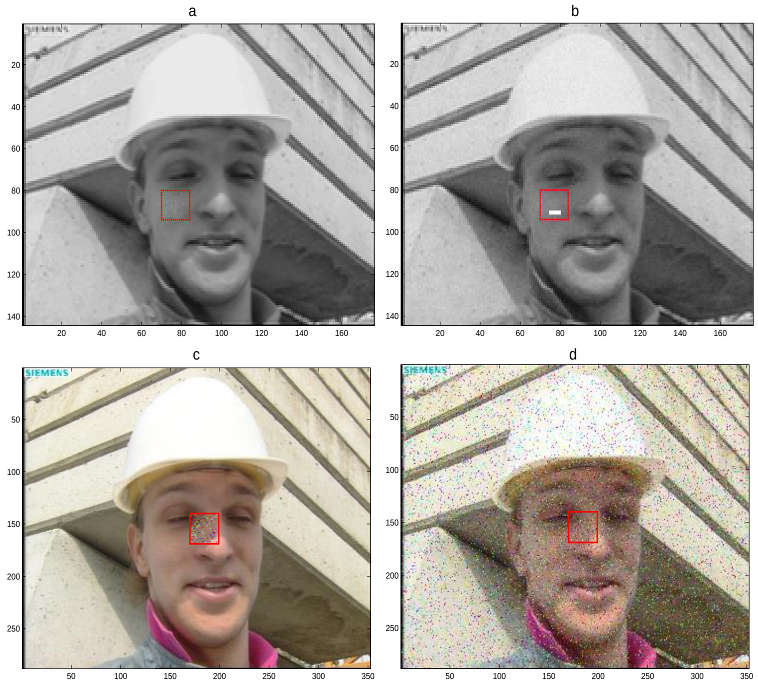


Fig. 1. *Images a, b (black and white, top), c, d (color, bottom).*

Devroye, 1994]) and the value k for the kNN estimator is set to 5. The true value of the parameters of interest is $(80, 70)$. Table 1 contains the means of the estimates obtained for 100 runs of the experiment described above. The two ME estimators estimate $\bar{\theta}$ without error in 100% of the runs and thus appear insensitive to the presence of outliers (the white patch).

In the second example, images (c) and (d), we compare the kNN-ME estimator (6) with a piME-o estimator using the optimal bandwidth $h^{\star}$, a second piME-e estimator using the estimated bandwidth $\hat{h}$ defined by $\hat{h}_j = \widehat{\sigma_j}(2n)^{-1/(d+4)}$ for each component of the observations (with $\widehat{\sigma_j} = \widehat{\sigma_j}(\theta)$ the estimated value of the standard deviation of the $e_i^j(\theta)$, $i = 1, \ldots, n$), see [Scott, 1992], and the standard LS estimator. Figure 2 shows (clockwise from top left) a typical plot of the respective criteria as functions of θ; here $\bar{\theta} = (140,\ 170)$. Note the good behavior of the kNN and piME-o estimators, and the loss of accuracy due to the estimation of the smoothing parameter h for piME-e. The LS criterion gives $\hat{\theta}_{\mathtt{LS}} = (136,\ 173)$ and its shape suggests that it is not suitable for such problems. The value of the entropy of the symmetrized residuals estimated by (6) is 9.99 for $\hat{\theta}_{\mathtt{LS}}$, as opposed to -0.23 for $\hat{\theta}_{\mathtt{kNN}} = \bar{\theta}$.

Table 1. Mean values of the estimates for 100 runs on a black and white picture; $\bar{\theta} = (80, 70)$.

LS	LAV	kNN	piME
$(94.25, 67.51)$	$(94.17, 68.26)$	$(80.00, 70.00)$	$(80.00, 70.00)$

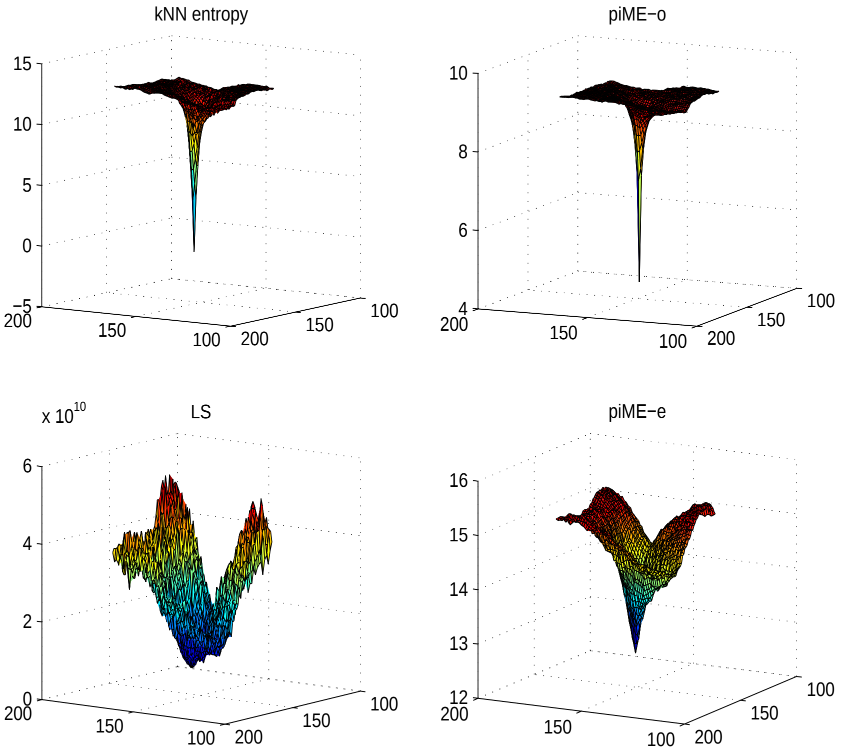


Fig. 2. *criteria vs θ in Example 2, clockwise from top-left : kNN, piME-o, piME-e, LS.* $\bar{\theta} = (140, 170)$.

4 Acknowledgements

We are thankful to Professor Michel Barlaud and Thomas André from Laboratoire I3S for their suggestion of applications in image processing, and for having provided us with the images used in the examples.

References

[Berlinet and Devroye, 1994]A. Berlinet and L. Devroye. A comparison of kernel density estimates. *Publications de l'Institut de Statistique de l'Université de Paris*, 38(3):3–59, 1994.

[Bickel, 1982]P.J. Bickel. On adaptive estimation. *Ann. Stat.*, 10:647–671, 1982.

[Devroye and Lugosi, 2000]L.P. Devroye and G. Lugosi. Variable kernel estimates: on the impossibility of tuning the parameters. In D. Mason E. Giné and J.A. Wellner, editors, *High-Dimensional Probability II*, pages 405–424. Springer-Verlag, New York, 2000.

[Devroye and Wagner, 1977]L.P. Devroye and T.J. Wagner. The strong uniform consistency of nearest neighbor density estimates. *Ann. Stat.*, 5(3):536–540, 1977.

[Goria *et al.*, 2005]M.N. Goria, N.N. Leonenko, V.V. Mergel, and P.L. Novi Inverardi. A new class of random vector entropy estimators and its applications in testing statistical hypotheses. *Jour. Nonparam. Stat.*, 2005. Submitted.

[Härdle and Linton, 1994]W. Härdle and O. Linton. *Applied Nonparametric methods. In "Hanbook of Econometrics"*, volume IV. Elsevier Science B.V., 1994.

[Kozachenko and Leonenko, 1987]L. Kozachenko and N. Leonenko. On statistical estimation of entropy of random vector. *Problems Infor. Transmiss.*, 23(2):95–101, 1987.

[Manski, 1984]C. Manski. Adaptive estimation of nonlinear regression models. *Econometric Reviews*, 3(2):145–194, 1984.

[Pronzato *et al.*, 2004]L. Pronzato, E. Thierry, and E. Wolsztynski. Minimum entropy estimation in semi-parametric models: a candidate for adaptive estimation? In Di Bucchianico, Läuter, and Wynn, editors, *mODa'7 – Advances in Model–Oriented Design and Analysis, Heeze (Netehrlands)*, pages 125–132, Heidelberg, 2004. Physica Springer-Verlag.

[Scott, 1992]D.W. Scott. *Multivariate Density Estimation: Theory, Practice, and Visualization*. Wiley, 1992.

[Stein, 1956]C. Stein. Efficient nonparametric testing and estimation. In *Proc. 3rd Berkeley Symp. Math. Stat. Prob.*, volume 1, pages 187–196. University of California Press, Berkeley, 1956.

[Türlach, 1994]B.A. Türlach. Fast implementation of density-weighted average derivative estimation. *Computationally Intensive Statistical Methods*, 26:28–33, 1994.

[Wolsztynski *et al.*, 2004]E. Wolsztynski, E. Thierry, and L. Pronzato. Consistency of a minimum-entropy estimator of location. Internal Report No I3S/RR-2004-38-FR, 30 pages, http://www.i3s.unice.fr/˜mh/RR/rapports.html, 2004.

[Wolsztynski *et al.*, 2005]E. Wolsztynski, E. Thierry, and L. Pronzato. Minimum-entropy estimation in semi-parametric models. *Signal Processing, Special Issue on Information Theoretic Signal Processing*, 2005.

Part IX

Finance and Insurance

Fair valuation schemes for life annuity contracts

Mariarosaria Coppola[1], Emilia Di Lorenzo[2], and Marilena Sibillo[1]

[1] Department of Economic and Statistical Sciences
Universitá di Salerno,
Complesso universitario, 84084 Fisciano (Sa), Italy
(e-mail: mcoppola@unisa.it, msibillo@unisa.it)

[2] Department of Mathematics and Statistics
Universitá di Napoli "Federico II"
Complesso Monte S. Angelo, via Cintia,
80126 Napoli, Italy
(e-mail: emilia.dilorenzo@unina.it)

Abstract. The paper focuses on the fair valuation of the stochastic reserve of a life policy portfolio. The method, presented for life annuities because of their particular importance in the life insurance market, substantially fits any kind of life policy portfolio. The quantitative approach starts from regulatory and managerial outlines aimed to indicate the reserve quantification as a mark-to-market valuation of the outstanding liabilities. Numerical examples clarify the valuation scheme, comparing the current values of projected cash-flows and the corresponding ones calculated at the contractual rate.

Keywords: Life insurance, Reserve, Fair valuation, Financial risk, Demographic risk.

1 Introduction

The international accounting standards for insurance have been partly defined during 2004 and are partly in course of definitive settlement after revision. Life insurance companies in EU have to follow the new standards and consequently the consolidated financial statements will have to be drawn up in conformity with them (cf. [Jorgensen, 2004]).

The basic idea emerging from the new instructions is to depict the firm as much as possible in its realistic economic profile. In particular, as regards the solvency assessment, the guidance takes shape in the request of the reserve quantification as a mark-to-market valuation of the outstanding liabilities, the so called *fair value.* Several definitions of fair value have been proposed, the strongest line converging toward *"the market value, if a sufficiently active market exists, or an estimated market value otherwise"* (cf. [CAS, 2002]) and the following ([FASB, 2004]): *"the price at which an asset or a liability could be exchanged in a current transaction between knowledgeable unrelated willing persons".*

It's evident that the traditional priciple basing the accounting system on historical cost is now substituted by a new standard founded on current values. The evolution of the international accounting standards reveals that the insurance business is *inside* the market and no more "preferred" rules will make possible to write out no troubled balance sheets.

The valuation techniques established in [FASB, 2004] are classified in a fair value hierarchy, in which the strenght of the connection of insurance cash flows to products traded in active markets is the ordering criterium.

As a consequence, it becomes necessary at the same time to satisfy the fair value principle, imposing the use of the market inputs as much as possible, and to overcome the lack of financial product identical to insurance assets and liabilities traded in a market.

Great importance Buhlmann's idea assumes in this question (cf. [Buhlmann, 2002]): he proposes to measure the liabilities of the insurer resulting from a single policy or a portfolio of policies, as a portfolio of financial instruments, so introducing the Valuation Portfolio (VaPo). The tool originates by the question of the stochastic discounting factor characterization, usable to price securities in arbitrage free markets, analysed by Long (cf. [Long,1990]): in that paper the author introduces the *numeraire portfolio*, defining it as a self financing trading strategy with positive value such that, if prices and cash flows are expressed in its units, the current net cash flow prices are the best prediction of the next period cash-flow prices. Buhlmann observes that the simple circumstance that the insurer sells the insurance contract or the portfolio of insurance contracts, involves that the financial instruments composing the VaPo exist in the economic reality, even if not traded on an existing market, in this way posing in evidence the character of "fair valuation" of the procedure (cf. [Buhlmann, 2004]). The financial component in Buhlmann's valuation process is properly faced by a numeraire approach, considering the cash flow generated by the policy or the portfolio of policies as expressed in "units" of Zero Coupon Bonds, since this methodology makes the valuations comparable each other.

Therefore, the current valuations, if connected with relatively simple insurance liabilities, can be estimated using prices for similar liabilities traded in active markets, in this case being in part of *level two* of the hierarchy proposed by [FASB, 2004].

The case we want to focus refers to a portfolio of life annuities, the interest in this kind of policies being due to their diffusion in the general life insurance outline and in the theoretical implications they have in the pension field. Moreover, this contractual case turns up characterised by a composite risk identity, being affected by the longevity risk, besides the technical (mortality) risk component and the financial risk one (cf. [Coppola *et al.*, 2002]).

In this paper we study the valuation of the stochastic reserve in the case of a portfolio of life annuities by means of a stochastic pricing model based on the no-arbitrage principle applied to the cash flow structure of future assets and liabilities. The valuation technique for estimating fair value will consist in expected valuation connected to cash flows discounted taking in account the systematic financial risk (cf. [FASB, 2004]) together with the demographic risk in its two displays, the accidental and the systematic components. In particular, to capture the effects of the systematic component due to the betterment of the mortality trend, we will use an opportune projected mortality table.

A comparison between a *fair valuation* and a classical procedure based on fixed valuation rates is presented.

2 The valuation scheme

Let us introduce two probability spaces $(\Omega, \mathfrak{F}', P')$, $(\Omega, \mathfrak{F}'', P'')$, where $\mathfrak{F}$' and $\mathfrak{F}$" are the σ-algebras containing, respectively, the *financial events* and the *life duration events* (referring to the unsystematic aspect of mortality). We assume that the randomness in mortality is independent on the fluctuations of interest rates. Let us consider the probability space $(\Omega, \mathfrak{F}, P)$ generated by the preceding two by means of canonical procedures; $\mathfrak{F}$ contains the information flow about mortality and financial history, represented by the filtration $\{\mathfrak{F}_k\} \subset \mathfrak{F}$, with $\mathfrak{F}_k = \mathfrak{F}'_k \cup \mathfrak{F}''_k$ and $\{\mathfrak{F}'_k\} \subset \mathfrak{F}'$, $\{\mathfrak{F}_k\}'' \subset \mathfrak{F}''$. Let us denote by N_j the number of claims (survivors or dieds according to the kind of life contract) at time j within a portfolio of identical policies. We are interested in evaluating at time t the stochastic stream of cash-flows $\hat{\mathbf{X}}^{\mathbf{t}} = N_{t+1}X_{t+1}, N_{t+2}X_{t+2}, \ldots, N_nX_n$, that is the stochastic loss at time t, referring to a portfolio perspective. In a *fair valuation* framework, we assume a frictionless market with continuous trading, no restrictions on borrowing or short-sales, the zero-bonds and the stocks are both infinitely divisible. In a risk-neutral valuation, the fair value at time t is given by

$$\mathfrak{V}_t = \mathbb{E}\left[\sum_{j>t} N_j X_j v(t,j) | \mathfrak{F}_t\right] \tag{1}$$

where $v(t,j)$ is the present value at time t of one monetary unit due at time j, and $\mathbb{E}$ represents the expectation under the risk-neutral probability measure, whose existence derives by well known results, based on the completeness of the market. For a deeper understanding, it is necessary to remark that the demographic valuation is not supported by the hypotesis of the completeness of the market. In any case it is possible to introduce an appropriate probability measure, as suggested in [De Felice and Moriconi, 2004].

Equivalently, indicating by c the number of policies at time 0, in the specific case of surviving benefits, we can write

$$\mathfrak{V}_t = \mathbb{E}\left[\sum_{j>t} c\mathbf{1}_{\{K_{x,t}>j\}} X_j v(t,j) | \mathfrak{F}_t\right] \tag{2}$$

where the indicator function $\mathbf{1}_{\{K_{x,t}>j\}}$ takes the value 1 if the curtate future lifetime of the insured, aged x at issue, takes values greater than $t+j$ $(j = 1, 2, \ldots)$, that is if the insured aged $x+t$ survives up to the time $t+j$, 0 otherwise. By virtue of the basic assumptions on the risk sources, we get

$$\mathfrak{V}_t = \sum_{j>t} cX_j \mathbb{E}[\mathbf{1}_{\{K_{x,t}>j\}}|\mathfrak{F}_t]\mathbb{E}[v(t,j)|\mathfrak{F}_t] \tag{3}$$

$$= \sum_{j>t} cX_j \; {}_tp_x \; {}_jp_{x+t}\mathbb{E}[v(t,j)|\mathfrak{F}_t]$$

where ${}_kp_y$ denotes the probability that an insured aged y survives until the age $y+k$. The terms on the right hand side clearly show that the expected discounted value of the stochastic stream can be regarded as the valuation of a portfolio of zero coupon bonds with maturities in j. The price in t of such portfolio, in a fair valuation approach, can be regarded as the market price of the zero coupon bonds portfolio, and therefore the current value of it.

In order to provide a more concrete application, we consider a portfolio of c insureds aged x, each of whom having a deferred life annuity policy with premiums payable at the beginning of the first T years and benefits payable at the beginning of each year after T if the insured is alive.

According to the notations in [Cocozza *et al.*, 2004], we assume

- B_s=benefit payable to each insured at time s,
- P_s=premium payable by each insured at time s,
- $\hat{\mathbf{X}}^{\mathbf{s}}$= the flow at time s related to each insured, with the generic element represented by the following scheme:

$$X_s = \begin{cases} -P_s & \text{if } s < T \\ B_s & \text{if } s \geq T \end{cases}$$

According to Buhlmann, we'll valuate the financial component of the risk neutral expected value accordingly with a numeraire approach: we determine the value of each flow using a market based discount factor, expressing the current price of the default free unit discount bond issued at time t and maturing at time j $(j \geq t)$.

A representation of the portfolio of financial instruments at time 0 we refer to, that is the Valuation Portfolio, is here reported, having indicated by $Z(t,j)$ the Zero Coupon Bond issued at time t and maturing at time j and considering constant premiums and benefits

$$(VaPo)_0 = \begin{cases} \text{unit} & \text{number of units} \\ Z(0,0) & -cP \\ Z(0,1) & -N_1P \\ \dots & \dots \\ Z(0,T-1) & -N_{T-1}P \\ Z(0,T) & N_TB \\ \dots & \dots \end{cases}$$

while the generic element of the $(VaPo)_t$ results:

$$Z(t,j) = \begin{cases} -N_{t+j/t}P & \text{if } j < T \\ N_{t+j/t}B & \text{if } j \geq T \end{cases}$$

with $N_{t+j/t}$ the number of survivors at time $t+j$ belonging to the group of those, among the c initial insured at time 0, are living at time t.

The calculations in (3) can be replicated also for a non-homogeneous portfolio of life annuities. In fact (cf. [Parker, 1997]) in this case we can divide the porfolio in homogeneous sub-portfolios, say m their number, identified by common characteristics, such as age at issue, policy duration, and so on. Let us assume

- n_i = policy duration for the i-th group
- c_i= number of policies in the i-th group ($\sum_{i=1}^{n} = c$)
- x_i = age at issue of the insureds of the i-th group
- $\hat{\mathbf{X}}^{\mathbf{i,j}}$= the stochastic flow at time j related to the i-th group
- $n = \max_i \ n_i$
- $N_{i,j}$ = number of survivors in the i-th group at time j

Now we can write the value in t for the entire portfolio

$$\mathfrak{V}_t = \mathbb{E}\left[\sum_{i=1}^{m}\sum_{j>t} N_{i,j}X_{i,j}v(t,j)|\mathfrak{F}_t\right] = \tag{4}$$

$$= \mathbb{E}\left[\sum_{i=1}^{m}\sum_{j>t} X_{i,j}c_i\mathbf{1}_{\{K_{x_i,t}>j\}}v(t,j)|\mathfrak{F}_t\right] =$$

$$= \sum_{i=1}^{m}\sum_{j>t} c_iX_{i,j}\ {}_tp_{x_i}\ {}_jp_{x_i+t}\mathbb{E}[v(t,j)|\mathfrak{F}_t]$$

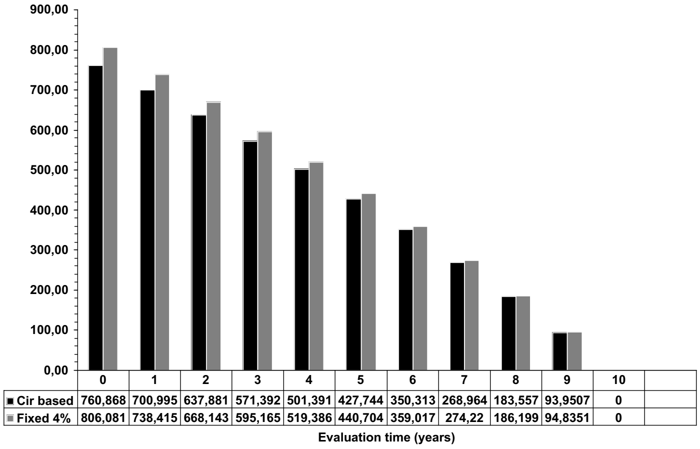
with obvious meaning of the symbol $\mathfrak{F}_t$ in this context.

3 Applications

We present an application of formula (3) for the case of an homogeneus portfolio of 100 immediate 10-years temporary unitary annuities, for policyholders aged 40 at issue. We assume a term structure of interest rates based on the Cox-Ingersoll-Ross square root process

$$dr_t = -k(r_t - \gamma)dt + \sigma\sqrt{r_t}dB_t \tag{5}$$

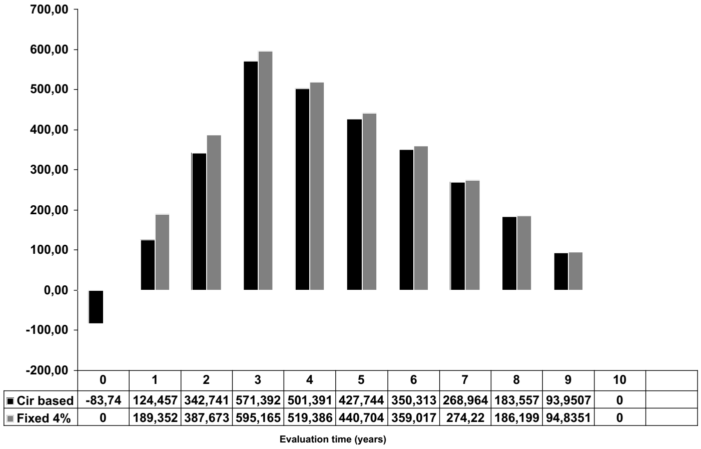
with k and σ positive constants, γ the long term mean and B_t a Brownian motion. In Fig.1 we report the fair values of the reserves and compare them with the corresponding values calculated at the contractual rate 0.04. In particular we assume for the CIR process γ=0,0452, σ=0,0053 and the initial value $r_0 = 0,0279$ (cf. [Cocozza *et al.*, 2004]). We use the survival probabilities deduced by the Italian Male Mortality called RG48, which take into account also the phenomenon concerning the improvement in the mortality trend.



	0	1	2	3	4	5	6	7	8	9	10	
■ Cir based	760,868	700,995	637,881	571,392	501,391	427,744	350,313	268,964	183,557	93,9507	0	
▨ Fixed 4%	806,081	738,415	668,143	595,165	519,386	440,704	359,017	274,22	186,199	94,8351	0	

Fig. 1. Reserves of life annuities
(t=0,1,. . . ,10)

In Fig.2 we present the results obtained, under the above hypotheses about survival and rates, for a portfolio of deferred (T=3 years) life annuities with the same characteristics mentioned above, but periodic premiums are paid at the beginning of each year of the deferment period. In $t = 0$ we have considered also the first premium paid. Since the premiums are calculated at 4%, in $t = 0$ the equity condition is obviously realized only for the contractual rate.



	0	1	2	3	4	5	6	7	8	9	10	
■ Cir based	-83,74	124,457	342,741	571,392	501,391	427,744	350,313	268,964	183,557	93,9507	0	
▨ Fixed 4%	0	189,352	387,673	595,165	519,386	440,704	359,017	274,22	186,199	94,8351	0	

Fig. 2. Reserves of deferred life annuities
(t=0,1,. . . ,10)

We can observe that the current values of the stochastic loss are always smaller than the corresponding calculated at the contractual rate; these differences decrease when the evaluation time increases.

This phenomenon is due to the relation between the impact on the reserve of the variations of the interest rate and the residual time of the policies (cf. [Cocozza *et al.*, 2004]). The above factors directly influence the reserve variation: it depends on the reserve amount and the reserve duration, and both these parameters decrease when the policy residual time decreases, so the impact of the fluctuations of the interest rates are stronger at the beginning of the evaluation time interval.

4 Concluding remarks

According to the general guide-lines provided by the international institutions, the reserve quantification will consist in a mark-to-market valuation of the outstanding liabilities, or, in other words, in its *fair value.* In this framework in the paper the fair value of the losses of a life annuity portfolio is analysed, on the basis of a risk-neutral valuation. The new appraisal will be beneficial to the transparency and for giving the actual economic profile of a life insurance business. More and clearer information will reach the policy holders and the investors, and a higher degree of comparison will be in force among EU life insurance companies (cf [Jorgensen 2004]).

From the point of view of actuarial valuations, the strong aspect of the question will disclose in a bigger volatility of the results, certainly very much greater than that ones gushing from the traditional historical cost based method.

A further development will be a sensitivity analysis concerning the parameters connected to the market fluctuations; moreover it could be interesting to implement simulation procedures aimed to determine the distribution of the stochastic loss as the evaluation time varies and this would provide useful information about solvency assessment.

References

[American Academy of Actuaries, 2003]American Academy of Actuaries. *Comment Letter on Exposure Draft 5 Insurance Contracts.* http://www.iasb.org/docs/ed05/ed5-cl92.pdf, 2003.

[Buhlmann, 2002]H. Buhlmann. New Math for Life Actuaries. *Astin Bulletin* 32, pages 209-211, 2002.

[Buhlmann, 2004]H. Buhlmann. Multidimensional Valuation. http://www.math.ethz.ch/ hbuhl/moskau4.pdf, 2004.

[CAS, 2002]CAS Fair Value Task Force, www.cassact.org/research/tffvl/012.pdf, 2002.

[Chan *et al.*, 1992]K.C. Chan, A.G. Karolyi, F.A. Longstaff and A.B. Sanders. An Empirical Comparison of Alternative Models of the Short-Term Interest Rate. *The Journal of Finance* 47, pages 1209-1227, 1992.

[Cocozza and Di Lorenzo, 2003]R. Cocozza and E. Di Lorenzo. Risk profiles of life insurance business: a combined approach. *Proceedings of the 6th Spanish-Italian Meeting on Financial Mathematics*, Dipartimento di Matematica Applicata Bruno de Finetti, Trieste, pages157-170, 2003.

[Cocozza *et al.*, 2004]R. Cocozza, E. Di Lorenzo and M. Sibillo. Life insurance risk indicators: a balance-sheet approach. *Proceedings of the IME conference*, Roma, 2004.

[Coppola *et al.*, 2002]M. Coppola, E. Di Lorenzo and M. Sibillo. Further Remarks on Risk Sources Measuring in the case of a Life Annuity Portfolio. *Journal of Actuarial Practice*, 10, pages 229-242, 2002.

[Deelstra and Parker, 1995]G. Deelstra and G. Parker. A Covariance Equivalent Discretisation of the CIR Model. *Proceedings of the V AFIR Colloquium*, pages 732-747, 1995.

[De Felice and Moriconi, 2004]M. De Felice and M. Moriconi. Market based tools for managing the life insurance company. http://www.math.ethz.ch/finance/Life_DFM.pdf, 2004.

[FASB, 2004]Financial Accounting Standard Board. *Project update: fair value measurements. March 5 2004*, www.fasb.org

[Frees, 1990]E. W. Frees. Stochastic life contingencies with solvency considerations. *Transactions of the Society of Actuaries*, XLII, pages 91-129, 1990.

[IAIS, 2000]IAIS Solvency & Actuarial Issues Subcommittee. *On solvency, solvency assessments and actuarial issues.* IAIS, March, 2000.

[IAIS, 2002]IAIS Solvency & Actuarial Issues Subcommittee. *Principles on capital adequacy and solvency*, IAIS, January, 2002.

[IASB, 2003]International Accounting Standards Board. *Exposure Draft 5 Insurance Contracts*. International Accounting Standards Board Committee Foundation, London, 2003.

[Jorgensen, 2004]P.L. Jorgensen. On Accounting Standards and Fair Valuation of Life Insurance and Pension Liabilities. *Scandinavian Actuarial Journal*, 5, pages 372-394, 2004.

[Long, 1990]J.B Long. The numeraire portfolio. *Journal of Financial Economics*, 26, pages 29-69, 1990.

[Parker, 1997]G. Parker. Stochastic analysis of the interaction between investment and insurance risks. *North American Actuarial Journal* 1 (2), pages 55-84, 1997.

[Vanderhoof and Altman, 2000]I. Vanderhoof, E.I. Altman Ed. *The fair value of Insurance Business*. Kluwer Academic Publishers, Boston, 2000.

The Fair Valuation of Life Insurance Participating Policies: The Mortality Risk Role

Massimiliano Politano

Department of Mathematics and Statistics
University of Naples "Federico II"
Via Cinthia, Monte S.Angelo - Napoli - Italy
(e-mail: politano@unina.it)

Abstract. This paper analyses the role of the term structure of interest and mortality rates for life insurance participating policies. In particular, aim of this work is to determine the fair valuation of such a policy by modelling mortality risk by means of a Lee Carter type methodology. Numerical results are investigated in order to determine the fair value accounting impact on reserve evaluations.

Keywords: Participating policies, Fair pricing, Lee Carter Methodology, CIR-Black and Scholes framework.

1 Introduction

In life insurance, actuaries have traditionally valued premiums and reserves using deterministic mortality intensity, which is a function of the age of the insured only, and some hypothesis on the dynamics of interest rates. However, since neither the interest rates nor the mortality intensity is deterministic, life insurance companies are essentially exposed to three kinds of risk: the financial risk, the systematic mortality risk, referring to the future development of the underlying mortality intensity, and unsystematic mortality risk, referring to a possible adverse development of the policyholders mortality. It must be pointed out that only the third kind of risk can be controlled by means of portfolio diversification. Since insurance contracts often run for a very long time, a mortality intensity which seems to be prudential at the time of issue, might turn out not to be so. An analogous phenomenon has been observed for the interest rates in the last two decades where we have experienced large drops in the stock prices and low returns on bonds. However, the systematic mortality risk is of different character than the financial risk. While the assets on the financial markets are very volatile, changes in the mortality intensity seems to occur more slowly. Thus, the financial market poses an immediate problem, whereas the level of mortality intensity poses a more long term, but also more permanent problem. This difference could be the reason why emphasis so far has been on the financial markets. In recent years, some of this attention has shifted towards valuation models that fully capture the interest and mortality rates dynamics. In this context, the contribute of the International Accounting standard Board was very important. It defines the Fair Value as "An estimate of an exit price determined by market interactions". At this proposal, it must be remembered that IASB allows for using stochastic models in order to estimate future cash flows. In practice, the problem of determine the market value of insurance liabilities is posed. In this field, it must be remembered the papers of Grosen-Jorgensen [Grosen and Jorgensen, 2000], Bacinello [Bacinello, 2001], Milevsky-Promislow [Milevsky and Promislow, 2001], Ballotta-Haberman [Ballotta and Haberman, 2003]. Here we analyse, in a Lee Carter mortality context, one of the most common life fe insurance policies present on the Italian insurance market, the so called revaluable policy. This policy, of endowment type, has the peculiarity that the insurance company, at the end of each year, grants a bonus which is credited to the mathematical reserve and depends on the performance of an investment portfolio. This bonus is determined in such a way that the total interest credited to the insured is equal to a give percentage of the annual return of the reference portfolio and anyway does not fall below the minimum interest rate guaranteed. Thus, the revaluable policy is of participating type. The paper is organised as follows: section 2 develops the framework for the valuation of the policy. in section 3, the Lee Carter model for the mortality risk is introduced, in section 4 the financial market model is presented. A numerical evidence is offered in section 5.

2 The Model

Let us consider an endowment policy issued at time 0 and maturing at time ξ, with initial sum insured C_0. Moreover, let us define $\{r_t; t = 1, ..., \xi\}$ and $\{\mu_{x+t}; t = 1, ..., \xi\}$ the random spot rate process and the mortality process respectively, both of them measurable with respect to the filtrations $\mathcal{F}^r$ and $\mathcal{F}^\mu$. The above mentioned processes are defined on a unique probability space $(\Omega, \mathcal{F}^{r,\mu}, P)$ such that $\mathcal{F}^{r,\mu} = \mathcal{F}^r \cup \mathcal{F}^\mu$. For the revaluable endowment policy, we assume that, in case of single premium, at the end of the t-th year, if the contract is still in force the mathematical reserve is adjusted at a rate ρ_t defined as follows [Bacinello, 2001]

$$\rho_t = max\left\{\frac{\eta S_t - i}{1+i}, 0\right\} \qquad t = 1, ...\xi \tag{1}$$

The parameter η, $0 \leq \eta \leq 1$, denotes the constant partecipating level, and S_t indicates the annual return of the reference portfolio. The relation (1) explains the fact that the total interest rate credited to the mathematical reserve during the t-th year, is the maximum between ηS_t and i, where i is the minimum rate guaranteed to the policyholder. Since we are dealing with a single premium contract, the bonus credited to the mathematical reserve implies a proportional adjustment at the rate ρ_t, also of the sum insured.

Denoting by C_t, $t = 1, ..., \xi$, the benefit paid at time t if the insured dies between ages x+t-1, x+t or, in case of survival, for $t = \xi$, the following recursive relation holds for the benefits of successive years

$$C_t = C_{t-1} (1 + \rho_t) \qquad t = 1, ..., \xi$$

The iterative expression for them is instead

$$C_t = C_0 \prod_{j=1}^{t} (1 + \rho_j) \qquad t = 1, ..., \xi$$

where we have indicated by ϕ_t the readjustment factor

$$\phi_t = \prod_{j=1}^{t} (1 + \rho_j) \qquad t = 1, ..., \xi$$

In this context, as the elimination of the policyholder can happen in case of death in the year $t \in [0, \xi\ [$ or in case of survival $t = \xi$ the liability borne out by the insurance company can be expressed in this manner

$$W_0^L = \sum_{t=0}^{\xi} C_t\ {}_{t-1/1}Y_x + C_\xi\ {}_\xi J_x \tag{2}$$

where

$${}_{t-1/1}Y_x = \begin{cases} e^{-\Delta(t)} & if\ t-1 < T_x \leq t \\ 0 & \text{otherwise} \end{cases} \qquad {}_\xi J_x = \begin{cases} 0 & if\ 0 < T_x < \xi \\ e^{-\Delta(\xi)} & T_x \geq \xi \end{cases}$$

In the previous expression T_x is a random variable which represents the remaining lifetime of a insured aged x, $\Delta(t) = \int_0^t r_u du$ is the accumulation function of the spot rate.

3 Mortality Risk modeling

for the dynamics of the process $\{\mu_{x+t}; t = 1, 2, ...\}$, we propose to choose a model based on the Lee Carter methodology. According to the traditional actuarial approach, the survival function of the random variable T_x is given by [Milevsky and Promislow, 2001]

$${}_\xi p_x = P\left(T_x > \xi / \mathcal{F}_0^\mu\right)$$

where $\mathcal{F}_0^\mu$ represents the mortality informative structure available at time 0. If we make the hypothesis of the time dependence of the mortality intensity and we define $\mu_{x+t:t}$ the mortality intensity for an individual aged x+t, observed in the year t, it is possible to express the previous formula as follows [Ballotta and Haberman, 2003]

$$ {}_{\xi}p_x = E\left(exp\left\{ -\int_0^{\xi} \mu_{x+t:t} dt \right\} / \mathcal{F}_0^{\mu} \right) \tag{3} $$

A widely used actuarial model for projecting mortality rates is the reduction factor model. This model has traditionally been formulated with respect to the conditional probability of dying in a year

$$ q\,(y,t) = q\,(y,0)\, RF\,(y,t) $$

where $q\,(y,0)$ represents the probability that a person aged y will die in the next year, based on the mortality experience for the base year 0 and correspondingly $q\,(y,t)$ relates to the future year t. Given the form of (3), it is considered a reduction factor approach for the mortality intensity so that

$$ \mu_{y:t} = \mu_{y:0} RF\,(y,t) \tag{4} $$

where $\mu_{y:0}$ is the mortality intensity of a person aged y in the base year 0, $\mu_{y:t}$ is the mortality intensity for a person attaining age y in the future year t, and the reduction factor is the ratio of the mortality intensity. It is possible to target RF, in a Lee Carter approach, $\mu_{y:0}$ being completely specified. Thus, $\mu_{y:0}$ is estimated

$$ \widehat{\mu}_{y:0} = \sum_t d_{y:t} / \sum_t e_{y:t} $$

where $d_{y:t}$ denotes the number of deaths at age y and time t and $e_{y:t}$ indicates the matching person years of exposure to the risk of death. Taking the logarithm of equation (4) and defining

$$ \alpha_y = \log\left(\mu_{y:0}\right) \qquad \log\left\{RF\,(y,t)\right\} = \beta_y k_t $$

the Lee Carter structure is reproduced [Renshaw and Haberman, 2003]. In fact the Lee Carter model for death rates is given by

$$ \ln\left(m_{yt}\right) = \alpha_y + \beta_y k_t + \epsilon_{yt} \tag{5} $$

where m_{yt} denotes the central mortality rate for age y at time t, α_y describes the shape of the age profile averaged over time, k_t is an index of the general level of mortality while β_y describes the tendency of mortality at age y to change when the general level of mortality k_t changes. ϵ_{yt} denotes the error. For this model, the strategy is to estimate the values for α_y, β_y, k_t on the historical data for the population in question, the difficulty concerns the fact that the quantities on the right hand of (5) are not directly observable. therefore, denoting by n the number of observable periods and $t = t_1, t_2, ..., t_n$, the parameters are normalized requiring that [Lee, 2000]

$$\sum_t k_t = 0 \qquad \sum_y \beta_y = 1$$

so that

$$\widehat{\alpha}_y = \ln\left[\left(\prod_t \widehat{\mu}_{yt}\right)^{\frac{1}{n}}\right] \qquad \widehat{k}_t = \sum_{y=0}^{\omega}\left[\ln\left(\widehat{m}_{yt}\right) - \widehat{\alpha}_x\right]$$

The parameter β_y can be estimated by an ordinary regression between $\widehat{k}_t$ and $\ln\left(\widehat{m}_{yt}\right)$ In this framework, for our purposes, with $y = x + t$, one can use the following model for the time evolution of the hazard rate

$$\mu_{x+t:t} = \mu_{x+t:0}\; e^{\beta_{x+t}k_t}$$

4 Financial Risk modeling

For the process $\{r_t; t = 1, 2, ...\}$, we assume a mean reverting square root dynamics

$$dr_t = f^r\left(r_t, t\right)dt + l^r\left(r_t, t\right)dZ_t^r$$

where $f^r\left(r_t, t\right)$ is the drift of the process, $l^r\left(r_t, t\right)$ is the diffusion coefficient, Z_t^r is a standard Brownian Motion; in particular, in the CIR model the drift function and the diffusion coefficient are defined respectively as [Cox *et al.*, 1985]

$$f^r\left(r_t, t\right) = k\left(\theta - r_t\right) \qquad l^r\left(r_t, t\right) = \sigma_r\sqrt{r_t}$$

where k is the mean reverting coefficient, θ the long term period "normal" rate, σ_r the spot rate volatility. It must be pointed out that for pricing interest rate derivatives the Vasicek model is widely used. Nevertheless, this model assigns positive probability to negative values of the spot rate; for long maturities, this can have relevant effects and therefore the vasicek model appears inadequate to value life insurance policies. Clearly on the fair pricing of our policy, it is very important the specification of the reference portfolio dynamics. The diffusion process for this dynamics is given by the stochastic differential equation

$$dS_t = f^S\left(S_t, t\right) + g^S\left(S_t, t\right)dZ_t^S$$

where S_t denotes the price at time t of the reference portfolio Z_t^S is a Standard Brownian Motion with the property

$$Cov\left(dZ_t^r, dZ_t^S\right) = \varphi dt \qquad \varphi \in R$$

Since we assume a BS type model [Black and Scholes, 1973], we have

$$f^S(S_t,t) = \mu_S S_t \qquad g^S(S_t,t) = \sigma_S S_t$$

where μ_S is the continously compounded market rate, assumed to be deterministic and constant and σ_S is the constant volatility paremeter.

In this context, for the policy under consideration, the unit price in t with maturity ξ is given by

$$u(t,\xi) = E\left[(exp\left\{-\int_t^{\xi} r_u du\right\}\phi_t/\mathcal{F}_t^r\right]$$

where $\mathcal{F}_t^r$ represents the financial informative structure available on the market at time t.

5 Some Applications

The described model has been applied in order to analyse the temporal profile of the insurance liability. The next table compares the value of the mathematical reserve using, for an insured aged 40 with time to maturity 20 years, a technical basis given by a constant interest rate of 3% and the life table SIM92 (Statistics Italian Males 1992) with the values obtained in a fair value context using the mortality Italian data for the period 1947-1999 for evaluate the projection of the mortality factor. Moreover we use the 3 month T-bill January 1996 - January 2004 for determine the interest rate factor, the parameters $\mu_S = 0.03$ $\sigma_S = 0.20$ for the stochastic evolution of the reference fund. For the correlation coefficient φ, we have adopted a slightly negative value ($\varphi = -0.06$) coherently with the literature for the Italian stock market.

The table 1 puts in evidence that the introduction of a fair value accounting system determines a reduction in the level of the liability borne out by the fund specially in the first years. This is mainly caused by the historical trend of the bond market where we have experienced a continuous decrease of interest rates. About the influence of the demographic factor, we have performed a comparison of the reserve value using the fair value hypothesis for the financial factors, and the life tables SIM92, RG48 (projected General Accountancy 1948) and the one obtained with the LC methodology for the mortality rates.

The fourth column of table 2 puts in evidence that the life table SIM92 underestimates the reserve values and don't capture the improvements of the human life. The life table RG48 accomplishes the second function but slightly overestimates the reserve value with respect to LC forecasts (i.e fifth column of table 2). Finally, we have calculated the variation coefficient of the contract value (column six of table 2), depending on demographic component in order to offer a measure of riskiness in reference to the problem to calculate an adequate margin for mortality risk in a fair value context.

Year	WtL	FVWtL	Δ WtL	Δ WtL/WtL
0	0,00	0,00	0,00	0,00%
1	1168,66	987,29	181,37	15,52%
2	1199,34	1028,77	170,57	14,22%
3	1231,14	1072,07	159,07	12,92%
4	1264,11	1117,28	146,83	11,62%
5	1295,36	1164,48	130,88	10,10%
6	1333,78	1213,73	120,05	9,00%
7	1370,64	1265,10	105,54	7,70%
8	1408,94	1318,64	90,30	6,41%
9	1448,92	1374,38	74,54	5,14%
10	1490,69	1432,30	58,39	3,92%
11	1534,02	1492,36	41,66	2,72%
12	1579,38	1554,45	24,93	1,58%
13	1620,72	1618,32	2,40	0,15%
14	1683,43	1683,43	0,00	0,00%
15	1735,89	1749,78	-13,89	-0,80%
16	1791,20	1815,98	-24,78	-1,38%
17	1879,40	1880,91	-1,51	-0,08%
18	1911,16	1942,83	-31,67	-1,66%
19	1976,96	1992,16	-15,20	-0,77%
20	2047,04	2047,04	0,00	0,00%

Table 1. Reserves temporal profile

Year	SIM92	RG48	LC	LC vs. SIM92	LC vs. RG48	CV
0	0,00	0,00	0,00	0,00%	0,00%	0,00000
1	991,75	986,80	987,29	-0,45%	0,05%	0,00276
2	1031,16	1028,78	1028,77	-0,23%	0,00%	0,00134
3	1072,35	1072,63	1072,07	-0,03%	-0,05%	0,00026
4	1115,43	1118,44	1117,28	0,17%	-0,10%	0,00136
5	1160,48	1166,29	1164,48	0,34%	-0,16%	0,00255
6	1207,59	1216,23	1213,73	0,51%	-0,21%	0,00367
7	1256,84	1268,36	1265,10	0,66%	-0,26%	0,00470
8	1308,28	1322,69	1318,64	0,79%	-0,31%	0,00564
9	1362,09	1379,23	1374,38	0,90%	-0,35%	0,00644
10	1418,32	1437,94	1432,30	0,99%	-0,39%	0,00707
11	1476,97	1498,74	1492,36	1,04%	-0,43%	0,00751
12	1537,32	1561,47	1554,45	1,11%	-0,45%	0,00801
13	1600,02	1625,86	1618,32	1,14%	-0,46%	0,00823
14	1664,60	1691,48	1683,43	1,13%	-0,48%	0,00821
15	1730,66	1757,71	1749,78	1,10%	-0,45%	0,00796
16	1797,43	1823,66	1815,98	1,03%	-0,42%	0,00744
17	1863,52	1888,04	1880,91	0,93%	-0,38%	0,00672
18	1927,68	1950,99	1942,83	0,79%	-0,42%	0,00610
19	1980,64	1996,91	1992,16	0,58%	-0,24%	0,00420
20	2047,04	2047,04	2047,04	0,00%	0,00%	0,00000

Table 2. Reserves mortality profile

References

[Bacinello, 2001]A.R. Bacinello. Fair pricing of life insurance partecipating policies with a minimum interest rate guaranteed. *Astin Bullettin*, pages 275–297, 2001.

[Ballotta and Haberman, 2003]L. Ballotta and S. Haberman. The fair valuation problem of guaranteed annuity option: the stochastic mortality environment case. *working paper*, 2003.

[Black and Scholes, 1973]F. Black and M. Scholes. The pricing of option and corporate liabilities. *Journal of Political Economy*, pages 637–654, 1973.

[Cox *et al.*, 1985]J. Cox, J. Ingersoll, and S. Ross. A theory of the term structure of interest rates. *Econometrica*, pages 385–408, 1985.

[Grosen and Jorgensen, 2000]A. Grosen and P.L. Jorgensen. Fair valuation of life insurance liabilities: the impact of interest rate guarantees, surrender options and bonus policies. *Insurance: Mathematics and Economics*, pages 37–52, 2000.

[Lee, 2000]R. Lee. The lee-carter method of forecasting mortality with various extensions and applicantions. *North American Actuarial Journal*, pages 80–93, 2000.

[Milevsky and Promislow, 2001]M.A. Milevsky and S.D. Promislow. Mortality derivatives and the option to annuitise. *Insurance: Mathematics and Economics*, pages 299–318, 2001.

[Renshaw and Haberman, 2003]A.E. Renshaw and S. Haberman. On the forecasting of mortality reduction factor. *Insurance: Mathematics and Economics*, pages 379–401, 2003.

The relationship between Stock Market Returns and Inflation: An econometric investigation using Greek data

Dimitris Ioannides[1], Costas Katrakilidis[2], and Andreas Lake[2]

[1] Department of Economics
University of Macedonia
54006 Thessaloniki, Greece
(e-mail: dimioan@uom.gr)

[2] Department of Economics
Aristotle University of Thessaloniki
54006 Thessaloniki, Greece
(e-mail: katrak@econ.auth.gr)

Abstract. Since the theory establishes a relationship between stock market returns and inflation rate, this study examines whether this holds for Greece, over the period 1985 - 2000. Taking a step further, we re-examine the above relationship taking into account the existence of possible structural breaks over the considered time horizon. The empirical methodology uses ARDL cointegration technique in conjunction with Granger causality tests to detect possible long-run and short-run effects between the involved variables as well as the direction of these effects. The results provide evidence in favour of a negative long-run causal relationship between the considered series after 1992.

Keywords: Inflation, stock market returns, ARDL cointegration, causality.

1 Intoduction

The Greek economy suffered from high inflation rates since the late 70's. During the 80's the government followed a loose monetary policy which increased inflation even more. In 1992, a tight monetary policy was introduced, and Greece attempted to decrease the level of inflation in order to achieve the Maastricht criteria. On the other hand, the Greek stock market followed an upward trend from 1985 to 2000, with some fluctuations. According to the generalized Fisher hypothesis, equity stocks, which represent claims against the real assets of a business, may serve as a hedge against inflation. Consequently, investors would sell financial assets in exchange for real assets when expected inflation is pronounced. In such a case, stock prices in nominal terms should fully reflect expected inflation and the relationship between these two variables should be found positively correlated ex ante. According to [Bodie, 1976], equities are a hedge against the increase of the price level due to the fact that they represent a claim to real assets and, hence, the real change on the price of the equities should not be affected. If we consider that firms are in a position to predict their profit margins and since equities are claims on current and future earnings, it also follows that the stock market operates as a hedge against inflation, at least in the long run. The earnings should be consistent with the inflation rate, and hence the real value of the stock market should remain unaltered in the long run. The argument that stock market serves as a hedge against inflation, implies that investors are fully compensated for increases in the general price level through corresponding increases in nominal stock market returns and thus the real returns remain unaffected. In other words the argument is that the real value of the stock market is immune to inflation pressures. This has been tested in the literature numerous times. The hedge hypothesis has been examined extensively in the literature. Empirical evidence is rather mixed and could be classified into the following three categories: a) Research findings which provide support in favor of a positive relationship between inflation and stock market returns. [Firth, 1979], and [Gultekin, 1983], conclude that the relationship between nominal stock returns and inflation in the United Kingdom is relative positive, a finding consistent with the generalized Fisher hypothesis. [Boudhouch and Richarson, 1993], employed data sets covering the period from 1802 to 1990 for the U.S and from 1820 to 1988 for Britain. The results that they obtained suggest a positive relationship between inflation and nominal stock returns over long horizons. [Ioannidis *et al.*, 2004], found evidence of positive correlation between inflation and stock market returns in Greece between 1985 and 2003. [Kessel, 1956], suggests that unexpected inflation increases the firm's equity values if the firm is a net debtor. b) Studies which provide evidence of a negative relationship between the inflation rate and the stock market returns. [Fama, 1981], suggests that there is a negative correlation between stock returns and the level of inflation. The negative relationship exists due to the correlation between inflation and future output. In particular, since stock prices reflect firms' future earnings potential, an economic downturn predicted by a rise in inflation will depress stock prices. [Spyrou, 2001], suggests that there is a negative relationship between stock market returns and inflation in Greece for the period 1990 to 1995. c) Studies which provide mixed results. [Pearce and Roley, 1988], found mixed empirical evidence on the subject. [Anari and Kolari, 2001], report negative correlations between stock prices and inflation in the short run which are followed by positive correlations in the long run.

Our research focuses on the relationship between inflation and stock market returns. The question we attempt to answer through the investigation of the above relationship is whether the stock market has been a safe place for investors in Greece. The empirical analysis is carried out by means of an ARDL cointegration, which permits the detection of long run as well as short run [Granger, 1969] causal effects. The remainder of the paper is organized as follows. Section 2 presents the methodology followed, section 3 presents the data and reports the empirical findings. Finally, section 4 presents a brief summary with some concluding remarks.

2 Methodological Issues

The autoregressive distributed lag approach to cointegration (ARDL) following the methodology outlined in [Pesaran and Shin, 1995] is employed in this paper. The main advantage of this procedure is that it can be applied regardless of the stationary properties of the variables in the sample and allows for inferences on long-run estimates, which is not possible under alternative cointegration procedures. In other words this strategy may applied irrespective of whether the series are $I(0)$ or $I(1)$, and this avoids the pre-testing in the model may be large. It is worth mentioning that the VAR models are not in position to allow for large number variables. The ARDL model in the [Pesaran and Shin, 1995] context is defined as:

$$\varPhi(L)yt = \alpha_0 + \alpha_1 w + \beta'(L)xI + ut \tag{1}$$

where $\varPhi(L) = 1 - \sum_{i=1}^{\infty} \varPhi_i L^i$,
$\beta(L) = \sum_{j=1}^{\infty} \beta_j L^j$
and $L$is the lag operator and wt is a vector of deterministic variables such as the intercept term, seasonal dummies, time trends or exogenous variables with fixed lags. Most of the standard model specifications can be easily derived by imposing restrictions on the parameters. The standard static model can be obtained by imposing the restriction $\beta_1 = \phi_1 = 0$. The restrictions $\beta_1 = 0$and $\phi_1 \neq 1$, on the other hand, implies the partial adjustment mechanism. The corresponding long run solution to equation (1) adjustment mechanism. The corresponding long run solution to equation (1)

$$\delta = \alpha_1/\varphi(1), \theta = \beta/\varphi(1) \tag{2}$$

is invalid but they provide an alternative method, which yields consistent estimates of the parameters and their standard errors. There are three steps that must be followed for the ARDL approach to cointegration. In particular in the first step the existence of a long run relationship between the variables is established by testing for the significance of lagged variables in an error correction mechanism regression. In this paper the regression estimated in this step is defined as:

$$DLSN = \alpha_0 + \sum_{i=0}^{p} cDLSN_{t-1} + \sum_{i=0}^{p} cDLSN_{t-1} + e_i \tag{3}$$

Where DLSN is the first log difference of the stock market index and DLP is the first log difference of the consumer price index (inflation)

In this step, the first lag of the levels of each variable are added to the equation to create the error correction mechanism equation and a variable addition test is performed by computing an f-test on the significance of all the added lagged variables.

$$DLSN = \alpha_0 + \sum_{i=0}^{p} cDLSN_{t-i} + \sum_{i=0}^{p} cDLSN_{t-i} + \delta_1 LSN_{t-1} + \delta_2 LP_{t-1} + e_i \tag{4}$$

The null hypothesis of non-existence of a long-run relationship is defined by

$H_0 : \delta_1 = \delta_2 = 0$ while $H_1 : \delta_1 \neq 0, \delta_2 neq 0$

The relevant statistic is the F-statistic for the joint significance of 1 and 2. The tests are distributed according to a non-standard F-statistic irrespective of whether the explanatory variables are stationary or non-stationary. The critical value bounds for these tests were computed by [Pesaran *et al.*, 1996]. In the case where the F-statistic lies below the lower bound, the long run relationship may be rejected. On the other hand if the F-statistic is higher than the upper bound of the critical value band the null of no long run relationship between the variables can be rejected irrespective of their order integration. In the case that the F-statistic is between the two bounds then a unit root test should be applied. The second step of this approach involves estimating the ARDL form of 1 where the optimal lag length is chosen according to one of the standard criteria such as the Akaike Information Criterion (AIC) or the Schwartz Bayesian Criterion (SBC). Then the restricted version of the equation is solved for the long run solution. The third step involves the estimation of the error correction equation using the differences of the variables and the lagged long run solution and determines the speed of adjustment of employment equilibrium.

3 Data and Empirical Results

Data

For the empirical analysis we use monthly data collected from the OECD data bank and covering the period between 1/1985 and 1/2000. In particular, we use the General Index of the Greek stock market (S) and the Consumer Price Index (P). The inflation rate (DLP) and the stock market returns (DLSN) were calculated as the first differences of the logarithmic price levels of the respective series. We do not expand the data sample beyond 1/2000 since after that date Greece joined the EMU.

Empirical Results

Since the ARDL methodology does not require pre-testing for the integration properties of the individual series used in the empirical analysis, we proceed by applying the bounds testing-ARDL procedure to equation (4). The joint significance of the lagged levels of the variables in (4) was next tested by computing an F-test and comparing it with the appropriate critical value tabulated by [Pesaran *et al.*, 1996]. The findings of the empirical analysis are reported in tables 1,2 and 3 in the appendix. Initially the analysis covered the period 1/985 to 1/2000. The results suggested cointegration with long run causality running from inflation to stock market returns. Nevertheless, the application of CUSUM and CUSUMSQ test, indicated lack of stability of the coefficients for the sample period. Based on the respective graphs, presented in the appendix, as well as the LS ([Lee and Strazicich, 1999a] and [Lee and Strazicich, 1999b]) stationarity test which accounts for possible structural breaks, we split the sample period into two sub-periods (1/1985-5/1992 and 6/1992-1/2000). The date of the break suggested by the above tests coincides with the 1992 Athens Stock Market Crisis. The results of the bounds test are reported in Table 1. For the shake of robustness, we report the f-tests for $p = q = 6$ and 12. The evidence is in favor of the existence of cointegration between the stock market returns and inflation only over the second sub-period. With regard to the whole period, as was mentioned earlier, the evidence is unreliable. The empirical findings from the application of the ARDL cointegration methodology is presented in Table 2 in the appendix. Considering only the results obtained from the examination of the two sub-periods the evidence is as follows. Over the first sub-period there is evidence of a long run relationship running from LP towards LSN. In the second sub-period the results indicate bidirectional long-run causality. Finally, the paper addresses the issue of possible short-run causal relationships by means of Granger causality tests. The results reported in table 3 in the appendix indicate that over the first sub-period there is a causal effect running from returns to inflation while over the second sub-period we found evidence of a causal effect running from inflation to returns. The results regarding the whole period are ignored as it was explained earlier.

4 Concluding Remarks

In this paper we have examined the relationship between inflation and stock market returns in Greece. The causal effects among the considered variables were explored by means of ARDL cointegration and Granger causality tests. The evidence is in favor of a bidirectional negative long-run causal relationship which is consistent with [Fama, 1981] and [Spyrou, 2001]. Besides, we report short run causal effects running from returns to inflation for the period between 1/1985 and 5/1992, while for the period 6/1992 to 1/2000 the direction is from inflation towards returns.

5 Appendix

In this section we present our numerical results.

Period: 1/1985 - 1/2000		
Dependent Variable	lag length (p=q)	F-values
DLSN	6	6.78
	12	7.45
DLP	6	3.28
	12	4.76
Period: 1/1985 - 5/1992		
Dependent Variable	lag length (p=q)	F-values
DLSN	6	2.45
	12	3.26
DLP	6	2.25
	12	2.87
Period: 6/1992 - 1/2000		
Dependent Variable	lag length (p=q)	F-values
DLSN	6	6.65
	12	6.93
DLP	6	7.15
	12	7.85

Table 1. Table 1. Critical values bounds testing-ARDL, for 0.05 significance levels (4.94 - 5.73)

Period: 1/1985 - 1/2000		
LSN	LP	P-value
1	1.8677	0.06
Period: 6/1992 - 1/2000		
LSN	LP	P-value
1	-6.2965	0.00
-0.1	1	0.005

Table 2. Table 2. Long-Run Causality based on ARDL selected model

References

[Anari and Kolari, 2001]A. Anari and J. Kolari. Stock prices and inflation. *Journal of Financial Research*, pages 587–602, 2001.

[Bodie, 1976]Z. Bodie. Common stocks as a hedge against inflation. *Journal of Finance*, pages 459–470, 1976.

[Boudhouch and Richarson, 1993]J. Boudhouch and M. Richarson. Stock returns and inflation: a long-horizon perspective. *American Economic Review*, pages 1346–1355, 1993.

[Fama, 1981]E.F. Fama. Stock returns, real activity, inflation and money. *American Economic Review*, pages 545–565, 1981.

[Firth, 1979]M. Firth. The relationship between stock market returns and rates of inflation. *Journal of Finance*, pages 743–749, 1979.

[Granger, 1969]C.W.I. Granger. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica*, pages 423–438, 1969.

[Gultekin, 1983]N.B. Gultekin. Stock market returns and inflation: evidence from other countries. *Journal of Finance*, pages 49–65, 1983.

[Ioannidis *et al.*, 2004]D. Ioannidis, K. Katrakilidis, and A.E. Lake. Inflation, uncertainty and stock market returns evidence using greek data. *Under Publication*, 2004.

[Kessel, 1956]R.A. Kessel. Inflation caused wealth redistribution: a test of hypothesis. *American Economics Review*, pages 128–141, 1956.

[Lee and Strazicich, 1999a]J. Lee and M. Strazicich. Minimum lm unit root test. 1999a.

[Lee and Strazicich, 1999b]J. Lee and M. Strazicich. Minimum lm unit root test with two structural breaks. 1999b.

[Pearce and Roley, 1988]D.K. Pearce and V.V. Roley. Firm characteristics, unanticipated inflation, and stock returns. pages 965–981, 1988.

[Pesaran and Shin, 1995]M.H. Pesaran and Y. Shin. An autoregressive distributed lag modelling approach to cointegration analysis. *DAE Working Paper, No. 9514*, 1995.

[Pesaran *et al.*, 1996]M.H. Pesaran, Y. Shin, and R.J. Smith. Testing for the existence of a long-run relationship. *DAE Working Paper, No. 9622*, 1996.

[Spyrou, 2001]S. Spyrou. Stock returns and inflation: Evidence from an emerging market. *Applied Economics Letter*, pages 447–450, 2001.

An Application Of The Stochastic Mcshane'S Equations In Financial Modelling

G. Constantin

West Univesity of Timisoara,
300 223 TIMISOARA, Romania
(e-mail: const@math.uvt.ro)

Abstract. We prove a result for the solvability of linear forward-backward stochastic differential equations of McShane type. The motivation for the study is a similar Black-Scholes type model in mathematical finance.

Keywords: forward-backward stochastic differential equations, McShane type integral, Black-Scholes type model.

1 Introduction

One of the most remarkable applications of stocahstic analysis is in mathematical finance. In particular, the Black-Scholes model enjoys great popularity (see, for example, [Musiela and Rutkowski, 1997]). Recently (see [Ma and Yong, 1999]), this model was derived by means of the theory of forward-backward stochastic differential equations of Itô type a setting appropriate for the case in which the filtration of the undelying probability space is given by Brownian motion. It appears that for a filtration induced by a finite variation process with a.s. continous sample paths, the Itô type stochastic integral is no longer appropriate. In this case one can use a McShane type integral (introduced by McShane in [McShane, 1969], [McShane, 1974] and further developed by Srinivasan in [Srinivasan, 1978] and, from a different point of wiev, by Protter in [Protter, 1992]; so a stochastic calculus could be called "unified calculus" since it includes ordinary calculus as a special case and also Itô Calculus) to construct a suitable model. This leads us to forward-backward stochastic differential equations of McShane type. Despite many investigation related to McShane type stochastic differential equations (see [Angulo Ibanez and Gutierrez Jaimez, 1988], [Constantin, 1998], [Ladde and Seikkala, 1986], [McShane, 1974] for theoretical approaches and [Srinivasan, 1978], [Srinivasan, 1984], [Hangii, 1980] for applications of McShane stochastic calculus to problems in physics) a study of forward-backward stochastic differential equations of McShane type has not been undertaken, to the best of our knowledge.

Stochastic calculus appears to be one of the natural tools for the study of models of those phenomena having some non-deterministic elements. For example, in the description of brownian motion the stochastic nature is adequately described by a linear differential equation with a random forcing term which is identified as a white noise process or has a formal derivative of the Wiener process.

However, when the results of the stochastic calculus were applied to other types phenomena, certain difficulties arose in the process of interpretation of stochastic differentials and approximation process. In many models, white noise process is explicitly introduced and the basic physical process in question is visualised as an approximation. Hence it is reasonable to expect some kind of a stability in the sense that the solutions that are obtained by approximating the white noise process should themselves approximate the process in question.

Ito stochastic calculus failed to satisfy this requirement of stability (see [McShane, 1974]). Moreover, in choosing the type of stochastic processes that we shall use us models of the noises we meet a dilema. On the one hand, there is no physical bases for considering an example considering any simple functions $W_j(t)$ exccept those of a rather simple structure. In fact, the noise input $W_j(t)-W_j(s)$ is measured be some sort of indicator and if this is mechanical it cannot move faster than the velocity of light, if it is electrical, it cannot suport more than some limited current or voltage difference without destruction and also some similiraties are in the financial modeling case.

In McShane's Calculus, the standard equations

$$(I) \qquad X^i(t,\omega) = X^i(0,\omega) + \int_0^t f^i(s, X(s,\omega))ds + \sum_{j=1}^r \int_0^t g^i_j(s, X(s,\omega))dW_j(s,\omega)$$

are replaced by what he calls a **canonical extension** (or canonical form or canonical system) of equation (I):

$$(II) \qquad X^i(t,\omega) = X^i(0,\omega) + \int_0^t f^i(s, X(s,\omega))ds +$$

$$\sum_{j=1}^r \int_0^t g^i_j(s, X(s,\omega))dW_j(s,\omega) + \frac{1}{2}\sum_{j,k=1}^r \int_0^t g^i_{j,k}(s, X(s,\omega))dW_j(s,\omega)dW_k(s,\omega)$$

in which

$$g^i_{j,k}(t,x,\omega) = \sum_{m=1}^n [\partial g^i_j(t,x,\omega)/\partial x^m] g^m_k(t,x,\omega)$$

$i = 1,2,...,n;\ j,k = 1,2,...,r;\ t \in [0,a];\ x \in \mathbf{R}^n$.

We are now able to describe the method by which we shall construct stochastic models of physical systems which in the physically realizable case of lipschitzian noises are known to satisfy the integral equation (I).

If $g^i_{j,k}(t,x,\omega)$ are functions defined for $t \in [0,a]$ and $x \in \mathbf{R}^n$ and bounded on bounded sets of (t,x), then the solution $X^i(t,\omega)$ of (I) is also a solution of (II) since the last integral vanishes for lipschitzian noises.

The McShane Calculus is better suited modeling dynamical phenomena described typically by McShane systems where $W_j(t,\omega)$ are noises processes.

McShane stochastic integral systems enjoy the following three important properties:

(i) The property of inclusiveness: the model must apply to systems in which the permitted noises are processes belonging to some family large enough to include processes with sample paths having lipschitzian property, all brownian motion processes, and such modifications as have proved convenient in applications;

(ii) The property of consistency: for lipschitzian noises, the solutions of the equations should coincide with the solutions of the equations that are normally believed to be applicable to physical systems;

(iii) The property of stability: the model must be such that if the noise process $W_j(t,\omega)$ is replaced by another permissible process $W_j^0(t,\omega)$ close to it, then the corresponding solutions $X^i(t,\omega)$, $X_0^i(t,\omega)$ are also close to each other (in the sense that an extreme degree of closeness corresponds to practical imposibility of distinguishing the process by means of available experimental procedures).

In section 2 we pursue the study of the solvability of a class of linear forward-backward stochastic differential equations of McShane type and we point out some drastic differences from the case of Itô type stochastic equations.

The approach developed in Section 2 is applied in Section 3 to a similar Black-Scholes type model in mathematical finance.

2 The main result

Consider the following forward-backward stochastic differential equations on $[0,T]$,

$$\begin{cases} dX(t) = [a(t)X(t) + b(t)]dt + [c(t)X(t) + d(t)]dW(t) \\ dY(t) = [f(t)X(t) + g(t)Y(t) + h(t)Z(t) + k(t)]dt + Z(t)dW(t) \\ X(0) = x_0, \;\; Y(T) = \alpha(X(T)) \end{cases} \tag{1}$$

where $T > 0$, $x_0 \in \mathbf{R}$ and $a, b, c, d, f, g, h, k : [0,T] \to \mathbf{R}$ are continuous functions, while $\alpha : \mathbf{R} \to \mathbf{R}$ is a function of class C^1. In (1), $(X(t), Y(t), Z(t))$ is a triplet of adapted stochastic processes on a complete filtered probability space $(\Omega, \mathcal{F}, \{\mathcal{F}_t\}_{t\in[0,T]}, \mathcal{P})$ such that $\{\mathcal{F}_t\}_{t\in[0,T]}$ is the natural filtration of a given stochastic process $\{W(t)\}_{t\in[0,T]}$, augmented with all $\mathcal{P}$-null sets. Throughout this paper, the process $\{W(t)\}_{t\in[0,T]}$ inducing the filtration is a finite variation process with continuous paths i.e. for almost all $\omega \in \Omega$ the sample path $t \to W(t,\omega)$ is continuous and of finite variation on $[0,T]$ as a particular noise of McShane type. For example, a process satisfying a.s. a Lipschitz condition

$$W(t,\omega) - W(s,\omega)| \leq L|t-s|, \;\; 0 \leq s \leq t \leq T$$

for some constant $L > 0$, is admissible. Let $C_{\mathcal{F}}[0,T]$ be the set of all $\{\mathcal{F}_t\}_{t\in[0,T]}$-progresively measurable continuous processes $X : [0,T] \times \Omega \to \mathbf{R}$ (that is, for almost all $\omega \in \Omega$ the sample paths $t \to X(t,\omega)$ is continuous on $[0,T]$), such that $E \sup\limits_{t\in[0,T]} |X(t)|^2 < \infty$. Observe that the space

$$M_{\mathcal{F}}[0,T] = C_{\mathcal{F}}[0,T] \times C_{\mathcal{F}}[0,T] \times C_{\mathcal{F}}[0,T]$$

is a Banach space under the norm

$$\|(X,Y,Z)\| = \{E \sup_{t\in[0,T]} |X(t)|^2 + E \sup_{t\in[0,T]} |Y(t)|^2 + E \sup_{t\in[0,T]} |Z(t)|^2\}^{\frac{1}{2}}.$$

Given $a,b,c,d,f,g,h,k \in C([0,T],\mathbf{R})$, $\alpha \in C^1(\mathbf{R},\mathbf{R})$, $x_0 \in \mathbf{R}$, and the finite variation continuous process $\{W(t)\}_{t\in[0,T]}$ inducing the filtration on the probability space, a process $(X,Y,Z) \in M_{\mathcal{F}}[0,T]$ is called an *adapted solution* of (1) if the following holds for any $t \in [0,T]$, almost surely:

$$\begin{cases} X(t) = x_0 + \int\limits_0^t [a(s)X(s) + b(s)]ds + \underset{0}{\overset{t}{\imath nt}}[c(s)X(s) + d(s)]dW(s), \\ Y(t) = \alpha(X(T)) - \int\limits_t^T [f(s)X(s) + g(s)Y(s) + h(s)Z(s) + k(s)]ds- \\ \qquad - \int\limits_t^T Z(s)dW(s), \end{cases} \tag{2}$$

where the stochastic integrals are McShane type integrals (see [Protter, 1992] for an approach to this integral close in spirit to the original one by McShane [McShane, 1969]). In Section 3 we will give an example in mathematical finance that motivates the study of (1). Let us now prove the solvability of (1).

Theorem. The system (1) admits an adapted solution $(X,Y,Z) \in M_{\mathcal{F}}[0,T]$.

Proof. To show the existence of a solution we introduce a direct method, similar to the scheme developed in [Ma *et al.*, 1994] for Itô type forward-backward stochastic differential equations. We will prove that the following three-step scheme is realizable:

(A) let $\theta : [0,T] \times \mathbf{R} \to \mathbf{R}$ be the C^1-solution of the following first-order linear partial differential equation

$$\begin{cases} \theta_t + ([a(t) - c(t)h(t)]x + b(t) - d(t)h(t))\theta_x = \\ = g(t)\theta + f(t)x + k(t),\ t \in [0,T],\ x \in \mathbf{R}, \\ \theta(T,x) = \alpha(x),\ \ x \in \mathbf{R}; \end{cases} \tag{3}$$

(B) let $X \in C_{\mathcal{F}}[0,T]$ be the solution of the following forward stochastic differential equation of McShane type:

$$\begin{cases} dX(t) = & [a(t)X(t) + b(t)]dt + \\ & +[c(t)X(t) + d(t)]dW(t), \quad t \in [0,T], \\ X(0) = & x_0; \end{cases} \tag{4}$$

(C) then X together with

$$Y(t) = \theta(t, X(t)), \quad Z(t) = \theta_x(t, X(t))(c(t)X(t) + d(t)), \quad t \in [0,T] \tag{5}$$

is an adapted solution to (1).

The special relations (5) among the components of the adapted solution $(X, Y, Z) \in M_{\mathcal{F}}[0,T]$ to (1) are suggested by the change of variables formula for McShane type stochastic integrals (see [McShane, 1974], p.146): if $X \in C_{\mathcal{F}}[0,T]$ solves (4), then

$$\begin{aligned} d\theta(t, X(t)) = & [\theta_t(t, X(t)) + \theta_x(t, X(t))(a(t)X(t) + b(t))]dt + \\ & +\theta_x(t, X(t))(c(t)X(t) + d(t))dW(t) \end{aligned}$$

and a comparison with the backward stochastic equation (for Y) in (2) confirms that the problem (3) for θ is precisely what is needed for the effectiveness of the solution scheme. Therefore, the existence part is proved if we show that steps (A) and (B) can be performed. Both problems can be explicitely solved. Indeed, the solution of (4) is (see [McShane, 1974], p.129-130)

$$X(t) = [x_0 + \Psi(t)]e^{\Phi(t)}, \quad t \in [0,T], \tag{6}$$

with

$$\Phi(t) = \int_0^t a(s)ds + \int_0^t c(s)dW(s), \quad t \in [0,T],$$

and

$$\Psi = \int_0^t e^{-\phi(s)}b(s)ds + \int_0^t e^{-\Phi(s)}dsdW(s), \quad t \in [0,T].$$

On the other hand, the method of characteristics enables us [John, 1962] to write down the explicit C^1-solution of the Cauchy problem (3). However, taking into account the intricacy of the resulting formula, we refrain from further details- we shall do the full details of the solution in Section 3 for the choice of coefficients in (1) dictated by a model in mathematical finance. The proof of the theorem is completed.

The statement of the theorem leaves open the question of uniqueness. Our solution scheme was constructed in analogy with the four step scheme (see [Ma *et al.*, 1994]) for the Itô type problem (2)- case in which $\{W(t)\}_{t\in[0,T]}$ is

Brownian motion (a process with a.s. continuous sample paths but a.s. the sample paths are of unbounded variation functions [Protter, 1992]) and all stochastic integrals in (2) are of Itô type.

For the Itô type problem (2), uniqueness holds (see [Ma and Yong, 1999], p.82) so that it is not unreasonable to expect uniqueness in the McShane type problem (2) that we are investigating. However, let us note an essential difference in the two solution schemes (the four step scheme from [Ma and Yong, 1999] and our three step scheme) which indicates that the McShane type problem is not a perfect replicate to the Itô type problem. In both problems the forward stochastic differential equation are replaced by a forward stochastic differential equation coupled with a Cauchy problem for a partial differential equation: in the Itô type problem we have a parabolic partial differential equation while in the McShane scheme type problem we have a linear first order partial differential equation. For parabolic partial differential equations, time-reversibility is not to be expected whereas for linear first-order partial differential equations this is not an issue. Here lies an essential difference between the schemes adapted in the Itô type case, respectively in the McShane type case. An example illustrates that the uniqueness for the McShane type case isn't assured.

3 Applications

In this section we analyse a model in mathematical finance that motivates the study of forward-backward stochastic differential equations of McShane type.

Consider a market that contains one bond and one stock. Their prices at time t are denoted by $P(t)$ and $X(t)$, respectively. An investor trades continuously, the wealth of the investor at time t being denoted by $Y(t)$ and the amount of money invested into the stock at time t is denoted by $\pi(t)$, called portfolio, while the rest of the money at time t, $Y(t) - \pi(t)$, is put into the bond. In a stochastic model (model with uncertainly) one assumes that both prices are stochastic processes, defined on some filtered probability space $(\Omega, \mathcal{F}, \{\mathcal{F}_t\}_{t\geq 0}, \mathcal{P})$. The fact that both prices can only be determined by the information up to time t is expressed mathematically by requiring the processes $P(t),\ X(t)$ to be both adapted to the filtration $\{\mathcal{F}_t\}_{t\geq 0}$. We assume that the filtration is generated by a given continuous process $\{W(t)\}_{t\geq 0}$ with sample paths of bounded variation on compact intervals. If the market is assumed to be Markovian, that is, the interes rate $r(t)$ of the bond and the appreciation rate and volatility of the stock $b(t)$, respectively σ, are deterministic (the time-dependence is assumed to be continuous), then the prices are subject to the following system of stochastic differential equations

$$\begin{cases} dP(t) = r(t)P(t)dt, & \text{(bond)} \\ dX(t) = X(t)b(t)dt + \sigma X(t)dW(t), & \text{(stock)} \\ P(0) = \ 1, \ \ X(0) = x_0, \end{cases} \tag{7}$$

where $x_0 > 0$ is a constant. The change of wealth $dY(t)$ follows therefore the dynamics

$$dY(t) = \frac{\pi(t)}{X(t)}dX(t) + \frac{Y(t)-\pi(t)}{P(t)}dP(t). \tag{8}$$

An option with maturity date $T > 0$ is an $\mathcal{F}_T$-measurable random variable $\alpha(X(T))$, where $\alpha : \mathbf{R} \to \mathbf{R}$ is a function of class C^1. Let us rewrite (7)-(8) as

$$\begin{cases} P(t) = & e^{\int\limits_0^t r(s)ds}, \\ X(t) = & x_0 + \int\limits_0^t b(s)X(s)ds + \sigma \overset{t}{\underset{0}{\in}} t\, X(s)dW(s), \\ dY(t) = [\pi(t)b(t) + r(t)(Y(t) - \pi(t))]dt + \sigma\pi(t)dW(t), \end{cases}$$

for $t \in [0,T]$. The interaction between the investor's wealth/strategy and the stock price is described by the following forward-backward stochastic differential equations of McShane type

$$\begin{cases} X(t) = x_0 + \int\limits_0^t b(s)X(s)ds + \sigma \overset{t}{\underset{0}{\in}} t\, X(s)dW(s); t \in [0,T], \\ Y(t) = \alpha(X(T)) - \int\limits_t^T [r(s)Y(s) + (b(s) - r(s))\pi(s)]ds- \\ \qquad -\sigma \int\limits_t^T \pi(s)dW(s),\ t \in [0,T]. \end{cases} \tag{9}$$

The purpose of the investor is to find an adapted solution (X, Y, π) to (9); this amounts to choosing a strategy π allowing the realization of the option $Y(T) = \alpha(X(T))$.

The problem (9) is of type (2) so that we may apply our three step scheme developed in Section 2 to find an explicit solution. Relation (6) ensures that the solution of the equation for X in (9) is precisely

$$X(t) = x_0 e^{\sigma[W(t)-W(0)]+\int\limits_0^t b(s)ds},\quad t \in [0,T]. \tag{10}$$

To find the explicit formula for the wealth $Y(t)$, we have to solve the problem (3), i.e.

$$\begin{cases} \theta_t + r(t)x\theta_x = r(t)\theta,\ t \in [0,T], & x \in \mathbf{R}, \\ \theta(T,x) = \alpha(x) & x \in \mathbf{R}. \end{cases} \tag{11}$$

In accordance to the study pursued in Section 2, we apply the method of characteristics to solve (11). The characteristic curves are given by the system of ordinary differential equations (with parameter s)

$$\begin{cases} \dfrac{dt}{ds} = 1, \\ \dfrac{dx}{ds} = r(t)x, \\ \dfrac{d\theta}{ds} = r(t)\theta, \end{cases} \tag{12}$$

and the Cauchy data corresponds to (at $s = 0$)

$$t = T, \;\; x = \xi, \;\; \theta = \alpha(\xi). \tag{13}$$

The solution to (12)-(13) has the parametric representation

$$t = s + T, \quad x = \xi e^{\int\limits_0^s r(\tau+T)d\tau}, \quad \theta = \alpha(\xi) e^{\int\limits_0^s r(\tau+T)d\tau},$$

as it can be easily verified. Eliminating s, ξ we find for the C^1-solution of the Cauchy problem (11) the representation

$$\theta(t,x) = \alpha(xe^{\int\limits_t^T r(\tau)d\tau})e^{-\int limits_t^T r(\tau)d\tau} \tag{14}$$

since

$$s = t - T, \quad \xi = xe^{\int\limits_t^T r(\tau)d\tau}.$$

We can check directly that (14) solves (11).

As a consequence of our theorem, taking into account relations (5), (10) and (14), we find that a solution of the problem (9) is given by

$$\begin{cases} X(t) = x_0 e^{\sigma[W(t)-W(0)]+\int\limits_0^t b(s)ds}, & t \in [0,T], \\ Y(t) = \theta(t, X(t)), & t \in [0,T], \\ Z(t) = X(t)\theta_x(t, X(t)), & t \in [0,T]. \end{cases}$$

Remark. The model presented above is of Black-Scholes type because if we consider $b(t)$, $r(t)$ to be positive constants and the process $\{W(t)\}_{t\in[0,T]}$ to be a Brownian motion, interpreting (9) as an Itô type problem, we end up with a parabolic problem instead of (11): the Black- Scholes partial differential equation (see [Ma and Yong, 1999], p.227).

References

[Angulo Ibanez and Gutierrez Jaimez, 1988]J.M. Angulo Ibanez and R. Gutierrez Jaimez. On the existence and uniqueness of the soltuion processes in mcshane's stochastic integral equation systems. *Ann.Sci.Univ.Blaise Pascal*, pages 1–9, 1988.

[Constantin, 1998]A. Constantin. On the existence, uniqueness and parametric dependence on the coefficients of the solution processes in mcshane's stochastic integral equations. *Publ.Math.*, pages 11–24, 1998.

[Hangii, 1980]P. Hangii. Langevin description of markovian integro-differential master equations. *Z.Phyzik B.*, pages 271–282, 1980.

[John, 1962]F. John. *Partial Differential Equations.* Springer Verlag, New York, 1962.

[Ladde and Seikkala, 1986]G.S. Ladde and S. Seikkala. Existence, uniqueness and upper estimates for solutions of mcshane type stochastic differential systems. *Stoch.Anal.Appl.*, pages 409–430, 1986.

[Ma and Yong, 1999]J. Ma and J. Yong. *Forward-Backward Stochastic Differential Equations and Their Applications*. Springer Verlag, Berlin, 1999.

[Ma *et al.*, 1994]J. Ma, P. Protter, and J. Yong. Solving forward-backward stochastic differential equations explicitly – a four step scheme. *Prob.Th.Rel.Fields.*, pages 339–359, 1994.

[McShane, 1969]E.J. McShane. Stochastic integral and stochastic functional equations. *SIAM J.Appl.Math.*, pages 287–306, 1969.

[McShane, 1974]E.J. McShane. *Stochastic Calculus and Stochastic Models*. Academic Press, New York, 1974.

[Musiela and Rutkowski, 1997]M. Musiela and M. Rutkowski. *Martingale Methods in Financial Modelling*. Springer Verlag, Berlin, 1997.

[Protter, 1992]P. Protter. *Stochastic Integration and Differential Equations*. Springer Verlag, Berlin, 1992.

[Srinivasan, 1978]S.K. Srinivasan. Stochastic integrals, solid mechanics archives. *Solid Mechanics Archives*, pages 325–379, 1978.

[Srinivasan, 1984]S.K. Srinivasan. Stochastic calculus and models of physical phenomena. *J.Math.Phys.Sci.*, pages 163–168, 1984.

Brownian Laplace motion and its use in financial modelling

William J. Reed

Department of Mathematics and Statistics
University of Victoria,
Victoria, B.C., Canada, V8W 3P4
(e-mail: reed@math.uvic.ca)

Abstract. A new Lévy motion with both continuous (Brownian) and discontinuous (Laplace motion) components is introduced. The increments of the process follow a *generalized normal Laplace* (GNL) distribution, which exhibits positive kurtosis and can be either symmetrical or skewed. The degree of kurtosis in the increments increases as the length of the increment decreases. This and other properties of Brownian-Laplace motion refelect those of observed time series of logarithmic stock-price returns and thus render it a good model for fitting to financial data and for the calculation of the theoretical value of financial derivatives. A formuala for the value of European call options based on Brownian-Laplace motion is given.

Keywords: Laplace motion, generalized normal-Laplace (GNL) distribution, Black-Scholes.

1 Introduction.

The Black-Scholes theory of option pricing was originally based on the assumption that asset prices follow geometric Brownian motion (GBM). For such a process the logarithmic returns ($\log(P_{t+1}/P_t)$ on the price P_t are independent identically distributed (iid) normal random variables. However it has been recognized for some time now that the logarithmic returns do not behave quite like this, particulary over short intervals. Empirical distributions of the logarithmic returns in high-frequency data usually exhibit excess kurtosis with more probability mass near the origin and in the tails and less in the flanks than would occur for normally distributed data. Furthermore the degree of excess kurtosis is known to increase as the sampling interval decreases (see *e.g.* [Rydberg, 2000]). In addition skewness can sometimes be present. To accomodate for these facts new models for price movement based on Lévy motion have been developed (see *e.g.* [Schoutens, 2003]). For any infinitely divisible distribution a Lévy process can be contructed whose increments follow the given distribution. Thus in modelling financial data one needs to find an infinitely divisible distribution which fits well to observed logarithmic returns. A number of such distributions have been suggested including the gamma, inverse Gaussian, Laplace (or variance gamma), Meixner and generalized hyperbolic distributions (see [Schoutens, 2003] for details and references).

In this paper a new infinitely divisible distribution – the *generalized normal Laplace* (or GNL) distribution – which exhibits the properties seen in observed logarithmic returns, is introduced. This distribution arises as the sum of independent normal and generalized Laplace [Kotz *et al.*, 2001] random variables[1]1. A Lévy process based on the generalized Laplace (variance-gamma) distribution alone has no Brownian component, only linear deterministic and pure jump components *i.e.* its Lévy-Khintchine triplet is of the form $(\gamma, 0, \nu(dx))$ (see [Schoutens, 2003]). The new distribution of this paper in effect adds a Brownian component to this motion, leading to what will be called *Brownian-Laplace motion*[2]2.

In the following section the generalized normal Laplace (GNL) distribution is defined and some properties given. Brownian-Laplace motion is then defined as a Lévy process whose increments follow the GNL distribution. In Sec. 3 a pricing formula is developed for European call options on a stock whose logarithmic price follows Brownian-Laplace motion.

2 The generalized normal Laplace (GNL) distribution.

The *generalized normal Laplace* (GNL) distribution is defined as that of a random variable Y with characteristic function

$$\phi(s) = \left[\frac{\alpha\beta\exp(\mu is - \sigma^2 s^2/2)}{(\alpha - is)(\beta + is)}\right]^\rho \tag{1}$$

where α, β, ρ and σ are positive parameters and $-\infty < \mu < \infty$. We shall write

$$Y\mathrm{simGNL}(\mu, \sigma^2, \alpha, \beta, \rho)$$

to indicate that the random variable Y follows such a distribution.
Since the characteristic function (1) can be written

$$\exp(\rho\mu is - \rho\sigma^2 s^2/2)\left[\frac{\alpha}{\alpha - is}\right]^\rho \left[\frac{\beta}{\beta + is}\right]^\rho$$

it follows that Y can be represented as

$$Y \overset{d}{=} \rho\mu + \sigma\sqrt{\rho}Z + \frac{1}{\alpha}G_1 - \frac{1}{\beta}G_2 \tag{2}$$

[1] 1. The generalized asymmmetric Laplace distribution is better known as the variance-gamma distribution in the finance literature. It is also known as the Bessel K-function distribution (see [Kotz *et al.*, 2001], for a discussion of the terminology and history of this distribution).

[2] 2. An alternative name, which invokes two of the greatest names in the history of mathematics, would be *Gaussian-Laplace motion*

where Z, G_1 and G_2 are independent with Zsim N(0,1) and G_1, G_2 gamma random variables with scale parameter 1 and shape parameter ρ, *i.e.* with probability density function (pdf)

$$g(x) = \frac{1}{\Gamma(\rho)} x^{\rho-1} e^{-x}.$$

This representation provides a straightforward way to generate pseudo-random deviates following a GNL distribution. Note from (1) it is easily established that the GNL is infinitely divisible. In fact the n-fold convolution of a GNL random variable also follows a GNL distribution.

The mean and variance of the GNL$(\mu, \sigma^2, \alpha, \beta, \rho)$ distribution are

$$\mathrm{E}(Y) = \rho\left(\mu + \frac{1}{\alpha} - \frac{1}{\beta}\right); \qquad \mathrm{var}(Y) = \rho\left(\sigma^2 + \frac{1}{\alpha^2} + \frac{1}{\beta^2}\right)$$

while the higher order cumulants are (for $r > 2$)

$$\kappa_r = \rho(r-1)!\left(\frac{1}{\alpha^r} + (-1)^r \frac{1}{\beta^r}\right).$$

The parameters μ and σ^2 influence the central location and spread of the distribution, while α and β affect the lengths of the tails. *Ceteris paribus* decreasing α (or β) puts more weight into the upper (or lower) tail. The tail behaviour of the GNL distribution can be determined from the nature of the poles of its characteristic (or moment generating) function (see *e.g.* [Doetsch, 1970]). In the tails the generalized Laplace component of the GNL dominates - precisely $f(y)$sim$c_1 y^{\rho-1} e^{-\alpha y}$ $(y \to \infty)$ and $f(y)$sim$c_2(-y)^{\rho-1} e^{\beta y}$ $(y \to -\infty)$, (where c_1 and c_2 are constants). Thus for $\rho < 1$, both tails are fatter than exponential; for $\rho = 1$ they are exactly exponential and for $\rho > 1$ they are less fat than exponential.

The parameter ρ affects all moments. However the coefficients of skewness $(\gamma_1 = \kappa_3/\kappa_2^{3/2})$ and of kurtosis $(\gamma_2 = \kappa_4/\kappa_2^2)$ both decrease with increasing ρ (and converge to zero as $\rho \to \infty$) with the shape of the distribution becoming more normal with increasing ρ, (exemplifying the central limit effect since the sum of n iid GNL$(\mu, \sigma^2, \alpha, \beta, \rho)$ random variables has a GNL$(\mu, \sigma^2, \alpha, \beta, n\rho)$ distribution).

When $\alpha = \beta$ the distribution is symmetric. In the limiting case $\alpha = \beta = \infty$ the GNL reduces to a normal distribution.

3 A Lévy process based on the GNL distribution - Brownian-Lapace motion.

Consider now a Lévy process $\{X_t\}_{t \geq 0}$, say for which the increments $X_{t+\tau} - X_\tau$ have characteristic function $(\phi(s))^t$ where ϕ is the characteristic function (1) of the GNL$(\mu, \sigma^2, \alpha, \beta, \rho)$ distribution (such a construction is always possible for an infinitely divisible distribution - see [Schoutens, 2003]. It is not difficult to show that the Lévy-Khintchine triplet for this process is $(\rho\mu, \rho\sigma^2, \Lambda)$ where Λ is the Lévy measure of asymmetric Laplace motion (see Kotz *et al.*, 2001, p.196). Laplace motion has an infinite number of jumps in any finite time interval (a pure jump process). The extension considered here adds a continuous Brownian component to Laplace motion leading to the name *Brownian-Laplace motion.*

The increments $X_{t+\tau} - X_\tau$ of this process will follow a GNL$(\mu, \sigma^2, \alpha, \beta, \rho t)$ distribution and will have fatter tails than the normal – indeed fatter than exponential for $\rho t < 1$. However as t increases the kurtosis of the distribution drops, and approaches zero as $t \to \infty$. Exactly this sort of behaviour has been observed in various studies on high-frequency financial data (*e.g.* [Rydberg, 2000]) - very little kurtosis in the distribution of logarithmic returns over long intervals but increasingly fat tails as the reporting interval is shortened. Thus Brownian-Laplace motion seems to provide a good model for the movement of logarithmic prices.

3.1 Option pricing for assets with logarithmic prices following Brownian-Laplace motion.

We consider an asset whose price S_t is given by

$$S_t = S_0 \exp(X_t)$$

where $\{X_t\}_{t \geq 0}$ is a Brownian-Laplace motion with $X_0 = 0$ and parameters $\mu, \sigma^2, \alpha, \beta, \rho$. We wish to determine the risk-neutral valuation of a European call option on the asset with strike price K at time T and risk-free interest rate r.

It can be shown using the Esscher equivalent martingale measure (see *e.g.* [Schoutens, 2003]) that the option value can be expressed in a form similar to that of the Black-Scholes formula. Precisely

$$OV = S_0 \int_\gamma^\infty d^{*T}_{GNL}(x; \theta + 1)dx - e^{-rT} K \int_\gamma^\infty d^{*T}_{GNL}(x; \theta)dx \tag{3}$$

where $\gamma = \log(K/S_0)$ and

$$d^{*T}_{GNL}(x; \theta) = \frac{e^{\theta x} d^{*T}_{GNL}(x)}{\int_{-\infty}^{\infty} e^{\theta y} d^{*T}_{GNL}(y)dy} \tag{4}$$

is the pdf of X_T under the risk-neutral measure. Here d^{*T}_{GNL} is the pdf of the T-fold convolution of the generalized normal-Laplace, GNL$(\mu, \sigma^2, \alpha, \beta, \rho)$, distribution and θ is the unique solution to the following equation involving the moment generating function (mgf) $M(s) = \phi(-is)$

$$\log M(\theta + 1) - \log M(\theta) = r. \tag{5}$$

The T-fold convolution of GNL$(\mu, \sigma^2, \alpha, \beta, \rho)$ is GNL$(\mu, \sigma^2, \alpha, \beta, \rho T)$ and so its moment generating function is (from (1))

$$M(s) = \left[\frac{\alpha\beta \exp(\mu s + \sigma^2 s^2/2)}{(\alpha - s)(\beta + s)} \right]^{\rho T}.$$

This provides the denominator of the expression (4) for the risk-neutral pdf.

Now let

$$I_\theta = \int_\gamma^\infty d_{GNL}^{*T}(x;\theta)dx = \frac{1}{[M(\theta)]^T} \int_\gamma^\infty e^{\theta x} d_{GNL}^{*T}(x) \tag{6}$$

so that

$$OV = S_0 I_{\theta+1} - e^{-rT} K I_\theta.$$

Thus to evaluate the option value we need only evaluate the integral in (6). This can be done using the representation (2) of a GNL random variable as the sum of normal and positive and negative gamma components. Precisely the integral can be written

$$\int_0^\infty g(u;\alpha) \int_0^\infty g(v;\beta) \int_\gamma^\infty e^{\theta x} \frac{1}{\sigma\sqrt{\rho T}} \phi\left(\frac{x - u + v - \mu\rho T}{\sigma\sqrt{\rho T}} \right) dx dv du \tag{7}$$

where

$$g(x;a) = \frac{a^{\rho T}}{\Gamma(\rho T)} x^{\rho T - 1} e^{-ax}$$

is the pdf of a gamma random variable with scale parameter a and shape parameter ρT; and ϕ is the pdf of a standard normal deviate. After completing the square in x and evaluating the x integral in terms of Φ^c, the complementary cdf of a standard normal, (6) can be expressed

$$I_\theta = \int_0^\infty g(u;\alpha - \theta) \int_0^\infty g(v;\beta + \theta) \Phi^c \left(\frac{\gamma - u + v - \mu\rho T - \theta\sigma^2\rho T}{\sigma\sqrt{\rho T}} \right) dv du. \tag{8}$$

For given parameter values the double integral (8) can be evaluated numerically quite quickly and thence the option value computed. For an example see [Reed, 2005].

References

[Doetsch, 1970]Doetsch, G. (1970). *Introduction to the Theory and Application of the Laplace Transformation.* Springer-Verlag, New York, Heidelberg, Berlin.

[Kotz *et al.*, 2001]Kotz, S., Kozubowski, T. J. and Podgórski, K. (2001). *The Laplace Distribution and Generalizations.* Birkhäuser, Boston.

[Reed, 2005]Reed, W. J. The normal-Laplace distribution and its relatives. To appear in a special volume honouring Professor Barry Arnold on the occasion of his sixtieth birthday. Birkhäuser, Boston.

[Rydberg, 2000]Rydberg, T. H. (2000). Realistic statistical modelling of financial data. *Inter. Stat. Rev.*, **68**, 233–258.

[Schoutens, 2003]Schoutens, W. (2003). *Lévy Processes in Finance*, J. Wiley and Sons, Chichester.

Managing Value-at-Risk for a bond using bond put options

Griselda Deelstra[1], Ahmed Ezzine[1], Dries Heyman[2], and Michèle Vanmaele[3]

[1] Department of Mathematics, ISRO and ECARES,
Université Libre de Bruxelles,
CP 210, 1050 Brussels, Belgium
(e-mail: griselda.deelstra@ulb.ac.be, ahmed.ezzine@ulb.ac.be)

[2] Department of Financial Economics,
Ghent University,
Wilsonplein 5D, 9000 Gent, Belgium
(e-mail: Dries.Heyman@UGent.be)

[3] Department of Applied Mathematics and Computer Science,
Ghent University,
Krijgslaan 281, building S9, 9000 Gent, Belgium
(e-mail: Michele.Vanmaele@UGent.be)

Abstract. This paper studies a strategy that minimizes the Value-at-Risk (VaR) of a position in a zero coupon bond by buying a percentage of a put option, subject to a fixed budget available for hedging. We elaborate a formula for determining the optimal strike price for this put option in case of a Vasicek stochastic interest rate model. We demonstrate the relevance of searching the optimal strike price, since moving away from the optimum implies a loss, either due to an increased VaR or due to an increased hedging expenditure. In this way, we extend the results of [Ahn *et al.*, 1999], who minimize VaR for a position in a share. In addition, we look at the alternative risk measure Tail Value-at-Risk.

Keywords: Value-at-Risk, bond hedging, Vasicek interest rate model.

1 Introduction

Many financial institutions and non-financial firms nowadays publicly report Value-at-Risk (VaR), a risk measure for potential losses. Internal uses of VaR and other sophisticated risk measures are on the rise in many financial institutions, where, for example, a bank risk committee may set VaR limits, both amounts and probabilities, for trading operations and fund management. At the industrial level, supervisors use VaR as a standard summary of market risk exposure. An advantage of the VaR measure, following from extreme value theory, is that it can be computed without full knowledge of the return distribution. Semi-parametric or fully non-parametric estimation methods are available for downside risk estimation. Furthermore, at a sufficiently low confidence level the VaR measure explicitly focuses risk managers and regulators attention on infrequent but potentially catastrophic extreme losses.

Value-at-Risk (VaR) has become the standard criterion for assessing risk in the financial industry. Given the widespread use of VaR, it becomes increasingly important to study the effects of options on the VaR-based risk management.

The starting point of our analysis is the classical hedging example, where an institution has an exposure to the price risk of an underlying asset. This may be currency exchange rates in the case of a multinational corporation, oil prices in the case of an energy provider, gold prices in the case of a mining company, etc. The corporation chooses VaR as its measure of market risk. Faced with the unhedged VaR of the position, we assume that the institution chooses to use options and in particular put options to hedge a long position in the underlying.

[Ahn *et al.*, 1999] consider the problem of hedging the Value-at-Risk of a position in a single share by investing a fixed amount C in a put option.
The principal purpose of our study is to extend these results to the situation of a bond. We consider the well-known continuous-time stochastic interest rate model of [Vasicek, 1977] to investigate the optimal speculative and hedging strategy based on this framework by minimizing the Value-at-Risk of the bond, subject to the fixed amount C which is spent on put options. In addition, we consider an alternative risk measure Tail Value-at-Risk (TVaR), for which we solve the minimization problem and obtain the optimal hedging policy. In further versions, we will elaborate this part more deeply.

The discussion is divided as follows: Section 2 presents the general risk management model, introduces the Vasicek model and considers hedging with bond put options. Afterwards, Section 3 discusses the optimal hedging policy for VaR, considers the closely related risk measure TVaR and introduces comparative statics. Section 4 consists of a numerical illustration. Finally, Section 5 summarizes the paper, concludes and introduces further research possibilities.

2 The mathematical framework

Consider a portfolio with value W_t at time t. The Value-at-Risk of this portfolio is defined as the $(1-\alpha)$-quantile of the loss distribution depending on a time interval with length T. Common time periods that are taken into consideration are $T = 1$, 10, 20 days. A formal definition for the $\mathrm{VaR}_{\alpha,T}$ is

$$\Pr(W_0 - W_T^d \geq \mathrm{VaR}_{\alpha,T}) = \alpha,$$

with W_T^d the value of the portfolio at time T, discounted back until time zero by means of a zero coupon with maturity T.
In other words $\mathrm{VaR}_{\alpha,T}$ is the loss of the worst case scenario on the investment at a $1-\alpha$ confidence level during the period $[0,T]$. It is possible to define the $\mathrm{VaR}_{\alpha,T}$ in a more general way

$$\mathrm{VaR}_{\alpha,T} = \inf\left\{Y \mid \Pr(W_0 - W_T^d \geq Y) < \alpha\right\}.$$

In this study, we focus on the hedging problem of a zero-coupon bond. Therefore, we need to define a process that describes the evolution of the instantaneous interest rate, and enables us to value the zero-coupon bond. As term structure model, we consider the Vasicek model, which is a typical example of an affine term structure model.

2.1 The Vasicek model

[Vasicek, 1977] assumes that the instantaneous interest rate follows a mean reverting process also known as an Ornstein-Uhlenbeck process:

$$dr(t) = \kappa(\theta - r(t))dt + \sigma dZ(t) \tag{1}$$

for a standard Brownian motion $Z(t)$ under the risk-neutral measure Q, and with constants κ, θ and σ. The parameter κ controls the mean-reversion speed, θ is the long-term average level of the spot interest rate around which $r(t)$ moves, and σ is the volatility measure. The reason of the Vasicek model's popularity is its analytical and mathematical tractability. An often cited critique is that applying the model sometimes results in a negative interest rate.

It can be shown that the expectation and variance of the stochastic variable $r(t)$ are:

$$E\left[r(t)\right] = m = \theta + (r(0) - \theta)e^{-\kappa t} \tag{2}$$

$$\mathrm{Var}\left[r(t)\right] = s^2 = \frac{\sigma^2}{2\kappa}(1 - e^{-2\kappa t}). \tag{3}$$

Based on these results, Vasicek develops an analytical expression for the price of a zero-coupon bond with maturity date S

$$Y(t,S) = \exp[A(t,S) - B(t,S)r(t)], \tag{4}$$

where

$$B(t,S) = \frac{1 - e^{-\kappa(S-t)}}{\kappa}, \tag{5}$$

$$A(t,S) = (B(t,S) - (S-t))(\theta - \frac{\sigma^2}{2\kappa^2}) - \frac{\sigma^2}{4\kappa}B(t,S)^2. \tag{6}$$

Since $A(t,S)$ and $B(t,S)$ are independent of $r(t)$, the distribution of a bond price at any given time must be lognormal with parameters Π and Σ^2:

$$\Pi(t,S) = A(t,S) - B(t,S)m, \qquad \Sigma(t,S)^2 = B(t,S)^2 s^2, \tag{7}$$

with m and s^2 given by (2) and (3).

From the formulae (4)-(7), we can see that for $S \geq T$ the present value of the loss of the (unhedged) portfolio can be expressed as function of z

$$L_0 = W_0 - W_T^d = Y(0,S) - Y(0,T)e^{\Pi(T,S)+\Sigma(T,S)z} = f(z) \tag{8}$$

where f is a strictly decreasing function and z is a stochastic variable with a standard normal distribution. Therefore, the $\mathrm{VaR}_{\alpha,T}$ of such a portfolio is determined by the formula

$$\mathrm{VaR}_{\alpha,T} = f(c(\alpha)), \tag{9}$$

where $c(\alpha)$ is the cut off point for the standard normal distribution at a certain percent level i.e. $\Pr(z \leq c(\alpha)) = \alpha$.

Since the distribution of the unhedged position in the zero-coupon bond is lognormal in the Vasicek model, from the formulae (8)-(9) we observe that the Value-at-Risk measure for the zero-coupon bond can be expressed as

$$\mathrm{VaR}_{\alpha,T} = Y(0,S) - Y(0,T)e^{\theta_B(\alpha)},$$

where

$$\theta_B(\alpha) = \Pi(T,S) + \Sigma(T,S)c(\alpha) \tag{10}$$

and $c(\cdot)$ is the percentile of the standard normal distribution.

2.2 Put options and hedging

We recall from [Ahn *et al.*, 1999] the classical hedging example, where an institution has an exposure to the price risk of an underlying asset S_T. The hedged future value of this portfolio at time T is given by

$$H_T = \max(hX + (1-h)S_T, S_T), \tag{11}$$

where $0 \leq h \leq 1$, represents the hedge ratio, that is, the percentage of put option P used in the hedge and X is the strike price of the option.

In our setup, the underlying security is a bond and the hedging tool is a bond put option, the price of which will be worked out hereafter.
We recall that the price of a European call option with the zero-coupon bond which matures at time S as the underlying security and with strike price X and exercise date T (with $T \leq S$) is at date t given by:

$$C(t,T,X) = Y(t,S)\Phi(d_1) - XY(t,T)\Phi(d_2), \tag{12}$$

where

$$d_1 = \frac{1}{\sigma_p}\log(\frac{Y(t,S)}{XY(t,T)}) + \frac{\sigma_p}{2}, \qquad d_2 = d_1 - \sigma_p,$$

$$\sigma_p = \frac{\sigma}{\kappa}(1 - e^{-\kappa(S-T)})\sqrt{\frac{1-e^{-2\kappa(T-t)}}{2\kappa}},$$

and $\Phi(z)$ is the cumulative distribution function of a standard normal random variable. The Put-Call parity model is designed to determine the value of a put option from a corresponding call option and provides in this case the following European put option price corresponding to (12):

$$P(t,T,X) = -Y(t,S)\Phi(-d_1) + XY(t,T)\Phi(-d_2). \tag{13}$$

3 The bond hedging problem

3.1 VaR minimization

Analogously to [Ahn *et al.*, 1999], we assume that we have one bond and we use only a percentage of a put option on the bond to hedge. We will find the optimal strike price which minimizes VaR for a given hedging cost.
Indeed, let us assume that the institution has an exposure to a bond, $Y(0,S)$, which matures at time S, and that the company has decided to hedge the bond value by using a percentage of one put option $P(0,T,X)$ with strike price X and exercise date T (with $T \leq S$). Then we can look at the future value of the hedged portfolio (which is composed of the bond Y and the put option $P(0,T,X)$) at time T as a function, analogously to (11), of the form

$$H_T = \max(hX + (1-h)Y(T,S), Y(T,S)).$$

If the put option finishes in-the-money (a case which is of interest to us), then the discounted value of the future value of the portfolio is

$$H_T^d = ((1-h)Y(T,S) + hX)Y(0,T).$$

Taking into account the cost of setting up our hedged portfolio, which is given by the sum of the bond price $Y(0,S)$ and the cost C of the position in the put option, we get for the present value of the loss

$$L_0 = Y(0,S) + C - ((1-h)Y(T,S) + hX)Y(0,T),$$

and this under the assumption that the put option finishes in-the-money.
We recall that $Y(T,S)$ has a lognormal distribution with parameters $\varPi$ and $\varSigma^2$, given by (7). Therefore the loss function equals

$$Y(0,S) + C - ((1-h)e^{\varPi(T,S)+\varSigma(T,S)z} + hX)Y(0,T),$$

where z again denotes a stochastic variable with a standard normal distribution. The Value-at-Risk at an α percent level of a position $H = \{Y, h, P\}$ consisting of a bond Y and h put options P (which are assumed to be in-the-money) with a strike price X and an expiry date T is equal to

$$\mathrm{VaR}_{\alpha,T}(L_0) = Y(0,S) + C - ((1-h)e^{\theta_B(\alpha)} + hX)Y(0,T), \qquad (14)$$

where we recall that $\theta_B(\alpha) = \varPi + \varSigma c(\alpha)$ and $c(\alpha)$ is the percentile of the standard normal distribution.

Similar to the Ahn et al. problem, we would like to minimize the risk of the future value of the hedged bond H_T, given a maximum hedging expenditure C. More precisely,

$$\min_X Y(0,S) + C - ((1-h)e^{\theta_B(\alpha)} + hX)Y(0,T)$$

subject to the restrictions $C = hP(0,T,X)$ and $h \in (0,1)$.
Solving this constrained optimization problem, we find that the optimal strike price X^* satisfies the following equation

$$P(0,T,X) - (X^* - e^{\theta_B(\alpha)})\frac{\partial P(0,T,X)}{\partial X} = 0.$$

or equivalently, when taking (13) into account,

$$e^{\theta_B(\alpha)} = \frac{Y(0,S)\Phi(-d_1)}{Y(0,T)\Phi(-d_2)}. \tag{15}$$

We note that the optimal strike price is independent of the hedging cost.

3.2 Tail VaR minimization

In this section, we introduce the concept of Tail Value-at-Risk, TVaR, also known as mean excess loss, mean shortfall or Conditional VaR. We further demonstrate the ease of extending our analysis to this alternative risk measure.

A drawback of the traditional Value-at-Risk measure is that it does not care about the tail behaviour of the losses. In other words, by focusing on the VaR at, let's say a 5% level, we ignore the potential severity of the losses below that 5% threshold. In other words, we have no information on how bad things can become in a real stress situation. Therefore, the important question of 'how bad is bad' is left unanswered. TVaR is trying to capture this problem by considering the possible losses, once the VaR threshold is crossed.

Formally,

$$\mathrm{TVaR}_{\alpha,T} = \frac{1}{\alpha}\int_{1-\alpha}^{1} \mathrm{VaR}_{1-\beta,T}\, d\beta.$$

This formula boils down to taking the arithmetic average of the quantiles of our loss, from $1-\alpha$ to 1 on, where we recall that $\mathrm{VaR}_{\alpha,T}$ stands for the quantile at the level $1-\alpha$.

If the cumulative distribution function of the loss is continuous, which is the case in our problem, TVaR is equal to the Conditional Tail Expectation (CTE) which for the loss L_0 is calculated as:

$$\mathrm{CTE}_{\alpha,T}(L_0) = E[L_0 \mid L_0 > \mathrm{VaR}_{\alpha,T}(L_0)].$$

A closely related risk measure concerns Expected Shortfall (ESF). It is defined as:

$$\mathrm{ESF}(L_0) = E\left[(L_0 - \mathrm{VaR}_{\alpha,T}(L_0))_+\right].$$

In order to determine $\mathrm{TVaR}_{\alpha,T}(L_0)$, we can make use of the following equality:

$$\begin{aligned}\mathrm{TVaR}_{\alpha,T}(L_0) &= \mathrm{VaR}_{\alpha,T}(L_0) + \frac{1}{\alpha}\mathrm{ESF}(L_0)\\ &= \mathrm{VaR}_{\alpha,T}(L_0) + \frac{1}{\alpha}E\left[(L_0 - \mathrm{VaR}_{\alpha,T}(L_0))_+\right].\end{aligned}$$

This formula already makes clear that $\mathrm{TVaR}_{\alpha,T}(L_0)$ will always be larger than $\mathrm{VaR}_{\alpha,T}(L_0)$.
In our case, the loss has a lognormal distribution, because of the lognormality of our bond prices. This allows us to write the ESF as

$$\mathrm{ESF}(L_0) = (1-h)Y(0,T)e^{\Pi(T,S)}\Big[\alpha e^{\Sigma(T,S)c(\alpha)} - e^{\frac{1}{2}\Sigma^2(T,S)}\Phi(c(\alpha) - \Sigma(T,S))\Big].$$

This reduces our $\mathrm{TVaR}_{\alpha,T}(L_0)$ to:

$$\begin{aligned}\mathrm{TVaR}_{\alpha,T}(L_0) &= Y(0,S) + C - hXY(0,T)\\ &\quad -\frac{1}{\alpha}(1-h)e^{\Pi(T,S)+\frac{1}{2}\Sigma^2(T,S)}\Phi(c(\alpha) - \Sigma(T,S))Y(0,T).\end{aligned}$$

We again seek to minimize this TVaR, in order to minimize potential losses. The procedure for minimizing this TVaR is analogue to the VaR minimization procedure. The resulting optimal strike price can thus be determined from the formula below:

$$\frac{1}{\alpha}e^{\Pi(T,S)+\frac{1}{2}\Sigma^2(T,S)}\Phi(c(\alpha) - \Sigma(T,S)) = \frac{Y(0,S)\Phi(-d_1)}{Y(0,T)\Phi(-d_2)}.$$

3.3 Comparative statics

We examine how changes in the parameters of the Vasicek model influence the optimal strike price, by means of the derivatives of the optimal strike price with respect to these parameters.

For both $\mathrm{VaR}_{\alpha,T}$ and $\mathrm{TVaR}_{\alpha,T}$, the optimal strike price is implicitly defined by

$$F(X,\beta) = \mathrm{FAC}\cdot Y(0,T)\Phi(-d_2) - Y(0,S)\Phi(-d_1) = 0,$$

with β the vector including the Vasicek parameters, that is θ, κ and the volatility σ, see Section 2.1, and with FAC representing $e^{\theta_B(\alpha)}$ in the case of $\mathrm{VaR}_{\alpha,T}$ and $\frac{1}{\alpha}e^{\Pi(T,S)+\frac{1}{2}\Sigma^2(T,S)}\Phi(c(\alpha) - \Sigma(T,S))$ in the case of $\mathrm{TVaR}_{\alpha,T}$.
Taking into account the implicit function theorem, we obtain the required derivatives as follows:

$$\frac{\partial F}{\partial X}dX + \frac{\partial F}{\partial \beta}d\beta = 0 \Longleftrightarrow \frac{dX}{d\beta} = -\frac{\frac{\partial F}{\partial \beta}}{\frac{\partial F}{\partial X}}. \tag{16}$$

The denominator of (16) is equal for the different derivatives, and is given by

$$\frac{\partial F}{\partial X} = \frac{\mathrm{FAC} \cdot Y(0,T)\varphi(d_2) - Y(0,S)\varphi(d_1)}{X\sigma_p}, \tag{17}$$

with φ being the density function of a standard normal random variable, while the numerator of (16) can be obtained by applying the following formula,

$$\begin{aligned} \frac{\partial F}{\partial \beta} = {} & \frac{\partial \mathrm{FAC}}{\partial \beta} Y(0,T)\Phi(-d_2) + \mathrm{FAC} \cdot \frac{\partial Y(0,T)}{\partial \beta}\Phi(-d_2) \\ & - \mathrm{FAC} \cdot Y(0,T)\varphi(d_2)\frac{\partial d_2}{\partial \beta} - \frac{\partial Y(0,S)}{\partial \beta}\Phi(-d_1) + Y(0,S)\varphi(d_1)\frac{\partial d_1}{\partial \beta}. \end{aligned} \tag{18}$$

These derivatives are rather involved and do not lead to a straightforward interpretation of their sign and magnitude. Therefore, we will describe the derivatives in the next paragraph using a numerical illustration.

Further relevant derivatives are $\frac{dX}{dS}$ and $\frac{dX}{dT}$ to study the response of the optimal strike price to a change in the maturity of both the underlying bond and the maturity of the bond option used to hedge the exposure. They follow from formulae (16)-(18), after having replaced β by S and T respectively, and taking into account the simplification due to the fact that $Y(0,T)$ is independent of S, and $Y(0,S)$ is independent of T. Again, we leave the interpretation of these derivatives to the next section.

A last derivative of interest is the one with respect to α, formally $\frac{dX}{d\alpha}$:

$$\frac{dX}{d\alpha} = -\frac{1}{\frac{\partial F}{\partial X}} \cdot \frac{\partial \mathrm{FAC}}{\partial \alpha} Y(0,T)\Phi(-d_2),$$

where $\frac{\partial \mathrm{FAC}}{\partial \alpha}$ is respectively given by

$$\frac{e^{\theta_B(\alpha)}\Sigma(T,S)}{\varphi(c(\alpha))} \qquad \text{(VaR)}$$

$$\frac{e^{\Pi(T,S)+\frac{1}{2}\Sigma^2(T,S)}}{\alpha^2}\left[\frac{\alpha\varphi(c(\alpha)-\Sigma(T,S))}{\varphi(c(\alpha))} - \Phi(c(\alpha)-\Sigma(T,S))\right] \quad \text{(TVaR)}.$$

4 Numerical results

We illustrate the usefulness of the above results for the VaR case (TVaR case is ongoing research). In order to provide a credible numerical illustration, we take the parameter estimates for the Vasicek model from [Chan *et al.*, 1992], who compare a variety of continuous-time models of the short term interest rate with respect to their ability to fit the U.S. Treasury bill yield. This results in the following parameter values: $\sigma = 0.02$, $\theta = 0.0866$, $\kappa = 0.1779$, $r(0) = 0.06715$. Next, we should consider the budget the financial institution is willing to spend on the hedging. Standardising the nominal value of the bond at issuance to 1, we start with a hedging budget of 0.05, so $C = 0.05$. We also assume the bank is considering the VaR at the five percent level, meaning that $\alpha = 5\%$.

We considered two situations, one in which the bank wishes to hedge a bond with a maturity of one year ($S = 1$), and one for a bond with a maturity of ten year ($S = 10$).
We observe that our strategy is successful in decreasing the risk, while, since we use options, still providing us with upward potential. In the one year bond case, the mean reduction in VaR (calculated as the difference between the VaR of the hedged position and the VaR of the unhedged position, divided by VaR of the unhedged position) over the holding period amounts to 6.25%. The maximum reduction is 26.23%, whereas the lowest reduction is 3.25%. In the ten year bond case, the mean VaR reduction over the holding period is 5.36%. The maximum reduction that can be achieved amounts to 26.15%. The minimum reduction is 2.59%.

As already mentioned above, we are also interested in the effect of changes in the parameter estimates of the Vasicek model on the optimal strike price. We examine these effects using the first example, in which the bond matures in one year. An increase in one of these parameters always leads to a lower optimal strike price. The influence of a 1% increase in κ only marginally effects the strike price. Changes in θ also have a moderate impact on the optimal strike. The most influential parameter of the Vasicek model undoubtedly is the volatility. Whereas for κ and θ the impact constantly decreases as the holding period comes closer to the maturity of the bond, we find a non-monotonic relationship between the derivative (with respect to the volatility) and the difference between the holding period T and the maturity S of the bond.

Increasing the maturity of the bond decreases the strike price, while increasing the holding period (meaning that the holding period moves closer to the maturity of the bond) increases the strike price. Reducing the certainty with which a bank wishes to know the value it can lose, or in other words, increasing α leads to a increased strike price. This increase again depends on the holding period in a non monotonic way.

5 Conclusion

In this paper, we studied the optimal risk control for one bond using a percentage of a put option by means of Value-at-Risk and Tail Value-at-Risk, widespread concepts in the financial world. The interest model we use for valuation, is the Vasicek model. The optimal strategy corresponds to buying a put option with optimal strike price in order to have a minimal VaR or TVaR given a fixed hedging cost. We did not obtain an explicit result, but numerical methods can be easily implemented to solve for the optimal strategy. For the VaR case, we demonstrate the relevance of searching for this optimal strike price, since moving away from this optimum implies a loss, either because of an increased VaR, or an increased hedging expenditure. For TVaR, the numerical illustration is part of ongoing research.

Further analysis is oriented in a number of directions. First of all, we plan to examine the implications of assuming a different interest rate model e.g. Hull-White. We will further turn to a deeper study of the effects in the optimal hedging policy of using either VaR or TVaR.

Acknowledgements

D. Heyman and M. Vanmaele would like to acknowledge the financial support by the BOF-project 001104599 of the Ghent University. A. Ezzine would like to thank prof. dr. J. Janssen for valuable discussions at the start of this research. The authors also would like to thank an anonymous referee for valuable suggestions.

References

[Ahn *et al.*, 1999]D.H. Ahn, J. Boudoukh, M. Richardson, and R. Whitelaw. Optimal risk management using options. *Journal of Finance*, pages 359–375, 1999.

[Chan *et al.*, 1992]K.C. Chan, G.A. Karolyi, F.A. Longstaff, and A.B. Sanders. An empirical comparison of alternative models of the short-term interest rate. *Journal of Finance*, pages 1209–1227, 1992.

[Vasicek, 1977]O. Vasicek. An equilibrium characterization of the term structure. *Journal of Financial Economics*, pages 177–188, 1977.

Valuing Credit Default Swap in a non-Homogeneous Semi-Markovian Rating-Based Model.

Guglielmo D'Amico[1], Giovanna De Medici[1], and Jacques Janssen[2]

[1] Dipartimento di Matematica per le Decisioni Economiche, Finanziarie ed Assicurative Universitá "La Sapienza",
via del Castro Laurenziano, 9, 00161 Roma, Italy.
(e-mail: guglielmo.damico@uniroma1.it, giovanna.demedici@uniroma1.it)

[2] CESIAF, EURIA
Université de Bretagne Occidentale
6 avenue le Gorgeu, CS 93837,
29238 BREST Cedex 3, France
(e-mail: cesiaf@belgacom.net)

Abstract. In this paper we use a discrete time non-homogeneous semi-Markov model for the rating evolution of the credit quality of a firm C and we determine the credit default swap spread for a contract between two parties, A and B that, respectively, sell and buy a protection about the failure of the firm C. We work in both the case of deterministic and stochastic recovery rate. We highlight the link between credit risk and reliability theory too.

Keywords: backward recurrence times processes, random recovery rate, reliability.

1 Introduction

The credit default swap (CDS) is a derivative that can be seen as default insurance on loans and bonds. These contracts are instruments that provide insurance against a particular company (that we will call company "C") defaulting on its debt. In this paper we present an evaluation procedure of credit default swap in a rating based model. We assume that the rating credit quality evolution of the company "C" that issue the bond follows a discrete time non-homogeneous semi-Markov process, so to consider the reference default risk we use the non-homogeneous semi-Markov reliability credit risk model [D'Amico *et al.*, 2004a]. In this way, how it is showed in [D'Amico *et al.*, 2004a], we solve all the non-markovianity problems highlighted by some empirical works in this area such [Carty and Fons, 1994] and [Nickell *et al.*, 2002].

We fix the credit default swap spread $U^*(s)$ imposing a fair game condition on the wealth balance equation for the swap contract. We compute $U^*(s)$ first considering a fixed recovery rate ρ and successively extending the computation to the case of a random recovery rate. Considering the non-homogeneity of the process we give the same definition of stochastic recovery rate as in [D'Amico *et al.*, 2004b] linking the random recovery rate in general on the last n states visited by the process first of the random default time τ_s.

In both the cases of deterministic and stochastic recovery rate, we express the price and the value of the swap as a function of the C's reliability.

2 The discrete time non-homogeneous semi-Markov reliability credit risk model

First of all we give some basic results on the theory of discrete time non-homogeneous semi-Markov processes. Let (Ω, F, P) be a probability space and let E be a finite state space. On our probability space we define two stochastic processes: $X_n : \Omega \longrightarrow E,\ T_n : \Omega \longrightarrow \mathbb{N}$.

X_n represents the state occupied at the n-th transition and T_n is the time of the n-th transition. The process (X, T) is a non-homogeneous Markov Renewal Process if $\forall i, j \in E$ and $\forall t \in \mathbb{N}$ the following condition holds:

$$P[X_{n+1}=j, T_{n+1}\le t|\sigma(X_h, T_h), X_n=i, T_n=s, 0\le h\le n]=$$

$$P[X_{n+1}=j, T_{n+1}\le t|X_n=i, T_n=s]\equiv Q_{ij}(s,t). \tag{1}$$

The transition matrix $P(s)$ of the non-homogeneous embedded Markov chain X_n is obtained as $p_{ij}(s) = \lim_{t\to\infty} Q_{ij}(s,t)\ \ \forall i, j \in E$.

We introduce also the following probabilities:

$$q_{ij}(s,t) = P[X_{n+1} = j, T_{n+1} = t|X_n = i, T_n = s], \tag{2}$$

$$H_i(t) = P[T_{n+1} \le t|X_n = i, T_n = s], \tag{3}$$

Let $N(t) = sup\{n : T_n \le t\}\ \forall t \in \mathbb{N}$; we define the non-homogeneous discrete time semi-Markov process $Z = (Z(t), t \in \mathbb{N})$ as $Z(t) = X_{N(t)}$, that represents, for each waiting time, the state occupied by the process.

We define, $\forall i, j \in E$, and $(s,t) \in \mathbb{N} \times \mathbb{N}$, the semi-Markov's transition probabilities as $\phi_{ij}(s,t) = P[Z(t) = j|Z(s) = i]$ satisfying the following system of equations:

$$\phi_{ij}(s,t) = \delta_{ij}(1 - H_i(s,t)) + \sum_{k\in E}\sum_{\tau=1}^{t} q_{ik}(s,\tau)\phi_{kj}(\tau,t). \tag{4}$$

At this time we explain briefly the non-homogeneous semi-Markov reliability credit risk model, see [D'Amico *et al.*, 2004a] to study in depth.

Let the state space E indicate the different rating classes that give a reliability degree of a firm bond. We partition this state space in two subset: $D = \{N + 1\}$ and $Up = \{1, 2, ..., N\}$, that we call respectively "Down" (default) and "Up" states. We assume that the set D is absorbing. The most important variable to compute is the reliability $R(s, \cdot)$ of the firm that is defined $\forall t \ge s$ as $R(s,t) = P[Z(u) \in Up,\ \forall u \in \{s, s + 1, ..., t\}]$. The reliability function $R_i(s,t)$ conditional on the starting state i at time s, is given by $R_i(s,t)=\sum_{j\in Up}\phi_{ij}(s,t)$, then solving the system of equations (4) (see [Blasi *et al.*, 2003]) and summing on the "Up" states we obtain the conditional reliability. Obviously $R(s,t)=\sum_{i\in Up}\sum_{j\in Up}\beta_i(s)\phi_{ij}(s,t)$ where $\beta(s)=(\beta_i(s))_{i\in E}$ denotes the random starting distribution at time s. In our model the reliability is equal to the availability, that give us the probability that the system is "Up" at the generic time t, because the only one defaulting state is absorbing.

3 The price and the value of the swap: the fixed recovery rate case

In this section we consider a CDS contract starting at time s with maturity T. We denote with $\tau_s=\inf\{t>s: Z(t)\in D\}$ and with v the deterministic discount factor. We write the wealth balance equation (w.b.e.) for the seller B of the protection about a failure of C that is given by:

$$\varDelta W|_s^T=\sum_{i=s+1}^{T\wedge\tau_s}U(s)\cdot v^{i-s}-(100-Y(T\wedge\tau_s))\cdot v^{(T\wedge\tau_s)-s}. \tag{5}$$

The term $(\sum_{i=s+1}^{T\wedge\tau_s}U(s)\cdot v^{i-s})$ is the random discounted amount of money that B will obtain writing the CDS contract and $(100-Y(T\wedge\tau_s))\cdot v^{(T\wedge\tau_s)-s}$ is the potential loss in case of a C's default.

We assume that $Y(T\wedge\tau_s)=100\cdot\rho\cdot 1_{\{s<\tau_s\leq T\}}+100\cdot 1_{\{\tau_s>T\}}$ where $\rho\in[0,1]$ is the deterministic recovery rate. This choice implies that the potential loss will be zero if there is no default up to time T whereas if a default occurs first of T the potential loss becomes a real loss equal to $100(1-\rho)$ discounted from default time to starting time s. Then the w.b.e. becomes:

$$\varDelta W|_s^T=\sum_{i=1}^{T\wedge\tau_s}U(s)\cdot v^{i-s}-(100[1-\rho])\cdot v^{(T\wedge\tau_s)-s}1_{\{s<\tau_s\leq T\}}.$$

Fixing the credit default swap spread $U(s)$ imposing a fair game condition so that the expectation of the w.b.e. is zero, we get in:

$$U^*(s)=\frac{(1-v)[100\times(1-\rho)]E[v^{(T\wedge\tau_s)-s}1_{\{s<\tau\leq T\}}]}{(v)\times[1-E[v^{(T\wedge\tau_s)-s}]]}. \tag{6}$$

Now having

$$E[v^{(T\wedge\tau_s)-s}]=\sum_{h=s+1}^{T}v^{h-s}\{R(s,h-1)-R(s,h)\}+v^{T-s}R(s,T) \tag{7}$$

$$E[v^{(T\wedge\tau_s)-s}1_{\{\tau_s\leq T\}}]=\sum_{h=s+1}^{T}v^{h-s}\{R(s,h-1)-R(s,h)\} \tag{8}$$

substituting in equation (6) we obtain:

$$U^*(s) = \frac{(1-v)[100 \times (1-\rho)][\sum_{h=s+1}^{T} v^{h-s}\{R(s,h-1) - R(s,h)\}]}{(v) \times [1 - \sum_{h=s+1}^{T} v^{h-s}\{R(s,h-1) - R(s,h)\} - v^{T-s}R(s,T)]}. \tag{9}$$

Now we turn our attention to the valuation procedure. The value of the swap at time t (conditional on no default first of time t) is given by the difference between the expected present value (at time t) of the future inflows minus the expected present value (at time t) of the future outflows. Let $V(s,t)$ the value of the swap and let $I(s,t) = \sum_{h=t+1}^{T\wedge\tau_s} U^*(s)v^{h-t}$ and $O(s,t) = 100(1-\rho)v^{(T\wedge\tau_s)-t}1_{\{s<\tau_s\leq T\}}$ be, respectively, the future inflows and the future outflows then by definition we have

$$V(s,t) = E[I(s,t) - O(s,t)|\tau_s > t] = E[I(s,t)|\tau_s > t] - E[O(s,t)|\tau_s > t]. \tag{10}$$

We obtain:

$$E[I(s,t)|\tau > t] = U^*(s)\Big\{\sum_{m=t+1}^{T} \frac{R(s,m-1) - R(s,m)}{R(s,t)}(\sum_{h=t+1}^{m} v^{h-t}) + \sum_{h=t+1}^{T} v^{h-t}\frac{R(s,T)}{R(s,t)}\Big\}. \tag{11}$$

$$E[O(s,t)|\tau_s > t] = 100(1-\rho)\Big\{\sum_{h=t+1}^{T} v^{h-t}\frac{R(s,h-1) - R(s,h)}{R(s,t)}\Big\}. \tag{12}$$

substituting in formula (10) we get the value of the swap at time t as a function of the reliability of the firm.

4 The price and the value of the swap: the random recovery rate case

In this section we extend our model considering a stochastic recovery rate ρ. [Berthault *et al.*, 2001] noted that the higher is the rating the lower is the loss in case of default. From this empirical evidence [Millossovich, 2002] linked the recovery rate to the last credit rating evaluation of the company first of the default time τ_s in a markovian time homogeneous environment. That extension was carried out enlarging the state space, considering multiple default classes, one for each possible recovery rate. [D'Amico *et al.*, 2004b] proposed a new way to allows for stochastic recovery rate that depends on the last (possibly n-last) rating evaluation, obtained first of the default time, without enlarging the state space E.

In this paper we use the same definition given in [D'Amico *et al.*, 2004b] being careful on the non-homogeneity of the rating process, so we define the one period stochastic recovery rate at time τ_s, "$\rho_1(\tau_s)$" in the following way:

$$\rho_1(\tau_s) = \begin{cases} r_j \text{ if } s < \tau_s \leq T \text{ and } Z(\tau_s - 1) = j,\ \forall j \neq D \\ 1 \text{ if } \tau_s > T > s \end{cases} \tag{13}$$

We proceed to compute the credit default swap spread $U^*(s)$ starting from equation (6) and imposing a fair game condition such that the expectation of the wealth balance equation is zero. In this case we get:

$$U_1^*(s) = \frac{(100[E[v^{(T\wedge\tau_s)-s}1_{\{s<\tau_s\leq T\}}] - E[\rho_1(\tau_s)v^{(T\wedge\tau_s)-s}1_{\{s<\tau_s\leq T\}}]]) \times (1-v)}{v \times (1 - E[v^{(T\wedge\tau_s)-s}])}. \tag{14}$$

The unique new component to evaluate is $E[\rho_1(\tau_s)v^{(T\wedge\tau_s)-s}1_{\{s<\tau_s\leq T\}}]$.

But $E[\rho_1(\tau_s)v^{T\wedge\tau_s}1_{\{s<\tau_s\leq T\}}] = E[E[\rho_1(\tau_s)v^{(T\wedge\tau_s)-s}1_{\{s<\tau_s\leq T\}}|\tau_s]]$ where

$$E[\rho_1(\tau_s)v^{(T\wedge\tau_s)-s}1_{\{s<\tau_s\leq T\}}|\tau_s] = \sum_{h=s+1}^{\infty} v^{(T\wedge h)-s}\rho_1(h)P[\tau_s = h]1_{\{h\leq T\}} =$$

$$\sum_{h=s+1}^{T} v^{h-s}\rho_1(h)P[\tau_s = h] = \sum_{h=s+1}^{T} v^{h-s}\rho_1(h)\{R(s,h-1) - R(s,h)\} \tag{15}$$

consequently $E[\rho_1(\tau_s)v^{T\wedge\tau_s-s}1_{\{s<\tau_s\leq T\}}] =$

$$\sum_{h=s+1}^{T} v^{h-s} \sum_{j\in Up} r_j P[Z(h-1) = j|Z(h) = D]\{R(s,h-1) - R(s,h)\} \tag{16}$$

To compute $P[Z(h-1) = j|Z(h) = D]$ we have to introduce the non-homogeneous discrete backward recurrence time process $B(t)$ defined as:

$$B(t) = \begin{cases} t + T_0 & \text{if } t < T_1 \\ t - T_{N(t)} & \text{if } t \geq T_1 \end{cases} \tag{17}$$

We know that the stochastic process $(Z(t), B(t))$ with values in $E \times \mathbb{N}$ is a markovian process and $\forall h \in \{1, 2, ...T\}$ and $j \in E$ conditioning on all possible values for $B(h-1)$ and from Bayes formula we have that

$$P[Z(h-1) = j|Z(h) = D] =$$

$$\frac{\sum_{l=0}^{h-1-s} P[Z(h)=D|Z(h-1)=j, B(h-1)=l]P[Z(h-1)=j, B(h-1)=l]}{\sum_{k\in Up}\sum_{l=0}^{h-1-s} P[Z(h)=D|Z(h-1)=k, B(h-1)=l]P[Z(h-1)=k, B(h-1)=l]} =$$

$$\frac{\sum_{i\in E}\beta_i(s)\sum_{l=0}^{h-1-s} L_{ij}(s,h-1,l)\varDelta_{jD}(h-1,l,h)}{\sum_{i\in E}\beta_i(s)\sum_{k\in Up}\sum_{l=0}^{h-1-s} L_{ik}(s,h-1,l)\varDelta_{kD}(h-1,l,h)} \tag{18}$$

where $\varDelta_{ij}(h,l,t) = P[Z(t) = j|Z(h) = i, B(h) = l]$ and

$L_{ij}(s,h,l) = P[Z(h)=j, B(h)=l|Z(s)=i, B(s)=0] = P_{(i,s)}[Z(h)=j, B(h)=l]$.

These probabilities can be computed from the knowledge of the semi-Markov kernel **Q** in fact we have, see [D'Amico *et al.*, 2004c], that

$$\Delta_{ij}(h,l,t)=\frac{\delta_{ij}(1-H_i(h-l,t))}{(1-H_i(h-l,h))}+\frac{1}{(1-H_i(h-l,h))}\sum_{k\in E}\sum_{m=h+1}^{t}q_{ik}(h-l,m)\phi_{kj}(m,t), \tag{19}$$

and $L_{ij}(s,h,l)$ satisfies the following system of equations:

$$L_{ij}(s,h,l)=1_{\{l=h-s\}}\delta_{ij}[1-H_i(s,h)]+\sum_{k\in E}\sum_{m=s+1}^{h-l}q_{ik}(s,m)L_{kj}(m,h,l), \tag{20}$$

Applying these results we get:

$$E[\rho_1(\tau_s)v^{T\wedge\tau_s-s}1_{\{s<\tau_s\le T\}}]=\sum_{h=s+1}^{T}v^{h-s}\{R(s,h-1)-R(s,h)\}\times$$

$$\times\sum_{j\in Up}r_j\frac{\sum_{i\in E}\beta_i(s)\sum_{l=0}^{h-1-s}L_{ij}(s,h-1,l)\frac{q_{jD}(h-1-l,h)}{(1-H_j(h-1-l,h-1))}}{\sum_{i\in E}\beta_i(s)\sum_{k\in Up}\sum_{l=0}^{h-1-s}L_{ik}(s,h-1,l)\frac{q_{kD}(h-1-l,h)}{(1-H_k(h-1-l,h-1))}} \tag{21}$$

finally putting (21) in equation (14) we obtain the credit default swap spread:

$$U_1^*(s)=\frac{100(1-v)\Big\{\sum_{h=s+1}^{T}v^{h-s}\{R(s,h-1)-R(s,h)\}\times}{v[1-\sum_{h=s+1}^{T}v^{h-s}\{R(s,h-1)-R(s,h)\}-v^{T-s}R(s,T)]}\times$$

$$[1-\sum_{j\in Up}r_j\frac{\sum_{i\in E}\beta_i(s)\sum_{l=0}^{h-1-s}L_{ij}(s,h-1,l)\frac{q_{jD}(h-1-l,h)}{(1-H_j(h-1-l,h-1))}}{\sum_{i\in E}\beta_i(s)\sum_{k\in Up}\sum_{l=0}^{h-1-s}L_{ik}(s,h-1,l)\frac{q_{kD}(h-1-l,h)}{(1-H_k(h-1-l,h-1))}}]\Big\} \tag{22}$$

Note that we can assume a dependence of the recovery rate on the last n states visited by the process first of default time τ_s. We define the n-period stochastic recovery rate as

$$\rho_n(\tau_s)=\begin{cases}1 & \text{if } s<T<\tau_s\\ \sum_{i=1}^{n}\alpha_{in}^{\tau_s}\rho_1(\tau_s-i+1) & \text{if } n+s\le\tau_s\le T\\ \rho_{\tau_s}(\tau_s) & \text{if } \tau_s<n+s\end{cases} \tag{23}$$

where $\rho_1(\tau_s-i+1)=r_j$ if $Z(\tau_s-i)=j$ and $Z(\tau_s)=D$, whereas $\alpha_{in}^{\tau_s}$ denote the proportion of the n period recovery rate with default time τ_s that depends on the one period recovery rate at time τ_s-i+1.

In such case to obtain the credit default swap spread we substitute the one period recovery rate in the equation (14) with the n-period one obtaining:

$$U_n^*(s)=\frac{(100[E[v^{T\wedge\tau_s-s}1_{\{s<\tau_s\le T\}}]-E[\rho_n(\tau_s)v^{T\wedge\tau_s-s}1_{\{s<\tau_s\le T\}}]])\times(1-v)}{v\times(1-E[v^{T\wedge\tau_s-s}])}. \tag{24}$$

If we choose $\underline{\alpha}$ such that $\alpha_{1n}^{\tau_s}{=}1$, $\alpha_{in}^{\tau_s}{=}0$ $\forall i \neq 1$ we obtain $U_n^*(s){=}U_1^*(s)$. All we need is to compute the unique new component $E[\rho_n(\tau_s)v^{T\wedge\tau_s-s}1_{\{s<\tau_s\leq T\}}]$.

$$E[\rho_n(\tau_s)v^{T\wedge\tau_s-s}1_{\{s<\tau_s\leq T\}}] = E[E[\rho_n(\tau_s)v^{T\wedge\tau_s-s}1_{\{s<\tau_s\leq T\}}|\tau_s]] \tag{25}$$

so we start computing the conditional expectation.

$$E[\rho_n(\tau_s)v^{(T\wedge\tau_s)-s}1_{\{s<\tau_s\leq T\}}|\tau_s] = \sum_{h=s+1}^{\infty}\rho_n(h)v^{(T\wedge h)-s}P[\tau_s = h]1_{\{h\leq T\}} =$$

$$\sum_{h=s+1}^{n+s-1}\rho_n(h)v^{h-s}P[\tau_s = h] + \sum_{h=n+s}^{T}\rho_n(h)v^{h-s}P[\tau_s = h] \tag{26}$$

now we apply the definition (23) and we get in

$$= \sum_{h=s+1}^{n+s-1}\rho_h(h)v^{h-s}P[\tau_s = h] + \sum_{h=n+s}^{T}\sum_{i=1}^{n}\alpha_{in}^{h}\rho_1(h-i+1)v^{h-s}P[\tau_s = h]$$

consequently $E[\rho_n(\tau_s)v^{(T\wedge\tau_s)-s}1_{\{s<\tau_s\leq T\}}] =$

$$\sum_{h=s+1}^{n+s-1}E[\rho_h(h)]v^{h-s}P[\tau_s = h] + \sum_{h=n+s}^{T}\sum_{i=1}^{n}\alpha_{in}^{h}E[\rho_1(h-i+1)]v^{h-s}P[\tau_s = h] \tag{27}$$

Now $E[\rho_1(h-i+1)] = \sum_{j\in Up} r_j P[Z(h-i) = j|Z(h) = D]$ and

$$E[\rho_h(h)] = \sum_{i=1}^{h}\alpha_{ih}^{h}E[\rho_1(h-i+1)] = \sum_{i=1}^{h}\alpha_{ih}^{h}\sum_{j\in Up} r_j P[Z(h-i) = j|Z(h) = D] \tag{28}$$

finally we obtain $E[\rho_n(\tau_s)v^{(T\wedge\tau_s)-s}1_{\{s<\tau_s\leq T\}}] =$

$$\sum_{h=s+1}^{n+s-1}\sum_{i=1}^{h}\alpha_{ih}^{h}\sum_{j\in Up} r_j P[Z(h-i) = j|Z(h) = D]v^{h-s}P[\tau_s = h] +$$

$$+ \sum_{h=s+n}^{T}\sum_{i=1}^{n}\alpha_{in}^{h}\sum_{j\in Up} r_j P[Z(h-i) = j|Z(h) = D]v^{h-s}P[\tau_s = h] \tag{29}$$

The probabilities $P[Z(h-i) = j|Z(h) = D]$ can be evaluated by the Bayes formula, in fact $\forall h, i \in \mathbb{N}$ such that $h - i \geq s$

$$P[Z(h-i) = j|Z(h) = D] =$$

$$\frac{\sum_{l=0}^{h-i-s} P[Z(h){=}D|Z(h{-}i){=}j, B(h{-}i){=}l]P[Z(h{-}i){=}j, B(h{-}i){=}l]}{\sum_{k\in Up}\sum_{l=0}^{h-i-s}P[Z(h){=}D|Z(h{-}i){=}k, B(h{-}i){=}l]P[Z(h{-}i){=}k, B(h{-}i){=}l]} =$$

$$\frac{\sum_{i\in E}\beta_i(s)\sum_{l=0}^{h-i-s}L_{ij}(s,h-i,l)\Delta_{jD}(h-i,l,h)}{\sum_{i\in E}\beta_i(s)\sum_{k\in Up}\sum_{l=0}^{h-i-s}L_{ik}(s,h-i,l)\Delta_{kD}(h-i,l,h)}. \qquad (30)$$

At this point we substitute (30) in (29) and the obtained (29) in (27). Finally we insert (27) together with (7) in (24) and we obtain $U_n^*(s)$.

We conclude noting that the evaluation procedure in case of random recovery rate doesn't present problems, in fact we have only to change the outflow's definition that in this case is

$$O(s,t) = 100(1-\rho_n(\tau_s))v^{(T\wedge\tau_s)-t}1_{\{t<\tau_s\leq T\}}$$

which expectation can be evaluated using computations similar as those used to determine the credit default swap spread corresponding to a random recovery rate.

References

[Berthault *et al.*, 2001]Berthault, A. and Gupton, G. and Hamilton, D.T. (2001) “Default and recovery rates of corporate bond issuers: 2000.”. *Moody's Investors Service special comment, february.*

[Blasi *et al.*, 2003]Blasi, A. and Janssen, J. and Manca, R. (2003) “Numerical treatment of homogeneous and non-homogeneous semi-Markov reliability models”. *Communications in Statistics, Theory and Method.* v. 33, 697-714 2003.

[Carty and Fons, 1994]Carty, L. and Fons, J. (1994) “Measuring changes in corporate credit quality”. *The Journal of Fixed Income.* June, 66-78 1994.

[D'Amico *et al.*, 2004a]D'Amico, G. and Janssen, J. and Manca, R.(2004a) “Downward credit risk problem and non-homogeneous semi-markov reliability transition models”. *Proceedings of the 8-th International Congress on Insurance: Mathematics and Economics.* 2004.

[D'Amico *et al.*, 2004b]D'Amico, G. and Janssen, J. and Manca, R. (2004b) “Credit Default Swap: a semi-Markov approach”. *Proceedings of the 8-th International Congress on Insurance: Mathematics and Economics.* 2004.

[D'Amico *et al.*, 2004c]D'Amico, G. and Janssen, J. and Manca, R.(2004c) “Transient Analysis of non-Homogeneous Recurrence Times Processes and Application in Option Pricing”. *To be submitted.*2004.

[Millossovich, 2002]Millossovich, P. (2002) “An Extension of the Jarrow-Lando-Turnbull Model to Random Recovery Rate”. *Working Paper Dipartimento di Matematica Applicata, Universitá di Trieste.* 2002.

[Nickell *et al.*, 2002]Nickell, P. and Perraudin, W. and Varotto, S. (2002). “Stability of rating transitions”. *Journal of Banking and Finance,* v.24,203-227.

Credit risk migration semi-Markov models: a reliability approach.

Guglielmo D'Amico[1], Jacques Janssen[2], and Raimondo Manca[1]

[1] Universitá di Roma La Sapienza
via del Castro Laurenziano, 9
00161 Roma Italy
(e-mail: guglielmo.damico@uniroma1.it, raimondo.manca@uniroma1.it)

[2] CESIAF, EURIA
Université de Bretagne Occidentale
6 avenue le Gorgeu, CS 93837,
29238 BREST Cedex 3, France
(e-mail: cesiaf@belgacom.net)

Abstract. Credit risk problem is one of the most important financial topics in this period because of the Basel II rules. In 1997 a seminal paper Jarrow Lando and Turnbull showed that this problem could be approached by means of a Markov chain tool. Subsequently in many papers it was shown that the Markov approach can give some problems, more precisely: In some previous papers the authors showed how it is possible by means of a reliability semi-Markov approach to solve the three problems. In this paper will be summarized the results obtained by the authors to give a complete overview of the proposed approach.

Keywords: Credit risk, semi-Markov, reliability.

1 Introduction

Homogeneous semi-Markov processes were defined in the fifties in [Levy, 1954]. Non-homogeneous semi-Markov processes were defined in [Iosifescu Manu, 1972]. A detailed theoretical analysis of semi-Markov processes was given in [Howard, 1971]. The importance of the Engineering applications of this kind of processes is highlighted in this book. As specified in [Howard, 1971] and more recently in [Limnios and Oprisan, 2000] book, one of the most important applications of semi-Markov processes is in reliability of mechanical systems. Putting the hypothesis that the next transition depends only on the last one (the future depends only on the present) the problem can be faced by means of Markov processes. In discrete time Markov chain environment the time transition is given. But in the reality, the transition between two states in a mechanical system usually happens after a random duration. This is the reason why the semi-Markov environment fits better than the Markov one in reliability problems. Another relevant phenomenon in the time evolution of a system can be the system age. The introduction of non-homogeneity gives the possibility to take into account this problem. All the highlighted aspects can be faced using non-homogeneous semi-Markov models. In the paper [Blasi *et al.*, 2003] how it is possible to apply non-homogeneous semi-Markov processes in reliability problems is described. Credit risk problem is one of the most important problems that are faced in the financial literature. Fundamentally it consists in computing the default probability of a firm that do a debt. The literature on this topic is very wide, but the interested lector can refer to the [Duffie and Singleton, 2003] book. Big interest in this field is given to the firms that issue bonds. For the credit risk evaluation there are international organisations, Fitch, Moodys and Standard & Poors, that give different ranks to the examined firms. At each firm is given a "rating" that is a vote to the "reliability" on the capacity to reimburse the debt. The rating level changes in the time and one way to follow the time evolution of ratings is by means of Markov processes [Jarrow *et al.*, 1997]. In this environment Markov models are called "migration models". Other papers, see for example [Nickell *et al.*, 2000], followed this approach working mainly on the generation of transition matrix. In some papers the problem of the unfitting of Markov process in credit risk environment was outlined, see [Carty and Fons, 1994], [Nickell *et al.*, 2000]. The problems of non-markovianity that are highlighted mainly are the following:

i - the duration inside a state. The probability to change rating depends on the time that a firm remains in the same rating [Carty and Fons, 1994];

ii - the time dependence of the rating evaluation (aging). This means that in general the rating evaluation depends on the time in which is done, see [Nickell *et al.*, 2000] The rating evaluation done at time t generally is different from the one done at time s, if $s \neq t$;

iii - the dependence of the new rating on the previous ones, not only on the last evaluated, [Carty and Fons, 1994], [Nickell *et al.*, 2000].

The first problem can be well solved by means of semi-Markov processes (SMP). In fact in SMP the transition probabilities are function of the waiting time spent in a state of the system. The second problem can be faced in a general approach by means of a non-homogeneous environment. The third effect exists in the downward cases but not in the upward ratings. More precisely if a firm got a lower rating then has a higher probability that the next rating will be lower than the preceding one. The first two are automatically solved applying the non-homogeneous semi-Markov environment. The third problem is solved increasing the number of states to differentiate the case in which the system arrives in a state from a lower or a higher rating evaluation. In a previous article [D'Amico *et al.*, 2003] presented a model based on the homogeneous semi-Markov processes (HSMP) in a reliability environment. The duration problem was fully solved for the first time, at authors knowing, in that paper. The other two credit risk problems were not faced. A second paper [D'Amico *et al.*, 2004a] presenting a non-homogeneous semi-Markov process (NHSMP) model takes into accounts the duration and the aging problem. In a third paper [D'Amico *et al.*, 2004b] also the third problem was solved. The non-homogeneous semi-Markov reliability model, presented together the homogeneous one in [Blasi *et al.*, 2003], will be applied, to solve the credit risk problem.

This paper will present a summary of the three papers and will expose the approach that was made to solve the Markov migration problems. The next part will present a short description of NHSMP. After this the reliability non-homogeneous semi-Markov model will be shown. In the successive paragraph the relation between the reliability model and the credit risk problem will be described. The model enlarges the number of states in this way the downward problem can be solved.

2 Non-homogeneous semi-Markov processes

In this part the NHSMP will be described; we follow the notation given in [Janssen and Manca, 2005]. First the stochastic process is defined. In SMP environment two random variables (r.v.) run together. J_n $n \in \mathbb{N}$ with state space $I = \{1, 2, \ldots, m\}$ represents the state at the n-th transition. T_n $n \in \mathbb{N}$ with state space equal to $\mathbb{R}^+$ represents the time of the n-th transition,

$$J_n : \Omega \to I \;\; T_n : \Omega \to \mathbb{R}^+.$$

We suppose that the process (J_n, T_n) is a non-homogeneous markovian renewal process. The kernel $\mathbf{Q} = [Q_{ij}(s,t)]$ associated to the process is defined in the following way:

$$Q_{ij}(s,t) = \mathrm{P}[J_{n+1} = j, T_{n+1} \leq t | J_n = j, T_n = s]$$

and it results:

$$p_{ij}(s) = \lim_{t \to \infty} Q_{ij}(s,t), \;\; i, j \in I, s, t \in \mathbb{R}^+, s \leq t$$

where $\mathbf{P}(s) = [p_{ij}(s)]$ is the transition matrix of the embedded non-homogeneous Markov chain in the process. Furthermore it is necessary to introduce the probability that process will leave the state i from the time s up to the time t:

$$S_i(s,t) = \mathrm{P}[T_{n+1} \leq t | J_n = j, T_n = s]$$

Obviously it results that:

$$S_i(s,t) = \sum_{j=1}^{m} Q_{ij}(s,t)$$

Now it is possible to define the distribution function of the waiting time in each state i, given that the state successively occupied is known:

$$G_{ij}(s,t) = \mathrm{P}[T_{n+1} \leq t | J_n = j, J_{n+1} = j, T_n = s]$$

Obviously the related probabilities can be obtained by means of the following formula:

$$G_{ij}(s,t) \;=\; \begin{cases} Q_{ij}(s,t)/p_{ij}(s) & \text{if } p_{ij}(s) \neq 0 \\ 1 & \text{if } p_{ij}(s) = 0 \end{cases}$$

The main difference between a continuous time non-homogeneous Markov process and a NHSMP is in the increasing distribution functions $G_{ij}(s,t)$. In Markov environment this function has to be a negative exponential function. Instead in the semi-Markov case the distribution functions $G_{ij}(s,t)$ can be of any type. If we apply the semi-Markov model in the credit risk environment we can take into account, by means of the $G_{ij}(s,t)$ the problem given by the duration of the rating inside the states. Now the NHSMP $Z = (Z_t,\, t \in \mathbb{R}^+)$ can be defined. It represents, for each waiting time, the state occupied by the process. The transition probabilities are defined in the following way:

$$\phi_{ij}(s,t) = \mathrm{P}[Z_t = j | Z_s = i]$$

They are obtained solving the following evolution equations:

$$\phi_{ij}(s,t) \;=\; \delta_{ij}(1 - S_i(s,t)) + \sum_{\beta=1}^{m} \int_{s}^{t} \dot{Q}_{i\beta}(s,\vartheta)\phi_{\beta j}(\vartheta,t)d\vartheta \tag{1}$$

where δ_{ij} represents the Kronecker symbol. The first part of relation (1)

$$\delta_{ij}(1 - S_i(s,t)) \tag{2}$$

gives the probability that the system doesn't have transitions up to the time t given that it was in the state i at time s. The (2) formula in rating migration case represents the probability that the rating organisation doesn't give any new rating evaluation from the time s up to the time t. This part has sense if and only if $i = j$. In the second part

$$\sum_{\beta=1}^{m} \int_{s}^{t} \dot{Q}_{i\beta}(s,\vartheta)\phi_{\beta j}(\vartheta,t)d\vartheta$$

$\dot{Q}_{i\beta}(s,\vartheta)$ is the derivative at time ϑ of $Q_{i\beta}(s,\vartheta)$ and represents the probability intensity that the system was at time s in the state i and remained in this state up to the time ϑ and that it went to the state β just at time ϑ. After the transition the system will go to the state j following one of the possible trajectories that go from the state β at the time ϑ to the state j within the time t. In the credit risk environment it means that from the time s up the time ϑ the rating company doesn't give any other evaluation of the firm; at time ϑ the rating company gave the new rating β at the evaluating firm. After this the rating will arrive to the state j within the time t following one of the possible rating trajectories.

3 Non-homogeneous semi-Markov reliability model

There are a lot of semi-Markov models in reliability theory see for example [Limnios and Oprisan, 2000]. The non-homogeneous case was presented in [Blasi *et al.*, 2003]. Let us consider a reliability system S that can be at every time t in one of the states of $I = \{1, \ldots, m\}$. The stochastic process of the successive states of S is $Z = \{Z(t),\ t \geq 0\}$. The state set is partitioned into sets U and D, so that:

$$I = U \cup D,\ \ \emptyset = U \cap D,\ \ U \neq \emptyset,\ \ U \neq I$$

The subset U contains all "good" states in which the system is working and subset D all "bad" states in which the system is not working well or is failed. The classical indicators used in reliability theory are the following ones:
(i) *the non-homogeneous reliability function R* giving the probability that the system was always working from time s to time t:

$$R(s,t) = P\left[Z(u) \in U : \forall u \in (s,t]\right] \tag{3}$$

(ii) *the point wise non-homogeneous availability function A* giving the probability that the system is working on time t whatever happens on $(s,t]$:

$$A(s,t) = P\left[Z(t) \in U\right], \tag{4}$$

(iii) *the non-homogeneous maintainability function M* giving the probability that the system will leave the set D within the time t being in D at time s:

$$M(s,t) = 1 - P\left[Z(u) \in D,\ \ \forall u \in (s,t]\right]. \tag{5}$$

It is shown in [Blasi *et al.*, 2003] that these three probabilities can be computed in the following way if the process is a non-homogeneous semi-Markov process of kernel $\mathbf{Q}$.
(i) *the point wise availability function A_i* given that $Z_s = i$.

$$A_i(s,t) = \sum_{j \in U} \phi_{ij}(s,t) \tag{6}$$

(ii) *the reliability function R_i* given that $Z_s = i$. To compute these probabilities all the states of the subset D are changed in absorbing states. $R_i(s,t)$ is given by solving the evolution equation of HSMP but now with the embedded Markov chain having:

$$p_{ij}(s) = \delta_{ij} \ \text{ if } \ i \in D$$

The related formula will be:

$$R_i(s,t) = \sum_{j \in U} \phi^r_{ij}(s,t) \tag{7}$$

where $\phi^r_{ij}(s,t)$ is the solution of equation (1) with all the states in D that are absorbing;
(iii) *the maintainability function* M_i given that $Z_s = i$:
in this case all the states of the subset U are changed in absorbing states. $M_i(s,t)$ is given by solving the evolution equation of HSMP with the embedded Markov chain having:

$$p_{ij}(s) = \delta_{ij} \text{ if } i \in U.$$

The related formula will be:

$$M_i(s,t) = \sum_{j \in U} \phi^m_{ij}(s,t) \tag{8}$$

where $\phi^m_{ij}(s,t)$ is the solution of equation (1) with all the states in U that are absorbing.

4 Non-homogeneous semi-Markov reliability credit risk model

The credit risk problem can be situated in the reliability environment. The rating process, done by the rating agency, gives a reliability degree of a firm bond. In the Standard & Poors case there are the 8 different classes of rating that means to have the following set of states:

$$I = \{\text{AAA, AA, A, BBB, BB, B, CCC, D}\}$$

To take into account the downward problem we introduce other 6 states. The set of the states becomes the following:

$$I = \{\text{AAA, AA,AA - , A,A - , BBB,BBB - , BB,BB - , B,B - , CCC,CCC - , D}\}$$

For example the state BBB is divided in BBB and BBB-. The system will be in the state BBB if it arrived from a lower rating, instead it will be in the state BBB- if it arrived in the state from a better rating (a downward transition). It is also possible to suppose that if there is a virtual transition than if the system is in the BBB- state it will go to the BBB state.

The first 13 states are working states (good states) and the last one is the only bad state. The two subsets are the following:

$$U = \{\text{AAA, AA,AA - , A,A - , BBB,BBB - , BB,BB - , B,B - , CCC,CCC - }\}$$
$$D = \{\text{ D}\}$$

In this case the maintainability function M doesn't have sense because the default state D is absorbing and once that the system went in this state

it is not possible to leave it. Furthermore the fact that the only bad state is an absorbing state implies that the availability function A and the reliability function R correspond. In this case the reliability model is substantially simplified. In fact to get all the results that are relevant in the credit risk case it is enough to solve only once the system (2.1). Solving this system we will obtain the following results:
1) $\phi_{ij}(s,t)$, that represents the probabilities to be in the state j after a time t starting in the state i at time s. These results take into account the different probabilities to change state during the permanence of the system in the same state (duration problem) and the different probabilities to change state in function of the different time of evaluation (aging problem). The different probability values given for the two states that are obtained because of downward problem solve the third Markovian model problem.
2) $R_i(s,t) = A_i(s,t) = \sum\limits_{j \in U} \phi_{ij}(s,t)$, that represents the probability that the system never goes in the default state from the time s up to the time t.
3) $1 - S_i(s,t)$, that represents the probability that from the time s up to the time t no one new rating evaluation was done for the firm.

Before to give another result that can be obtained in a SMP environment, we have to introduce the concept of the first transition after the time t. More precisely we suppose that the system at time s was in the state i. We know that with probability $1 - S_i(s,t)$ the system doesn't move from the state i. Under these hypotheses we would know the probability that the next transition will be to the state j. This probability will be denoted by $\varphi_{ij}(s,t)$. That has the following meaning:

$$\varphi_{ij}(s,t) = \ P\left[X_{n+1} = j | X_n = i,\ T_{n+1} > t, T_n = s\right] \tag{9}$$

This probability can be obtained by means of the following formula:

$$\varphi_{ij}(s,t) = \frac{p_{ij}(s) - Q_{ij}(s,t)}{1 - S_i(s,t)}$$

After the definition (9) by means of SMP it is possible to get the following result:
4) $\varphi_{ij}(s,t)$ represents the probability to get the rank j at next rating if the previous state was i and no one rating evaluation was done from the time s up to the time t. In this way, for example, if the transition to the default state is possible and if the system doesn't move from the time s up to the time t from the state i, we know the probability that in the next transition the system will go to the default state.

5 Conclusions

This paper summarizes the three theoretical step that the authors did to improve the so called migration models in the credit risk environment. The first step solved the problem of different probability transactions because of the time duration inside a rating state by means of introduction of SMP in credit risk environment. The second step, by means of non-homogeneity introduction in the SMP environment, gave the way to consider also the system time dependence problem. The third step solved the credit risk downward problem. The three models start from the idea that credit risk problem can be considered a special case of reliability problem and this idea allows the application of some non-homogeneous semi-Markov reliability results in the credit risk environment. The downward problem was solved enlarging the state number. Authors in the next future hope to be able to get data from rating companies. In this case they will apply to real data their credit risk models.

References

[Blasi *et al.*, 2003]Blasi, A., Janssen, J., Manca, R. (2003). Numerical treatment of homogeneous and non-homogeneous reliability semi-Markov models. *Communications in Statistics*, Theory and Models .

[Carty and Fons, 1994]Carty, L., Fons, J. (1994).Measuring changes in corporate credit quality. The Journal of Fixed Income v. 4, 27-41.

[D'Amico *et al.*, 2003]D'Amico, G., Janssen, J., Manca, R. (2003), Homogeneous discrete time semi-Markov reliability models for credit risk Management. Proceedings of the XXVI AMASES Conference.

[D'Amico *et al.*, 2004a]D'Amico, G., Janssen, J., Manca, R. (2004), Non-homogeneous semi-Markov reliability transition credit risk models. Proceedings of the II International Workshop in Applied Probability.

[D'Amico *et al.*, 2004b]D'Amico, G., Janssen, J., Manca, R. (2004), Downward credit risk problem and non-homogeneous semi-Markov reliability transition models. Proceedings of IME 2004.

[Duffie and Singleton, 2003]Duffie, D., Singleton, K.J. (2003). *Credit Risk*. Princeton

[Howard, 1971]Howard, R. , (1971). *Dynamic probabilistic systems*, vol I & II, Wiley.

[Iosifescu Manu, 1972]Iosifescu Manu, A.(1972), Non homogeneous semi-Markov processes, *Stud. Lere. Mat.* 24, 529-533.

[Jarrow *et al.*, 1997]Jarrow, A. J., Lando, D., Turnbull, S. M. (1997). A Markov model for the term structure of credit risk spreads. The Review of Financial Studies, v. 10, 481-523.

[Janssen and Manca, 2005]Janssen, J., Manca, R. (2005). *Applied semi-Markov processes for Finance, Insurance and Reliability.* to be published.

[Levy, 1954]Levy, P. (1954). Processus semi-Markoviens. *Proc. of International Congress of Mathematics*, Amsterdam.

[Limnios and Oprisan, 2000]Limnios, N., Oprisan, G. (2000). *Semi-Markov Processes and Reliability modeling.* World Scientific, Singapore.

[Nickell *et al.*, 2000]Nickell, P., Perraudin, W., Varotto, S. (2000). Stability of rating transitions. *Journal of Banking and Finance*, v. 24, 203-227.

An application of comonotonicity in multiple state models

Jaap Spreeuw

City University, Cass Business School
106 Bunhill Row
EC1Y8TZ, London, UK
(e-mail: j.spreeuw@city.ac.uk)

Abstract. In this paper, comonotonicity will be applied to multiple state models under Markov assumptions in a continuous time framework. We deal with one application in actuarial science. It involves deriving the distribution of the present value of benefits less premiums of a disability annuity, also known as an income protection policy. The quality of the approximation is investigated by comparing the distribution obtained with the one derived from the algorithm presented in the paper by [Hesselager and Norberg, 1996].

Keywords: comonotonicity, multiple state, disability.

References

[Hesselager and Norberg, 1996]Hesselager, O., Norberg, R., 1996. On probability distributions of present values in life insurance. *Insurance: Mathematics and Economics,* 18, 35-42.

Homogeneous Backward Semi-Markov Reward Models for Insurance Contracts

Raimondo Manca[1], Dmitrii Silvestrov[2], and Fredrik Stenberg[2]

[1] Universitá di Roma La Sapienza
via del Castro Laurenziano, 9
00161 Roma Italy
(e-mail: raimondo.manca@uniroma1.it)

[2] Malardalen University
Box 883
SE-721 23 Vasteras Sweden (e-mail: dmitrii.silvestrov@mdh.se,
fredrik.stenberg@mdh.se

Abstract. Semi-Markov reward processes are a very important tool for the solution of insurance problems. In disability problems can assume great relevance the date of the disability accident. In fact the mortality probability of a disabled person of a given age is higher respect the one of a person of the same age that is healthy. But the difference decreases with the running of the time after the instant of the disability. By means of backward semi-Markov processes it is possible to take in account the duration of the disability for an insured person. In this paper is shown for the first time, at authors' knowing, how to apply backward semi-Markov reward processes in insurance environment. An application will be shown.

Keywords: backward semi-Markov processes, reward processes, disability insurance.

1 Introduction

Semi-Markov processes was first defined by [Levy, 1954] in the fifties. At the beginning their application was in engineering, mainly where the application were linked to ageing. The use of so called multiple state models have long been used in the actuarial world for dealing with disability and illness among other things, see for example the book by [Haberman and Pitacco, 1999]. These models can be described by the use of semi-Markov processes and semi-Markov reward processes. An insurance contract ensures the holder benefits in the future from some random event(s) occurring at some random moment(s). The holder of the insurance contract pays a premium for the contract. Denote the discounted cash flow that occurs between the counter parties as the discounted accumulated reward where both the premiums and benefits are considered to be rewards. When developing an insurance contract between the writer and receiver the following questions must be asked. How shall the reward structure of the contract be determined? The fee can depend on the individuals exposure to becoming disabled in different states, and the benefits can be of two types, either instant rewards associated with transition between states or permanence rewards associated with maintaining in a state. In time evolution of insurance problems it is necessary to consider two different kind of randomness. One is originated by the accumulation during the time of the premiums and benefits paid or received (the financial evolution); the other is given by the time of the state change of the insured person, usually in insurance problem the transition among the states are effected at a random time. A semi-Markov environment can naturally take into account of both the two random aspects. This property was outlined for example in [Janssen and Manca, 2003] and [Janssen and Manca, 2004]. Another problem in insurance mainly in disability is the fact that the probability to change state is function of the distance from the moment of the disability. For example the probability to die in a disabled person of a given age is higher respect the one of a person of the same age that is healthy. But the difference decreases with the running of the time. In this paper the authors will consider also this duration effect using a bacward homogeneous semi-Markov reward process. By means of semi-Markov reward both financial and transition time randomness will be considered. By means of the backward environment also the duration phenomenon can be taken into account. It is to remark that, at authors' knowing, it is the first time that this last problem is faced by means of SMP in insurance field.

2 Homogenous Model.

Given the probability space (Ω, F, P) consider a homogenous Markov renewal process (X_n, T_n), $T_0 \leq T_1 \leq T_2 \leq ...$. Let the stochastic process $X_n, n \in \mathbb{N}$ have state space $E = \{1, 2, ..., m\}$ representing the state at the n-th transition. Let T_n represent the random time of the n-th transition with state space $\mathbb{N}$. For the combined process (X_n, T_n) define $Q_{ij}(t), b_{ij}(t), S_i(t)$ as,

$$Q_{ij}(t) = P(X_{n+1} = j, T_{n+1} - T_n \leq t | X_n = i) \tag{1}$$

$$b_{ij}(t) = P(X_{n+1} = j, T_{n+1} - T_n = t | X_n = i) \tag{2}$$

$$S_i(t) = P(T_{n+1} - T_n \leq t | X_n = i). \tag{3}$$

We allow for $Q_{ii}(t) \neq 0$, $t = 1, 2, ...$, i.e., artificial jumps from state i to itself, this is due to that sometimes this possibility makes sense in insurance applications. Impose $Q_{ij}(0) = 0$ for all $i, j \in E$, i.e., no instantaneously jumps in our process. Obviously,

$$S_i(t) = \sum_j Q_{ij}(t) \tag{4}$$

and

$$b_{ij}(t) = Q_{ij}(t) - Q_{ij}(t-1). \tag{5}$$

It is well known that,

$$p_{ij} = \lim_{t \to \infty} Q_{ij}(t) \quad i, j \in E$$

where $\mathbf{P} = [p_{ij}]$ is the transition probability of the embedded Markov chain. The conditional distribution functions for waiting time in each state i is given the state subsequently is j is given by,

$$G_{ij}(t) = P(T_{n+1} - T_n \leq t | X_n = i, X_{n+1} = j) = \begin{cases} \frac{Q_{ij}(t)}{p_{ij}}, & \text{if } p_{ij} \neq 0, \\ 1, & \text{if } p_{ij} = 0. \end{cases}$$

Define $\kappa(t)$ in the following way,

$$\kappa(t) = t - \max_{T_n \leq t} \{T_n\}. \tag{6}$$

$\kappa(t)$ describes the time already spent in the current state at time t.

3 Homogenous Rewards

The notation of rewards is given by;

ψ_i, $\psi_i(\kappa(t))$, $\psi_i(\kappa(t), t)$ denotes the rewards that are given for the permanence in the i-th state. The first reward doesn't change with the time and the future transition. The second changes with the time spent in the state. The third changes with the time spent in the state and is function of $\kappa(t)$ and t. They represent the flows of annuity that is paid during the presence in state i.

γ_{ij}, $\gamma_{ij}(\kappa(t))$, $\gamma_{ij}(\kappa(t), t)$ denote the rewards that are given for the transition from the i-th state to the j-th one. The distinctions among the three impulse rewards is the same given previuosly for the permanence rewards.

We will in this paper focus on constant rewards but our result can be extended into the other cases on the expense of more notation and indexes.

Let $e^{-t\delta}$ denote the discount factor for t periods with common fixed intensity of interest rate δ. Let $\xi_{i,u}(s,t)$, $s \leq t$ denote the accumulated discounted reward from s excluding s up to and including t given that the at time s the process is at state $i \in E$ and the previous jump accursed u moments ago. Here we apply the convention that $\xi_{i,u}(t,t) = 0$ for all t.

Theorem 1 *The reward process $\xi_{i,u}(s,t)$ is homogenous*

$$\xi_{i,u}(s,t) \overset{d}{=} \xi_{i,u}(0, t-s) \quad \forall i, u, s, t. \tag{7}$$

if the underlying process is a homogenous semi-Markov process and if the rewards only depends on $\kappa(t)$.

Introduce $T_{i,u}$ with the following distribution,

$$P(T_{i,u} > t) = \frac{1 - S_i(t+u)}{1 - S_i(u)} \tag{8}$$

and

$$P(T_{i,u} = s, X_{i,u} = j) = \frac{b_{ij}(u+s)}{1 - S_i(u)}. \tag{9}$$

Then $T_{i,u}$ describes the time to the next jump given that the process already have visited the state i for u units of time and let $X_{i,u}$ denote the corresponding state we end up in after the jump.

Let us assume $u = 0$, and first find a recursive relation for $\xi_{i,0}(0,t)$. We will have to consider two cases, if no jump occurs before moment t, or if at least one jump occurs between moment 0 up to moment t. If we introduce the indicator variables for these events we fill find the following relationship for $\xi_{i,0}(0,t)$,

$$\begin{aligned}\xi_{i,0}(0,t) &\overset{d}{=} \chi(T_{i,0} > t)\sum_{s=1}^{t}\psi_i e^{-\delta s}\\ &\quad+\sum_j\sum_{s=1}^{t}(\chi(T_{i,0} = s, X_{i,0} = j)(e^{-\delta s}\gamma_{ij}(s) + \sum_{u=1}^{s}\psi_i(s)e^{-\delta u})) \qquad (10)\\ &\quad+\sum_j\sum_{s=1}^{t}(\chi(T_{i,0} = s, X_{i,0} = j)e^{-\delta s}\xi_{j,0}(0,t-s)) \quad i \in E, \quad t = 1, 2, ...\end{aligned}$$

where $\xi_{j,0}(0, t-s)$ are independent of indicators $\chi(T_{i,0} = s, X_{i,0} = j)$ and $\chi(T_{i,0} > t)$. The first term represents the discounted reward we receive at moment u to jump from state i to state j, the second term is due to the fact that the process restarts and is Markov at the moment of jump together with the assumption of homogeneities. The last term consists of the rewards we receive for the presence in state i between the moment 0 and u. This defines a closed system of equations which recursively can be solved.

To simplify the expression we can introduce some notation,

$$a_i(t) = \sum_{s=1}^{t}\psi_i(s)e^{-\delta s} \tag{11}$$

$$\tilde{a}_{ij}(t) = a_i(t) + e^{-\delta t}\gamma_{ij}(t). \tag{12}$$

Here $a_i(t)$ corresponds to the discounted accumulated reward for persistence in state i for t moments of time and $\tilde{a}_{ij}(t)$ the discounted accumulated reward for the persistence in state i for t moments of time plus the discounted instant reward for transition from state i to j at time t.

Then,

$$\begin{aligned}\xi_{i,0}(0,t) \overset{d}{=} \chi(T_{i,0} > t)a_i(t) + \sum_j \sum_{s=1}^{t} \chi(T_{i,0} = s, X_{i,0} = j)\tilde{a}_{ij}(s)\\ + \sum_j \sum_{s=1}^{t} (\chi(T_{i,0} = s, X_{i,0} = j)e^{-\delta s}\xi_{j,0}(0, t-s) \quad i \in E, \quad t = 1, 2, ...\end{aligned}$$

In the case $u \neq 0$, i.e., if we are interested in finding the accumulated reward toward a moment in time not associated with a jump;

$$\begin{aligned}\xi_{i,u}(0,t) \overset{d}{=} \chi(T_{i,u} > t)a_i(t) + \sum_j \sum_{s=1}^{t} \chi(T_{i,u} = s, X_{i,u} = j)\tilde{a}_{ij}(s)\\ + \sum_j \sum_{s=1}^{t} \chi(T_{i,u} = s, X_{i,u} = j)e^{-\delta s}\xi_{j,0}(0, t-s) \quad i \in E, \quad t = 1, 2, ...\end{aligned}$$

The only difference from the previous expressions is that we have to remember that our first jump-time depends on u, i.e., our surjeon time in the initial state is at least $u + 1$.

The first moment can now be calculated using these relationships, first consider the case $u = 0$,

$$\begin{aligned}E[\xi_{i,0}(0,t)] &= E[\chi(T_{i,0} > t)]a_i(t) + \sum_j \sum_{s=1}^{t} E[\chi(T_{i,0} = s, X_{i,0} = j)]\tilde{a}_{ij}(s)\\ &+ \sum_j \sum_{s=1}^{t} E[\chi(T_{i,0} = s, X_{i,u} = j)]E[\xi_{j,0}(0, t-s)]e^{-\delta s}\\ &= (1 - S_i(t))a_i(t) + \sum_j \sum_{s=1}^{t} b_{ij}(s)\tilde{a}_{ij}(s) \qquad (13)\\ &+ \sum_j \sum_{s=1}^{t} b_{ij}(s)E[\xi_{j,0}(0, t-s)]e^{-\delta s} \quad i \in E, \quad t = 1, 2, ...\end{aligned}$$

which follows from independence mentioned earlier. This set of equations can recursively be solved. Let $V_i(t) = E[\xi_{i,0}(0,t)], i \in E, t = 1, 2, ...$ then

$$\begin{aligned}
&V_i(0) = 0 && \forall i \in E\\
&V_i(1) = (1 - S_i(1))a_i(1) + \sum_j b_{ij}(1)\tilde{a}_{ij}(1) && \forall i \in E\\
&V_i(2) = (1 - S_i(2))a_i(2) + \sum_j b_{ij}(1)\tilde{a}_{ij}(1) + \sum_j b_{ij}(2)\tilde{a}_{ij}(2)\\
&\qquad + \sum_j e^{-\delta} b_{ij}(1)V_j(1) \quad i \in E. && (14)
\end{aligned}$$

And in general,

$$V_i(t) = (1 - S_i(t))a_i(t) + \sum_j \sum_{s=1}^{t} b_{ij}(s)\tilde{a}_{ij}(s) + \sum_j \sum_{s=1}^{t} e^{-\delta s} b_{ij}(s)V_j(t-s)$$
$$\forall i \in E$$

The values of $S_i(t),\ a_i(t),\ \tilde{a}_{ij}(t)$ are known and the only unknown parameters are $V_i(t)$. Above we see how we can recursively determine $V_i(t)$ by recursively solving $V_j(1), V_j(2), ..., V_j(t-1)$ for all $j \in E$.

And in the general case with $u \neq 0$,

$$\begin{aligned}
E[\xi_{i,u}(0,t)] &= E[\chi(T_{i,u} > t)]a_i(t) + \sum_j \sum_{s=1}^{t} E[\chi(T_{i,u} = s, X_{i,u} = j)]\tilde{a}_{ij}(s)\\
&\quad + \sum_j \sum_{s=1}^{t} E[\chi(T_{i,u} = s, X_{i,u} = j)]E[\xi_{j,0}(0, t-s)]e^{-\delta s}\\
&= \frac{1 - S_i(t+u)}{1 - S_i(u)} a_i(t) + \sum_j \sum_{s=1}^{t} \frac{b_{ij}(u+s)}{1 - S_i(u)} \tilde{a}_{ij}(s) \qquad (15)\\
&\quad + \sum_j \sum_{s=1}^{t} \frac{b_{ij}(u+s)}{1 - S_i(u)} E[\xi_{j,0}(0, t-s)]e^{-\delta s} \quad i \in E,\ \ t = 1, 2, ...
\end{aligned}$$

Note here that its enough to determine all $E[\xi_{j,0}(0,s)]$ for all $j \in E, s = 0, 1, ..., t-1$ to determine $E[\xi_{i,u}(0,t)]$. We are thereby back to our basic case $u = 0$.

4 Disability

In the papers by [Janssen and Manca, 2003]and [Janssen *et al.*, 2004] it is shown how to apply continuous time semi-Markov reward processes in multiple life insurance. In the paper by Blasi et al (2004), a real case study using real disability data is given. We will extend the example given in this paper using the backward homogeneous semi-Markov reward process that can take into account the duration of disability.

The model is a 5-state model. The considered states are the following:

states	disability degree	reward
1	$[0, .1)$	1000
2	$[.1, .3)$	1500
3	$[.3, .5)$	2000
4	$[.5, .7)$	2500
5	$[.7, 1]$	3000

The data gives the disability history of 840 persons that had silicosis problems and that live in Campania, a region in Italy. The reward is given to construct the example, it represents the money amount that is paid for each time period to the disable in function of its degree of illness. The transition occurs after a doctor visit that can be seen as the check to decide in which state the disable person is in. This gives naturally an example where virtual transitions are possible.

To be able to apply the technique developed in this paper for homogenous semi-Markov processes, we must first construct the embedded Markov-chain. The transition matrix is constructed from real data and is reported in the following table.

	0-10	10-30	30-50	50-70	70-100
0-10	0	1	0	0	0
10-30	0	0.811	0.180	0.005	0.004
30-50	0	0.017	0.75	0.21	0.02
50-70	0	0.023	0.03	0.72	0.22
70-100	0	0	0	0	1

Next step is to construct the matrix valued waiting time distribution $G(t)$.

To show the difference due to the introduction of the backward process the results with $u = 0$ (that means that the person entered in the state i when we begin the study of the system) and with $u = 2$ are reported.

s	0-10	10-30	30-50	50-70	70-100
1	970,87	1456,31	1941,74	2427,18	2912,62
2	1913,47	2876,75	3831,08	4780,34	5740,40
3	2993,34	4278,51	5684,64	7081,26	8485,83
4	4283,87	5657,51	7510,09	9330,86	11151,29
5	5547,74	7015,79	9310,84	11536,94	13739,12
6	6792,15	8352,65	11078,77	13684,65	16251,57
7	8015,84	9667,20	12817,61	15778,89	18690,84
8	9219,65	10959,30	14522,48	17822,21	21059,07
9	10403,01	12229,95	16193,96	19816,49	23358,32
10	11565,50	13479,60	17830,89	21759,34	25590,60

mean total reward with $u = 0$

s	0-10	10-30	30-50	50-70	70-100
1	970,87	1456,31	1941,74	2427,18	2912,62
2	2346,91	2904,34	3880,52	4827,55	5740,40
3	3688,40	4340,42	5812,47	7207,04	8485,83
4	5009,97	5759,51	7714,30	9527,12	11151,29
5	6308,27	7159,74	9590,50	11792,99	13739,12
6	6792,15	8352,65	11078,77	13684,65	16251,57
7	8840,21	9892,87	13237,77	16171,26	18690,84
8	10072,70	11227,54	15006,35	18279,73	21059,07

mean total reward with $u = 2$

Few words to describe the results. We present an example only to show that taking into account the permanence into the state before the beginning of the study of the system changes the results. We did not change the transition probabilities changing the backward variable u. The different dead probability means different transition probability. But also without changing the probabilities the results were different. It is only to observe that the last state is absorbing and from (15) it follows that the results do not change. Furthermore the payments of the first year are always the same because they are equal to the corresponding first discounted rewards as it was proved in [Janssen and Manca, 2005].

5 Conclusions

In this paper a first step for the application of the backward semi-Markov reward in insurance field was done. In future works the authors would generalize this approach in non homogeneous environment. Reward processes represent the first moment of the total revenues that are given in a stochastic financial operation. The author would also find models and algorithms useful to compute the higher moments.

References

[Janssen and Manca, 2003]Janssen, J. and Manca, R., Multi-state insurance model description by means of continuous time homogeneous semi-Markov reward processes. *Mimeo.*

[Janssen and Manca, 2004]Janssen, J. and Manca, R., Discrete Time Non-Homogeneous Semi-Markov Reward Processes, Generalized Stochastic Annuities and Multi-State Insurance Model. *Proc. of XXVIII AMASES* Modena 2004.

[Janssen and Manca, 2005]Janssen, J., Manca, R., *Applied semi-Markov processes for Finance, Insurance and Reliability* to be published 2005.

[Janssen *et al.*, 2004]Janssen, J., Manca, R. and Volpe di Prignano E., Continuous time non homogeneous semi-Markov reward processes and multi-state insurance application. *Proc. of IME 2004* 2004.

[Haberman and Pitacco, 1999]Haberman, S. and Pitacco, E., *"Actuarial models for disability Insurance"*, 1999, Chapman and Hall. 1999

[Levy, 1954]Levy, P., Processus semi-Markoviens, *Proc. of International Congress of Mathematics*, Amsterdam, 1954.

Discrete Time Reward Processes, Stochastic Annuities and Insurance Models

Jaqcues Janssen[1], Raimondo Manca[2], and Giuseppina Ventura[2]

[1] CESIAF, EURIA
Université de Bretagne Occidentale
6 avenue le Gorgeu, CS 93837,
29238 BREST Cedex 3, France
(e-mail: cesiaf@belgacom.net)

[2] Universitá di Roma La Sapienza
via del Castro Laurenziano, 9
00161 Roma Italy
(e-mail: raimondo.manca@uniroma1.it, giuseppina.ventura@uniroma1.it)

Abstract. The Markov and semi-Markov reward processes are a very powerful tool. They can be applied in many different fields, like mechanical systems, evaluation of computer systems etc. But in authors' opinion the most fruitful and natural application environment of these tools is the insurance field. In the paper will be given the definition of stochastic annuity and of its generalization their strict relation to the homogeneous and non-homogeneous semi-Markov reward processes. At last it will be shown how is natural to apply these rewards processes in the insurance environment.

Keywords: semi-Markov processes, Markov processes, stochastic annuities.

1 Introduction

Homogeneous semi-Markov processes (HSMP) were defined in the fifties. At beginning their applications were in engineering field, mainly in problem of reliability and maintenance see for example [Howard, 1971]. Non-homogeneous semi-Markov processes were defined in [Iosifescu Manu, 1972]. Applications of the semi-Markov processes were presented in finance and insurance see for example in [Janssen, 1966], [CMIR12, 1991]. Some of these applications were done attaching a reward structure to the process. This structure can be thought as a random variable associated with the state occupancies and transitions [Howard, 1971]. Non-homogeneous semi Markov reward processes were defined in [De Dominicis *et al.*, 1986]. The non-homogeneity results of great relevance in actuarial field, because in this way it is possible to take into account the different behaviour in function of the age. A stochastic approach to the annuity was given in [Wolthuis, 2003]. In this book the continuous time non-homogeneous Markov processes were used to generalize the annuity concept. This approach did not use the reward environment. The non-homogeneous Markov model was used to solve the multiple state insurance problems see [Wolthuis, 2003]. The aim of this paper is to investigate the strict correlation existing between insurance models and reward processes in a Markov or semi-Markov environment. In this light the financial concepts of stochastic annuities and generalized stochastic annuities are given. It is shown how these concepts correspond from financial view point to the reward processes. For an applicative intent, the paper is in discrete time environment. The paper begins introducing the semi-Markov processes both in homogeneous and non-homogeneous case. In section 3 the Discrete Time Homogeneous and Non-Homogeneous Markov and Semi-Markov ReWard Processes are presented. The subsequent section defines the concepts of stochastic annuity and of generalized stochastic annuity. In this part it is also explained the strict connection between the reward processes presented before and the annuities. In the last section the relations between multiple states insurance models and stochastic annuity are given.

2 Discrete time homogeneous and non-homogeneous semi-Markov processes.

In this part will be shortly described both the DTHSMP and DTNHSMP.

Let $E = \{1, 2, \ldots, m\}$ be the set of states of our system, $J_n \in E$ the random variable (r.v.) representing the state at the n th transition and $T_n \in \mathbb{N}$ an other r.v. with set of states equal to $\mathbb{N}$ where T_n represents the time of the n th transition. It results:

$$J_n : \Omega \to E \quad T_n : \Omega \to \mathbb{N}$$

The process (J_n, T_n) is a homogeneous (non-homogeneous) markovian renewal process if the kernel $\mathbf{Q} = [Q_{ij}(t)](\mathbf{Q} = [Q_{ij}(s,t)]])$ associated to the process is defined in the following way:

$$Q_{ij}(t) = P[J_{n+1} = j, T_{n+1} - T_n \le t | J_n = i]$$
$$(Q_{ij}(s,t) = P[J_{n+1} = j, T_{n+1} \le t | J_n = i, T_n = s])$$

Furthermore it is necessary to introduce the probability that the process will leave the state i in a time t:

$$H_i(t) = \sum_{j=1}^{m} Q_{ij}(t) \quad \left(H_i(s,t) = \sum_{j=1}^{m} Q_{ij}(s,t) \right)$$

Furthermore the probabilities that there is a transtion at time t are considered:

$$b_{ij}(t) = \begin{cases} Q_{\mathrm{ij}}(t) = 0 \text{ if } t = 0 \\ Q_{ij}(t) - Q_{ij}(t-1) \text{ if } t > 0 \end{cases}$$

$$\left(b_{ij}(s,t) \; = \; \begin{cases} Q_{ij}(s,t)=0 \;\text{ if }\; t \le s \\ Q_{ij}(s,t)-Q_{ij}(s,t-1) \;\text{ if } t > s \end{cases} \right)$$

Now it is possible to define the probability distribution of the waiting time in each state i, given that the state successively occupied is known:

$$F_{ij}(t) = P[T_{n+1} - T_n \le t | J_n = i, J_{n+1} = j]$$
$$(F_{ij}(s,t) = P[T_{n+1} \le t | J_n = i, J_{n+1} = j, T_n = s]).$$

Now the DTHSMP (DTNHSMP) $Z = (Z_t, t \in \mathbb{N})$ can be defined. It represents, for each waiting time, the state occupied by the process. The transition probabilities are defined in the following way:

$$\phi_{ij}(t) = \mathrm{P}\left[Z_t = j | Z_0 = i\right]$$
$$\left(\phi_{ij}(s,t) = \mathrm{P}\left[Z_t = j | Z_s = i\right]\right).$$

They are obtained solving the following evolution equations:

$$\phi_{ij}(t) \; = \; \delta_{ij}(1 - H_i(t)) + \sum_{\beta=1}^{m} \sum_{\vartheta=1}^{t} b_{i\beta}(\vartheta)\phi_{\beta j}(t-\vartheta)$$

$$\left(\phi_{ij}(s,t) \; = \; \delta_{ij}(1 - H_i(s,t)) + \sum_{\beta=1}^{m} \sum_{\vartheta=1}^{t} b_{i\beta}(s,\vartheta)\phi_{\beta j}(\vartheta,t) \right)$$

where δ_{ij} represents the Kronecker symbol.

3 The discrete time homogeneous and non-homogeneous Markov and semi-Markov reward processes

Now a reward structure will be introduced, this structure is connected with the Z process. The reward process, both in Markov and semi-Markov cases, can be considered a class of stochastic processes in which, depending on the hypotheses, the evolution equation varies. In non-homogeneous case the rewards can depend also on the time of entrance in the state. Furthermore the non-homogeneity can involve the interest law in the sense that the interest rate can depend on the time of beginning of the operation and the time in which the operation ends (non-homogeneous time interest rate laws). This fact implies that in the non-homogeneous environment should be considered more cases than in homogeneous one. There are permanence rewards and transition rewards. In the literature they are also called respectively rate rewards and impulse rewards;; the first represents the reward given for the permanence in a state and the second the one paid because of a transition. The reward processes can be discounted or non-discounted. We are dealing with financial phenomena and we will present only the discounted cases. As already we told, there many different evolution equations (in non-homogeneous case more than three hundreds) but we will present the general cases. We will distinguish only between the immediate and the due cases. In the first the instalment is paid at the end of each period in the second at the beginning. This distinction seems to be trivial but from computational point of view it assumes great relevance. This time we present before the Markov relation that in the immediate case has the following structure:

$$\begin{aligned} V_i^{(n)} &= V_i^{(n-1)} + \nu(n) \\ &\cdot \sum_{k=1}^{m} p_{ik}^{(n-1)} \left((1-swpe)\psi_k(n) + \sum_{j=1}^{m} p_{kj} \left(\gamma_{kj}(n) + swpe \cdot \psi_{kj}(n)\right) \right), \end{aligned}$$

$$\left(\begin{array}{l} V_i^{(n)}(s) = V_i^{(n-1)}(s) + \nu(s,s+n) \sum_{k=1}^{m} p_{ik}^{(n-1)}(s) \cdot \left((1-swpe)\psi_k(s,s+n)\right. \\ \left. + \sum_{j=1}^{m} p_{kj}(s+n) \left(\gamma_{kj}(s,s+n) + swpe \cdot \psi_{kj}(s,s+n)\right)\right), \end{array} \right)$$

where $V_i^{(n)}$ $\left(V_i^{(n)}(s)\right)$ represents the mean present value of all the rewards paid from 0 to n (s to s+n) and $\nu(s)$ $(\nu(s,s+n))$ the corresponding discount factor. Furthermore *swpe* represents a variable that will have value 1 if the permanence rewards depend on the next transition and 0 if they do not depend on the transition. In Markov due case we have the following relation:

$$\begin{aligned} &\ddot{V}_i^{(n)} = \ddot{V}_i^{(n-1)} + \nu(n) \sum_{k=1}^{m} p_{ik}^{(n-1)} \sum_{j=1}^{m} p_{kj}\gamma_{kj}(n) + \\ &\nu(n-1) \left((1-swpe) \sum_{k=1}^{m} p_{ik}^{(n-1)} \psi_k(n-1) + swpe \sum_{k=1}^{m} p_{ik}^{(n-2)} \sum_{j=1}^{m} p_{kj}\psi_{kj}(n-1) \right) \end{aligned}$$

$$\left(\begin{array}{l} \ddot{V}_i^{(n)} = \ddot{V}_i^{(n-1)} + \nu(s,s+n) \sum_{k=1}^{m} p_{ik}^{(n-1)} \sum_{j=1}^{m} p_{kj}\gamma_{kj}(s,s+n) + \nu(s,s+n-1) \\ \cdot \left((1-swpe) \sum_{k=1}^{m} p_{ik}^{(n-1)} \psi_k(s,s+n-1) + swpe \sum_{k=1}^{m} p_{ik}^{(n-2)} \sum_{j=1}^{m} p_{kj}\psi_{kj}(s,s+n-1) \right) \end{array} \right)$$

In the semi-Markov immediate case the relation is the following:

$$\begin{array}{l} V_i(t) = (1 - swpe)(1 - H_i(t)) \sum\limits_{\tau=1}^{t} \psi_i(\tau)\nu(\tau) + swpe(1 - H_i(t)) \sum\limits_{k=1}^{m} \varphi_{ik}(t) \sum\limits_{\tau=1}^{t} \psi_{ik}(\tau)\nu(\tau) \\ + \sum\limits_{k=1}^{m} \sum\limits_{\vartheta=1}^{t} b_{ik}(\vartheta) \sum\limits_{\tau=1}^{\vartheta} \psi_{ik}(\tau)\nu(\tau) + \sum\limits_{k=1}^{m} \sum\limits_{\vartheta=1}^{t} b_{ik}(\vartheta)\gamma_{ik}(\vartheta)\nu\,(\vartheta) + \sum\limits_{k=1}^{m} \sum\limits_{\vartheta=1}^{t} b_{ik}(\vartheta)V_k(t-\vartheta)\nu\,(\vartheta) \end{array}$$

$$\left(\begin{array}{l} V_i(s,t) = (1 - swpe)(1 - H_i(s,t)) \sum\limits_{\tau=s+1}^{t} \psi_i(s,\tau)\nu(s,\tau) \\ + swpe(1 - H_i(s,t)) \sum\limits_{k=1}^{m} \varphi_{ik}(s,t) \sum\limits_{\tau=s+1}^{t} \psi_{ik}(s,\tau)\nu(s,\tau) \\ + \sum\limits_{k=1}^{m} \sum\limits_{\vartheta=s+1}^{t} b_{ik}(s,\vartheta) \sum\limits_{\tau=s+1}^{\vartheta} \psi_{ik}(s,\tau)\nu(s,\tau) \\ + \sum\limits_{k=1}^{m} \sum\limits_{\vartheta=s+1}^{t} b_{ik}(s,\vartheta)\gamma_{ik}(s,\vartheta)\nu\,(s,\vartheta) + \sum\limits_{k=1}^{m} \sum\limits_{\vartheta=s+1}^{t} b_{ik}(s,\vartheta)V_k(\vartheta,t)\nu\,(s,\vartheta) \end{array} \right)$$

where:

$$\varphi_{ij}(t) = \frac{p_{ij} - Q_{ij}(t)}{1 - H_i(t)} \quad \left(\varphi_{ij}(s,t) = \frac{p_{ij}(s) - Q_{ij}(s,t)}{1 - H_i(s,t)} \right)$$

In the due case we have:

$$\begin{array}{l} \ddot{V}_i(t) = (1 - swpe)(1 - H_i(t)) \sum\limits_{\tau=0}^{t-1} \psi_i(\tau)\nu(\tau) + swpe(1 - H_i(t)) \sum\limits_{k=1}^{m} \sum\limits_{\tau=0}^{t-1} \varphi_{ik}(t)\psi_{ik}(\tau)\nu(\tau) \\ + \sum\limits_{k=1}^{m} \sum\limits_{\vartheta=1}^{t} b_{ik}(\vartheta) \sum\limits_{\tau=0}^{\vartheta-1} \psi_{ik}(\tau)\nu(\tau) + \sum\limits_{k=1}^{m} \sum\limits_{\vartheta=1}^{t} \nu(\vartheta)b_{ik}(\vartheta)\gamma_{ik}(\vartheta) + \sum\limits_{k=1}^{m} \sum\limits_{\vartheta=1}^{t} \nu(\vartheta-1)b_{ik}(\vartheta)\ddot{V}_k(t-\vartheta) \end{array}$$

$$\left(\begin{array}{l} \ddot{V}_i(s,t) = \sum\limits_{k=1}^{m} \sum\limits_{\vartheta=s+1}^{t} b_{ik}(s,\vartheta) \sum\limits_{\tau=s}^{\vartheta-1} \psi_{ik}(s,\tau)\nu(s,\tau) \\ + \sum\limits_{k=1}^{m} \sum\limits_{\vartheta=s+1}^{t} \nu(s,\vartheta)b_{ik}(s,\vartheta)\gamma_{ik}(s,\vartheta) + \sum\limits_{k=1}^{m} \sum\limits_{\vartheta=s+1}^{t} \nu(s,\vartheta-1)b_{ik}(s,\vartheta)\ddot{V}_k(\vartheta,t) \\ + (1 - swpe)(1 - H_i(s,t)) \sum\limits_{\tau=s}^{t-1} \psi_i(s,\tau)\nu(s,\tau) \\ + swpe(1 - H_i(s,t)) \sum\limits_{k=1}^{m} \sum\limits_{\tau=s}^{t-1} \varphi_{ik}(s,t)\psi_{ik}(s,\tau)\nu(s,\tau). \end{array} \right)$$

4 Stochastic annuities

Definition 1 *Let:*

$$E = \{1, 2, \dots, m\}$$

be the states of a system and A, B two persons. Furthermore, let

$$\mathbf{S} = \{S_1, S_2, \dots, S_m\}\ ,\ S_i \in \mathbb{R}$$

be sums. The sums **S** *represent the instalments of the annuity. The instalment S_i will be paid or received from it A to it B if the system is in the state i. The instalment will be given for each period of the contractual time horizon. We say that this financial operation is a discrete time homogeneous constant stochastic annuity if:*

i) the transitions among the states are governed by a homogeneous discrete time Markov Chain $\mathbf{P} = [p_{ij}]$

ii) when there is a transition from i to j it is possible that is paid or received a sum γ_{ij}.

The sums γ_{ij} are named transition payments.

- The annuity will be respectively *immediate* if the payments of the ψ_i are scheduled at the *end* of the period and *due* at the *beginning.*

- The annuity is *non-homogenous* if the Markov chain is non-homogeneous. In this case it results $\mathbf{P}(t) = [p_{ij}(t)]$

- The annuity can be *variable* if the instalments and/or the transition payments change during the time horizon. In the non-homogeneous case the sums paid or received can vary also in function of the *starting* time of the financial operation.

It is useful to report the following

Remark 1 *If there is a single state then the discrete time stochastic annuity corresponds to the usual concept of discrete time annuity.*

Remark 2 *The concepts of homogeneous and non-homogeneous discrete time stochastic annuity correspond respectively to the ones of discrete time homogeneous and non-homogeneous Markov reward processes.*

Definition 2 *Under the same condition of* **Definition 1** *we have the generalized case if the statement i) becomes:*

i') the transitions among the states are governed by a homogeneous discrete time semi-Markov Chain with kernel $\mathbf{Q}(t) = [Q_{ij}(t)]$

- The generalized stochastic annuity is *non-homogenous* if the semi-Markov chain is non-homogeneous, and the kernel becomes $\mathbf{Q}(s,t) = [Q_{ij}(s,t)]$.

Remark 3 *The concepts of homogeneous and non-homogeneous discrete time generalized stochastic annuity correspond respectively to the ones of discrete time homogeneous and non-homogeneous semi-Markov reward processes.*

5 Multiple state insurance models and discrete time Markov and semi-Markov reward processes

The definition of multiple state insurance models corresponds with the definition of graph see [Haberman and Pitacco, 1999]. A multiple state model corresponds with a graph that describes the transitions among the states of the considered problem. The transition matrix describes the multiple state insurance models in the homogeneous case. In non homogeneous case a sequence of transition matrices describes the multiple state model. Premiums and benefits can be considered as rewards. The evolution of a general multiple state insurance model could be studied by means of Markov or semi-Markov models under the property that future is function only of the present. As well known, in discrete time the Markov process has the property that the time interval between two subsequent transitions is always the same. In the semi-Markov case the time between two transitions is a random variable. Some times an insurance contract can be studied well by means of a Markov process, some times the Markov environment is necessary because the transition are scheduled at each period (there is no randomness in the transition times), i.e. motorcar insurance. But in general in insurance problems the semi-Markov environment fits better than Markov one. In fact in the most part of insurance contracts the time of transition is stochastic. It is clear that in this light a multiple state insurance problem should be dealt in a better way by semi-Markov models. The reward processes gives the possibility to take into account directly the benefits and premiums that are considered in the multiple state models. Furthermore, usually, the insurance models are non-homogeneous respect the age of the insured person. It could be possible to use continuous time semi-Markov processes see [CMIR12, 1991] to construct multiple state insurance models. The problem in this case is that the solution of evolution equation is a very difficult task and that the analytical solution, excluding few particular cases not useful in the real problems, is impossible to find. The way could be the numerical solution of the evolution equation. But as it was shown in [Janssen and Manca, 2001], the numerical discretization corresponds to the discrete time processes. Summarizing we think that the best way to solve the multiple state insurance problem under Markov hypotheses is given by the application of DTNHSMRWP. In some cases the Markov environment suffices or it is necessary. Usually the problem should be faced in non-homogeneous environment. To construct non-homogeneous Markov or semi-Markov chains it is necessary to have huge amount of data that some times are not available, in these cases homogeneous environment should be used. Now considering what we state in the previous section we can affirm that *each multiple state insurance model can be considered a stochastic or a generalized stochastic annuity* depending on the insurance contract to be modelled. This statement confirms the fact that insurance problems should be considered as a generalization of financial problems in which the stochastic aspects assume great relevance.

6 Conclusions

In the paper the description of discrete time homogeneous and non-homogeneous semi Markov processes was given. After the concepts of Markov and semi-Markov reward processes were presented. The definitions of stochastic annuity and generalized stochastic annuity have been presented. The strict relation between the annuities and the reward processes was outlined.

All the paper moved in a discrete time approach because the applications are more suitable in this environment.

The paper should be seen as a theoretic step of these topics, for this reason there are no applications. The applications were presented in some less general paper see [Janssen and Manca, 2004]. In a near future the authors hope to study in depth the applicative aspects of the concepts given in this paper.

References

[CMIR12, 1991]CMIR12 (Continuous Mortality Investigation Report, number 12).,*The analysis of permanent health insurance data.* 1991 The Institute of Actuaries and the Faculty of Actuaries.

[De Dominicis *et al.*, 1986]De Dominicis, R, Manca, R Some new results on the transient behaviour of semi-Markov reward processes.*Methods of Operations Research*, 1986 54.

[Haberman and Pitacco, 1999]Haberman, S. and Pitacco, E., *"Actuarial models for disability Insurance"*, 1999, Chapman and Hall.

[Howard, 1971]Howard, R.,*Dynamic probabilistic systems*, 1971 vol I II, Wiley.

[Iosifescu Manu, 1972]Iosifescu Manu A., Non homogeneous semi-Markov processes,*Stud. Lere. Mat.* 1972 24, 529-533.

[Janssen, 1966]Janssen, J., Application des processus semi-markoviens à un probléme d'invalidité,*Bulletin de l'Association Royale des Actuaries Belges*, 1966 63, 35-52.

[Janssen and Manca, 2001]Janssen J. and Manca R., Numerical solution of non homogeneous semi-Markov processes in transient case.*Methodology and Computing in Applied Probability.* 2001 Kluwer, 3, 271-293.

[Janssen and Manca, 2004]Janssen, J. and Manca, R., Discrete Time Non-Homogeneous Semi-Markov Reward Processes, Generalized Stochastic Annuities and Multi-State Insurance Model. *Proc. of XXVIII AMASES* Modena 2004.

[Wolthuis, 2003]Wolthuis H.,*Life Insurance Mathematics (the Markovian Model)*, II edition, 2003 Peeters Publishers, Herent.

Part X

Health

Product-limit estimators of the survival function with left or right censored data

Valentin Patilea[1] and Jean-Marie Rolin[2]

[1] CREST-ENSAI
Campus de Ker-Lann
Rue Blaise Pascal - BP 37203
35172 Bruz cedex, France
(e-mail: patilea@ensai.fr)
[2] Institut de Statistique
Université Catholique de Louvain
20, voie du Roman Pays
1348 Louvain-la-Neuve, Belgique
(e-mail: rolin@stat.ucl.ac.be)

Abstract. The problem of estimating the distribution of a lifetime when data may be left or right censored is considered. Two models are introduced and the corresponding product-limit estimators are derived. Strong uniform convergence and asymptotic normality are proved for the product-limit estimators on the whole range of the observations. A bootstrap procedure that can be applied to confidence intervals construction is proposed.
Keywords: bootstrap, delta-method, martingales, left and right censoring, strong convergence, weak convergence.

1 Introduction

A great deal of recent attention in survival analysis has focused on estimating the survivor distributions in the presence of various and complex censoring mechanisms. The goal of this paper is to analyze simple models for lifetime data that may be left or right censored. Typically, a lifetime T is left or right censored if, instead of observing T we observe a finite nonnegative random variable Y, and a discrete random variable with values 0, 1 or 2. By definition, when $A = 0$, $Y = T$, when $A = 1$, $Y < T$ and, when $A = 2$, $Y > T$. Models for left or right censored data were proposed by [Turnbull, 1974], [Sampath and Chandra, 1990] and [Huang, 1999]. See also [Gu and Zhang, 1993], [Kim, 1994].

Assume that the sample consists of n independent copies of (Y, A) and let F_T be the distribution of the lifetime of interest T. Using the plug-in (or substitution) principle, the nonparametric estimation of F_T is straight as soon as F_T can be expressed as an explicit function of the distribution of (Y, A). The existence of such a function requires a precise description of the censoring mechanism that is generally achieved by introducing 'latent' variables and by making assumptions on their distributions. In this paper, two latent models allowing for explicit inversion formula, that is closed-form function relating F_T to the distribution of (Y, A), are proposed.

In some sense, our first latent model lies between the classical right-censorship model and the current-status data model. It may be applied to the following framework. Consider a study where T the age at onset for a disease is analyzed. The individuals are examined only one time and they belong to one of the following categories: (i) evidence of the disease is present and the age at onset is known (from medical records, interviews with the patient or family members, ...); (ii) the disease is diagnosed but the age at onset is unknown or the accuracy of the information about this is questionable; and (iii) the disease is not diagnosed at the examination time. Let C denote the age of the individual at the examination time. In the first case the exact failure time T (age at onset) is observed, that is $Y = T$. In case (ii) the failure time T is left-censored by C and thus $Y = C$, $A = 2$. Finally, the onset time T is right-censored by C for the individuals who have not yet developed the disease; in this case $Y = C$, $A = 1$. If no observation as in (ii) occurs, we are in the classical right-censorship framework, while if no uncensored observation is recorded we have current-status data. Our first latent model can be applied, for instance, with the data sets analyzed by [Turnbull and Weiss, 1978].

The second latent model proposed is closely related to the first one. It lies between the left-censorship model and the current-status data model. Consider the example of a reliability experiment where the failure time of a type of device is analyzed. A sample of devices is considered and a single inspection for each device in the sample is undertaken. Some of them already failed without knowing when (left censored observations). To increase the precision of the estimates, a proportion of the devices still working is selected randomly and followed until failure (uncensored observations). For the remaining working devices the failure time is right censored by the inspection time.

Let us point out that, without any model assumption, given a distribution for the observed variables (Y, A) with $Y \geq 0$ and $A \in \{0, 1, 2\}$, we can always apply our two inversion formulae. In this way we build two pseudo-true distribution functions of the lifetime of interest which are functionals of the observed distribution. If the experiment under observation is compatible with the hypothesis of one of our latent models, the true F_T can be exactly recovered from the observed distribution. Otherwise, we can only approximate the true lifetime distribution.

The paper is organized as follows. Section 2 introduces our two latent models through the equations relating the distribution of the observations to those of the latent variables. Solving these equations for F_T we deduce the inversion formulae. The product-limit estimators are obtained by applying the inversion formulae to the empirical distribution. Section 2 is ended with some remarks and comments on related models. Section 3 contains the asymptotic results for the first latent model (similar arguments apply for the second model). We prove strong uniform convergence for the product-limit estimator on the whole range of the observations. Our proof extends and simplifies the results of [Stute and Wang, 1994] and [Gill, 1994] provided in the case of the Kaplan-Meier estimator. Next, the asymptotic normality of our product-limit estimator is obtained. The variance of the limit Gaussian process being complicated, a bootstrap procedure for which the asymptotic validity is a direct consequence of the delta-method is proposed.

2 The latent models

2.1 Model 1

The survival time of interest is T (e.g., the age at onset). Let C be a censoring time (e.g., the age of the individual at the examination time) and $\varDelta$ be a Bernoulli random variable. Assume that the latent variables T, C and $\varDelta$ are independent. The observations are independent copies of the variables (Y, A), with $Y \geq 0$ and $A \in \{0, 1, 2\}$. These variables are defined as

$$Y = \min(T, C) + (1 - \varDelta) \max(C - T, 0) = C + \varDelta \min(T - C, 0)$$

and $A = 2(1 - \varDelta)\mathbf{1}_{\{T \leq C\}} + \mathbf{1}_{\{C < T\}}$, where $\mathbf{1}_A$ denotes the indicator function of the set A. With this censoring mechanism the lifetime T is observed, right censored or left censored. In view of the definitions of Y and A, note that if $\varDelta$ is constant and equal to one (resp. zero), we obtain right censored (resp. current status) data.

Let F_T and F_C denote the distributions of T and C, respectively. Let $p = P\left(\varDelta = 1\right)$. Define the observed subdistributions of Y as

$$H_k\left(B\right) = P\left(Y \in B,\ A = k\right), \qquad k = 0, 1, 2, \tag{1}$$

for any B Borel subset of $[0, \infty]$. As usually in survival analysis, the censoring mechanism defines a map $\varPhi$ between the distributions of the latent variables and the observed distributions. For the censoring mechanism we consider, the relationship $(H_0, H_1, H_2) = \varPhi\left(F_T, F_C, p\right)$ between the subdistributions of Y and the distributions of the latent variables T, C and $\varDelta$ is the following:

$$\begin{cases} H_0(dt) = p\, F_C\left([t, \infty]\right) F_T(dt) \\ H_1(dt) = F_T\left((t, \infty]\right) F_C(dt) \\ H_2(dt) = (1 - p)\, F_T\left([0, t]\right) F_C(dt) \end{cases} . \tag{2}$$

Remark that when $p = 1$ (resp. $p = 0$) the equations (2) boil down to the equations of the classical independent right-censoring (resp. current status) model.

By plug-in applied with the empirical distribution, the nonparametric estimation of the distribution of T is straight as soon as the map $\varPhi$ is invertible and F_T can be written as an explicit function of the observed subdistributions H_k, $k = 0, 1, 2$. The model considered allows us an explicit inversion formula for F_T. In order to derive this inversion formula, integrate the first and the second equation in (2) on $[t, \infty]$ and deduce

$$H_0([t,\infty]) + pH_1([t,\infty]) = pF_T\left([t,\infty]\right) F_C\left([t,\infty]\right). \tag{3}$$

For $t = 0$, it follows that

$$p = \frac{H_0([0,\infty])}{1 - H_1([0,\infty])} = \frac{H_0([0,\infty])}{H_0([0,\infty]) + H_2([0,\infty])}. \tag{4}$$

Recall that the hazard measure associated to a distribution F is $\Lambda(dt) = F(dt)/F([t,\infty])$. In our case, use (2)-(3) to deduce that the hazard function corresponding to F_T can be written as

$$\Lambda_T(dt) = \frac{H_0(dt)}{H_0([t,\infty]) + pH_1([t,\infty])}. \tag{5}$$

Finally, the distribution F_T can be expressed as

$$F_T\left((t,\infty]\right) = \prod_{[0,t]} (1 - \Lambda_T(ds)), \tag{6}$$

where $\prod_{[0,t]}$ is the product-integral on $[0,t]$. Note that there is no explicit formula for F_T if $p = 0$ in equations (2), that is with current status data.

Given the explicit relationship between the distribution of T and the observed subdistributions, to obtain the product-limit estimator of F_T, we simply replace H_k, $k = 0, 1, 2$ by their empirical counterparts. Let $\widehat{F}_T$ denote the product-limit estimator of F_T.

2.2 Model 2

As in Model 1, assume that T, C and Δ are independent. The observations are independent copies of the variables (Y, A), with $Y \geq 0$ and $A \in \{0, 1, 2\}$ where

$$\begin{cases} Y = T, & A = 0 \;\; if \quad 0 \leq C \leq T \quad and \quad \Delta = 1; \\ Y = C, & A = 1 \;\; if \quad 0 \leq C \leq T \quad and \quad \Delta = 0; \\ Y = C, & A = 2 \;\; if \quad \;\; 0 \leq T < C. \end{cases} \tag{7}$$

The equations of this model are

$$\begin{cases} H_0(dt) = p\, F_C\left([0,t]\right) F_T(dt) \\ H_1(dt) = (1-p)\, F_T\left([t,\infty]\right) F_C(dt) \\ H_2(dt) = F_T\left([0,t)\right) F_C(dt) \end{cases}. \tag{8}$$

Remark that when $p = 1$ (resp. $p = 0$) the equations (8) boil down to the equations of the classical independent left-censoring (resp. current status) model. This model also allows for an explicit inversion formula for F_T. By integration in the first and the third equation in (8), $H_0([0,t])+pH_2([0,t]) = pF_T([0,t])F_C([0,t])$. Deduce

$$p = \frac{H_0([0,\infty])}{1-H_2([0,\infty])}.$$

Recall that given a distribution F, the associated reverse hazard measure is $M(dt) = F(dt)/F([0,t])$. By equations (8) deduce that the reverse hazard function M_T associated to F_T can be written as

$$M_T(dt) = \frac{H_0(dt)}{H_0([0,t]) + pH_2([0,t])}.$$

Finally, the distribution F_T can be expressed as

$$F_T\left([0,t]\right) = \prod_{(t,\infty]} (1 - M_T(ds)).$$

Applying the inversion formula with the empirical subdistributions, we get the product-limit estimator of F_T in Model 2.

Note that if $\widetilde{T} = h(T)$ and $\widetilde{C} = h(C)$, with $h \geq 0$ a decreasing transformation, then $\widetilde{T}$, $\widetilde{C}$ and $\varDelta$ are the variables of Model 1 applied to the left or right censored lifetime $h(Y)$. In other words, Model 2 is equivalent to Model 1, up to a time reversal transformation.

2.3 Related models

[Huang, 1999] introduced a model for the so-called partly interval-censored data, Case 1; see also [Kim, 1994]. In such data, for some subjects, the exact failure time of interest T is observed. For the remaining subjects, only the information on their current status at the examination time is available. [Huang, 1999] considered the nonparametric maximum likelihood estimator (NPMLE) of F_T. Unfortunately, NPMLE does not have an explicit form and therefore Huang needs strong assumptions for deriving its asymptotic properties and a numerical algorithm for the applications. Let us point out that, on contrary to our Model 1 (resp. Model 2), in Huang's model one may observe exact failure times even if failure occurs after (resp. before) the examination time. Moreover, in Huang's model one may still obtain a $\sqrt{n}-$consistent estimator of the distribution F_T if one simply considers the empirical distribution of the uncensored lifetimes. This is no longer true in our models.

Perhaps, the most popular model for left or right-censored data is the one introduced by [Turnbull, 1974]; see also [Gu and Zhang, 1993]. In Turnbull's model there are three latent lifetimes L (left-censoring), T (lifetime of interest) and R (right-censoring) with $L \leq R$. The observed variables are $Y = \max(L, \min(T, R)) = \min(\max(L, T), R)$ and A defines as follows: $A = 0$ if $L < T \leq R$; $A = 1$ if $R < T$; and $A = 2$ if $T \leq L$. The equations of this model are

$$\begin{cases} H_0(dt) = \{F_R([t,\infty]) - F_L([t,\infty])\}\ F_T(dt) \\ H_1(dt) = F_T((t,\infty])\ F_R(dt) \\ H_2(dt) = F_T([0,t])\ F_L(dt) \end{cases},$$

where $H_k,\ k = 0, 1, 2$ are defined as in (1) and F_T, F_L and F_R are the distributions of T, L and R, respectively. The NPMLE of the distribution of the failure time T is not explicit but it can be computed, for instance, by iterations based on the so-called self-consistency equation. Note that imposing $F_C(dt) = (1-p)^{-1}F_L(dt) = F_R(dt)$, one recovers the equations of Model 1. However, for the applications we have in mind, there is no natural interpretation for such a constraint in Turnbull's model. Moreover, we derive a product-limit estimator for our Model 1. Finally, the asymptotic results below are much simpler and they are obtained under weaker conditions than in Turnbull's model.

3 Asymptotic results

In this section the strong uniform convergence and the asymptotic normality for the estimator of the distribution F_T in Model 1 are derived. Moreover, we propose a bootstrap procedure that can be used to build confidence intervals for F_T. As in the previous sections, the distributions F_T and F_C need not be continuous. For simpler notation, hereafter, the subscript T is suppressed when there is no possible confusion. We write $\widehat{F}$ (resp. F, $\widehat{\Lambda}$ and Λ) instead of $\widehat{F}_T$ (resp. F_T, $\widehat{\Lambda}_T$ and Λ_T).

3.1 Strong uniform convergence

Let H_{nk} be the empirical counterparts of the subdistributions H_k, $k = 0, 1, 2,,$ that is

$$H_{nk}([0,t]) = \sum_{i=1}^{n} \mathbf{1}_{\{Y_i \leq t,\ A_i = k\}}, \qquad k = 0, 1, 2.$$

Clearly, $\sup_{t \geq 0} |H_{nk}([0,t]) - H_k([0,t])| \to 0$, almost surely. We want to prove the strong uniform convergence of the distribution $\widehat{F}$, that is

$$\sup_{t \in I} \left| \widehat{F}([0,t]) - F([0,t]) \right| \to 0, \quad asn \to \infty, \qquad almost\ surely,$$

where $I = \{t : H_0([t,\infty]) + pH_1([t,\infty]) > 0\}$. For this purpose, first we prove the almost sure convergence of the hazard function.

Theorem 1 *Assume that* $p \in (0,1]$ *and let* $t_* = \sup I$. *For any* $\sigma \in I$,

$$\sup_{0 \le t \le \sigma} \left| \widehat{\Lambda}([0,t]) - \Lambda([0,t]) \right| \to 0, \qquad \text{as } n \to \infty, \qquad \text{almost surely.}$$

Moreover, if $t_* \notin I$ *and* $\Lambda([0,t_*)) < \infty$, *then* $\widehat{\Lambda}([0,t_*)) \to \Lambda([0,t_*))$, *almost surely.*

The strong uniform convergence of the distribution $\widehat{F}$ follows without any additional assumption.

Theorem 2 *Assume that* $p \in (0,1]$. *Then*

$$\sup_{t \in I} \left| \widehat{F}([0,t]) - F([0,t]) \right| \to 0, \qquad \text{as} \quad n \to \infty, \quad \text{almost surely.}$$

With $p = 1$ one recovers the strong uniform convergence result for the Kaplan-Meier estimator obtained by [Stute and Wang, 1994], [Gill, 1994]. Our alternative proof is simpler.

3.2 Asymptotic normality

Here we study the weak convergence of the process $\sqrt{n}(\widehat{F} - F)$ where $\widehat{F}$ is the product-limit estimator of Model 1. In this case, $\widehat{\Lambda}$ does no longer have a martingale structure (in t) as in the case of the Nelson-Aalen estimator, that is when $p = 1$. However, a continuous time *submartingale* property for $\widehat{\Lambda}$ can be obtained. This suffices us to extend the techniques of Gill (1983) and to use them in combination with the functional delta-method in order to establish the weak convergence of $\sqrt{n}(\widehat{F} - F)$ to a Gaussian process. Here, the weak convergence is denoted by $\Rightarrow$. The space $D[a,b]$ of càdlàg functions defined on $[a,b]$ is endowed with the supremum norm and the ball $\sigma-$field.

Theorem 3 *Assume that* $p \in (0,1]$ *and define* $U(t) = \sqrt{n}(\widehat{F}([0,t]) - F([0,t]))$, $t \ge 0$. *Let* $t_* = \sup I$.

a) Let τ *be a point in* I. *Then,* $U \Rightarrow \mathbf{G}$ *in* $D[0,\tau]$, *where* $\mathbf{G}$ *is a Gaussian process.*

b) If $t_* \notin I$, *but*

$$\int_{[0,t_*)} \frac{H_0(dt)}{\{H_0([t,\infty]) + pH_1([t,\infty])\}^2} < \infty, \tag{1}$$

then $\mathbf{G}$ *can be extended to a Gaussian process on* $[0,t_*]$ *and* $U \Rightarrow \mathbf{G}$ *in* $D[0,t_*]$.

The proof of the weak convergence is postponed to the appendix. Note that when $t_* \notin I$, condition (1) is equivalent to

$$F_T([t_*,\infty]) > 0 \qquad \text{and} \qquad \int_{[0,t_*)} \frac{F_T(dt)}{F_C([t,\infty])} < \infty. \tag{2}$$

3.3 Bootstrapping the product-limit estimator

Theorem 3 may be used to obtain confidence intervals and confidence bands for F. However, the law of the process $\widetilde{\mathbf{G}}(t) = \mathbf{G}(t)/F\,((t,\infty])$ being complicated, one may prefer a bootstrap method in order to avoid handling this process in practical applications. Here, a bootstrap sample is obtained by simple random sample with replacement from the set of observations $\{(Y_i, A_i) : 1 \leq i \leq n\}$. Let $\{(Y_i^*, A_i^*) : 1 \leq i \leq n\}$ denote a bootstrap sample and let H_k^* be the corresponding subdistributions. Apply equations (4) to (6) to obtain the bootstrap estimator $\widehat{F}^*$. The following theorem state that the bootstrap works almost surely for our product-limit estimator on any interval $[0,\tau]$ such that $H_0([\tau,\infty]) + pH_1([\tau,\infty]) > 0$. This result is a simple corollary of Theorem 3.9.13 of [Van der Vaart and Wellner, 1996] and it is based on the uniform Hadamard differentiability of the maps involved in the inversion formula of Model 1.

Theorem 4 *Let $\tau \in I$ and let $\widetilde{\mathbf{G}}(t)$ be the limit of $U(t)/F\,((t,\infty])$ in $D[0,\tau]$, as obtained from Theorem 3. Then, the process*

$$\sqrt{n}\{\widehat{F}^*([0,t]) - \widehat{F}([0,t])\}/\widehat{F}\,((t,\infty])$$

converges to $\widetilde{\mathbf{G}}$ in $D[0,\tau]$, almost surely.

References

[Gill, 1994]R. Gill. Lectures on survival analysis. *Lecture Notes in Mathematics (Ecole d'été de Probabilités de Saint-Flour XXII 1992)*, pages 115–241, 1994.

[Gu and Zhang, 1993]M.G. Gu and C.-H. Zhang. Asymptotic properties of self-consistent estimators based on doubly censored data. *Ann. Statist.*, pages 611–624, 1993.

[Huang, 1999]J. Huang. Asymptotic properties of nonparametric estimation based on partly interval-censored data. *Statist. Sinica*, pages 501–519, 1999.

[Kim, 2003]J.S. Kim. Maximum likelihood estimation for the proportional hazards models with partly interval-censored data. *J. R. Stat. Soc. Ser B*, pages 489–502, 2003.

[Samuelsen, 1989]S.O. Samuelsen. Asymptotic theory for non-parametric estimators from doubly censored data. *Scand. J. Statist.*, pages 1–21, 1989.

[Stute and Wang, 1994]W. Stute and J. Wang. The strong law under random censorship. *Ann. Statist.*, pages 1591–1607, 1994.

[Turnbull and Weiss, 1978]B.W. Turnbull and L. Weiss. A likelihood ratio statistics for testing goodness of fit with randomly censored data. *Biometrics*, pages 367–375, 1978.

[Turnbull, 1974]B.W. Turnbull. Nonparametric estimation of a survivorship function with doubly censored data. *J. Amer. Statist. Assoc.*, pages 169–173, 1974.

[Van der Vaart and Wellner, 1996]A.W. Van der Vaart and J.A. Wellner. *Weak convergence and empirical processes.* Springer Verlag, New-York, 1996.

Stochastic Models Applied in Health Care and Medical Education

Augustin Prodan[1], Madalina Rusu[1],
Rodica Prodan[2], and Remus Campean[1]

[1] Iuliu Hatieganu University
Str. Emil Isac 13, 400023 Cluj-Napoca, Romania
(e-mail: {aprodan,mrusu,rcampean}@umfcluj.ro)

[2] MedFam Group
Str. Constanta 5, 400158 Cluj-Napoca, Romania
(e-mail: familiaprodan@yahoo.com)

Abstract. The paper presents a Java framework for stochastic modelling and simulation, used as an infrastructure to model and simulate real world activities, phenomena and processes, particularly in health care, patient monitoring and medical education. We modelled the flow of patients through medical units, considering both their arrivals and their stay in the hospital. Also, we implemented bootstrapping methods, which are quite useful in simulation studies. We created bootstrapping e-tools for simulating laboratory works and experiments, to be used in both didactic and research activities.

Keywords: stochastic model, distributional model, simulation, bootstrapping methods, Java framework.

1 Introduction

Previous research has shown that stochastic models are advantageous tools for representation of real world activities, phenomena and processes. Due to actual spread of fast and inexpensive computational power everywhere in the world, the best approach is to model a real phenomenon as faithfully as possible, and then rely on a simulation study to analyze it. Based on theoretical fundamentals in stochastic modelling and simulation [Ross, 1990], we implemented an object-oriented Java framework for stochastic modelling, analysis and simulation of problems arising in a practical context, particularly in medicine and pharmacy. We created an infrastructure, consisting of a collection of Java class libraries, which are used to model and to simulate distributional models, stochastic processes and Monte Carlo methods. The basic design philosophy of our object-oriented approach to simulation of the random variables by means of distributional models is presented in [Prodan *et al.*, 1999]. The object-oriented Java framework containing the set of baseline classes for stochastic modelling and simulation is presented in [Prodan and Prodan, 2001]. We implemented *bootstrapping methods* [Hesterberg *et al.*, 2003], then we created and implemented bootstrapping e-tools with the purpose of simulating laboratory works and experiments, in both didactic and research activities [Prodan and Campean, 2004]. There are two reasons for creating such e-tools: (a) to reduce the number of animals (guinea pigs, frogs, etc.) used in experimentation (an ethical reason), and (b) to reduce the consumption of substances and reactants (an economical reason). Using a bootstrapping e-tool, the experimenter can repeat the original experiment on computer, obtaining pseudo-data as plausible as those obtained from the original experiment.

2 The infrastructure

We created an infrastructure consisting of a set of Java class libraries for stochastic modelling and simulation. The classical random variables are the simplest stochastic models, also called *distributional models*, which enter into the composition of other complex models. We propose a hierarchy of Java classes for modelling the classical distributions. Each distribution class encapsulates a particular *simValue*() method (see Figure 1) incorporating a simulation algorithm, able to generate a specific value for that distribution. In other words, the simulation algorithms for distributional models are implemented via a polymorphic method called *simValue*(). The particular implementation in case of each simulation algorithm is based on one or more of the following techniques: the Inverse Transform Technique, the Acceptance-Rejection Technique and the Composition Technique (see [Ross, 1990] and [Prodan *et al.*, 1999]).

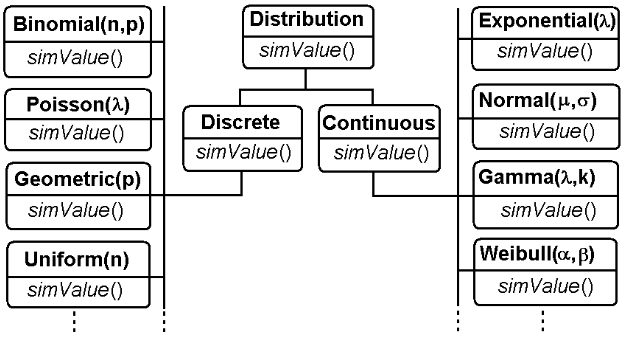


Fig. 1. The hierarchy of Java classes for distributional models

We consider three levels of simulation (Figure 2). The first level consists of simulating random numbers, as they are the basis of any stochastic simulation study. Using this first level, we build the second level, applied for classical distributions, for stochastic processes and for Monte Carlo methods. The third level of simulation is devoted to applications. As applications, we modelled activities, processes and phenomena from health care, patient monitoring and pharmacy, we created e-learning tools (see [Prodan and Prodan, 2003] and [Prodan, 2004]) and we implemented bootstrapping methods [Prodan and Campean, 2004].

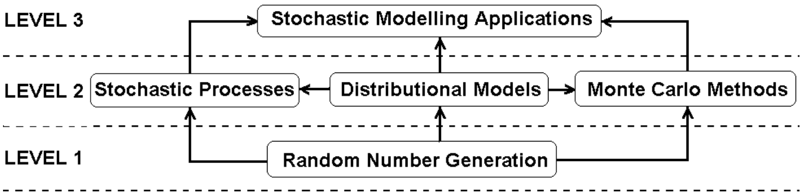


Fig. 2. Levels of simulation

To make a simulation study, it is necessary to generate more values, a sequence of values. One may choose to continually generate additional values, stopping when the efficiency of the simulation is good enough. Generally, one may use the variance of the estimators obtained during the simulation study to decide when to stop the generation of additional values. For example, if the objective is to estimate the mean value $\mu = E(X), i = 0, 1, 2, ...$, one may continue to generate new data values until one has generated n data values for which the estimate of the standard error (i.e. the standard deviation of the mean) is less than an acceptable value. We implemented a general simulation class as a *canvas*, which encapsulates the methods *doSimulation*() and *doVisualisation*(). These methods are inherited by specific simulation classes for specific distributions (binomial, exponential, etc), which encapsulates specific *redraw*() methods, able to show the results of specific simulations. Figure 3 shows the results of simulations from some discrete and continuous distributional models.

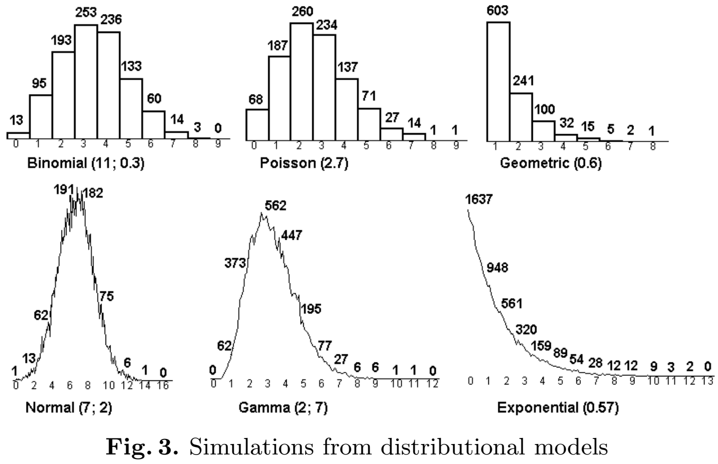
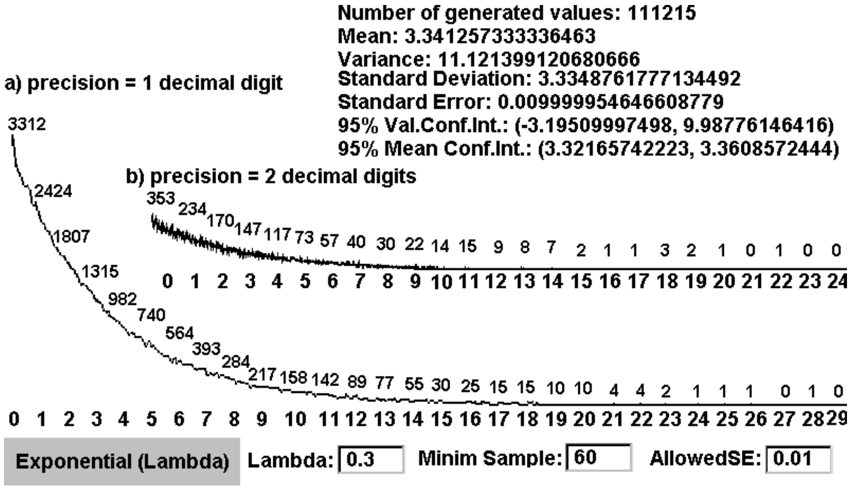


Fig. 3. Simulations from distributional models

When simulate from a continuous random variable X, a generated value $x \in X$ is approximated with a given *precision* expressed by the number of decimal digits to be considered. The user has the possibility to choose a precision of one, two, or more decimal digits. If no decimals are considered, the real value x is approximated by integer part of the x, the continuous random variable X is rudely approximated by a discrete one, and the results of a simulation can be graphically expressed in a segmented line format. If a precision of one decimal digit is selected, the results of the same simulation is more precisely visualized by a more refined segmented line. With a precision of two decimal digits, a more refined visualization is obtained. The higher this precision is, the higher is the *resolution* realized in visualization [Prodan and Prodan, 2002]. Figure 4 compares two visualizations for the same set of generated values from the exponential distribution with parameter $\lambda = 0.3$, the first visualization being with a precision of one decimal digit (Figure 4, graph a), and the second with a precision of two decimal digits (Figure 4, graph b).



```
Number of generated values: 111215
Mean: 3.341257333336463
Variance: 11.121399120680666
Standard Deviation: 3.3348761777134492
Standard Error: 0.009999954646608779
95% Val.Conf.Int.: (-3.19509997498, 9.98776146416)
95% Mean Conf.Int.: (3.32165742223, 3.3608572444)
```

Lambda: 0.3 Minim Sample: 60 AllowedSE: 0.01

Fig. 4. Visualization with precision of one decimal digit, versus visualization with precision of two decimal digits, for the same set of generated values from the distribution *Exponential*(0.3)

As can be seen in this figure, when the precision grows with one decimal digit, the resolution grows ten times. With a precision of one decimal digit, ten numbers are considered between two successive integers, while if the precision is of two decimal digits, one hundred numbers are considered between two successive integers. When necessary, intermediate resolutions can be considered.

3 A model for patient flow simulation

Chronic patients may generally be thought of as progressing through two standard stages: firstly, the *acute care*, consisting of diagnosis, assessment and rehabilitation and secondly, the *long-stay care* where a small proportion of them remains in hospital for months or even years. Obviously, these patients may be very consuming of resources, situation which implies a serious analysis of health care costs in order to avoid the distortion of the performance statistics.

We applied stochastic processes to model the flow of patients through chronic diseases departments. The science of best designing the movement of patients through hospitals has not yet been discovered, but the use of the queuing theory models may provide a good enough solution to the problem. We intend to use results from both the queuing theory, particularly, and stochastic modelling, generally, in order to optimize the bed inventory and the cost-effectiveness of a hospital system. We describe patients arrivals by a Poisson process, hospital beds by the servers and the lengths of stay are modelled using *phase-type distributions.* In queuing terminology this is known as a $M/Ph/c/N$ queue, where M denotes Poisson (Markov) arrivals, the service distribution is phase-type, c is the number of servers (i.e. beds) and N represents the finite capacity of the system, comprising both waiting patients and patients being served [Gorunescu and Prodan, 2001]. It is also assumed that the queuing system is in steady state which, in practical terms, means that we assume that the hospital system has been running, in its present form, for a few years. This model enables us to study the whole system of geriatric medicine and is used to look either at the time patients spend in hospital, or at the subsequent time patients spend in the community.

In order to simulate the model, we have split it into two parts: the arrival of patients and the in-patient care. We modelled the arrival of patients as a Poisson process with a parameter λ estimated by using the inter-arrival times. These times are independent exponential random variables, each with the parameter λ and with the corresponding density function $f(t) = \lambda e^{-\lambda t}$. Figure 5 shows the results of a simulation, considering the arrival of patients as a Poisson process at rate $\lambda = 7.25$ patients per day.

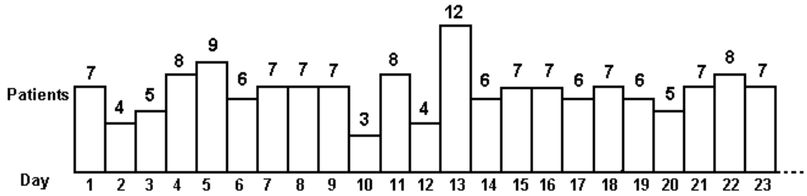


Fig. 5. Poisson arrivals at rate $\lambda = 7.25$ patients per day

The care time is modelled by the application of a mixed-exponential distribution, where the number of terms in the mixture corresponds to the number of stages of patient care. A common scenario is that there are two stages for in-patient care: *acute* and *long-stay*. In this case we compose two exponential distributions with parameters α and β, representing the access rates for the corresponding stages. The mixed-exponential phase-type distribution has the probability density function $f(t) = \rho\alpha e^{-\alpha t} + (1-\rho)\beta e^{-\beta t}$, which imply a mean care time of $\frac{\rho}{\alpha} + \frac{(1-\rho)}{\beta}$ days per patient. Figure 6 shows the results of a simulation with parameters $\rho = 0.07$, $\alpha = \frac{1}{77.18}$ and $\beta = \frac{1}{33.3}$.

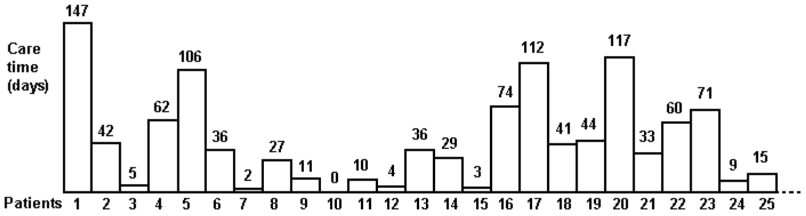


Fig. 6. The simulated results for in-patient care time

4 E-learning tools for medical education

Based on the infrastructure presented in section 2, we created e-learning tools and incorporated them into an e-learning environment, to be used by both the students and the teaching staff in their didactic and research activities. We implemented e-learning scenarios by looking at problems that can be put in a probabilistic framework. Every new concept is developed systematically through completely worked out examples from current medical and pharmaceutical problems. In addition, we introduced in each e-learning scenario specific probability models that fit out some real life problems, by assessing the probabilities of certain events from actual past databases.

As an example, we propose an e-learning scenario for students in Medicine. A learner that traverses such a scenario, will be able to apply a binomial distributional model in studying the chance of patients suffering from a particular type of cancer, to survive for at least a six month period after diagnosis. We would have to appeal to previous studies and information from actual databases to assess the chances of a patient surviving. This might indicate, for instance, that the probability of survival is $p = 0.3$, and consequently the complementary probability of death is $q = 1 - p = 0.7$. In real life, we are frequently interested what might happen to a group of patients we are studying. Therefore, we may formulate the following problem, as a piece of the current e-learning scenario:

Of the 11 patients in a particular cancer program, what is the chance of 7 or more of them surviving at least six months past diagnosis?

If p_k is the probability that k patients survives ($k \leq 11$), the solution is given by the sum $P = p_7 + p_8 + p_9 + p_{10} + p_{11}$. The distributional model *Binomial*(11, 0.3) gives the values for p_k ($p_7 \approx 0.017$, $p_8 \approx 0.003$ and $p_9 \approx p_{10} \approx p_{11} \approx 0$), hence the solution is $P \approx 0.017 + 0.003 = 0.02$. The e-learning scenario may be resourceful in showing additional information about probabilities and statistics. We prepare and configure suggestive visualizations, based on a friendly and efficient dialogue with the learner. As an example, for any learner may be useful to see the previous probabilities p_k in a suggestive column format, and to recognize the solution to previous problem shown with dashed columns (see Figure 7).

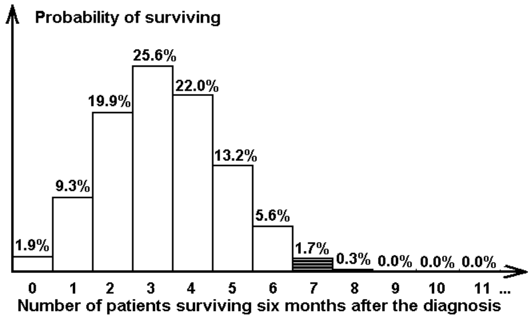


Fig. 7. A graphical solution for patient surviving problem (dashed columns)

An e-learning scenario combines simulation with interactive visualization and allows the learners to explore the knowledge bases with some well-defined learning purposes. We define a simulation class and a visualization class for each application object. These classes are then configured to obtain a particular simulation with a specific visualization. In an e-learning scenario, visualization is an active part of the system, serving as an additional interface for modifying dynamically some parameters. For example, the same distributional model *Binomial*(11, 0.3) may be applied in an e-learning scenario for students in Pharmacy, studying the effect of digitalis on frogs. Suppose we know from previous studies and experiments, that injection of a certain dose of digitalis per unit of body weight into a large number of frogs, causes death of 30% of them. We may propose the following problem, as a piece of the current e-learning scenario for students in Pharmacy (similar with that proposed for students in Medicine):

If this dose of digitalis is injected into each of a group of 11 frogs, what is the probability that the number of deaths will be 7 or more?

To answer this question, it is used the same distributional model as for students in Medicine, and the solution is given by the same sum of probabilities. The numerical result is the same, because the binomial template is the same, with the same values for parameters, but with specific texts (see Figure 8).

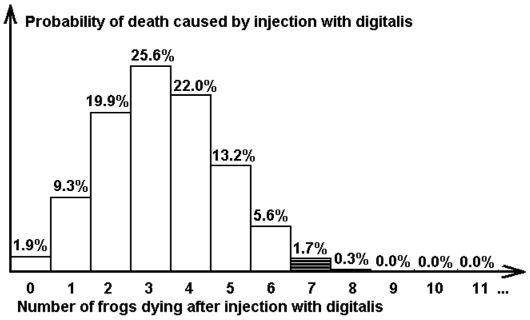


Fig. 8. A graphical solution for students in Pharmacy (dashed columns)

5 The implementation of the bootstrapping e-tools

We implemented *bootstrapping methods* and we created bootstrapping e-tools [Prodan and Prodan, 2002] for simulating laboratory works and experiments. Both the students and the teaching staff use traditional statistical methods to infer the truth from sample data gathered in laboratory experiments. However, the repeated laboratory experiments mean the consumption of a great deal of substances and reactants. At the same time, there are some ethically motivated reasons to reduce the number of animals (guinea pigs, frogs, etc.) used in experimentation. Using a bootstrapping tool and the computer power, the experimenter can repeat the original experiment on computer, obtaining pseudo-data as plausible as those obtained from the original experiment.

Based on distributional models available in JAR (Java ARchive) libraries as infrastructure, we implement in bootstrapping e-tools both parametric and non-parametric bootstrapping methods. When we can not assume the distribution of the population from which the original sample v is taken from, we use the non-parametric bootstrap. When we can make safely assumptions about the distribution of v, we may use the parametric bootstrap. We may use the sample v to calculate a statistic of interest θ^* that is an estimator of some population parameter θ. If we could obtain more samples, we evaluate the estimator on each of these samples. In fact, we only have the one actual sample to work with, so the idea of bootstrapping is to simulate not from the population, but from the single actual sample which we have available.

We have to simulate the population and to generate more so called *bootstrap samples*, or *re-samples*, then to calculate the statistic of interest θ^* for each re-sample, named the *bootstrap replication.* The bootstrapping e-tools provide a set of procedures and functions for re-sampling, for hypothesis testing and for obtaining standard errors, confidence intervals and other measures of uncertainty. The basic bootstrap reassigns randomly the original data and recalculates the estimates. As a computer-intensive method, the bootstrap repeats these reassignments and recalculations thousands of times, treating them as repeated experiments. Using a bootstrapping tool and the computer power, one can repeat the original experiment as many times as necessary to satisfy didactic and research activities.

Generally, we implemented the following algorithm for a bootstrapping e-tool:

i) Read the actual sample $v = (v_1, v_2, \ldots, v_n)$ and evaluate the empirical distribution function F_e.

ii) Simulate N independent bootstrap samples $v^{*1}, v^{*2}, \ldots, v^{*N}$, each containing n data values drawn with replacement from v, based on distribution F_e.

iii) Evaluate the bootstrap replication $\theta^*(k)$ corresponding to each bootstrap sample v^{*k}, for $k = 1, 2, ..., N$.

iv) Estimate the standard error by the sample standard deviation of the N replicates.

The distribution of the statistic of interest θ^* is called *bootstrap distribution.* The bootstrap distribution gives information about the shape, center, and spread of the corresponding population parameter θ.

6 Conclusions and future work

We presented a Java framework for stochastic modelling and simulation, used as an infrastructure to create models and to simulate real world activities, phenomena and processes, particularly in health care, patient monitoring and medical education. As future work, we will combine stochastic modelling with new AI (Artificial Intelligence) paradigms, such as Bayesian inference, intelligent agents and case based reasoning, for simulations and for incorporating intelligent strategies in e-learning scenarios. We will write all simulation and visualization classes in Java and will use the XML (eXtensible Markup Language) format to describe the configurations.

In cooperation with Pharmaceutical Technologies Department of the our university, we have to apply bootstrapping methods in modelling and simulation of some drug design experiments. Real experimental data and simulated pseudo-data refer to some tests made for the characterization of drugs with delayed action, so called retard drugs. This approach is useful, the purpose being to improve real data with simulated valid pseudo-data and to reduce the number of actual tests.

References

[Gorunescu and Prodan, 2001]F. Gorunescu and A. Prodan. *Modelare Stochastica si Simulare*. Microinformatica, Cluj-Napoca, 1st edition, 2001.

[Hesterberg *et al.*, 2003]T. Hesterberg, S. Monaghan, D. S. Moore, A. Clipson, and R. Epstein. *Bootstrap Methods and Permutation Tests*. W. H. Freeman and Company, New York, 1st edition, 2003.

[Prodan and Campean, 2004]A. Prodan and R. Campean. Bootstrapping e-tools for simulating laboratory works. In *ICICTE'2004, International Conference on Information Communication Technologies in Education*, pages 489–494, Samos Island, 2004. University of Athens.

[Prodan and Prodan, 2001]A. Prodan and Rodica Prodan. Stochastic simulation and modelling. In *ETK-NTTS'2001, Exchange of Technology and Know-how - New Techniques and Technologies for Statistics*, pages 461–466, Crete Island, 2001. JRC-ISIS, European Commission.

[Prodan and Prodan, 2002]A. Prodan and R. Prodan. A collection of Java class libraries for stochastic modelling and simulation. In *ICCS'2002, International Conference on Computational Science*, volume 1, pages 1040–1048, Amsterdam, 2002. Springer.

[Prodan and Prodan, 2003]A. Prodan and R. Prodan. A Java framework for intelligent and practical e-learning tools. In *ICICTE'2003, International Conference on Information Communication Technologies in Education*, pages 45–52, Samos Island, 2003. University of Athens.

[Prodan *et al.*, 1999]A. Prodan, F. Gorunescu, and R. Prodan. Simulating and modelling in Java. In *POOSC'99, Workshop on Parallel/High-Performance Object-Oriented Scientific Computing*, pages 55–64. Zentralinstitut für Angewandte Mathematik, 1999.

[Prodan, 2004]A. Prodan. An intelligent and practical educational environment. In *CBLIS'2003, Computer Based Learning in Science*, pages 229–239, Nicosia, 2004. University of Cyprus.

[Ross, 1990]S. M. Ross. *A Course in Simulation*. Maxwell Macmillan, New York, 1st edition, 1990.

Parametric and Non Homogeneous semi-Markov Process for HIV Control

Eve Mathieu[1], Yohann Foucher[1], Pierre Dellamonica[2], and Jean-Pierre Daures[1]

[1] Clinical Research University Institute. Biostatistics Laboratory.
641 av. D.G. Giraud.
34093 Montpellier , France (e-mail: emathieu@iurc.montp.inserm.fr)

[2] Infectious Disease Department
Archet Hospital
BP 3079. 06202 Nice , France

Abstract. In AIDS control, physicians have a growing need to use pragmatically useful and interpretable tools in their daily medical taking care of patients. In that sense, semi-Markov process seems to be well adapted to model the evolution of HIV-1 infected patients. In this study, we introduce and define a Non Homogeneous semi-Markov Model (NHSMM) in continuous time. Then the problem of finding the equations that describe the biological evolution of patient is studied and the interval transition probabilities are computed. A parametric approach is used and the maximum likelihood estimators of the process are given. As results, follow-up time has an impact on the evolution of patients and interval transition probabilities are computed.

Keywords: Semi-Markov process, Non homogeneity, Maximum likelihood estimation, Right censored data, interval transition probabilities.

1 Introduction

The CD4 count and the VL measurement are both fundamental markers of the state of an HIV-1 infected patient. The potential of these immunological and virological reservoirs determines the way the patients are handled. In the context of HIV, it seems reasonable to think that the probability of a patient's transition from one state to another depends on how long he has spent in this state. Therefore the semi-Markov Models (SMM) seem to be appropriated [Janssen and Limnios, 1999].

The SMM have been considered in the HIV modelling [Wilson and Solomon, 1994], [Satten and Sternberg, 1999], [Joly and Commenges, 1999]. These models were time Homogeneous semi-Markov Models (HSMM) and unidirectional. Nowadays it seems to be appropriated to take into account the impact of the follow-up time on the patients' evolution. The goal of this paper is to formulate a Non Homogeneous semi-Markov Model (NHSMM) of the HIV biological process and to compute its interval transition probabilities. The NHSMM have found many applications, in breast cancer [Davidov, 1999], [Davidov and Zelen, 2000] in manpower system [Papadopoulou and Vassiliou, 1999],

[Vassiliou and Papadopoulou, 1992],[Papadopoulou, 1998], [McClean *et al.*, 1998] and [Janssen and Manca, 2001].

This paper is organized as follows. In the next section, the model and associated notation are introduced. Section 3 defines the semi-Markovian interval transition probabilities and solves integral equations. In section 4, the emi-Markov process is parametrically modelled and the likelihood function is built. Section 5 illustrates an application to HIV control. Finally, section 6 is a summary and discussion.

2 Model description and Notation

The natural history of HIV infection can be considered as a series of stages through which a patient progresses. Based both on currently information and physicians' opinion, we have taken four immunological and virological states: state 1 ($VL \leq 400$ and $CD4 \leq 200$), state 2 ($VL \leq 400$ and $CD4 > 200$), state 3 ($VL > 400$ and $CD4 > 200$), state 4 ($VL > 400$ and $CD4 \leq 200$). Patients move thought these four states according ten transitions given in figure 1.

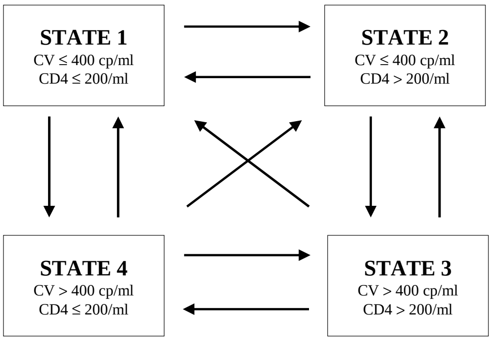


Fig. 1. An HIV Multi-state model, with 4 immunological and virological states and 10 transitions.

More formally, let $E = \{1, 2, 3, 4\}$ be the state space and $(\Omega, \mathcal{F}, P)$ be a probability space. We define the following random variables [Janssen and Manca, 2001]:

$$J_n : \Omega \rightarrow E, \qquad S_n : \Omega \rightarrow [0, +\infty),$$

where J_n represents the state at the n-th transition and S_n represents the chronological time of the n-th transition. Let $N(t)$ be the counting process

$(N(t), t \geq 0)$ associated to the point process $(S_n)_{n \in \mathrm{N}}$ defined for any time $t \geq 0$ by :

$$N(t) = \sup\{n : S_n \leq t\}.$$

The random variable $N(t)$ represents the number of transitions occured in the interval of time $[0, t]$. Let us define the $(X_n)_{n \in \mathrm{N}}$ '*duration process*' by :

$$\begin{aligned} &X_0 = 0, \\ &X_{n+1} = S_{n+1} - S_n, \quad n \in \mathrm{N}^* \end{aligned}$$

where X_{n+1} represents the duration time spent in state J_n.

The $(J_n, S_n)_{n \in \mathrm{N}}$ process is called '*non-homogeneous Markov renewal process*' if :

$$P(J_{n+1} = j, S_{n+1} \leq t|\ J_n = i, S_n = s, J_{n-1}, S_{n-1}, ..., J_0, S_0) = P(J_{n+1} = j, S_{n+1} \leq t|J_n = i, S_n = s),$$

and for $j \neq i$

$$Q_{ij}(s,t) = P(J_{N(s)+1} = j, S_{N(s)+1} \leq t|J_{N(s)} = i, S_{N(s)} = s),$$

is the associated *non-homogeneous semi-Markov kernel* Q. The semi-Markov kernel is written again :

$$Q_{ij}(s,x) = P(J_{N(s)+1} = j, X_{N(s)+1} \leq x|J_{N(s)} = i, S_{N(s)} = s).$$

The second composant of Q, namely x, represents a duration time whereas s represents a chronological time.

As is well known [Wadjda, 1992],

$$\begin{aligned} p_{ij}(s) &= \lim_{x \to \infty} Q_{ij}(s,x), \quad i, j \in E, j \neq i \\ &= P(J_{N(s)+1} = j|J_{N(s)} = i, S_{N(s)} = s), \end{aligned}$$

represents the probability of a patient making its next transition to state j, given that he entered state i at time s and $\mathbf{P}(s) = [p_{ij}(s)]_{i,j}$ is the (4×4) transition probability matrix of the *embedded non-homogeneous Markov chain* $(J_n)_{n \in \mathrm{N}}$.

However, before the entrance into j, the patients 'holds' for a time x in state i. The conditional cumulative distribution function of the waiting time in each state, given the state subsequently occupied, is defined by :

$$F_{ij}(s,x) = P(X_{N(s)+1} \leq x|J_{N(s)+1} = j, J_{N(s)} = i, S_{N(s)} = s).$$

This probability function is obtained by :

$$F_{ij}(s,x) = \begin{cases} \frac{Q_{ij}(s,x)}{p_{ij}(s)} & \text{if } p_{ij}(s) \neq 0 \\ 1 & \text{if } p_{ij}(s) = 0 \end{cases}$$

and for more feasability, it is supposed free of the chronological time s, namely $F_{ij}(x)$. Without loss of generality, the waiting time also has a probability density function, namely $f_{ij}(x)$ and $\mathbf{D}(x) = [f_{ij}(x)]_{i,j}$ represents the (4×4) duration matrix.

Let introduce the probability that the process stays in state i for at least a duration time x, given state i is entered at chronological time s :

$$H_i(s,x) = P(X_{N(s)+1} \leq x | J_{N(s)} = i, S_{N(s)} = s).$$

Of course,

$$H_i(s,x) = \sum_{j \neq i}^{4} Q_{ij}(s,x) = \sum_{j \neq i}^{4} p_{ij}(s) F_{ij}(x).$$

Therefore, the marginal cumulative distribution functions of the waiting time in each state depend on both time. Let us define $S_i\,(s,x) = 1 - H_i(s,x)$.

Now it is possible to define the continuous time *non homogeneous semi Markov process* $Z(t)$, which represents, for each time t, the state occupied by the process [Cox and Isham, 1980], [Janssen, 1986], as :

$$Z(t) = J_{N(t)}, \quad t \in \mathrm{R}_+.$$

with :

$$P[Z(t) = j] = P[S_{N(t)} \leq t < S_{N(t)+1}, J_{N(t)} = j].$$

This SM process is both characterized by a set of Markov transition matrices $\{\mathbf{P}(t)\}_{t \geq 0}$, and a set of duration matrices $\{\mathbf{D}(x)\}_{x \geq 0}$. Note that two time scales arise, the chronological time and the internal time scales. The chronological time, namely t, is relative to an arbitrary origin. In our case, $t = 0$ represents the first immunological and virological measurement experimented by the patient in hospital. The internal time, namely x, is relative to the duration time in each state [Davidov and Zelen, 2000]. Our model is quite simple and completely defined by both the jump and duration processes. The advantage of semi-Markov model is their mathematical tractability and simple interpretation. The SMM presented in this section is non homogeneous with time since the jump process $(p_{ij}(t))_{i,j,t \geq 0}$ depends on the chronological time.

3 Interval transition probabilities

In the perspective of a more and more effective taking care of patients, physicians need tools of prediction and reference points. Let us define, $\forall i, j = 1, ..., 4$, $\phi_{ij}(t,x)$ as the following probability [Papadopoulou and Vassiliou, 1999] :

$$\begin{aligned} \phi_{ij}(t,x) &= P\left[\text{a patient is in state } j \text{ at time } t+x \mid \text{he entered state } i \text{ at time } t\right]. \\ &= P\left[Z(t+x) = j \mid J_{N(t)} = i \ ; \ S_{N(t)} = t\right] \end{aligned}$$

These probabilities are real quantities of interest in the medical practice. Let us precise that $\phi_{ij}(t,x) \neq \phi_{ij}(t+h, x+h), \forall h > 0$. We now turn on the question of developing a functional relationship between the probabilities $\phi_{ij}(t,x)$, which from now on we call the interval transition probabilities of the SM process, and the probabilities $p_{ij}(t)$ and $d_{ij}(x)$.This could be done by taking all the possible mutually exclusive ways in which it is possible for the event of interest to take place. With careful reasoning we could prove that $\forall t, x \geq 0$:

$$\phi_{ij}(t,x) = \delta_{ij} \times S_{i.}(t,x) + \sum_{\substack{l=1 \\ l \neq i}}^{4} \int_{0}^{x} p_{il}(t) d_{il}(u) \phi_{lj}(t+u, x-u) du.$$

This equation represents the evolution equation of a continuous NHSMM. Let $c_{il}(t,x)$ be the product $p_{il}(t)d_{il}(x)$. Then the previous equation is written:

$$\phi_{ij}(t,x) = \delta_{ij} \times S_{i.}(t,x) + \sum_{\substack{l=1 \\ l \neq i}}^{4} \int_{0}^{x} c_{il}(t,u) \phi_{lj}(t+u, x-u) du. \tag{1}$$

Obviously $\phi_{ij}(t,0)$=0 for $j \neq i$, 1 otherwise. Using probabilistic arguments, we could find probabilities $\phi_{ij}(t,x)$ in closed analytic form. Let k be the index of the number of transitions in the interval of time $]t, t+x[$, and let $t+x_1, t+x_1+x_2, ..., t+ x_1+x_2+...+x_k$ be the chronological times where they successively occur. Then the equation (1) is written as follows

$$\begin{aligned}
\phi_{ij}(t,x) = {} & \delta_{ij} \times S_{i.}(t,x) \\
& + \textstyle\int_0^x c_{ij}(t,x_1) S_{j.}(t+x_1, x-x_1) dx_1 \\
& + \textstyle\sum_{\substack{l=1 \\ l\neq i, l\neq j}}^4 \int_0^x \int_0^{x-x_1} c_{il}(t,x_1) c_{lj}(t+x_1,x_2) S_{j.}(t+x_1+x_2, x-x_1-x_2) dx_2 dx_1 \\
& + \textstyle\sum_{k=3}^{\infty} \sum_{\substack{l=1 \\ l\neq i}}^4 \sum_{\substack{m=1 \\ m\neq l}}^4 \cdots \sum_{\substack{w=1 \\ w\neq v}}^4 \int_0^x \int_0^{x-x_1} \cdots \int_0^{x-x_1-x_2-...-x_{k-1}} \\
& \qquad\qquad c_{il}(t,x_1) c_{lm}(t+x_1,x_2) ... c_{wj}(t+x_1+x_2+...+x_{k-1}, x_k) \\
& \times S_{j.}(t+x_1+x_2+...+x_{k-1}+x_k, x-x_1-x_2-...-x_{k-1}-x_k) dx_k ... dx_2 dx_1.
\end{aligned} \tag{2}$$

This previous expression formalizes the fact that the event of interest {a patient of the NHSMM is in in state j at time $t+x$, given he entered state i at time t} may be derived from no transition ($k=0$) or from exactly one transition ($k=1$) or from exactly two transitions ($k=2$) or more ($k \geq 3$). Let us define $\phi_{ij}^k(t,x)$ by the following probability:

$$\begin{aligned}
\phi_{ij}^k(t,x) = P[\text{patient in state } j \text{ at } t+x; k \text{ transitions during }]t, t+x[\\
\mid \text{he entered state } i \text{ at time } t \ ] \ .
\end{aligned}$$

Finally the equation (2) can be written $\forall i, j \in \{1,2,3,4\}$

$$\phi_{ij}(t,x) = \sum_{k=0}^{\infty} \phi_{ij}^k(t,x) \tag{3}$$

and in matrix form, with $\mathbf{\Phi}(t,x) = \left(\phi_{ij}(t,x)\right)_{i,j}$ and $\mathbf{\Phi}^k(t,x) = \left(\phi_{ij}^k(t,x)\right)_{i,j}$

$$\mathbf{\Phi}(t,x) = \sum_{k=0}^{\infty} \mathbf{\Phi}^k(t,x). \tag{4}$$

4 The likelihood function

Over a period of time, M patients are observed ($p = 1, ..., M$). Each patient begins his immunological and virological trajectory in any state, which is revealed by the first measurement at time $s = 0$. Let us assume that the p^{th} subject changes state $(n_p - 1)$ times in the instants $s_{p,1} < s_{p,2} < ... < s_{p,n_p-1}$ and successively occupies states $J_{p,1}$, $J_{p,2}$,...,J_{p,n_p-1} with $J_{p,n} \neq J_{p,n+1}$, $\forall n \geq 1$. At the last observed time of the follow-up, namely s_{p,n_p} , the patient either may enter a new state J_{p,n_p} or stay in the state J_{p,n_p-1}. In the last case, the last duration time in state J_{p,n_p-1} is right censored. More generally, the contribution for an observed transition $i \to j$, after a duration time x spent in state i, equals $p_{ij}(t) f_{ij}(x)$, namely the probability P[duration time $= x; next = j|$ state i is entered at time t]. If the transition from state i is right censored, after a staying time x, then the contribution is the function $S_i(t,x)$. The likelihood function for all times and transition times observed, is written as follows

$$L = \prod_{p=1}^{M} \prod_{n=1}^{n_p} [p_{J_{p,n-1},J_{p,n}}(s_{p,n-1}) f_{J_{p,n-1},J_{p,n}}(s_{p,n} - s_{p,n-1})]^{\xi_{p,n}} [S_{J_{p,n-1}}(s_{p,n-1},\ s_{p,n} - s_{p,n-1})]^{1-\xi_{p,n}}$$

where $\xi_{p,n} = 1$, if the n^{th} transition is observed for the individual p, and $\xi_{p,n} = 0$ if censored. Our parametric approach for both jump and duration processes consists respectively in a linear and a Weibull modelings

$$p_{ij}(t|\theta_{ij}) = a_{ij} t + b_{ij} \qquad \forall j \neq i \tag{5}$$

$$p_{ii}(t) = 0 \qquad \forall i = 1, ..., 4$$

$$f_{ij}(x|\gamma_{ij}) = \nu_{ij} \sigma_{ij}^{\nu_{ij}} x^{\nu_{ij}-1} Exp[-(\sigma_{ij} x)^{\nu_{ij}}] \qquad \forall j \neq i \tag{6}$$

5 Application to HIV control

In this section, we apply the previous parametric NHSMM to an HIV-1 infected patients database. The database NADIS is made of patients followed in the Nice Hospital, France. The study sample is made of 1313 patients and 17888 virologic and immunologic measurements. The chronological time is measured from the first biological measurement. From the modelings (5) and (6), we test several restrictions in order to select the parametric model which offers the best adequacy (Likelihood Ratio Test). The selected parametric NHSMM is based on both exponential and Weibull duration times, but also on time linear and constant probabilities.The estimations of the NHSMM parameters are given in Table 1.

Transition	Estimators of the duration process	Estimators of the jump process
$i \to j$	$d_{ij}(x)$	$p_{ij}(t)$
$1 \to 2$	*Weibull* (1.1069 , 1.6795)	$(0.0450 \times t)+$ 0.4748
$1 \to 3$	*Weibull* (1.4460 , 1.8283)	0.1111
$1 \to 4$	*Weibull* $(1.0971, 1.7254)$	$(-0.0450 \times t) - 0.4141$
$2 \to 1$	*Weibull* (0.5878 , 0.0940)	$(-0.0213 \times t)+$ 0.3148
$2 \to 3$	*Weibull* $(1.0500, 0.8844)$	$(0.0213 \times t) + 0.6852$
$3 \to 2$	*Expo* (1.0841)	0.8496
$3 \to 4$	*Weibull* $(0.7842, 0.7597$)	0.1504
$4 \to 1$	*Weibull* $(0.9095, 1.0556)$	$(-0.0276 \times t)+$ 0.4779
$4 \to 2$	*Weibull* $(1.1866, 1.5765)$	0.1605
$4 \to 3$	*Expo* (1.8410)	$(0.0276 \times t)+$ 0.3616

Table 1. Estimations of parameters in the NHSMM defined by the linear jump process $\{p_{ij}(t)\}_{i,j}$ and the duration process $\{d_{ij}(x)\}_{i,j}$

Mathematical computing was preformed on R software version 1.9.1. The standard error deviations are not presented for more lisibility. The real quantities of interest are the semi-Markovian interval transition probabilities defined in Section 3. Indeed in medical practice, physicians are often interested in predictions. In this view, the 4×4 matrix of the interval transition probabilities for fixed chronological time t and duration time x,given by equation (4) in section 3, are useful. For exemple, the estimations of $\mathbf{\Phi}(0,1)$ are given in Table 2.

state j state i	1	2	3	4
1	0.2059	0.3212	0.2182	0.1935
2	0.0336	0.6530	0.2623	0.0210
3	0.0219	0.4311	0.4673	0.0094
4	0.1464	0.2557	0.2266	0.2649

Table 2. The 4×4 interval transition matrix $(\phi_{ij}(0,1))_{i,j}$

Given a patient enters state 2 at $t = 0$, he has a 0.652 probability to be 1-year later in state 2; Given a patient enters state 3 at $t = 0$, there is a quasi equiprobability to be 1-year later in state 2 or 3. Lastly, given a patient enters state 4 at $t = 0$, there is a quasi equiprobability to be 1-year later in state 2, 3 or 4.

6 Discussion

The HIV model considered in this study clearly relates to a 'macroscopic' view of the disease process and it is based both on the CD4 count and VL measurement. This multi-state model is made of 4 immunological and virological states and 10 transitions. The non homogeneous semi-Markov model captures the main features of the disease process and therefore provides a reasonable approximation of a very complicated process. The homogeneity hypothesis reveals to be too restrictive in the HIV context which nowadays becomes a chronic disease. The follow-up time has a significant impact on the disease process. We use a parametric approach and compute the maximum likelihood estimators of the NHSMM. The integral evolution equations of the continuous NHSMM are solved and the interval transition probabilities are computed. Therefore physicians have interesting reference points and some predictions can be made as regards the biological evolution of patients. Here are the three characteristics of a good model which should be mathematically tractable, pragmatically useful and interpretable.

References

[Cox and Isham, 1980]D.R. Cox and V. Isham. *Point Processes.* Chapman and hall, 1980.

[Davidov and Zelen, 2000]O. Davidov and M. Zelen. Designing cancer prevention trials: a stochastic model approach. *Statistics in Medicine*, pages 1983–1995, 2000.

[Davidov, 1999]O. Davidov. The steady-state probabilities for regenerative semi-markov processes with application to prevention and screening. *Applied Stochastic Models and Data Analysis*, pages 55–63, 1999.

[Janssen and Limnios, 1999]J. Janssen and N. Limnios. *Semi-Markov Models and Applications.* Kluwer Academic publishers, 1999.

[Janssen and Manca, 2001]J. Janssen and R. Manca. Numerical solution of non-homogeneous semi-markov processes in transient case. *Methodology and Computing in Applied Probability*, pages 271–293, 2001.

[Janssen, 1986]J. Janssen. *Semi Markov Models. Theory and Applications.* Plenum press, 1986.

[Joly and Commenges, 1999]P. Joly and D. Commenges. A penalized likelihood approach for a progressive three-state model with censored and truncated data: application to aids. *Biometrics*, pages 887–890, 1999.

[McClean *et al.*, 1998]S. McClean, E. Montgomery, and F. Ugwuowo. Non-homogeneous continuous-time markov and semi-markov manpower models. *Applied Stochastic Models and Data Analysis*, pages 191–198, 1998.

[Papadopoulou and Vassiliou, 1999]A.A. Papadopoulou and P.C.G. Vassiliou. Continuous time non homogeneous semi-markov systems. In J. Janssen and N. Limnios, editors, *Semi-Markov Models and Applications*, pages 241–251, 1999.

[Papadopoulou, 1998]A.A. Papadopoulou. Counting transitions-entrance probabilities in non-homogeneous semi-markov systems. *Applied Stochastic Models and Data Analysis*, pages 199–206, 1998.

[Satten and Sternberg, 1999]G.A. Satten and M.R. Sternberg. Fitting semi-markov models to interval-censored data with unknown initiation times. *Biometrics*, pages 507–513, 1999.

[Vassiliou and Papadopoulou, 1992]P.C.G. Vassiliou and A.A. Papadopoulou. Non-homogeneous semi-markov systems and maintainability of the state sizes. *J. Appl. Prob.*, pages 519–534, 1992.

[Wadjda, 1992]W. Wadjda. Uniformly strong ergodicity for non-homogeneous semi-markov processes. *Demonstration Mathematica*, pages 755–764, 1992.

[Wilson and Solomon, 1994]S.R. Wilson and P.J. Solomon. Estimates for different stages of hiv/aids disease. *Comput Appl. Biosci*, pages 681–683, 1994.

Modelling the recurrence of bladder cancer

Gregorio Rubio[1], Cristina Santamaría[1], Belén García[1], and José Luis Pontones[2]

[1] Matemática multidisciplinar
Universidad Politécnica
Valencia (Spain)
(e-mail: crisanna@imm.upv.es, magarmo5@imm.upv.es, grubio@imm.upv.es)

[2] Hospital Universitario La Fe
Valencia (Spain)
(e-mail: pontones_jos@gva.es)

Abstract. The aim of this paper is to evaluate the risk of tumor recurrence of bladder cancer after surgical operation (TUR: Trans-urethral Resection). The prognostic significance of some clinical features in 454 patients with primary superficial bladder carcinoma is studied. The modelling procedure is featured within interval censored and right censored framework.

Keywords: bladder carcinoma, prognostic factors, recurrence, interval–censored survival data, generalized non-linear model, Cox model.

1 Introduction

Transitional bladder cancer represents about 2% of all human tumors. It supposes an important public health problem because it is biologically very aggressive and causes more than 130.000 deaths by year all around the world. Superficial bladder tumors are characterized by *recurrence* (reappearance of a new tumor) in 50-70% of cases. Although most recurrences are still superficial, *progression* to muscle invasive disease occurs in 10-30% of patients. Therefore, when superficial bladder tumor is diagnosed, it is important to identify patients who are at risk of disease recurrence and progression. If it were possible to define exactly which subset of superficial bladder tumors have more risk to recur and to progress, preemptive therapy could be used. Identifying the prognostic factors that determine that risk in each patient remains a subject of extensive research [Jaemal *et al.*, 2003], [Black *et al.*, 2002] and [Royston *et al.*, 2002].

Biotechnological advances have allowed us to use different therapeutic procedures (surgery, radiotherapy, chemotherapy, immunotherapy) successfully but still many patients suffer an unfavorable outcome without control of disease.

Multiple clinical and pathological variables are important in predicting outcome in patients with transitional bladder cancer, among which pathological stage and grade of differentiation are recognized as the most important [Zieger *et al.*, 1998], [Kurth *et al.*, 1995]. Therefore, an ideal prediction model should combine stage, and grade, along with any other features shown to be associated with outcome in a multivariate model (histological characteristics, size, number of tumors, etc).

The TNM system (classification of 1997) is generally used to establish the stage of the bladder tumors [Hermanek and Sobin, 1998]:

Tis : tumor is limited to the mucosa and is flat (a carcinoma in situ).
Ta : tumor is papillary and it is limited to the mucosa.
T1 : tumor penetrates the lamina propia but not the muscle layer.
T2-T4 : tumor invades muscle and is staged from T2 to T4 according to the depth of infiltration of muscle tissue or the extent to which the surrounding tissue is affected.

Superficial bladder tumors (stages Ta and T1) have trend to produce recurrences (generally with similar stage). Tumors that invade the bladder muscle are highly aggressive and have a strong potential metastasize preferentially to regional lymph nodes, lungs, liver, and bone.

The *histologic grade* establishes according to the WHO (World Health Organization) 1999 classification [Hermanek and Sobin, 1998]:
G1: Urothelial carcinoma grade I (differentiated)
G2: Urothelial carcinoma grade II (intermediate differentiation)
G3: Urothelial carcinoma grade III (poor differentiated)

Well differentiated tumors (G1 grade) have generally low agressivity while poor differentiated tumors (G3 grade) are highly aggressive (cause many recurrences) [Millan *et al.*, 2000].

Prediction models can be used to counsel patients, determine the need for adjuvant therapy, stratify patients in risk groups, and develop appropriate postoperative surveillance programs tailored to risk for cancer progression. There are quite a few models in the medical literature, see [Millan *et al.*, 2000] for a little account. Nevertheless, many studies are based only on univariate analysis. Even if multivariate analysis is performed, usually the event of interest, for instance tumor recurrence, is recorded at scheduled screening times. It may be more convenient to consider arbitrarily interval-censored survival data because the exact time of the event of interest is not known. Our aim is to construct a prognostic model for predicting the outcome of superficial bladder cancer of transitional cells, within this framework. Then we perform the usual Cox model approach in order to compare.

In our study the time origin concern to the so called TUR (trans-urethral resection): a surgical endoscopic technique used to remove the macroscopic tumor from the inner of the bladder. The end-point is the first tumor recurrence.

The paper is organized as follows: in Section 2 the data on the survival times of 454 patients and their characteristics (explanatory variables) are described. In Section 3 we give a brief description of a method for analyzing interval–censored data proposed by Farrington [Farrington, 1996], and we apply the method to our data base. In Section 4, a multivariate analysis is performed by using the *Cox proportional hazards* model.

We have used the packages S-PLUS ([Venables and Ripley, 2002]), SPSS and SAS ([Delwiche and Slaughter, 1998]).

2 Data and selection of variables

In this research, 454 patients from *La Fe* University Hospital from Valencia (Spain) were examined. They had primary superficial transitional cell carcinoma of the bladder initially treated with transurethral resection (TUR). The variable of interest was time (in days) from TUR to the first appearance of recurrence. The exact time of the recurrence will be unknown and the only information available concerns whether or not recurrence is identified when a patient visits the clinic. So, each individual may have a different time interval in which the recurrence has occurred and data are referred to as arbitrarily *interval-censored data.*

The period goes from 1973 to 2003. Variables considered for this study were: sex, age, tumor stage (pTa and pT1), tumor grade (G1, G2 and G3), number of tumors (one or more than one), tumor size (≤ 3 cm or > 3 cm) and treatment (Thiotepa, Adriamicine, Cisplatine, BCG and others treatments), see Table 1

3 Interval–censored analysis

The method for analyzing such data, assuming proportional hazards, is based on a non-linear model for binary data. The model is known as a *generalized non-linear model*[Farrington, 1996], see [Collett, 2003]:

The likelihood function for n observations may be expressed as:

$$\prod_{i=1}^{n+c} p_i^{y_i}(1-p_i)^{1-y_i} \tag{1}$$

where $y_1, y_2, \ldots, y_{n+c}$ are observations from a Bernoulli distribution with response probability p_i, $i = 1, 2, \ldots, n+c$, where c is the number of confined observations.

The survivor function is given by:

$$S_i(t) = S_0(t)^{exp(\beta' x_i)} \tag{2}$$

where $S_0(t)$ is the baseline survivor function and x_i is the vector of values of p explanatory variables for the i*th* individual, $i = 1, 2, \ldots, n$. The baseline survivor function will be modelled as a step function, where the steps occur at the k ordered censoring times, $t_1, t_2, \ldots, t_k$, where $t_1 < t_2 < \ldots < t_k$ (subset of times at which observations are interval–censored).

Variable	N patients	(%)
Stage		
pTa	114	25.1
pT1	340	74.9
Grade		
G1	260	57.3
G2	162	35.7
G3	32	7.0
Sex		
Men	383	84.4
Women	71	15.6
Number		
One	380	83.7
Two or more	74	16.3
Size		
$\leq$ 3 cm	357	78.6
$>$ 3 cm	97	21.4
Age		
$\leq$ 40 years	20	4.4
between 41 y 60 years	150	33
$>$ 61 years	284	62.6
Treatment		
Thiotepa	257	56.6
Adriamicine	33	7.3
Cisplatine	21	4.6
BCG	62	13.7
Others treatments	81	17.8

Table 1. Patients characteristics.

This methodology defines the following baseline survivor function:

$$S_0(t) = exp\ (-\sum_{j=1}^{k} \theta_j\ d_{ij}) \tag{3}$$

where $d_{ij} = 1$ if $t_j \leq t_i$, $d_{ij} = 0$ if $t_j > t_i$ and θ_j are given by:

$$\theta_j = \log \frac{S_0(t_{(j-1)})}{S_0(t_{(j)})} \tag{4}$$

Then it follows that the response probability can be expressed in the form:

$$p_i = 1 - exp\ (-exp\ (\beta' x_i) \sum_{j=1}^{k} \theta_j\ d_{ij}) \tag{5}$$

This leads to a *generalized non-linear model* for a binary response variables, with values y_i, and corresponding probabilities p_i, for $i = 1, 2, \ldots, n+c$. The model contains $k+p$ unknown parameters. After fitting the model, the statistic $-2 \log \hat{L}$ can be used to compare alternative manner.

Patients were followed up at clinic visits, generating observations as follows: 69 *left–censored*; 216 *right–censored* and the remaining patients are confined.

For the survivor function model, a minimal set of censoring times was chosen. The set of ordered censoring times is 50, 171, 261, 343, 399, 579, 674, 851, 1046, 1290, 1427, 1524, 1750, 2069, 2290, 2633, 2953, 3365, 3768, 5287.

We use the statistic $-2 \log \hat{L}$ in a strategy of selection of variables. We obtain number, tumor size and treatment as prognostic factors.

On fitting the model with tumor size and number of tumors the value of the statistic $-2 \log \hat{L}$ is 1498.4. On adding *Treatment* to the model, the value of this statistics is reduces to 1471.7. This reduction is significant at the 1% level.

Parameter	$\hat{\beta}$	Exp($\hat{\beta}$)	se($\hat{\beta}$)
two or more	0.2933	1.3408	0.1719
$>$ 3 cm	0.3651	1.4406	0.1524
Cisplatine	0.3212	1.3787	0.2327
BCG	0.1279	1.1364	0.3314
ADR	0.6428	1.9017	0.1872
Others treatments	0.0733	1.0760	0.2027

Table 2. Generalized non-linear model. Parameters estimates

Using this model we may conclude that the relative hazard of first recurrence after TUR is increased in a 90% if adriamicine is provided, relative to a patient on thiotepa alone. The relative hazards are 1.37 and 1.13 respect thiotepa, when ciplastine and BCG are applied. This hazard is increased in a 7.3% if others treatments are applied, relative to a patient on thiotepa treatment alone. Patients with two or more tumors have a risk of recurrence 34% higher than patients with only one tumor and individuals with tumors $>$ 3 cm have a risk 44% bigger than patients with tumor $\leq$ 3 cm.

We have checked the model by means of residuals proposed by Farrington in [Farrington, 2000]. It is assumed that the observation process that generates the interval censoring is independent of the survival times and the covariates. In that sense Figure 1 shows the distribution of interval lengths by observation number. The plots do not reveal any systematic differences in the observation process between treatment groups.

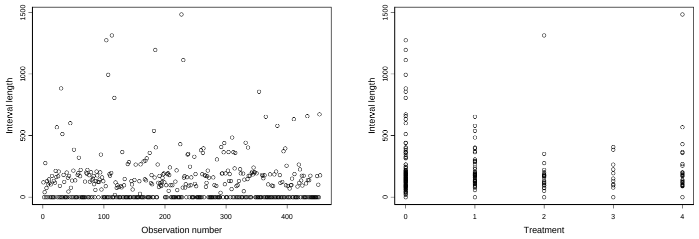


Fig. 1. Distribution of interval length: by observation number and treatment

Martingale residuals, in large samples, were shown to have zero mean under the correct model. That type of residuals reveal the existence of outliers. In Figure 2 patients 384 and 396 are separated from the bulk of the data. These patients belong to groups with the same features in size (> 3 cm), number (two or more) and treatment (Adriamicine).

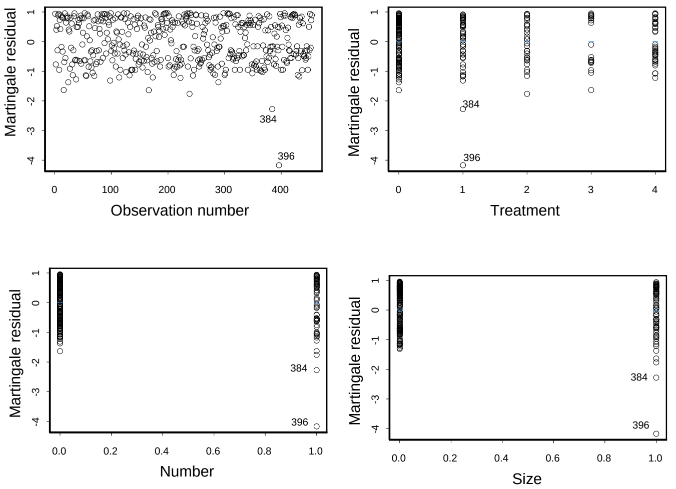


Fig. 2. Martingale residual by observation number, treatment, number and size

It would be useful to plot these residuals against log interval length and its analysis with deviance residuals as it is shown in [Farrington, 2000].

4 A Cox model of tumor recurrence

Let us consider now that time of recurrence is the time at which recurrence is detected.

The survival experience of the 454 patients depends on several variables, whose values have been recorded for each patient at the time origin. The aim of this Section is to determine which of explanatory variables have an impact on the free of disease time of the patients (survival time).

The focus is modelling the *recurrence hazard* (*risk of recurrence*)at time t. The *recurrence hazard* is obtained from the *hazard function* $h(t)$ and it is obtained from the basic model for survival data: *proportional hazard model* or *Cox regression model* given by:

$$h_i(t) = \exp(\beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_p x_{pi}) h_0(t)\,, \tag{6}$$

where $h_0(t)$ is the *baseline hazard function.*

On the other hand, the objective of this modelling procedure is to determine which combination of explanatory variables affects the form of the hazard function. In this process we use the statistic $-2Log\hat{L}$.

Indicator or *Dummies* variables are generated for the analysis. From *treatment* (five categories) four *Dummies* are defined: Adriamicine, Cisplatine, BCG and others treatments. From *grade* (three categories) two *Dummies*: G2 and G3. *Sex*, *number*, *size* and *stage* are dichotomic variables. *Age* is continuous. In this way the individual of reference is a 65 years old man (average patient), with only one tumor, of pTa stage, G1 grade, with a size minor or equal than 3 cm and with Thiotepa treatment after TUR.

Parameters estimates in the Cox regression model are presented in Table 3. The model allows us to compare risks among different groups of patients in a similar way of previous section.

Parameter	$\hat{\beta}$	Exp($\hat{\beta}$)	se($\hat{\beta}$)	**z**	**p-value**	**lower.95**	**upper.95**
$>$ **3 cm**	0.408	1.50	0.147	2.766	0.006	1.126	2.01
Cisplatine	0.418	1.52	0.224	1.866	0.062	0.979	2.36
BCG	0.201	1.22	0.329	0.611	0.540	0.642	2.33
ADR	0.725	2.07	0.181	4.017	0.000	1.450	2.94
Others treatments	0.147	1.16	0.201	0.731	0.460	0.781	1.72

Table 3. Cox regression model. Parameters estimates

Let us begin the model checking by testing the proportional hazards assumption. Grambsch and Therneau [Therneau and Grambsch, 2000] show that the expected value of the ith *scaled Schoenfeld residual* is given by $E\left(r^*_{Pji}\right) \approx \hat{\beta}_j\left(t_i\right) - \hat{\beta}_j$, and so a plot of the values of $r^*_{Pji} + \hat{\beta}_j$ against the death times should give information about the form of the time-dependent coefficient of X_j, $\beta_j\left(t\right)$.

The horizontal line in each graph of Figure 3 indicates no suggestion of non-proportional hazards and that the coefficients of these variables are constant.

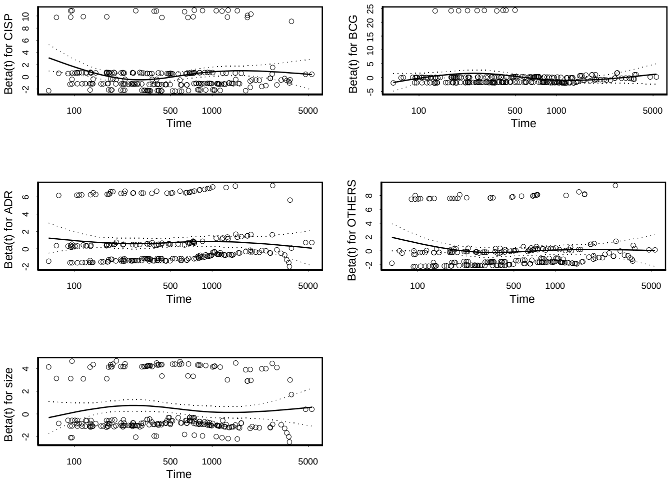


Fig. 3. Plots of scaled Schoenfeld residuals against time for each variable.

This graphical diagnostic is supplemented by a test for each variable, along with a global test for the model as a whole. In Table 4 it is showed the mentioned global test and the tests for each variable.

Here *rho* is the Pearson product-moment correlation between the *scaled Schoenfeld residuals* and time for each variable. The column *chisq* gives the tests statistics for each variable and the last row GLOBAL gives the global test for a χ^2 of 5 degree of freedom. With these results we may assume the proportional hazard hypothesis.

Validation and diagnostic of our model is based on *Martingale* and *Deviance* residuals. All results were consistent. The following graphics, see figure 4, show an Index plot of those residuals. In both plots the cluster of points is rather compact. We highlight patients 443 and 448 (they are patients 384 and 396 of section 3) whose survival times are larger than expected from the model.

variable	**rho**	**chisq**
Cisplatine	0.0182	0.0787
BCG	-0.0913	1.9992
Adriamicine	-0.0178	0.0748
Others treatments	-0.0524	0.6489
$\leq$ **3cm**	-0.0352	0.2944
GLOBAL		**2.8889**

Table 4. Test for the Proportional Hazards

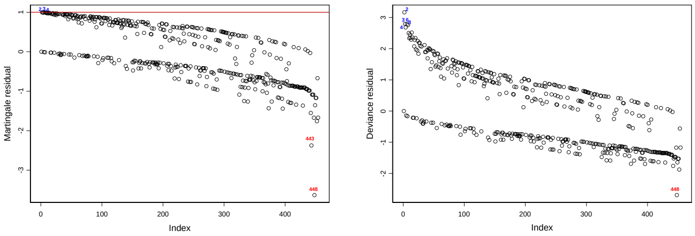


Fig. 4. Martingale and Deviance residuals

Identification of influential observations is performed by means of *Delta–Beta* test and examining the $-2\log\hat{L}$ changes. We found no alarming observations.

5 Conclusions

We have studied the prognostic factor for bladder cancer by means two different models: Cox regression and generalized non–linear models. In the first model, the prognostic factors are size and treatment; in the second model these factors are number, size and treatment. In the validation of both models the same two patients are detected and they belong to groups with the same features. Their characteristics correspond to the highest risk of recurrence and, however, they are among the patients with the longest time free of disease (what justify their behavior in our analysis). But this is not an important fact.

Acknowledgements: This work has been supported by the Generalitat Valenciana Grant GV04B-483

References

[Black *et al.*, 2002]R.J Black, R. Sankila, J. Ferlay, and D.M Parkin. Estimates of cancer incidence in europe for 1995. *Europe Journal Cancer*, pages 99–166, 2002.

[Collett, 2003]D. Collett. *Modelling Survival Data in Medical Research* 2^{th} *ed.* Chapman & Hall/CR, Boca Raton, Florida, 2003.

[Delwiche and Slaughter, 1998]L.D Delwiche and S.J Slaughter. *The Little SAS Book.* SAS Institute, Cary, NC, 1998.

[Farrington, 1996]C.P Farrington. Interval censored survival data: A generalized linear modelling approach. *Statistics in Medicine*, pages 283–292, 1996.

[Farrington, 2000]C.P. Farrington. Residuals for proportional hazards models with interval–censored survival data. *Biometrics*, pages 473–482, 2000.

[Hermanek and Sobin, 1998]P. Hermanek and L.H Sobin. *TNM Classification of malignant tumours* 4^{th} *ed.* Springer–Verlag, Berlin, 1998.

[Jaemal *et al.*, 2003]A. Jaemal, T. Murray, A. Samuels, E. Ghafoor, A. adn Ward, and M. Thun. Cancer statistics 2002. *CA Cancer J Clin*, pages 5–26, 2003.

[Kurth *et al.*, 1995]K.H Kurth, L. Denis, C. Bouffoux, R. Sylvester, F.M Debruyne, M. Pavone-Macaluso, and W. Oosterlinck. Factors affecting recurrence and progression in superficial bladder tumors. *Eug J Cancer*, pages 1840–46, 1995.

[Millan *et al.*, 2000]F. Millan, G. Chechile, J. Salvador, J. Palou, and J. Vicente. Multivariate analysis of the prognosis factors of primary superficial bladder cancer. *Journal Urology*, pages 73–78, 2000.

[Royston *et al.*, 2002]P. Royston, M. Parmar, and R. Sylvester. Flexible proportional–hazards and proportional–odds models for censored survival data, with application to prognostic modelling and estimation of treatment effects. *Statistics in Medicine*, pages 2175–2197, 2002.

[Therneau and Grambsch, 2000]T.M Therneau and P.M Grambsch. *Modeling Survival Data, Extending the Cox Model.* Springer, New York, 2000.

[Venables and Ripley, 2002]W.N Venables and B.D Ripley. *Modern Applied Statistics with S.* 4^{th} *ed.* Springer, New York, 2002.

[Zieger *et al.*, 1998]K. Zieger, H. Wolf, P.R Olsen, and K. Hojgaard. Long–term survival of patients with bladder tumours: the significance of risk factors. *Br. Journal Urology*, pages 667–72, 1998.

A heuristic approach in hepatic cancer diagnosis using a probabilistic neural network-based model

Marina Gorunescu[1], Florin Gorunescu[2], Marius Ene[2], and Elia El-Darzi[3]

[1] Faculty of Mathematics and Computer Science
University of Craiova,
13 A.I. Cuza, Craiova, Romania
(e-mail: gorun@euroweb.ro)

[2] Department of Mathematics, Biostatistics and Computer Science
University of Medicine and Pharmacy of Craiova,
2-4 Petru Rares, Craiova, Romania
(e-mail: gorun@umfcv.ro, enem@umfcv.ro)

[3] Harrow School of Computer Science
University of Westminster, London
Watford Road, Northwick Park, Harrow HA1 3TP, UK
(e-mail: eldarze@wmin.ac.uk)

Abstract. This paper is focusing on the application of a probabilistic neural network-based model in diagnosing hepatic diseases. In the diagnose process, the physicians compare numerical medical data against prior knowledge in order to determine the right diagnostic. Neural networks are ideal in recognizing diseases using representative examples since there is no need to provide a specific algorithm on how to identify the disease. The goal of this paper is to explore a PNN-based approach to determine the (near) optimum diagnosis for hepatic cancer. As concerns the concrete program, a Java implementation is provided as well.

Keywords: probabilistic neural networks, Monte Carlo approach, hepatic diseases diagnosis, Java implementation.

1 Introduction

Hepatocellular carcinoma (HCC), briefly hepatic cancer, represents a primary malignant tumor of the liver that ranks fifth in frequency among all malignancies in the world. HCC is increasing in many countries, especially in areas where hepatitis C virus (HCV) infection is more common than hepatitis B virus (HBV) infection. The diagnosis of HCC is difficult in the early stages, most of the patients being diagnosed in advanced stages. Although alpha-fetoprotein (AFP) is the most important tumor marker for the diagnosis of HCC, a considerable proportion of HCC's do not produce AFP, making early diagnosis difficult with this marker alone. Imaging modalities (power Doppler, harmonic imaging, pulse inversion, etc.), combined with micro bubble contrast agents and a better understanding of the importance of serum enzymes significantly improved the rate of detection for early (small)

HCCs'. Among these detection factors, the serum enzymes analysis is by far the fastest and simplest method, representing the first step in hepatic cancer diagnosis.

The probabilistic neural network (PNN) was developed by [Specht, 1988] [Specht, 1990]. This particular type of artificial neural networks (ANNs) provides a general solution to pattern classification problems by following the probabilistic approach based on the Bayes formula. The Bayes decision theory, emerged from his celebrated formula and developed in the 1950's, takes into account the relative likelihood of events and uses a priori information to improve prediction. The network paradigm uses the Parzen estimators to obtain the corresponding probability density functions (p.d.f.) to the classification categories. In his classic paper, Parzen [Parzen, 1962] showed that a class of p.d.f. estimators asymptotically approach the underlying density function, provided that it is continuous. Cacoulos [Cacoulos, 1966] extended Parzen's method to the multivariate case. PNN uses a supervised training set to develop probability density functions within a pattern layer. Training of a PNN is much simpler than other ANNs techniques. Key advantages of PNN are that training requires only an unique pass and that the decision hipersurfaces are guaranted to approach the Bayes-optimal decision boundaries as the number of training samples grows. On the other hand, the main criticism of PNN is that all training samples must be stored and used in classifying new patterns (very rapid increase in memory and computing time when the dimension of the input vector and the training set size increase). However, to reduce the computational cost, dimensionality reduction and clustering methods are usually applied, previous to the PNN construction.

ANNs in general and PNNs especially are currently a main research area in health care modelling and it is believed that they will receive extensive application to biomedical systems in the next years ([Lin *et al.*, 2002], [Norton *et al.*, 2001], [Taktak *et al.*, 2004]). Neural networks learn by examples so the details of how to recognize the disease is not needed. We only need a set of examples (patterns) that are representative of all the variations of the specific disease. To obtain a high accuracy level in the disease recognition the patterns generally need to be selected carefully.

2 Bayes decision rule for PNNs

Bayesian decision theory is a fundamental statistical approach to the problem of pattern classification. To illustrate the formalism of the Bayes decision rule, consider the sample space Ω and $B_1, B_2, ...B_n$ a partition of Ω. Then the celebrated reverend Bayes formula (1763) is given by:

$$P(B_i|A) = \frac{P(A|B_i)P(B_i)}{\sum\limits_{i=1}^{n} P(A|B_i)P(B_i)} \tag{1}$$

Usually, the Bayes formula becomes:

$$Posterior = \frac{likelihood \times prior}{evidence} \tag{2}$$

where $P(B_i|A)$ is known as *Posterior*, $P(B_i)$ -the *prior* probabilities, $P(A|B_i)$ -the *likelihood*, $P(A)$ -the *evidence*.

Formally, the Bayes decision rule in a simplified form is given by:

- Decision D_k: "State of nature is B_k";
- Given measurement x if the decision is D_k then the error is $P(error|x) = 1 - P(B_k|x)$;
- Minimize the probability error;
- Bayes decision rule: "*Decide* D_k *if* $P(B_k|x) > P(B_j|x),\ \forall j \neq k$" or, equivalently, "*Decide* D_k *if* $P(x|B_k)P(B_k) > P(x|B_j)P(B_j),\ \forall j \neq k$"

To illustrate the way the Bayes decision rule is applied to PNNs, consider the general case of the q-category classification problem, in which the states of nature will be denoted by $\Omega_1, \Omega_2, ..., \Omega_q$. The goal is to determine the class (category) membership of a multivariate sample data (i.e. a p-dimensional random vector $\mathbf{x}$) into one of the q possible groups $\Omega_1, \Omega_2, ..., \Omega_q$, that is, we have to make the decision $D(x) = \Omega_i,\ i = 1,\ 2, ...,\ q$, where $\mathbf{x}$ represents the sample (data vector). If we know the (multivariate) probability density functions $f_1(x),\ f_2(x), ...,\ f_q(x)$, associated with the categories $\Omega_1, \Omega_2, ..., \Omega_q$, the *a priori* probabilities $h_i = P(\Omega_i)$ of occurrence of patterns from categories Ω_i and the *loss* (or *cost*) parameters l_i associated with all incorrect decisions given $\Omega = \Omega_i$, then, according to the Bayes decision rule, we classify $\mathbf{x}$ into the category Ω_i if the following inequality holds true:

$$l_i h_i f_i(x) > l_j h_j f_j(x),\ i \neq j. \tag{3}$$

The boundaries between every two decision classes Ω_i and $\Omega_j,\ i \neq j$, are given by the hypersurfaces:

$$l_i h_i f_i(x) = l_j h_j f_j(x),\ i \neq j, \tag{4}$$

and the accuracy of the decision depends on the accuracy of estimating the corresponding p.d.f's.

The key to using the Bayes decision rule to PNNs is represented by the technique chosen to estimate the p.d.f's $f_i(x)$ corresponding to each decision class Ω_i, based upon the training patterns set. The classical approach uses a sum of small multivariate Gaussian distributions, centered at each training sample, that is:

$$f_i(x) = \frac{1}{\sigma_i^p (2\pi)^{p/2}} \cdot \frac{1}{m_i} \cdot \sum_{j=1}^{m_i} exp\left(-\frac{\|x - x_j\|^2}{2\sigma_i^2}\right),\ i = 1,\ 2, ...,\ q, \tag{5}$$

where m_i is the total number of training patterns in Ω_i, x_j is the j-th training pattern from category Ω_i, p is the input space dimension and σ is an adjustable "*smoothing*" parameter using the training procedure. The main issue in PNNs methodology is represented by the way to determine the value of σ, since this parameter needs to be estimated to cause reasonable amount of overlap. Commonly, the smoothing factor is chosen heuristically. If σ is too large or too small the corresponding probability density functions will lead to the increase in misclassification rate. Fortunately, PNNs are not too sensitive to the very precise choice of the smoothing factor.

3 Modified Specht algorithm (Monte Carlo approach)

The only control parameter that needs to be selected for probabilistic neural network training is the radial deviation of the Gaussian densities -the smoothing factor. This section deals with one of the simplest but most robust algorithm, straight related to the Parzen-Cacoulos window classifiers, using the sum of training patterns that are classified in the right way as cost function and the Monte Carlo method for searching for the best solution. Among other statistical or Artificial Intelligence techniques, the Monte Carlo method allows us to obtain the optimization of the smoothing factor for each category with a good accuracy and saving computational effort.

Algorithm (training)

Input. Consider q decision classes $\Omega_1, \Omega_2, ..., \Omega_q$, each decision class Ω_i containing a number of m_i training patterns.

i) For each class Ω_i, $i = 1, 2, ..., q$, compute the (Euclidian) distance between any pair of training patterns;

ii) For each class Ω_i, $i = 1, 2, ..., q$, compute the corresponding average distances and standard deviations, denoted by D_i, SD_i respectively.

iii) For each class Ω_i, $i = 1, 2, ..., q$, compute the corresponding confidence intervals $I_{\Omega_i} = (D_i - 3SD_i, D_i + 3SD_i)$ for the average distances. This intervals represent the domains of the smoothing factors σ_i.

iv) For each decision class Ω_i, $i = 1, 2, ..., q$, consider the Parzen-Cacoulos classifiers $f_i(x)$ as the corresponding parent densities. Assign (σ_i, D_i), $i = 1, 2, ..., q$.

v) In each decision class Ω_i (randomly) choose a certain vector x_i^0 and compute $f_i(x_i^0)$.

vi) (*Bayes decision rule*) Compare $f_i(x_i^0)$ and $f_j(x_i^0)$ for all $i \neq j$ following the algorithm: "IF $l_i h_i f_i(x_i^0) > l_j h_j f_j(x_i^0)$ (for all $j \neq i$) THEN $x_i^0 \in \Omega_i$ ELSE IF $l_i h_i f_i(x_i^0) \leq l_j h_j f_j(x_i^0)$ (for some $j \neq i$) THEN $x_i^0 \notin \Omega_i$".

vii) (*Measuring the classification accuracy. Updating counter*) For each (fixed) decision class Ω_i consider the 3-valued logic: TRUE -if $l_i h_i f_i > l_j h_j f_j$ (for all $j \neq i$), UNKNOWN -if $l_i h_i f_i = l_j h_j f_j$ (for some $j \neq i$) and FALSE -otherwise. Initially, each of the three variables is set to zero. Whenever a truth value is obtained, the corresponding variable is incremented with step size 1.

viii) The cost function is given by the sum of training patterns that are classified in the right way.

ix) Repeat step 5 for another choice for x_i^0 in Ω_i until all of them are chosen. Increment counter.

x) Repeat step 5 for all vectors x_j^0 in Ω_j for all $j \neq i$. Increment counter.

xi) (*Searching for optimal smoothing parameter*) Generate in each confidence interval I_{Ω_i} a number of N random dividing points $\{P_1,\ P_2, ...,\ P_N\}$, uniformly distributed in I_{Ω_i}. Repeat step 5 by assigning $\sigma_i = P_k,\ k = 1,\ 2, ...,\ N$ for each $i = 1,\ 2, ...,\ q$.

xii) Find the maximum of the cost function.

Output. $\sigma_i,\ i = 1,\ 2, ...,\ q$, corresponding to the maximum of the cost function, represent the optimal values of the smoothing parameters $\sigma' s$ for each decision category $\Omega_i,\ i = 1,\ 2, ...,\ q$.

Note. It is well-known that, on the one hand, the health care modelling domain frequently encounters situations of non-numeric data (e.g. nominal data, ordinal data, images, multimedia data, even data collected from WWW) and, on the other hand, PNNs do not tend to perform well with such a data. Moreover, in the use of complex patterns in the health care area, weights for the attributes may be incorporated, in order to highlight the importance of each attribute. Under these circumstances, the Euclidian distance used in the PNN algorithm does not work correctly any more. Fortunately, there are methods to deal with these problems [Bishop, 1995]. One of the simplest approaches consists in using a mixed-weighted measure of similarity instead of the Euclidian distance [Gorunescu, 2003]. Such a measure allows us to compute distances between training patters consisting in numerical and non-numerical attributes (e.g. images) and taking into account the significance of each attribute in the decision process.

4 PNN application to hepatic cancer diagnosis

The PNN-based decision model was applied to classify a group of individuals into a certain categories of diagnosis in the area of hepatic diseases. This application might be seen as a case-control study investigating a way of selecting people with liver cancer (HCC) -the cases, using comparable persons who do not have this disease (the controls). It has been suggested [Ibrahim and Spitzer, 1979] that a case-control study requires at least two control groups to minimize the possibility of accepting a false result; the rationale is that if the same result is not achieved in the two case-control comparisons, both the apparent results are suspect. In our application there is a case group (HCC) and three control groups (CH), (LC) and (HP). Since PNNs are particularly useful for classification problems with more than two outputs, we have enlarged the previous case-control study in order to classify people in four diagnosis group: healthy people (HP), chronic hepatitis (CH), liver cirrhosis (LC) and hepatic cancer (HCC), instead of persons developing hepatic cancer vs. persons who do not have the disease.

The PNN-based classification algorithm has been applied to data in order to classify the initial group of individuals into four categories, depending on the diagnosis type: Ω_1 = HCC, Ω_2 = LC, Ω_3 = CH and Ω_4 = HP. Each person in the data set is represented by a 15-dimensional vector $\mathbf{x}$ = $(x_1, x_2, ..., x_{15})$ where the components represent some of the most important characteristics leading to the right medical diagnosis. Concretely, x_1 = TB (total bilirubin), x_2 = DB (direct bilirubin), x_3 = IB (indirect bilirubin), x_4 = AP (alkaline phosphatase), x_5 = GGT (gamma glutamyl transpeptidase), x_6 = LAP (leucine amino peptidase), x_7 = AST (aspartate amino transferase), x_8 = ALT (alanine amino transferase), x_9 = LDH (lactic dehydrogenase), x_{10} = PI (prothrombin index), x_{11} = GAMMA, x_{12} = ALBUMIN, x_{13} = GLYCEMIA, x_{14} = CHOLESTEROL and x_{15} = AGE. An example of such training data vector related to hepatic cancer is the following one: (6.97, 3.04, 3.93, 438, 279, 182, 135, 52, 95, 450, 3.6, 80, 1.2, 56, 1).

The model was fitted to real data consisting of 299 individuals (both patients and healthy people) from the Department of Internal Medicine, Division of Gastroenterology, University Emergency Hospital of Craiova, Romania. This group of individuals consists of 60 patients with chronic hepatitis (CH), 179 patients with liver cirrhosis (LC), 30 patients with hepatocellular carcinoma (HCC) and 30 healthy people (HP).

5 Experimental results

It is worth to mention that we have used only raw data without any previous data checking or data preparation (some errors in recording data or the existence of certain outliers is thus possible); moreover, no data screening has been performed [Altman, 1990]. The goal of such an approach is to verify the robustness of the PNN technique to learn from raw data.

The key to obtain a good classification using PNNs is to optimally estimate the two parameters of the Bayes decision rule, the misclassification costs and the prior probabilities. Unfortunately, there is no definitive science to obtain them and must be assigned as a specific part of the problem definition. In our practical experiment we have estimate them heuristically. Thus, as concerns the costs parameters, we have considered them depending on the average distances D_i, inversely proportional, that is $l_i = 1/D_i$; in this case the accuracy rate for N = 450 was about 90%. As concerns the prior probabilities, they measure the membership probability in each group and, thus, we have considered them equal to each group size, that is $h_i = m_i$.

To avoid overfitting, the data set was randomly partitioned into two sets: the training set and the validation set. A number of 254 persons (85%) of the initial group was withheld from the initial group for the smoothing factor adjustment (the training process). Once optimal smoothing parameters $\sigma's$ for each decision category were obtained using the training set, the trained PNN was applied to the validation set (the remaining 45 persons). Since we have used raw data to perform the PNN algorithm and to avoid the criticism of some people against the Monte Carlo method due to the fact the smallness of the error of method is only ensured with a certain probability, we have repeated 10 times the above procedure to diminish the outliers influence and a possible Monte Carlo technique weakness.

We have use the Java package for the algorithm implementation. What is important about the Java implementation of the program is that all data about patients collected by physicians can, at any time, be added, modified or deleted, with no change in the source of the program whatsoever. That is so because for the processing of the data we have used JDBC (Java Database Connectivity). Thus the program is connected to a database and the records of the specific table of this database can always be updated by the users themselves (in MS Access or MS Excel) with no further worries concerning the applicability of the program.

The experimental results are shown in Table 1 and Table 2. Table 1 presents the accuracy rates for both the training process and for the validation process.

Training accuracy rate (%)	Validation accuracy rate (%)
97.32	92.22
85.28	88.88
85.61	95.55
95.65	93.88
91.60	92.00
89.62	93.66
89.28	93.77
90.26	92.22
87.94	91.22
88.29	93.33
Average accuracy 90.10	92.67

Table 1. PNN classifier: experimental results

When the PNN was applied to the training process, the sensitivity analysis indicated that the proportion of the patients correctly diagnosed was (average) 90.10%.

When the PNN was applied to the validation data set, which was not subjected to neural network training, the proportion was (average) 92.67%.

The general predictive abilities of the PNN with the validation data set is particularly positive, given the fact that the validation data were not used in the training of the neural network.

In Table 2 we have considered the 3-valued logic: TRUE, FALSE and UNKNOWN and we have displayed the accuracy rates obtained during the validation process related to this classification, that is the percentage of patients correctly classified, incorrectly classified and unclassified.

Correctly classified patients	Incorrectly classified patients	Unclassified patients
92.22	6.78	1.00
88.88	8.12	3.00
95.55	4.45	0.00
93.88	6.12	0.00
92.00	7.00	1.00
93.66	6.34	0.00
93.77	6.23	0.00
92.22	7.78	0.00
91.22	8.78	0.00
90.10	7.90	2.00

Table 2. PNN classifier: classification correctness

We see that the unclassified cases represent at most 3% of the whole number of patients and in 60% of the computer program running we obtained no unclassified cases.

6 Conclusion and further work

In this paper we have developed and demonstrated the applicability and suitability of a PNN-based model for decision-making in the hepatic diagnosis process. PNNs learn by examples so the details of how to recognize the disease are not needed. What is needed is a set of examples that are representative of all the variations of the disease. We used raw data (the only data available for the experiment) and we obtained reliable results proving the PNNs ability and flexibility to learn by raw examples.

A problem to deal with in PNNs applications is the data set size. The number of cases required for PNN training frequently presents difficulties. As the number of variables increases, the number of cases required increases nonlinearly, so that with a fairly small number of variables a huge number of cases are required. In our experiment we used 299 cases with 15 variables. Further works should perform a heuristic study relating the number of variables to the number of cases.

In comparison with other PNN approaches related to the diagnosis process, the accuracy of this technique is competitive. For instance, in predicting ascites in broilers based on minimally invasive inputs [Roush *et al.*, 1997], a validation rate accuracy of 95% was reported. At the same time, a validation accuracy rate of 92.3% was reported in estimating the mortality risk following cardiac surgery [Orr, 1997].

Although the early diagnosis of liver cancer in liver cirrhosis is based on biochemical tests, modern approaches also use imaging tests (i.e. transabdominal ultrasound and/or spiral computed tomography). Therefore, another way to enlarge this heuristic approach in medical research is represented by the replacement of the Euclidian distance with a general mixed-weighted measure of similarity. Such an approach will strengthen the decision process by using much more attributes of the training patterns.

Clearly, much work still needs to be done to improve this methodology and to apply it to other health care classification problems.

References

[Altman, 1990]D.G. Altman. *Practical statistics for medical research.* Chapman and Hall, 1990.

[Bishop, 1995]C. Bishop. *Neural Networks for Pattern Recognition.* Oxford University Press, 1995.

[Cacoulos, 1966]R. Cacoulos. Estimation of a multivariate density. *Ann. Inst. Stat. Math.(Tokyo)*, 18, pages 179–189, 1966.

[Gorunescu, 2003]F. Gorunescu. Measuring similarities: an application to the chromosomes supervised selection. *Research Notes in Artificial Intelligence and Data Communications*, 103, pages 56–66, 2003.

[Ibrahim and Spitzer, 1979]M.A. Ibrahim and W. Spitzer. The case-control study: the problem and the prospect. *J. chron. Dis.*, 32, pages 130–144, 1979.

[Lin *et al.*, 2002]F. Lin, C. Chiu and S. Wu. Using Bayesian Networks for Discovering Temporal-State Transition patterns in Hemodialysis. *Proceedings of the 35th Annual Hawaii International Conference on System Sciences (HICSS-35.02), 0-7695-1435-9/02 © 2002 IEEE*, 2002.

[Norton *et al.*, 2001]I.D. Norton, Y. Zheng, M.S. Wiersema, J. Greenleaf, J. Clain and E. Dimagno. Neural network analysis of EUS images to differentiate between pancreatic malignancy and pancreatitis. *Gastrointest Endosc.*, 54(5), pages 625–629, 2001.

[Orr, 1997]R.K. Orr. Use of a Probabilistic Neural Network to Estimate the Risk of Mortality after Cardiac Surgery. *Med. Decis. Making*, 17, pages 178–185, 1997.

[Parzen, 1962]E. Parzen. On estimation of a probability density function and mode. *Ann. Math. Stat.*, 33, pages 1065–1076, 1962.

[Roush *et al.*, 1997]W.B. Roush, T.L. Cravener, Y.K. Kirby and R.F. Wideman. Probabilistic Neural Network Prediction of Ascites in Broilers Based on Minimally Invasive Physiological Factors *Poultry Science*, 76, pages 1513–1516, 1997.

[Specht, 1988]D.F. Specht. Probabilistic neural networks for classification mapping or associative memory. *Proceedings IEEE International Conference on Neural Networks*, 1, pages 525–532, 1988.

[Specht, 1990]D.F. Specht. Probabilistic neural networks. *Neural Networks*, 3, pages 109–118, 1990.

[Taktak *et al.*, 2004]A. Taktak, A. Fisher and B. Damato. Modelling survival after treatment of intraocular melanoma using artificial neural networks and Bayes theorem. *Phys. Med. Biol.*, 49, pages 87–98, 2004.

Estimating Vaccine Efficacy From Data With Recruitments

Claude Lefèvre

Université Libre de Bruxelles
Institut de Statistique et de Recherche Opérationnelle, C.P. 210,
B-1050 Bruxelles, Belgique
(e-mail: clefevre@ulb.ac.be)

Abstract. Vaccine induced protection against infection is often random. A concept of protective vaccine efficacy, depending on the mean relative susceptibility of vaccinated individuals, is considered for a large vaccine trial in which participants are recruited over a period of time. Bounds are derived that make statistical inference possible under weak assumptions about the transmission process, irrespectively of the type of protection induced by the vaccine.

This is a joint work with Niels Becker (The Australian National University, Canberra, Australia) and Sergey Utev (University of Nottingham, Nottingham, United Kingdom).

Keywords: Protective vaccine efficacy, Bounds, Estimation.

1 Introduction

A standard concept of vaccine efficacy is defined as

$$VE_{\mathrm{P}} = 1 - \frac{c_{\mathrm{V}}}{c_{\mathrm{U}}}, \tag{1}$$

c_{U} and c_{V} representing the proportions of cases among unvaccinated and vaccinated individuals, respectively.

As pointed out by, e.g., [Smith and Fine, 1984], vaccine efficacy depends on the type of the protection induced by a vaccine. Two types of vaccine response are usually discussed. A first case is when a vaccinee receives either complete protection or no protection against infection (i.e. the vaccine confers a complete/no (CN in short) protection). A second case is when every vaccinee receives exactly the same partial protection (i.e. the vaccine confers a partial/uniform (PU in short) protection).

Recently, [Becker and Utev, 2002] introduced a class of vaccine responses that includes CN and PU protection as particular cases. Shortly, if at time t, the force of infection acting on an unvaccinated susceptible individual is $\lambda(t)$, then the force of infection acting on a vaccinated susceptible individual is reduced to $A\lambda(t)$, A denoting a discrete random variable with probability distribution

$$\Pr(A = a_j) = p_j, \quad j = 1, \ldots, r, \tag{2}$$

where the possible values a_j are in $[0,1]$. These authors proposed for this class a concept of protective vaccine efficacy given by

$$VE_{\rm P} = 1 - {\rm E}A. \tag{3}$$

When a vaccine induces CN protection, (3) yields $VE_{\rm P} = p_1$, i.e. the probability that the vaccinee is completely protected. For PU protection, (3) becomes $VE_{\rm P} = 1 - a_1$, i.e. the per-contact reduction in the probability of disease transmission.

Estimating $VE_{\rm P}$ from data on the eventual numbers of vaccinated and unvaccinated cases requires to specify assumptions about the type of vaccine response. [Becker and Utev, 2002] showed, however, that for a standard model of epidemics in a large uniformly mixing community, the inequality

$$1 - \frac{c_{\rm V}}{c_{\rm U}} \;\leq\; VE_{\rm P} \;\leq\; 1 - \frac{\ln(1 - c_{\rm V})}{\ln(1 - c_{\rm U})}, \tag{4}$$

holds independently of all types of protection induced. These bounds are estimable from data on the eventual numbers of vaccinated and unvaccinated cases, and seem to be close enough to be used for inference about $VE_{\rm P}$.

Our purpose in the present paper is to show how to extend the analysis made in [Becker and Utev, 2002] to a more general model (i) based on less restrictive assumptions about the force of infection and (ii) allowing for recruitments of participants over time (which is useful for large field trials and/or for rather rare diseases). As a key result, we will obtain lower and upper bounds that are analogous to (but different fom) those given in (4). Furthermore, we will then prove that if the vaccination coverage remains constant over time, the lower bound can provide a good estimate of the vaccine efficacy.

This is a joint work with Niels Becker (The Australian National University, Canberra, Australia) and Sergey Utev (University of Nottingham, Nottingham, United Kingdom).

2 An epidemic model with vaccination

Denote by A the relative susceptibility of a vaccinated individual, such as defined by (2). Vaccinated individuals for which $A = a_j$ are said to be of type j, and unvaccinated individuals are said to be of type U. In practice, only unvaccinated (U) and vaccinated (V) individuals can be distinguished.

The population sizes are described by a deterministic model (valid for large trials). Let $N_{\rm U}(t)$ be the number of unvaccinated trial participants recruited by time t, and let $N_{\rm V}(t)$ be the number of vaccinated trial members recruited by time t. Initially, there are n individuals of whom a fraction u are unvaccinated and a fraction v are vaccinated $(u + v = 1)$. In Section 3, the proportion of vaccinated trial participants will be assumed to be always as large as its initial level v. In particular, the vaccination coverage can then remain constant.

At time 0, the numbers of susceptible trial participants are given by

$$S_{\mathrm{U}}(0) = nu, \quad \text{and} \quad S_j(0) = nvp_j, \quad j = 1, \ldots, r.$$

Let $\lambda(t)$ be the force of infection on an unvaccinated individual. Then, the number of unvaccinated trial members who are susceptible to infection at time t is ruled by the differential equation

$$dS_{\mathrm{U}}(t) = -\lambda(t) S_{\mathrm{U}}(t)\, dt + dN_{\mathrm{U}}(t).$$

For the vaccinated members, these numbers are governed by the differential equations

$$dS_j(t) = -a_j \lambda(t) S_j(t)\, dt + p_j\, dN_{\mathrm{V}}(t), \quad j = 1, \ldots, r.$$

Putting $\Lambda(t) = \int_0^t \lambda(x)\, dx$, the solutions to these equations are respectively given by

$$S_{\mathrm{U}}(t) = \int_{0-}^{t} \exp[\Lambda(x) - \Lambda(t)]\, dN_{\mathrm{U}}(x), \tag{5}$$

and

$$S_j(t) = p_j \int_{0-}^{t} \exp[a_j \Lambda(x) - a_j \Lambda(t)]\, dN_{\mathrm{V}}(x), \quad j = 1, \ldots, r. \tag{6}$$

Let us fix any finite time interval $[0, T]$. The number of unvaccinated trial participants who are cases by time T is

$$C_{\mathrm{U}} = N_{\mathrm{U}}(T) - S_{\mathrm{U}}(T),$$

and the number of vaccinated cases by time T is

$$C_{\mathrm{V}} = N_{\mathrm{V}}(T) - \sum_{j=1}^{r} S_j(T).$$

3 An estimator for the vaccine efficacy

As a first step, we begin by showing how (4) can be generalized to the present framework.

Proposition 1 *Provided that the proportion of vaccinated trial participants remains in the course of time as large as its initial level v, then*

$$1 - \frac{u}{v} \frac{C_{\mathrm{V}}}{C_{\mathrm{U}}} \;\leq\; VE_{\mathrm{P}} \;\leq\; 1 - \frac{\ln[1 - C_{\mathrm{V}}/N_{\mathrm{V}}(T)]}{\ln[1 - C_{\mathrm{U}}/N_{\mathrm{U}}(0)]}. \tag{7}$$

Moreover, the lower bound is attained when the vaccine induces CN protection and the vaccine trial has a non-varying vaccination coverage.

In the proof, a central point is a simple inequality for the expectation of a concave function of a random variable (see, e.g., [Becker and Utev, 2002]): if g is a continuous concave function defined on a finite interval $[c_1, c_2]$, then for any random variable A taking values in $[c_1, c_2]$,

$$g(c_1)\frac{c_2 - \mathrm{E}A}{c_2 - c_1} \;+\; g(c_2)\frac{\mathrm{E}A - c_1}{c_2 - c_1} \;\leq\; \mathrm{E}\,g(A) \;\leq\; g(\mathrm{E}A).$$

Now, let us give some comments on this result. We see that (7) reduces to (4) when recruitment occurs only at time $t = 0$. We also observe that the bounds of (4) still apply when recruitment occurs after time 0, but they ignore data on individuals recruited after time $t = 0$. Obviously, (7) uses data on individuals recruited after time 0, but the upper bound does so only through C_U, C_V and $N_V(T)$. Finally, we indicate that the upper bound in (7) cannot be attained with recruitment after time 0, but it is attained when the vaccine induces PU protection and the only recruitment is at time 0.

As a second step, we are going to derive an approximate estimator for the vaccine efficacy. More precisely, let us assume that all vaccine trial participants are recruited at k different instants during $[0, T]$. Initially, in each group i, $i \in \{1, \ldots, k\}$, there are n_i participants, and an identical vaccination coverage v is applied to each group. In group i, an unvaccinated individual escapes the disease with probability $\pi_i = \exp(-\Lambda_i)$, Λ_i denoting a cumulative force of infection upon this group until time T. A vaccinated individual in group i escapes the disease with probability $\mathrm{E}[(\pi_i)^A]$ where the random variable A has a distribution given by (2).

It is well-known (see, e.g., [Smith and Fine, 1984]) that without recruitment (i.e. when $\pi_i = \pi$), and if the vaccine induces CN protection, the measure (1) constitutes a maximum likelihood estimator of the vaccine efficacy VE_{P}. Hereafter, we will consider the cases, rather frequent in reality, where the different cumulative forces of infection Λ_i are all relatively small. We will then show that the lower bound, $1 - uC_V/vC_U$, derived in (7) provides a good estimator for VE_{P}.

Proposition 2 *Under the condition that* $\max_i(1 - \pi_i) \downarrow 0$, *then*

$$\widehat{VE_{\mathrm{P}}} = 1 - \frac{u}{v}\frac{C_{\mathrm{V}}}{C_{\mathrm{U}}} \tag{8}$$

is asymptotically equivalent to a maximum likelihood estimator of VE_{P}.

In the proof, the starting point is an expression for the global likelihood function L as a function of the unknown parameters $\{\pi_i,\ i = 1, \ldots, k\}$, $\{p_j,\ j = 1, \ldots, r\}$ and $\{a_j,\ j = 1, \ldots, r\}$. To construct L, we will have to introduce the final number of cases among vaccinated and unvaccinated participants in each group.

It is important to underline, however, that the only data needed for this estimator are the final numbers of cases observed at time T.

An asymptotic distribution as $n \to \infty$ can also be derived by using standard statistical arguments. First, a central limit theorem allows us to show that the lower bound type estimator $\widehat{\mathrm{E}A} = uC_V/vC_U$ is approximately normal. Then, using inequalities between integrals of special functions of exponential type, we are able to prove that the asymptotic mean of $\widehat{\mathrm{E}A}$, denoted by a, is given by

$$a\mathrm{sim}\mathrm{E}A + \alpha \quad \text{with} \quad 0 \leq \alpha \leq \varepsilon/8(1-\varepsilon)^2, \tag{9}$$

where $\varepsilon = 1 - \exp[-\Lambda(T)]$ is small by the assumption made before. The variance can also be calculated in a similar way.

References

[Becker and Utev, 2002]N.G. Becker and S. Utev. Protective vaccine efficacy when vaccine response is random. *Biometrical Journal*, pages 29–42, 2002.

[Smith and Fine, 1984]P.G. Smith and P.E.M. Fine. Assessment of the protective efficacy of vaccines against common diseases using case-control and cohort studies. *International Journal of Epidemiology*, pages 87–93, 1984.

Number of segregating sites in a sample of genes
under the genetic instability hypothesis

Mathieu Emily and Olivier François

Laboratoire TIMC-IMAG
Institut d'Ingénierie de l'Information de Santé
Faculté de Médecine
38706 La Tronche cedex, France
(e-mail: mathieu.emily@imag.fr, olivier.francois@imag.fr)

Abstract. Early detection of (pre)tumor is a priority in the understanding of cancer development in tissues. Several hypotheses have been proposed to explain tumorigenesis. One of them, the *mutator phenotype*, postulates that the loss of mismatch repair (MMR) generates a raise in the mutation rate. Under this assumption estimating the increase in the mutation rate is a key step for detecting a tumor. In this paper an estimator of the raised mutation rate based on the number of segregating sites in a sample of cells is proposed. The bias and the mean squared error of this estimator have been assessed through a simulation study.
Keywords: Tumorigenesis, Genetic instability, Mutator phenotype, Coalescent theory, Number of segregating sites.

1 Introduction

Cancer is known to be a very complex phenomenon. Since early in the 20th century [Boveri, 1929], it is widely assumed that a normal cell is converted to a tumoral cell by a succession of genetic events. More precisely the genetic equilibrium of a cell is disrupted by an initiating event and then because of a cascade process the cell becomes tumoral. This is the so-called *genetic instability* hypothesis for tumorigenesis.

Three major competing hypotheses have been formulated concerning the initial event of tumorigenesis. The first one [Tomlinson and Bodmer, 1999] [Cairns, 1975]explains that a cell must exhibit a selective advantage to be converted into a pretumoral cell. Then by a selective clonal expansion the cell becomes malignant. The second hypothesis is based on the experimental results that most of tumoral cells are victims of aneuploidy [Duesberg *et al.*, 1998]. This chromosomal instability may be responsible for the multistep process that leads to cancer [Duesberg and Rasnick, 2000]. The third hypothesis is called the *mutator phenotype* [Loeb and Springgate, 1974]. Considering the high fidelity of DNA replication in normal cells and the large number of genetic alterations that are observable in cancer cells, it postulates that the initial event in tumorigenesis is a particular mutation. This mutation should take place in genes that control the fidelity of DNA replication and the efficacity of DNA repair. These genes are directly responsible for the genetic stability of a cell. An alteration of their functions called loss of Mismatch Repair (*loss of MMR*) may generate a deregulation of the apoptosis or a reduction of the cell cycle duration. As a result of loss of MMR, the mutation rate will be raised in all cells that are descendants of the cell affected by loss of MMR. It has been suggested that the loss of MMR is required to initiate tumorigenesis [Loeb, 1991].

It is still a matter of debate to know exactly which event is the initiating event of tumorigenesis. Several mathematical models have been studied to understand the *mutator phenotype* hypothesis. Some of them argued that selection prevails on the raise of the mutation rate [Tomlinson and Bodmer, 1995]. Other models study the effect of the loss of MMR and how it hastens tumorigenesis [Plotkin and Nowak, 2002, Michor *et al.*, 2003]. Evolutionary models had been developed to infer the age of the loss of MMR [Tsao *et al.*, 2000] [Calabrese *et al.*, 2004]. It is widely assumed ([Shibata *et al.*, 1994] and [Bhattacharyya *et al.*, 1994]) that after the loss of MMR the mutation rate increases 10^2- to 10^3-fold. The need for deeper mathematical studies has been formulated in a recent review [Michor *et al.*, 2004] to better understand the influence of the three hypothesis (selection, aneuploidy and mutator phenotype) in the evolution of a cell. So far no mathematical model has been developed to estimate the raised mutation rate, and this is the focus of this article. A classical method in population genetics for estimating a mutation rate consists in counting the number of segregration sites in a sample of genes. In this article we propose a correction of this estimator in the context of genetic instability based on the coalescent theory.

2 Model Description

We consider a sample of n copies of a particular gene taken from a (pre)tumoral tissue, and assume that the loss of MMR occurred once in the sample history. However, the date and place at which this event occurred are unknown. Loss of MMR can be considered as a particular deleterious mutation of a mismatch repair gene. We denote by μ_{LMMR} the rate of this particular mutation, and we assume that the rate of this event is very small (μ_{LMMR} goes to 0).

The sample is divided in two random subsamples $\mathcal{B}$ and $\mathcal{C}$ where $\mathcal{B}$ denotes the subset of descendants of the mutation and $\mathcal{C}$ its complement. Given the number $B = b$ of genes in $\mathcal{B}$, the number of genes in $\mathcal{C}$ is then equal to $n - b$ (see Figure 1). Genes are characterized by their DNA sequences. For instance, such data may arise from the FISH (Fluorescence In Situ Hybridation) technology [Pinkel *et al.*, 1986]. In our model, the evolution of genes is described by a two-rates model. We denote by $\mu_{\mathcal{C}}$ the *normal rate*, ie the mutation rate per base per generation in $\mathcal{C}$. On the other hand, we denote by $\mu_{\mathcal{B}}$ the *raised mutation rate* in $\mathcal{B}$.

Conditional on $B = b$, the genealogy of the n genes can be described by the so-called *conditional coalescent* [Wiuf and Donnelly, 1999] for which the genes in $\mathcal{B}$ share a common ancestor before any of them shares an ancestor with $\mathcal{C}$. In addition, loss of MMR occurred between the time of the most recent common ancestor (MRCA) of the subsample $\mathcal{B}$ and the time at which $\mathcal{B}$ coalesces with $\mathcal{C}$ (see Figure 1). The coalescent approximation was introduced by Kingman in the 80's [Kingman, 1982]. It is similar to the diffusion approximation in population genetics. Time is measured in units of N generations where N is the total population size. In this setting, mutation rates are rescaled as $\theta_{\mathcal{B}}/2 = 2N\mu_{\mathcal{B}}$ and $\theta_{\mathcal{C}}/2 = 2N\mu_{\mathcal{C}}$.

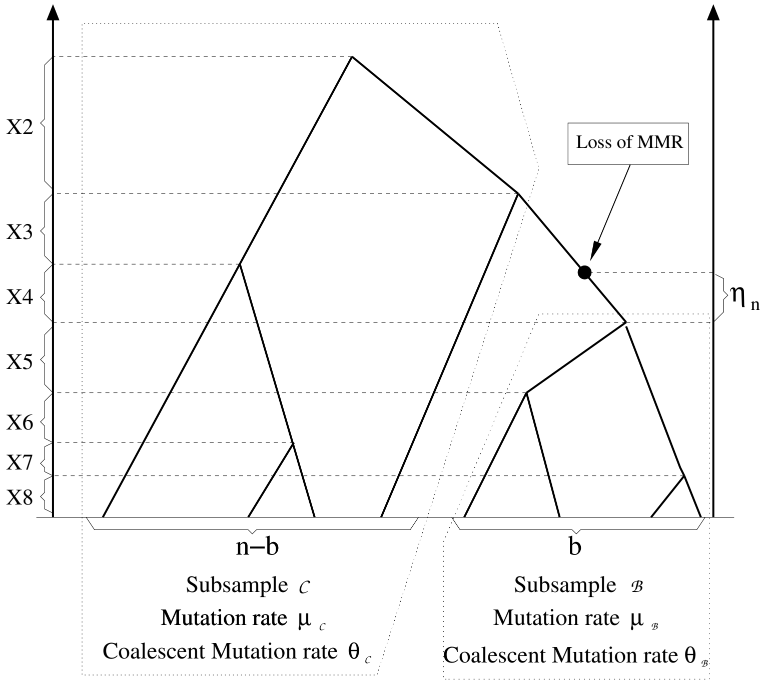

X2
X3
X4
X5
X6
X7
X8
Loss of MMR
η_n
n−b
Subsample $\mathcal{C}$
Mutation rate $\mu_{\mathcal{C}}$
Coalescent Mutation rate $\theta_{\mathcal{C}}$
b
Subsample $\mathcal{B}$
Mutation rate $\mu_{\mathcal{B}}$
Coalescent Mutation rate $\theta_{\mathcal{B}}$

Fig. 1. A coalescent tree of size $n = 8$ conditional on $B = 4$. Time is represented backward and the loss of MMR event is indicated.

In the coalescent, mutations occur according to independent Poisson processes of rate $\theta/2$ along the branches of the tree. Among the various models that describe the mutation types, the *infinitely-many sites* model may be one of the most appropriate [Watterson, 1975]. In this model, each DNA sequence consists of completely linked sites (ie, no recombination occurs).

Each mutation occurs at a site of the DNA sequence that had not been mutated before, so that a new segregating site arises. The number of segregating sites corresponds to the number of substitutions of ancestral bases since the MRCA.

3 Theoretical analysis

3.1 Background

In this section, we recall well-known results about the number of segregating sites under the infinitely-many sites model of mutation. These results are valid when loss of MMR do not occur which means that there is only one mutation rate , written θ [Watterson, 1975]. In the neutral coalescent, the gene lineages coalesce at random, and the times separating the coalescence events $X_i,\ i = 2 \cdots n$ are independent exponential random variables of parameter $i(i-1)/2$. The tree has total length $L_n = \sum_{i=2}^{n} iX_i$ of expectation

$$\mathbb{E}[L_n] = 2H_{n-1} \approx 2\log n$$

and variance

$$Var[L_n] = 4\sum_{i=1}^{n-1} \frac{1}{i^2} \approx \frac{2\pi^2}{3}$$

where H_n is the n^{th} harmonic number $H_n = \sum_{i=1}^{n} 1/i$.

The number of segregating sites $\hat{\theta} = S_n/H_{n-1}$ is frequently used as an unbiased estimator of the mutation rate θ. Using that

$$Var[S_n] = \sum_{i=1}^{n-1} \left(\frac{\theta^2}{i^2} + \frac{\theta}{i} \right)$$

we see that $\hat{\theta}$ converges to θ at a logarithmic rate.

3.2 Number of sites of segregation in the two-rates model

In the mutator phenotype hypothesis, a rare mutation is responsible for an increase in the DNA mutation rate from $\theta_{\mathcal{C}}$ to $\theta_{\mathcal{B}}$. In this section, we build an approximately unbiased estimator of $\theta_{\mathcal{B}}$.

First of all, the number B of genes that carry the mutator phenotype (the *frequency spectrum*) has a Yule distribution [Stephens, 2000]

$$P(B = b) = \frac{1}{bH_{n-1}}, \quad b = 1, \ldots, n-1 \tag{1}$$

Given $B = b$, the total length $\widetilde{L_n}$ of the genealogy of the subsample $\mathcal{B}$ has an expected value equal to [Griffiths and Tavaré, 2003]

$$\mathbb{E}[\widetilde{L_n}|B = b] = L_{n,b} = \binom{n-1}{b}^{-1} \sum_{j=2}^{n-b+1} \binom{n-j}{b-1} \sum_{k=j+1}^{n} \frac{2}{k(k-1)} c_{jk} \tag{2}$$

where $b = 2, \cdots, n-1$, and

$$c_{jk} = b - (b-1)\frac{n-k}{n-j} - \frac{(n-k)!(n-j-b+1)!}{(n-j)!(n-k-b+1)!} \tag{3}$$

Consider the time η_n that separates the MRCA of the $\mathcal{B}$ sample from the loss of MMR event. Wiuf and Donnelly [Wiuf and Donnelly, 1999] showed that

$$\mathbb{E}[\eta_n | B = b] = 2\binom{n-1}{b}^{-1} \sum_{j=2}^{n-b+1} \frac{1}{j}\binom{n-j}{b-1} \qquad b = 1, \cdots, n-1 \tag{4}$$

Now, consider the total length $\tilde{L}_n + \eta_n$ (see Figure 1), and take the expectation. We set

$$\beta_n = \mathbb{E}[\tilde{L}_n + \eta_n]/2$$

The average number of mutations in descendants of the loss of MMR event is given by

$$\mathbb{E}[S_n^{\mathcal{B}}] = \beta_n \theta_{\mathcal{B}}$$

In addition, the average number of mutations in the subsample $\mathcal{B}$ is

$$\mathbb{E}[S_n^{\mathcal{C}}] = \gamma_n \theta_{\mathcal{C}}$$

where

$$\gamma_n \approx H_{n-1} - \beta_n \tag{5}$$

Finally, consider the total number S_n of segregating sites. We obtain that

$$\mathbb{E}[S_n] = \beta_n \theta_{\mathcal{B}} + \gamma_n \theta_{\mathcal{C}} \tag{6}$$

An unbiased estimator of the raised mutation rate can be proposed as follows

$$\hat{\theta}_{\mathcal{B}} = \frac{S_n - \gamma_n \theta_{\mathcal{C}}}{\beta_n} \tag{7}$$

4 Results and discussion

In this section, we study the behaviour of the estimator given in equation (7) through simulations. Data were simulated as follows. The first step was the determination of $\mathcal{B}$ using the *frequency spectrum* distribution described in equation (1). Then, we built a conditional coalescent tree given $B = b$, with biased inter-coalescence times. We computed both the total length L_n of the tree and the length $\tilde{L}_n$ of the b-subtree, and we simulated the random variable η_n. Finally, we simulated the random variable S_n as a Poisson distributed variable of rate $\hat{\beta}_n \theta_{\mathcal{B}} + \hat{\gamma}_n \theta_{\mathcal{C}}$ where $\hat{\beta}_n = (\tilde{L}_n + \eta_n)/2$ and $\hat{\gamma}_n = L_n/2 - \hat{\beta}_n$. Biased inter-coalescence times were obtained from a rejection algorithm. The simulation procedure was validated by recovering various known quantities (such as $L_{n,b}$).

The experimental results regarding the estimator $\hat{\theta}_{\mathcal{B}}$ are presented in Table 1. These results were obtained from the procedure above described using the following experimental design. The parameter $\theta_{\mathcal{C}}$ was set equal to a small value $\theta_{\mathcal{C}} = 0.01$. This corresponds to the rough value of a mutation rate $\mu_{\mathcal{C}} \approx 10^{-10}$, the total number of cells $N \approx 10^8$. Four different values for the raised mutation rate $\theta_{\mathcal{B}} = 0.1$, 0.2, 1 and 10 were considered. Sample sizes of $n = 10$, $n = 20$ and $n = 50$ cells were considered. For each simulation we took two configurations of the mutation rate θ_{LMMR}, and observed that this had a weak influence on the result.

	$\theta_{\mathcal{B}} = 0.1$			$\theta_{\mathcal{B}} = 0.2$		
	$\mathbb{E}[\hat{\theta_{\mathcal{B}}}]$	$SD[\hat{\theta_{\mathcal{B}}}]$	$\sqrt{MSE[\hat{\theta_{\mathcal{B}}}]}$	$\mathbb{E}[\hat{\theta_{\mathcal{B}}}]$	$SD[\hat{\theta_{\mathcal{B}}}]$	$\sqrt{MSE[\hat{\theta_{\mathcal{B}}}]}$
$n = 10$						
	0.085	0.48	0.48	0.20	0.60	0.60
$n = 20$						
	0.057	0.38	0.38	0.19	0.71	0.71
$n = 50$						
	0.18	0.58	0.59	0.21	0.66	0.66

	$\theta_{\mathcal{B}} = 1$			$\theta_{\mathcal{B}} = 10$		
	$\mathbb{E}[\hat{\theta_{\mathcal{B}}}]$	$SD[\hat{\theta_{\mathcal{B}}}]$	$\sqrt{MSE[\hat{\theta_{\mathcal{B}}}]}$	$\mathbb{E}[\hat{\theta_{\mathcal{B}}}]$	$SD[\hat{\theta_{\mathcal{B}}}]$	$\sqrt{MSE[\hat{\theta_{\mathcal{B}}}]}$
$n = 10$						
	0.93	1.43	1.42	8.92	11.42	11.44
$n = 20$						
	0.94	1.56	1.56	6.80	11.43	11.84
$n = 50$						
	0.75	1.82	1.84	9.65	16.15	16.11

Table 1. Results of our estimator $\hat{\theta_{\mathcal{B}}}$ on simulations. This table summarises results obtained under various conditions. Simulations were made for a population of $n = 10$, $n = 20$ and $n = 50$ cells in total. For each n, 500 simulations were performed in each 4 cases : $\theta_{\mathcal{B}} = 0.1$ $\theta_{\mathcal{B}} = 0.2$, $\theta_{\mathcal{B}} = 1$ and $\theta_{\mathcal{B}} = 10$.

Table 1 gives the bias, variance and mean squared error estimated over 500 simulations. The results show that $\hat{\theta}_{\mathcal{B}}$ is indeed weakly biased. The major source of bias was the limit of a null mutation rate μ_{LMMR} considered in the theoretical analysis. Nevertheless, the mean squared error is very high, and the distribution of the estimator appeared to be positively skewed. In addition, the variance did not decrease with the sample size. This might be due the dependence of data within the $\mathcal{B}$ subsample, and the fact that the MRCA of the subsample is expected to be recent.

Although it is crucial in the fight against cancer, detection of the disease at a pretumoral stage is a very difficult issue. In this paper we showed that estimating the raised mutation rate (posterior to loss of MMR) based on the number of segregating sites may not be an efficient method while the use of this estimator is widely spread in more classical population genetics studies.

As well, we observed that the variance of the estimator did not decrease as the sample size increased (from $n = 10$ to $n = 50$). Consequently if DNA analyses of a (supposed tumoral) tissue are necessary, collecting a large number of DNA sequences may not be the best approach for inferring the raised mutation rate. This issue may be overcome by considering several chromosomal loci instead of a single locus as we did. Nevertheless the fact that empirical distributions of the estimator are positively skewed indicates that statistical testing using the number of segregating sites might be lacking power.

References

[Bhattacharyya *et al.*, 1994]N.P. Bhattacharyya, A. Skandalis, A. Ganesh, J. Groden, and M. Meuth. Mutator phenotypes in human colorectarl carcinoma cell lines. *Proc. Nat. Acad. Sc.*, 91:6319–6323, 1994.

[Boveri, 1929]T. Boveri. *Origin of the Malignant Tumors.* Williams & Williams Publising Co., 1929.

[Cairns, 1975]J. Cairns. Mutation selection and the natural history of cancer. *Nature*, 255:197–200, 1975.

[Calabrese *et al.*, 2004]P. Calabrese, J.L. Tsao, Y. Yatabe, R. Salovaara, J.P. Mecklin, H.J. Järvinen, L.A. Aaltonen, S. Tavaré, and Shibata D. Colorectal pretumor progression before and after loss of DNA mismatch repair. *Am. J. Pathol.*, 164:1447–1453, 2004.

[Duesberg and Rasnick, 2000]P. Duesberg and D. Rasnick. Aneuploidy, the somatic mutation that makes cancer species of its own. *Cell Motil Cytoskeleton*, 47:81–107, 2000.

[Duesberg *et al.*, 1998]P. Duesberg, C. Raush, D. Rasnik, and R. Hehlmann. Genetic instability of cancer cells is proportional to their degree of aneuploidy. *Proc. Nat. Acad Sci.*, 95:13692–13697, 1998.

[Griffiths and Tavaré, 2003]R.C. Griffiths and S. Tavaré. The genealogy of a neutral mutation. In P. Green, N. Hjort, and S. Richardson, editors, *Highly Structured Stochastic Systems*, pages 393–412, 2003.

[Kingman, 1982]J.F.C. Kingman. The coalescent. *Stoch. Process. Appl.*, pages 235–248, 1982.

[Loeb and Springgate, 1974]L.A. Loeb and Battula N. Springgate, C.F. and. Errors in DNA replication as a basis of malignant changes. *Cancer Res.*, pages 238–242, 1974.

[Loeb, 1991]L.A. Loeb. Mutator phenotype may be required for multistage carcinogenesis. *Cancer Res.*, 51:3075–3079, 1991.

[Michor *et al.*, 2003]F. Michor, M.A. Nowak, S.A. Franck, and Y. Iwasa. Stochastic elimination of cancer cells. *Proc. R. Soc. Lond.*, 270:2017–2024, 2003.

[Michor *et al.*, 2004]F. Michor, Y. Iwasa, and M.A. Nowak. Dynamics of cancer progression. *Nature Reviews*, 4:197, 2004.

[Pinkel *et al.*, 1986]D. Pinkel, T. Straume, and J.W. Gray. Cytogenetic analysis using quantitative, high sensitivity, fluorescence hybridation. *Proc. Nat. Acad. Sci.*, pages 2934–2938, 1986.

[Plotkin and Nowak, 2002]J.B. Plotkin and M.A. Nowak. The differnet effects of apoptosis and DNA repair on tumorigenesis. *J. Theor. Biol.*, 214:453–467, 2002.

[Shibata *et al.*, 1994]D. Shibata, M.A. Peinado, Y. Ionov, S. Malkhosyan, and M. Perucho. Genomic instability in repeated sequences is an early somatic event in colorectal tumorigenesis that persists after transformation. *Nat. Genet.*, 6:273–281, 1994.

[Stephens, 2000]M. Stephens. Time on trees and the age of an allele. *Theo. Pop. Biol.*, pages 109–119, 2000.

[Tomlinson and Bodmer, 1995]I.P.M. Tomlinson and W.F. Bodmer. Failure of programmed cell death and differentiation as causes of tumors : some simple mathematical models. *Proc. Nat. Acad. Sci.*, 92:11130–11134, 1995.

[Tomlinson and Bodmer, 1999]I.P.M. Tomlinson and W.F. Bodmer. Selection, the mutation rate and cancer: Ensuring that the tail does not wag the dog. *Nat. Med.*, 5(1):11–12, 1999.

[Tsao *et al.*, 2000]J.L. Tsao, Y. Yatabe, R. Salovaara, J.P. Mecklin, H.J. Järvinen, L.A. Aaltonen, S. Tavaré, and Shibata D. Genetic reconstruction of individual colorectal tumor histories. *Proc. Nat. Acad. Sc.*, 97(3):1236–1241, 2000.

[Watterson, 1975]G.A. Watterson. On the number of segregating sites in genetical models without recombination. *Theo. Pop. Biol.*, pages 256–276, 1975.

[Wiuf and Donnelly, 1999]C. Wiuf and P. Donnelly. Conditionnal genealogies and the age of a neutral mutant. *Theo. Pop. Biol.*, pages 183–201, 1999.

Play-the-winner rule in clinical trials: Models for adaptative designs and Bayesian methods

Bruno Lecoutre and Khadija Elqasyr

ERIS, Laboratoire de Mathématiques Raphael Salem,
UMR 6085, C.N.R.S. et Université de Rouen,
Mathématiques, Site Colbert, 76821 Mont-Saint-Aignan Cedex, France
(e-mail: bruno.lecoutre@univ-rouen.fr,
khadija.elqasyr@etu.univ-rouen.fr)

Abstract. Adaptative designs for clinical trials that are based on a generalization of the "play-the-winner" rule are considered as an alternative to previously developed models. Theoretical and numerical results show that these designs perform better for the usual criteria. Bayesian methods are proposed for the statistical analysis of these designs.

Keywords: Clinical trials, Adaptative designs, Play-the-winner rule, Generalized Friedman's urn, Bayesian methods.

1 Introduction

From ethical point of view, adaptative designs can be desirable for some clinical trials. In such designs subjects are assumed to arrive sequentially and they are assigned to a treatment with a probability that is updated as a function of the previous events. The intent is to favor the "most effective treatment" given available information. Originally, the *play-the-winner* allocation rule was designed for two treatments with a dichotomous (e.g. success/failure) outcome [Zelen, 1969]. It involves an "all-or-none" process: if subject $n-1$ is assigned to treatment t and if the outcome is a success, subject n is assigned to the same treatment; if on the contrary the outcome is a failure, subject n is assigned to the other treatment.

Later, different designs were developed to generalize the rule to the case of three or more treatments and/or to take into account the case of delayed responses (most clinical trials do not result in immediate outcomes and the subject's outcome can be not observable when the next subject arrives): see e.g. [Hoel and Sobel, 1998], [Wei and Durham, 1978], [Andersen and Tamura, 1994], [Bai *et al.*, 2002a], [Biswas, 2003]. These designs are generally presented as a *randomized play-the-winner* rule or as a modified version of this rule. We shall see that this name is misleading, because all these designs alter the original all-or-none rule by replacing it with a "linear" adaptive process.

In spite of its apparent determinism, the play-the-winner rule is a stochastic process, since it depends on the probabilities of success on each treatment. However many people believe that a "less deterministic" rule is better in practice. We shall see that it is not the case.

2 GFU models and extensions

The traditional approach is to depict the adaptive rule as a generalized Friedman's urn (also named as generalized Pólya urn) model (GFU model) [Freedman, 1965]. A typical GFU model for two treatments can be described as follows. When a new subject n arrives, the urn contains (Y_{n-1}^1, Y_{n-1}^2) balls (or "particles" since the number of balls in the urn can be non integer) that represents the two treatments. A ball is drawn at random and replaced. Then the subject is assigned to the corresponding treatment (say t). When the subject outcome is known, balls are added to the urn. For instance, for a dichotomous outcome $u+v$ balls are added: u type t balls and v balls of the other type in case of success; v type t balls and u balls of the other type in case of failure. Then, if we assume an initial urn composition (Y_0^1, Y_0^2) and immediate outcomes, the urn contains at step n (Y_n^1, Y_n^2) balls, with $Y_n^1 + Y_n^2 = Y_0^1 + Y_0^2 + n(u+v)$. Therefore the number of balls in the urn at step n is the same, whatever the previous events are.

Bai, Hu and Shen developed a general class of adaptative designs for T treatments and a dichotomous outcome [Bai *et al.*, 2002a] that extend in a straigthforward way the model above. They considered models with $u = 1$ and $v = 0$. Then the models in the class differ only with respect to the repartition of the balls when the response to treatment t is a failure. They proposed in particular the three following models. GFU model 1 consists of equally adding 1/(T-1) (fractional) balls of each of the other (T-1) types (see [Wei, 1979]); of course it is not very satisfactory. GFU model 2 consists of adding balls proportional to the "known" probabilities of success, but this theoretical model is not applicable in practice. Then in model 3 the unknown probabilities are replaced with the estimated probability of success; this looks more satisfactory, but the model is much more complex and is no longer a GFU.

They investigated the asymptotic properties of this class of models and found them to be "desirable" (see also [Bai *et al.*, 2002b]). It must be emphasized that the case of delayed outcomes is directly taking into account by the models, the urn being updated when outcomes become available; moreover this does not affect the limiting distribution, although the adaptation process can be considerably slowed.

"In order to demonstrate the performance of the new design" the authors gave numerical illustrations. Unfortunately, if we look through their numerical tables ([Bai *et al.*, 2002a], page 17), we can seriously questioned the real value of their asymptotic results for samples of moderate size, even with immediate outcomes.

For instance, let us consider three treatments with probabilities of success 0.50, 0.80 and 0.90. For the "best design" of the authors, the average allocation proportions in a trial of 100 subjects are respectively 0.165, 0.354 and 0.481, and they are very distant from the asymptotic values 0.089, 0.295 and 0.616. Even in a trial of 10 000 subjects the proportions – 0.099, 0.325 and 0.576 – are not what could be expected. So, we were induced to consider other designs that directly generalize the play-the-winner rule and appear to be preferable.

3 Alternative models and some basic results

We shall adopt here an equivalent but slightly different conceptualization. For simplification, we present only the case of two treatments. We represent the *state* of the investigator before subject n arrives by a vector $\mathbf{z}_{n-1} = (z^1_{n-1}, z^2_{n-1})$ where $0 \leq z^i_{n-1} \leq 1$ and $\sum z^i_{n-1} = 1$. For each subject n, there are two observable events: (1) the treatment t_n to which this subject is assigned; $t_n = t^i$ with probability z^i_{n-1}; and (2) the corresponding outcome r_n; $r_n = 1$ (success) with probability φ_1 for t^1 and probability φ_2 for t^2. We assume an initial state $\mathbf{z}_0 = (z^1_0, z^2_0)$.

The probability transition for the GFU model with two treatments described above (named here as Model I) is given in Table 1

t_n	r_n	**Model I**	**Model II**
t^1	1	$z^1_n = \frac{n_0+(n-1)(u+v)}{n_0+n(u+v)} z^1_{n-1} + \frac{u}{n_0+n(u+v)}$	$z^1_n = az^1_{n-1} + (1-a)b$
t^1	0	$z^1_n = \frac{n_0+(n-1)(u+v)}{n_0+n(u+v)} z^1_{n-1} + \frac{v}{n_0+n(u+v)}$	$z^1_n = az^1_{n-1} + (1-a)(1-b)$
t^2	1	$z^1_n = \frac{n_0+(n-1)(u+v)}{n_0+n(u+v)} z^1_{n-1} + \frac{v}{n_0+n(u+v)}$	$z^1_n = az^1_{n-1} + (1-a)(1-b)$
t^2	0	$z^1_n = \frac{n_0+(n-1)(u+v)}{n_0+n(u+v)} z^1_{n-1} + \frac{u}{n_0+n(u+v)}$	$z^1_n = az^1_{n-1} + (1-a)b$

Table 1. Probability transitions for the two classes of models

It must be noted that the initial urn composition (Y^1_0, Y^2_0) is here represented by two parameters with distinct status, on the one hand the initial state $\mathbf{z}_0$ $(z^1_0 = Y^1_0/(Y^1_0 + Y^2_0))$, and on the other hand the parameter n_0 $(= Y^1_0 + Y^2_0)$. Consequently, with the new conceptualization, one can let $n_0 = 0$, so that the initial state only intervenes for the assignment of the first subject, but does not intervene in the probability transition.

In that follows, we shall consider only, as usually done, the particular case $u = 1$ and $v = 0$.

It can be shown that, in order to improve the fastness of the adaptation process, the property of a constant number of balls in the urn at a given step must be relaxed . For this purpose we can then envisage a new class of models, named as Model II, where z_n^1 is again a linear function of z_{n-1}^1, but with constant coefficients. The corresponding probability transition is given in Table 1. It must be emphasized that, unlike Model I, Model II includes the original play-the-winner rule when $a = 0$ and $b = 1$. In this case z_n^1 takes only the values 0 and 1 ("all or none" model).

In that follows, we shall consider only the particular case $b = 1$ and we shall assume $z_0^1 = z_0^2 = 0.5$.

The two models can be characterized by the recurrence relation

$$\mathrm{E}(z_n^1) = A_n \mathrm{E}(z_{n-1}^1) + B_n$$

where A_n and B_n are constants that are function of the model parameters, and furthermore of n for Model I. It can be deduced that

$$\mathrm{E}(z_n^1) = z_0^1 \prod_{i=1}^{n} A_i + \sum_{j=1}^{n} B_j \prod_{i=j+1}^{n} A_i$$

For Model I ($u = 1$ and $v = 0$),

$$A_i = 1 - \frac{2 - \varphi_1 - \varphi_2}{n_0 + i} \quad \text{and} \quad B_i = \frac{1 - \varphi_2}{n_0 + i}$$

For Model II ($b = 0$) A_n and B_n does not depend on n

$$A_i = a + (1 - a)(\varphi_1 + \varphi_2 - 1) \quad \text{and} \quad B_i = (1 - a)(1 - \varphi_2)$$

hence

$$\mathrm{E}(z_n^1) - \psi_1 = (z_0^1 - \varphi_1)\Big(a + (1 - a)\big(1 - \frac{1 - \varphi_2}{\psi_1}\big)\Big)^n$$

For each of the two models, we have asymptoticaly

$$\text{when } n \to \infty, \ \mathrm{E}(z_n^1) \to \psi_1 = \frac{1 - \varphi_2}{1 - \varphi_1 + 1 - \varphi_2}$$

but the convergence is faster for Model II as shown by the two equalities

$$\text{Model I: } \mathrm{E}(z_n^1) - \psi_1 = (z_0^1 - \varphi_1) \prod_{i=1}^{n} \Big(1 - \frac{1 - \varphi_2}{(n_0 + i)\psi_1}\Big)^n$$

$$\text{Model II: } \mathrm{E}(z_n^1) - \psi_1 = (z_0^1 - \varphi_1)\Big(a + (1 - a)\big(1 - \frac{1 - \varphi_2}{\psi_1}\big)\Big)^n$$

Furthermore, for Model II we have the following properties. The smaller a, the smaller $|\mathrm{E}(z_n^1 - \psi_1|$ is, and when $a = 0$ the minimum is such that

$$\mathrm{E}(z_n^1) - \psi_1 = (z_0^1 - \varphi_1)(\varphi_1 + \varphi_2 - 1)^n$$

The closer to one $\varphi_1 + \varphi_2$, the smaller $|\mathrm{E}(z_n^1) - \psi_1|$ is, and for $\varphi_1 + \varphi_2 = 1$, $\mathrm{E}(z_n^1) = \psi_1$ $(\forall n\ \forall z_0^1)$.

Let T_N^1 be the number of subjects assigned to treatment t^1 in a trial of N subjects. It can be deduced that

$$\mathrm{E}(T_N^1) = \frac{1}{N}\sum_{n=0}^{N-1} \mathrm{E}(z_n^1) = \psi_1 + \frac{1}{N}(z_0^1 - \varphi_1)\frac{1-h^N}{1-h}$$

$$\text{where } h = a + (1-a)(\varphi_1 + \varphi_2 - 1)$$

$$= \varphi_1 + \varphi_2 - 1 \quad \text{if } a = 0$$

Table 2 illustrates the superiority of the all-or-none model for the probability of success $\varphi_1 = 0.60$ and $\varphi_2 = 0.80$. The possibility of setting $n_0 = 0$ in Model I improves the average allocation proportion, but notably increases the standard deviation.

$\varphi_1 = 0.60$ $\varphi_2 = 0.80$

$N = 50$ subjects

Model I $n_0 = 1$	Model I $n_0 = 0$	Model II $a = 0$	$N \to \infty$
0.618 (0.149)	0.649 (0.186)	0.661 (0.101)	0.667

Table 2. Comparison of Models I and II with two treatments: average allocation proportions (exact) for treatment t^2 (standard deviations estimated from 10^6 replications)

4 Generalizations

The two class of models can be easily generalized to the case of $\mathrm{T} > 2$ treatments. We can translate as a Model I each of the particular models (1, 2 and 3) considered by Bai *et al.* (and other related models proposed). We can also associate a Model II to each of these models; these models differ with respect to the probability transition in case of failure, while for $a = 0$ they comply with the original play-the-winner rule which is to repeat the treatment in case of success. As for Model I, delayed outcomes are directly taking into account. Moreover, it can be demonstrated that each particular Model II has the same asymptotic properties as the corresponding Model I. But it always perform better, the adaptation process being the fastest when $a = 0$.

$\varphi_1 = 0.50 \quad \varphi_2 = 0.80 \quad \varphi_3 = 0.90$

	Model I-1		Model II-1 ($a = 0$)		
	$N = 100$	$N = 300$	$N = 100$	$N = 300$	$N \to \infty$
t^1	0.181 (0.088)	0.135 (0.063)	0.122 (0.053)	0.089 (0.036)	0.118
t^2	0.355 (0.152)	0.349 (0.127)	0.299 (0.119)	0.296 (0.073)	0.294
t_3	0.464 (0.165)	0.516 (0.137)	0.579 (0.134)	0.615 (0.079)	0.588
	Model I-3		Model II-3 ($a = 0$)		
	$N = 100$	$N = 300$	$N = 100$	$N = 300$	$N \to \infty$
t^1	0.165 (0.092)	0.135 (0.063)	0.097 (0.056)	0.089 (0.036)	0.089
t^2	0.354 (0.157)	0.349 (0.127)	0.296 (0.127)	0.296 (0.073)	0.295
t_3	0.481 (0.167)	0.516 (0.137)	0.607 (0.136)	0.615 (0.079)	0.616

Table 3. Comparison of Models I and II with three treatments: average allocation proportions and standard deviations between parentheses (estimated from 10^6 replications)

This is illustrated in table 3 for the probability of success $\varphi_1 = 0.50$, $\varphi_2 = 0.80$ and $\varphi_3 = 0.90$.

We have also computed the proportions of the different orders of treatment allocations in each replication. Table 4 illustrates again the manifest superiority of Model II. For the same probability of success, when $N = 300$, for a given trials there is for instance about a 98% chance with Model II-3 that a majority of subjects is assigned to the most effective treatment against only about a 74% chance with Model I-3.

$\varphi_1 = 0.50 \quad \varphi_2 = 0.80 \quad \varphi_3 = 0.90$

	Model I-1		Model II-1	
	$N = 100$	$N = 300$	$N = 100$	$N = 300$
$t^1 \leq t^2 \leq t^3$ $+ + +$	0.489	0.676	0.801	0.975
$t^2 \leq t^1 \leq t^3$ $- - +$	0.136	0.060	0.064	0.007
$t^3 \leq t^2 \leq t^1$ $- + -$	0.015	0.001	0.000	0.000
$t^1 \leq t^3 \leq t^2$ $+ - -$	0.287	0.253	0.129	0.018
$t^3 \leq t^1 \leq t^2$ or $t^2 \leq t^3 \leq t^1$ $- - -$	0.073	0.010	0.003	0.000
	Model I-3		Model II-3	
	$N = 100$	$N = 300$	$N = 100$	$N = 300$
$t^1 \leq t^2 \leq t^3$ $+ + +$	0.511	0.675	0.814	0.975
$t^2 \leq t^1 \leq t^3$ $- - +$	0.136	0.061	0.071	0.006
$t^3 \leq t^2 \leq t^1$ $- + -$	0.012	0.001	0.000	0.000
$t^1 \leq t^3 \leq t^2$ $+ - -$	0.281	0.254	0.113	0.018
$t^3 \leq t^1 \leq t^2$ or $t^2 \leq t^3 \leq t^1$ $- - -$	0.061	0.010	0.003	0.000

Table 4. Comparison of Models I and II with three treatments: Proportions of the different orders of treatment allocations (estimated from 10^6 replications)

5 Bayesian methods

In conclusion, our results are very incentive: the simplest model is the best! This greatly facilitates both a thoughtful planning into the design phase and the use of efficient inference procedures.

For this purpose, the Bayesian statistical methodology can be used for designing the study (how many subjects?) and for comparing the treatments. A clinical trial is generally expected to bring evidence by itself. So it is desirable in clinical research to assume noninformative priors for objective report in publication, the posterior distribution being based solely on the data. But alternative choices of priors may be used to refining inference.

Moreover, the Bayesian predictive approach is a very appealing method for monitoring the study and in particular for stopping it early if necessary (e.g., [Spiegelhalter *et al.*, 1986, Lecoutre *et al.*, 1995, Lecoutre *et al.*, 2002]). It simulates the probability of achieving the trial target, conditionally on available data and simple conjectures about the future observations. The simulations can be explicitly based on either the hypotheses used to design the study, expressed in terms of the prior distribution, or on available data, or on both.

We shall briefly illustrate Bayesian methods for the basic situation of an adaptative design with two treatments using Model II (with $a = 0$). We also assume immediate outcomes. The sequel of treatment allocations (t_1, $t_2 \ldots t_n$, $t_{n+1} \ldots t_{N+1}$) contains all the information in the data. Indeed, $t_n = t_{n+1}$ implies that a success to t_n has been observed and $t_n \neq t_{n+1}$ implies that a failure to t_n has been observed. Moreover, the likelihood function is simply

$$l(\varphi_1, \varphi_2)|(t_1, \ldots t_{N+1}) = \frac{1}{2}\,\varphi_1^{n_{11}}(1-\varphi_1)^{n_{10}}\varphi_2^{n_{21}}(1-\varphi_2)^{n_{20}}$$

where n_{ij} is the number of pairs (t_n, t_{n+1}) equal to (t^i, t^j), so that n_{11} and n_{21} are the respective numbers of success to treatments t^1 and t^2, and n_{10} and n_{20} are the numbers of failure (1/2 is the probability of t_1).

Bayesian methods only involve the likelihood function and are immediately available. This results from the fact that the likelihood function is identical (up to a multiplicative constant) with the likelihood function associated with the comparison of two independent binomial proportions. Therefore we can apply the same Bayesian procedures. A simple and usual solution assumes two independent beta prior distributions for φ_1 and φ_2: respectively $\beta(\nu_{11}, \nu_{10})$ and $\beta(\nu_{21}, \nu_{20})$. The marginal posterior distribution are again two independent beta distributions: $\beta(\nu_{11} + n_{11}, \nu_{10} + n_{10})$ and $\beta(\nu_{21} + n_{21}, \nu_{20} + n_{20})$. The predictive distributions for future observations are two independent beta-binomial distributions ([Lecoutre *et al.*, 1995]).

Let us consider for illustration the results of a trial with $N = 100$ subjects. The observed rates of success are respectively 17 out of 31 attributions for treatment t^1 and 56 out of 69 attributions for treatment t^2.

A joint probability statement is, in a way, the best summary of the posterior distribution. For instance, if we conventionally adopt the Jeffreys prior ($\nu_{11} = \nu_{10} = \nu_{21} = \nu_{20} = .5$), the joint posterior probability that $\varphi_1 < 0.712$ and $\varphi_2 > 0.708$ is 0.95.

However, a statement that deals with the comparison of the two treatments directly would be preferable. So we have a probability 0.996 that $\varphi_2 > \varphi_1$. Moreover, the main classical criteria for comparing two proportions can be dealt with. This is easily solved in the Bayesian approach, since the distribution of any derived parameter of interest can be easily obtained from the joint posterior distribution using numerical methods. For instance, we find the 95% credible intervals $[+0.068, +0.453]$ for $\varphi_2 - \varphi_1$, $[1.10, 2.18]$ for φ_2/φ_1 and $[1.41, 9.07]$ for $(\varphi_2/(1-\varphi_2))/\varphi_1/(1-\varphi_1))$.

For the Jeffreys prior, Bayesian methods have fairly good frequentist coverage properties for interval estimates, even in the the cases of moderate sample sizes and small parameter values (see e.g., [Lecoutre and Charron, 2000]). As an illustration, 10^5 samples of size $N = 50$ were generated for different set of parameter values. We considered the inference about the difference $\varphi_2 - \varphi_1$. The proportion of samples for which respectively the 95% lower and 95% upper limits were respectively greater and smaller than the true difference are reported in Table 5.

	Lower limit	Upper limit
$\varphi2 = 0.80$ $\varphi1 = 0.80$	0.059	0.057
$\varphi2 = 0.60$ $\varphi1 = 0.60$	0.053	0.053
$\varphi2 = 0.50$ $\varphi1 = 0.50$	0.052	0.051
$\varphi2 = 0.80$ $\varphi1 = 0.70$	0.058	0.052
$\varphi2 = 0.70$ $\varphi1 = 0.60$	0.056	0.049
$\varphi2 = 0.60$ $\varphi1 = 0.50$	0.053	0.051
$\varphi2 = 0.80$ $\varphi1 = 0.60$	0.055	0.053
$\varphi2 = 0.70$ $\varphi1 = 0.50$	0.058	0.050
$\varphi2 = 0.60$ $\varphi1 = 0.40$	0.058	0.047

Table 5. Coverage properties of Bayesian credible intervals for the comparison of two treatments: proportions of errors for the 95% lower and 95% upper limits (10^5 replications)

These methods can be easily generalized with virtually no more conceptual difficulties to the case of several treatments and/or delayed outcomes. The Bayesian approach is appropriate as well for a definitely decisional trial (e.g., for selecting the best treatment) as for estimation (e.g., for assessing the difference in efficacy between two treatments). Moreover, the predictive approach enables the trial to be stopped early, or on the contrary to be extended to an adequate size, in a sequential perspective that fits with the methodological principle of adaptative designs.

References

[Andersen and Tamura, 1994]D. Andersen, J.and Faries and R. Tamura. A randomized play-the-winner design for multiarm clinical trials. *Communications in Statistics, Theory and Methods*, pages 309–323, 1994.

[Bai *et al.*, 2002a]Z. D. Bai, F. Hu, and W. F. Rosenberger. Asymptotic properties of adaptive designs for clinical trials with delayed response. *The Annals of Statistics*, pages 122–139, 2002.

[Bai *et al.*, 2002b]Z. D. Bai, F. Hu, and L. Shen. An adaptive design for multi-arm clinical trials. *Journal of Multivariate Analysis*, pages 1–18, 2002.

[Biswas, 2003]A. Biswas. Generalized delayed response in randomized play-the-winner rule. *Communications in Statistics, Simulation and Computation*, pages 259–274, 2003.

[Freedman, 1965]D. Freedman. Bernard friedman's urn. *The Annals of Mathematical Statistics*, pages 956–970, 1965.

[Hoel and Sobel, 1998]D. G. Hoel and M. Sobel. Comparison of sequential procedures for selecting the best binomial population. In *Proceedings of the Sixth Berkeley Symposium on Probability and Statistics*, pages 53–69, 1998.

[Lecoutre and Charron, 2000]B. Lecoutre and C. Charron. Bayesian procedures for prediction analysis of implication hypotheses in 2 × 2 contingency tables. *Journal of Educational and Behavioral Statistics*, pages 185–201, 2000.

[Lecoutre *et al.*, 1995]B. Lecoutre, G. Derzko, and J.-M. Grouin. Bayesian predictive approach for inference about proportions. *Statistics in Medicine*, pages 1057–1063, 1995.

[Lecoutre *et al.*, 2002]B. Lecoutre, B. Mabika, and G. Derzko. Assessment and monitoring in clinical trials when survival curves have distinct shapes in two groups: a bayesian approach with weibull modeling. *Statistics in Medicine*, pages 663–674, 2002.

[Spiegelhalter *et al.*, 1986]D.J. Spiegelhalter, L.S. Freedman, and P.R. Blackburn. Monitoring clinical trials: Conditional or predictive power? *Controlled Clinical Trials*, pages 8–17, 1986.

[Wei and Durham, 1978]L. J. Wei and S. Durham. The randomized play-the-winner rule in medical trial. *Journal of the American Statistical Association*, pages 840–843, 1978.

[Wei, 1979]L. J. Wei. The generalized pólya urn design for sequential medical trials. *Annals of Statistics*, pages 291–296, 1979.

[Zelen, 1969]M. Zelen. Play the winner rule and the controlled clinical trial. *Journal of the American Statistical Association*, pages 131–146, 1969.

Part XI

Markov Processes

Time-Average Optimality for Semi-Markov Control Processes with Feller Transition Probabilities

Anna Jaśkiewicz[1] and Andrzej S. Nowak[2]

[1] Instytut Matematyki
Politechnika Wrocławska
50-370 Wrocław, Poland
(e-mail: ajaskiew@im.pwr.wroc.pl)

[2] Wydział Matematyki, Informatyki i Ekonometrii
Uniwersytet Zielonogórski
65-516 Zielona Góra, Poland
(e-mail: a.nowak@wmie.uz.zgora.pl)

Abstract. Semi-Markov control processes with Borel state space and Feller transition probabilities are considered. We prove that under fairly general conditions the two expected average costs: the time-average and the ratio-average coincide for stationary policies. Moreover, the optimal stationary policy for the ratio-average cost criterion is also optimal for the time-average cost criterion.

Keywords: semi-Markov control models, average cost optimality equation.

1 The model

Let X and A be Borel spaces, the state and the action space, respectively. By $A(x)$ we denote the compact set of actions available in state x. Define

$$K := \{(x, a) : x \in X, a \in A(x)\},$$

the set of admissible pairs as a Borel subset of $X \times A$.

If the current state is x and an action $a \in A(x)$ is selected, then the immediate cost of $c_1(x, a)$ is incurred and the system remains in state $x_0 = x$ for a

random time T with the cumulative distribution $G(\cdot|x, a)$ depending only on x and a. The cost of $c_2(x, a)$ per unit time is incurred until the next transition occurs. Afterwards the system jumps to the state $x_1 = y$ according to the probability distribution (*transition law*) $q(\cdot|x, a)$. This procedure repeats itself and yields a trajectory $(x_0, a_0, t_1, x_1, a_1, t_2, \ldots)$ of some stochastic process, where x_n is the state, a_n is the control variable and t_n is the time of the nth transition, $n \geq 0$.

A *control policy* $\pi = \{\pi_n\}$ and a stationary policy $\pi = \{f, f, \ldots\}$ are defined in a usual way. By Π and F we denote the set of all policies and the set of all stationary policies, respectively. Further, we will identify any stationary policy $\pi = \{f, f, \ldots\}$ with $f \in F$.

Let $(\Omega, \mathcal{F})$ be the measurable space consisiting of the sample space $\Omega := (X \times A \times [0, +\infty))^\infty$ and the corresponding product σ-algebra $\mathcal{F}$. Obviously, any policy π, the transition law q, and the conditional cumulative distribution function G of the differences $\{T_{n+1} - T_n\}$ generate the stochastic process $\{x_n, a_n, T_n\}$, $n \geq 0$ on $(\Omega, \mathcal{F})$.

Let E_x^π be the expectation operator with respect to the probability measure P_x^π defined on the product space Ω.

Let $\pi \in \Pi$, $x \in X$ and $t \geq 0$ be fixed. Put

$$N(t) := \max\{n \geq 0 : T_n \leq t\}$$

as the counting process, and

$$\tau(x,a) := \int_0^\infty t P_x^a(dt) = \int_0^\infty t G(dt|x,a) = E_x^a T$$

as the mean holding (sojourn) time. By our assumptions $P_x^\pi(N(t) < \infty) = 1$

We shall consider the two average expected costs:
- the *ratio-average cost*

$$J(x,\pi) := \limsup_{n\to\infty} \frac{E_x^\pi \left(\sum_{k=0}^{n-1} c(x_k, a_k)\right)}{E_x^\pi T_n},$$

- the *time-average cost*

$$j(x,\pi) := \limsup_{t\to\infty} \frac{E_x^\pi \left(\sum_{k=0}^{N(t)} c(x_k, a_k)\right)}{t},$$

where

$$c(x,a) := c_1(x,a) + \tau(x,a)c_2(x,a)$$

for each $(x,a) \in K$.
We impose the following assumptions on the model.

(**B**) *Basic assumptions*:
(i) for each $x \in X$, $A(x)$ is a compact metric space and, moreover, the set-valued mapping $x \mapsto A(x)$ is upper semicontinuous, i.e. $\{x \in X : A(x) \cap B \neq \emptyset\}$ is closed for every closed set B in A;
(ii) the cost function c is lower semicontinuous on K;
(iii) the transition law q is weakly continuous on K, i.e.,

$$\int_X u(y) q(dy|x,a)$$

is continuous function of (x,a) for every bounded continuous function u on X;

(iv) the mean holding time τ is continuous on K, and there exist positive constants b and B such that

$$b \leq \tau(x,a) \leq B$$

for all $(x,a) \in K$;
(v) there exist a constant $L > 0$ and a continuous function $V : X \mapsto [1,\infty)$ such that $|c(x,a)| \leq LV(x)$ for every $(x,a) \in K$;
(vi) the function

$$\int_X V(y)(dy|x,a)$$

is continuous on K.
(**GE**) *Geometric ergodicity assumptions*:
(i) there exists a Borel set $C \subset X$ such that for some $\lambda \in (0,1)$ and $\eta > 0$, we have

$$\int_X V(y)q(dy|x,a) \leq \lambda V(x) + \eta 1_C(x)$$

for each $(x,a) \in K$; V is the function introduced in (**B**, v);
(ii) the function V is bounded on C, i.e.,

$$v_C := \sup_{x \in C} V(x) < \infty;$$

(iii) there exist some $\delta \in (0,1)$ and a probability measure μ concentrated on the Borel set C with the property that

$$q(D|x,a) \geq \delta\mu(D)$$

for each Borel set $D \subset C$, $x \in C$ and $a \in A(x)$.

For any function $u : X \mapsto R$ define the V-norm

$$\|u\|_V := \sup_{x \in X} \frac{|u(x)|}{V(x)}.$$

By L_V^∞ we denote the Banach space of all Borel measurable functions u for which $\|u\|_V$ is finite.

Let L_V denote the subset of L_V^∞ consisting of all lower semicontinuous functions.

Under (**GE**) the embedded state process $\{x_n\}$ governed by a stationary policy is a positive recurrent aperiodic Markov chain and for each stationary policy f, there exists a unique invariant probability measure, denoted by π_f (see Theorem 11.3.4 and page 116 in [Meyn and Tweedie, 1993]). Moreover, by Theorem 2.3 in [Meyn and Tweedie, 1994], $\{x_n\}$ is V-uniformly ergodic. Thi results in the following

$$J(f) := J(x,f) = \frac{\int_X c(x,f(x))\pi_f(dx)}{\int_X \tau(x,f(x))\pi_f(dx)}$$

for every $f \in F$.

We also make two additional assumptions on the sojourn time T.

(**R**) *Regularity condition*:
there exist $\epsilon > 0$ and $\beta < 1$ such that

$$P_x^a(T \leq \epsilon) \leq \beta$$

for all $x \in C$ and $a \in A(x)$.
(**I**) *Uniform integrability condition*:

$$\lim_{t \to \infty} \sup_{x \in C} \sup_{a \in A(x)} P_x^a(T > t) = 0.$$

For further and broad discussion of the assumptions the reader is referred to [Jaśkiewicz, 2001] and [Ross, 1970].

2 Main results

In this section we present two new theorems on SMCPs with Borel state spaces. Theorem 1 concerns the existence of the optimal stationary policy for the ratio-average criterion. The proof combines some ideas and tools used in [Jaśkiewicz, 2001].

For the ε-perturbed SMCPs, we prove that the associated with them the average cost optimality equation has a solution.

Next, taking into account slightly modified solutions, we obtain a certain optimality inequality, which is enough to obtain an average optimal policy. It is worth pointing out that compared with previous work [Jaśkiewicz, 2001] in the limit passage we need to use of Fatou's lemma for weakly convergent measures [Serfozo, 1982].

Theorem 1. Assume (**B**, **GE**). There exist a constatant g^*, a function $h_* \in L_V$ and $f^* \in F$ such that

$$\begin{aligned} h_*(x) &\geq \min_{a \in A(x)} \left[c(x,a) - g^* \tau(x,a) + \int_X h_*(y) q(dy|x,a) \right] \qquad (1) \\ &= c(x, f^*(x)) - g^* \tau(x, f^*(x)) + \int_X h_*(y) q(dy|x, f^*(x)) \end{aligned}$$

for all $x \in X$. Moreover, f^* is an average optimal policy and g^* is optimal cost with respect to the ratio-average criterion, i.e.,

$$g^* = \inf_{\pi \in \Pi} J(x, \pi) = J(f^*)$$

for every $x \in X$.

Theorem 2 deals with the equivalence of the two expected average cost criteria for SMCPs with Feller transition probabilities. Related result under the strong continuity of $q(\cdot|x,a)$ in $a \in A(x)$ is given in [Jaśkiewicz, 2004].

To obtain the mentioned equivalence we use two inequalities as the point of departure. Using them we define a supermartingale and submartingale, and then by Doob's theorem we obtain the equality of the two optimal costs according to the ratio-average and time-average cost criteria. To apply the optional sampling theorem we have to prove the uniform integrability of the supermartingale and submartingale involved. This issue is studied in [Jaśkiewicz, 2004]. The whole analysis relies on dealing with the consecutive returns of the process (induced by q, an arbitrary π, and the cumulative distribution G) to the small set C.

Theorem 2. Assume (**B**, **GE**, **R**, **I**). Then
(a) $g^* = \inf_{\pi \in \Pi} j(x,\pi)$;
(b) $j(x,f) = J(x,f)$ for any $f \in F$.

References

[Jaśkiewicz, 2001]A. Jaśkiewicz. An approximation approach to ergodic semi-markov control processes. *Math. Methods Oper. Res.*, pages 1–19, 2001.

[Jaśkiewicz, 2004]A. Jaśkiewicz. On the equivalence of two expected average cost criteria for semi-markov control processes. *Math. Oper. Res.*, pages 326–338, 2004.

[Meyn and Tweedie, 1993]S.P. Meyn and R.L. Tweedie. *Markov Chains and Stochastic Stability.* Springer-Verlag, New York, 1993.

[Meyn and Tweedie, 1994]S.P. Meyn and R.L. Tweedie. Computable bounds for geometric convergence rates of markov chains. *Ann. Appl. Prob.*, pages 981–1011, 1994.

[Ross, 1970]S.M. Ross. *Applied Probability Models with Optimization Applications.* Holden-Day, San Francisco, 1970.

[Serfozo, 1982]R. Serfozo. Convergence of lebesgue integrals with varying measures. *Sankhya, Ser. A*, pages 380–402, 1982.

Centered Semi-Markov Random Walk in Diffusion Approximation Scheme

Vladimir S. Koroliuk[1] and Nikolaos Limnios[2]

[1] Institute of Mathematics,
National Academy of Sciences,
Kiev 01001, Ukraine

[2] Laboratoire de Mathématiques Appliquées,
Université de Technologie de Compiègne,
B.P. 20529, 60205 Compiègne Cedex, France

Abstract. In this paper we present a diffusion approximation algorithm for a centered semi-Markov random walks in the series scheme with the small parameter series $\varepsilon \to 0$, $(\varepsilon > 0)$.

1 The semi-Markov random walk

The semi-Markov random walk (SMRW) is defined on the real line $\mathbf{R} = (-\infty, +\infty)$ by the superposition of two independent renewal processes of the i.i.d. sequences of nonnegative random variables $\alpha_k^{\pm}$, $k \geq 1$, and the two sequences of nonnegative independent i.i.d. random variables $\beta_k^{\pm}, k \geq 1$, as follows

$$\zeta(t) = u + \sum_{k=1}^{\nu^+(t)} \beta_k^+ - \sum_{k=1}^{\nu^-(t)} \beta_k^-, \quad t \geq 0. \tag{1}$$

The renewal process are

$$\nu^{\pm}(t) := \max \left\{ n : \sum_{k=1}^{n} \alpha_k^{\pm} \leq t \right\}, \quad t \geq 0. \tag{2}$$

The distribution functions

$$P_{\pm}(t) = P\{\alpha_k^{\pm} \leq t\}, \quad G_{\pm}(u) = P\{\beta_k^{\pm} \leq u\} \tag{3}$$

are given.

SMRW (1) was investigated in average, diffusion and Poisson approximation schemes under distinct assumption of semi-continuity [Korolyuk and Korolyuk, 1999], [Korolyuk, 1997], [Korolyuk, 1999], etc. This kind of processes are interesting for various applied problems. The number of customs in the queue system is described by (1) with the given distribution function of arrival and service time and with . The process (1) can be considered as a mathematical model of risk process with arbitrary distribution of interval between moments of payment of claims and the premium income.

In this paper we discuss a centered normalized in diffusion approximation scheme (see process (12) below).

2 The superposition of two renewal processes

Relation (2) can be described by the counting process

$$\nu(t) = \nu^{+}(t) + \nu^{-}(t), \quad t \geq 0, \tag{4}$$

for the Markov renewal process $x_n, \tau_n, n \geq 0$, on the phase space $E = E_{+} \cup E_{-}, E_{\pm} = \{\pm, x > 0\}$ by the formula of sojourn times $\theta_{n+1} := \tau_{n+1} - \tau_n, \quad n \geq 0$ [Korolyuk and Korolyuk, 1999]:

$$\theta_x^{\pm} = \alpha^{\pm} \wedge x. \tag{5}$$

The transition probabilities of the **embedded Markov chain (EMC)** $x_n, \theta_n, n \geq 0$, is defined by the matrix [Korolyuk and Limnios, (2004b]

$$P(x, dy) = \begin{pmatrix} P_{+}(x - dy) & P_{+}(x + dy) \\ P_{-}(x + dy) & P_{-}(x - dy) \end{pmatrix}.$$

The stationary distribution of the EMC has the density

$$\rho_{\pm}(t) = \overline{P}_{\mp}(t)/a, \quad a := a_{+} + a_{-}, \quad a_{\pm} := \mathbf{E}\alpha^{\pm}. \tag{7}$$

As usual, $\overline{P}_{\pm}(t) := 1 - P_{\pm}(t)$.

The embedded SMRW $\zeta_n := \zeta(\tau_n), n \geq 0$, is defined by the relations

$$\begin{gathered} \zeta_{n+1} = \zeta_n + \beta_{n+1}, \quad n \geq 0, \\ \beta_{n+1} := \beta_{n+1}^{+} I(x_{n+1} \in E_{+}) - \beta_{n+1}^{-} I(x_{n+1} \in E_{-}), \end{gathered} \tag{8}$$

where, as usual, $I(A)$ is the indicator of a random event A.

The SMRW (1) can be defined as follows: $\zeta(t) = \zeta_{\nu(t)}, t \geq 0$. It is worth noticing that the average drift per unit time of the SMRW (1) is defined by the value

$$b = b_{+}/a_{+} - b_{-}/a_{-}, \quad b_{\pm} := E\beta^{\pm}. \tag{9}$$

The average algorithm for SMRW (1) is realized in the following series scheme with the small series parameter $\varepsilon \to 0$ $(\varepsilon > 0)$:

$$\zeta_\varepsilon(t) = u + \varepsilon \sum_{k=1}^{\nu^{+}(t/\varepsilon)} \beta_k^{+} - \varepsilon \sum_{k=1}^{\nu^{-}(t/\varepsilon)} \beta_k^{-}, \quad t \geq 0. \tag{10}$$

Under the condition $b \neq 0$, the weak convergence takes place:

$$\zeta_\varepsilon(t) \Rightarrow \zeta_0(t) = u + bt, \quad \varepsilon \to 0. \tag{11}$$

3 The algorithm of diffusion approximation

The centered SMRW in the series scheme is considered as follows:

$$\zeta^{\varepsilon}(t) = u + \varepsilon \left[\sum_{k=1}^{\nu^{+}(t/\varepsilon^2)} \beta_k^{+} - \varepsilon \sum_{k=1}^{\nu^{-}(t/\varepsilon^2)} \beta_k^{-} \right]^{+} - b\tau(t/\varepsilon^2). \tag{12}$$

The renewal process $\tau(t) := \tau_{\nu(t)}, t \geq 0$, defines the last renewal moment before time t.

Introduce the random variables

$$\gamma_n := \beta_n - b\theta_n, \quad n \geq 1, \tag{13}$$

it is worth noticing that, for any $x \in E_\pm$,

$$b_\pm(x) := \mathbf{E}[\beta_{n+1}|x_n = x] = \pm[b_\pm P_\pm(x) - b_\mp \overline{P}_\pm(x)] \tag{14}$$

and

$$\widetilde{b}_\pm(x) := \mathbf{E}[\gamma_{n+1}|x_n = x] = b_\pm(x) - ba_\pm(x), \tag{15}$$

where

$$a_\pm(x) := \mathbf{E}[\theta_{n+1}|x_n = x] = \int_0^\infty \overline{P}_\pm(t)dt. \tag{16}$$

The centered SMRW (12) can be represented in the following form:

$$\zeta^\varepsilon(t) = u + \varepsilon \sum_{n=1}^{\nu(t/\varepsilon^2)} \gamma_n, \quad t \geq 0. \tag{17}$$

Theorem 1 *Let $b \neq 0$ defined in (9) and the third moments $E[\beta_n^\pm]^3 < \infty$. Then the weak convergence*

$$\zeta^\varepsilon(t) \Rightarrow \zeta^0(t) = u + \sigma w(t), \quad \varepsilon \to 0 \tag{18}$$

takes place. The variance σ^2 of the standard Wiener process $w(t)$ is calculated by the formulae:

$$\begin{array}{c} \sigma^2 = \sigma_0^2 + \sigma_1^2 - \sigma_2^2, \\ \sigma_0^2 = q\int_0^\infty [\rho_+(x)C_+(x) + \rho_-(x)C_-(x)]dx, \\ \sigma_1^2 = 2\int_0^\infty [\pi_+(x)h_+(x) + \pi_-(x)h_-(x)]dx, \\ \sigma_2^2 = 2q\int_0^\infty [\rho_+(x)\widetilde{b}_+^2(x) + \rho_-(x)\widetilde{b}_-^2(x)]dx. \end{array} \tag{19}$$

Here, by definition:

$$C_\pm := E[\gamma_{n+1}^2|x_n = x],$$

$$h_\pm := -\widetilde{b}_\pm^0(x)R_0^\pm \widetilde{b}_\pm^0(x), \widetilde{b}_\pm^0(x) := \widetilde{b}_\pm / a_\pm(x),$$

$$\pi_\pm(x) := q\rho_\pm(x)a_\pm(x), q := 1/a_+ + 1/a_-,$$

where $x \in E_\pm$.

The potential operator $R_0^\pm$ is defined for the generator of the Markov kernel

$$Q = q(x)[P - I].$$

Remark. It is worth noticing that $\sigma_1^2 - \sigma_2^2 \geq 0$.

4 Scheme of Proof

The construction of the algorithm of diffusion approximation is realized by the scheme introduced in our papers [Korolyuk and Limnios, 2004a] and [Korolyuk and Limnios, (2004b].

The compensating operator $\mathbf{L}^\varepsilon$ of the extended Markov renewal process

$$\zeta_n^\varepsilon := \zeta^\varepsilon(\tau_n^\varepsilon), \quad x_n, \quad \tau_n^\varepsilon := \varepsilon^2\tau_n, \quad n \geq 0 \tag{20}$$

on the test-function $\varphi(u, \cdot) \in C^3(R)$ admit the asymptotic representation

$$\mathbf{L}^\varepsilon\varphi(u,x) = [\varepsilon^{-2}Q + \varepsilon^{-1}Q_1(x) + Q_2(x)]\varphi(u,x) + \theta_l^\varepsilon\varphi(u,x) \tag{21}$$

where

$$Q_1(x)\varphi(u) = q(x)P\widetilde{b}(x)\varphi'(u), \tag{22}$$

$$Q_2(x)\varphi(u) = \frac{1}{2}q(x)PC(x)\varphi''(u), \tag{23}$$

and the remainder operator θ_l^ε satisfies the negligible condition:

$$||\theta_l^\varepsilon\varphi(u)|| \to 0, \varepsilon \to 0, \varphi(u) \in C^3(R). \tag{24}$$

The limit operator

$$\mathbf{L}\varphi(u) = \frac{1}{2}\sigma^2\varphi''(u)$$

is determined by a solution of the singular perturbation problem for the truncated operator

$$\mathbf{L}_0^\varepsilon\varphi^\varepsilon := [\varepsilon^{-2}Q+\varepsilon^{-1}Q_1+Q_2](\varphi(u)+\varepsilon\varphi_1(u,x)+\varepsilon^2\varphi_2(u,x)) = \mathbf{L}\varphi(u)+\theta_0^\varepsilon\varphi(u). \tag{25}$$

According to Lemma 3.3 [Korolyuk and Korolyuk, 1999] (p.51) the operator $\mathbf{L}$ in (25) is calculated by the formula

$$\mathbf{L}\Pi = \Pi Q_2\Pi - \Pi Q_1 R_0 Q_1\Pi, \tag{26}$$

where the projector Π is defined by the stationary distribution of the associated Markov process with the generator $Q = q(x)[P - I], q(x) = 1/m(x), m(x) := E\theta_x$.

After some computation we obtain the result of Theorem 1.

The verification of the algorithm of diffusion approximation follows some familiar procedure in the theory of convergence of stochastic processes [Ethier and Kurtz, 1986], [Jacod and Shiryaev, 1987], adapted to the semi-Markov switching process in [Korolyuk and Limnios, 2002a], [Korolyuk and Limnios, 2004a], [Korolyuk and Limnios, (2004b].

References

[Bratiichuk, 1995]N.S. Bratiichuk (1995). Limit theorems for some characteristics of system In *Exploring Stochastic Laws*, VSP, 77-90.

[Ethier and Kurtz, 1986]S.N. Ethier and T.G. Kurtz (1986). *Markov Processes: Characterization and convergence*, J. Wiley, New York.

[Jacod and Shiryaev, 1987]J. Jacod and A.N. Shiryaev (1987). *Limit Theorems for Stochastic Processes*, Springer-Verlang, Berlin.

[Korolyuk, 1997]V.S. Korolyuk (1997). Ruin Problems. Explicit and asymptotic approaches. *Theory of Stoch. Processes*, 2 (18), No.1-2, 4-14.

[Korolyuk, 1999]V. S. Korolyuk (1999). Semi-Markov random walk, In Semi-Markov Models and Applications, J. Janssen, N. Limnios (Eds.), pp 61– 75, Kluwer, Dordrecht.

[Korolyuk and Korolyuk, 1999]V. S. Korolyuk and V. V. Korolyuk (1999).

[Korolyuk and Limnios, 2002a]V. S. Korolyuk and N. Limnios, (2002). “Markov additive processes in a phase merging scheme”, *Theory Stochastic Processes*, vol. 8, no 24, pp 213–226.

[Korolyuk and Limnios, 2002b]V. S. Korolyuk, N. Limnios, (2002). “Poisson Approximation of Homogeneous Stochastic Additive Functionals with Semi-Markov Switching”, *Theory of Probability and Mathematical Statistics*, vol. 64, pp 75–84.

[Korolyuk and Limnios, 2004a]V. S. Korolyuk and N. Limnios, (2004). “Poisson approximation of increment processes with Markov switching”, *Theor. Probab. Appl.*, 49(4), 1-18.

[Korolyuk and Limnios, 2004c]V. S. Korolyuk and N. Limnios, (2004). “Average and diffusion approximation for evolutionary systems in an asymptotic split phase space”, *Annals Appl. Probab.*, 14(1), pp 489–516.

[Korolyuk and Limnios, (2004b]V. S. Korolyuk, N. Limnios, (2004). ”Semi-Markov random walk in Poisson approximation scheme”, *Communication in Statistics - Theory and Methods*, 33(3), pp 507–516.

[Prabhu, 1980]N.U. Prabhu (1980). *Stochastic Storage Processes*, Springer-Verlag, Berlin.

[Silvestrov, 2004]D. S. Silvestrov (2004). *Limit Theorems for Randomly Stopped Stochastic Processes.* Series: Probability and its Applications, Springer.

Nonparametric Estimation for Semi-Markov Processes Based on $\boldsymbol{K}$-Sample Paths with Application to Reliability

Nikolaos Limnios[1] and Brahim Ouhbi[2]

[1] Laboratoire de Mathématiques Appliquées,
Université de Technologie de Compiègne,
B.P. 20529, 60205 Compiègne Cedex, France
[2] Ecole Nationale Supérieure d'Arts et Métiers
Marjane II, Meknès Ismailia,
Béni M'Hamed, Meknès, Maroc,

Abstract. The problem concerned here is the estimation of ergodic finite semi-Markov processes from data observed by considering K independent censored sample paths with application in the reliability.

1 Introduction

Semi-Markov modeling, as a generalization of Markov modeling, is a an active area in research. See, e.g., [Alvarez, 2005]-[Voelkel and Cronwley, 1984].

In our previous work [Ouhbi and Limnios, 1999], we have considered one trajectory in the time interval $[0, T]$, and given the estimators and their asymptotic properties, as $T \to \infty$. In the present work, we consider K trajectories in the time interval $[0, T]$, generated by K independent semi-Markov processes having the same semi-Markov kernel Q and initial distribution α. We obtain asymptotic properties of the estimators when $K \to \infty$. In this case the time T is finite and fixed. This type of observation can be viewed as a generalization of the fixed (or type I) censoring of a single failure time. Our method, as in our previous works [Ouhbi and Limnios, 1999, Ouhbi and Limnios, 2003, Ouhbi and Limnios, 2001], consists in obtaining estimators of the semi-Markov kernel, by using a maximum likelihood estimator (MLE) of the hazard rate function of transitions between states, and then considering estimators of other quantities, as the semi-Markov transition function, Markov renewal function, and reliability functions as statistical functionals of the semi-Markov kernel via analytic explicit formula.

2 Estimation of the hazard rate function of transitions

We will consider in this paper a semi-Markov process with a finite state space, $E = \{1, 2, ..., s\}$ say, with irreducible embedded Markov chain and finite sojourn time in all states [Limnios and Oprişan, 2001].

In this section, we will derive and study the maximum likelihood estimator of the hazard rate functions of piecewise constant type estimator (PEXE).

Let us suppose that the semi-Markov kernel Q is absolutely continuous with respect to the Lebesgue measure on $\mathbf{R}_+$ and denote by q its Radon-Nikodym derivative, that is, for any $i, j \in E$,

$$\frac{Q_{ij}(dt)}{dt} =: q_{ij}(t). \tag{1}$$

So, we can write also $q_{ij}(t) = P(i,j) f_{ij}(t)$, where f_{ij} is the density function of the distribution F_{ij}.

For any i and j in E, let us define the hazard rate function of transition distributions between states, $\lambda_{ij}(t)$, $t \geq 0$, of a semi-Markov kernel by

$$\lambda_{ij}(t) = \begin{cases} \frac{q_{ij}(t)}{1-H_i(t)} & \text{if} \quad P(i,j) > 0 \quad \text{and } H_i(t) < 1, \\ 0 & \text{otherwise.} \end{cases} \tag{2}$$

Let us also define the cumulative hazard rate from state i to state j at time t by $\Lambda_{ij}(t) = \int_0^t \lambda_{ij}(u)du$ and the total cumulative hazard rate of state i at time t by $\Lambda_i(t) = \Sigma_{j\in E} \Lambda_{ij}(t)$. We have also

$$Q_{ij}(t) = \int_0^t \exp[-\Lambda_i(u)] \lambda_{ij}(u) du. \tag{3}$$

Let us consider now a family of K independent E-valued Markov renewal processes $(J_n^r, S_n^r, n \geq 0), 1 \leq r \leq K$, defined by the same semi-Markov kernel Q, and the initial distribution α, that is, for any r, $1 \leq r \leq K$,

$$Q_{ij}(t) := \mathbf{P}(J_{n+1}^r = j, S_{n+1}^r - S_n^r \leq t \mid J_n^r = i), \quad i, j \in E, t \in \mathbf{R}_+, n \in \mathbf{N},$$

$$\alpha(i) = \mathbf{P}(J_0^r = i), \quad i \in E.$$

For any r, let us denote by $N_i^r(t), N_{ij}^r(t), N^r(t), ...$ the corresponding quantities $N_i(t)$, $N_{ij}(t)$, $N(t), ...$, and define further

$$N_i(t,K) := \sum_{r=1}^K N_i^r(t), \quad N_{ij}(t,K) := \sum_{r=1}^K N_{ij}^r(t), \quad N(t) := \sum_{r=1}^K N^r(t). \tag{4}$$

If $t = T$ fixed, then we will note simply $N_i, N_{ij}, ...$

The maximum likelihood estimator of the hazard rate functions will be based upon the observation of the above K independent MRP $\{(J^r, S^r) = [(J_n^r, S_n^r)_{n\geq 0}], 1 \leq r \leq K\}$.

We assume hereafter that we observe each MRPs over the period of time $[0,T]$ for some finite and fixed T. A sample or history for the r-th MRP is given by

$$\mathcal{H}^r(K) = (J_0^r, J_1^r, ..., J_{N^r(T)}^r, X_1^r, X_2^r, ..., X_{N^r(T)}^r, U_T^r), \tag{5}$$

where $U_T^r = T - S_{N^r(T)}^r$ is the backward recurrence time.

The log-likelihood function associated to $(\mathcal{H}^r(T), 1 \leq r \leq K)$ is:

$$l(K) = \log L(K) = \sum_{r=1}^K \Big\{ \sum_{l=1}^{N^r(T)} [\log \lambda_{J_{l-1}^r, J_l^r}(X_l^r) - \Lambda_{J_{l-1}^r}(X_l^r)] - \Lambda_{J_{N^r(T)}^r}(U_T^r) \Big\}. \tag{6}$$

In the sequel of this paper, we will approximate the hazard rate function $\lambda_{ij}(t)$ by the piecewise constant function $\lambda_{ij}^*(t)$ defined by $\lambda_{ij}^*(t) = \lambda_{ij}(v_k) = \lambda_{ijk} \in \mathbf{R}_+$

for $t \in (v_{k-1}, v_k] = I_k$, $1 \leq k \leq M$, where $(v_k)_{0 \leq k \leq M}$ is a regular subdivision of $[0,T]$, that is, $v_k = k\Delta$, $0 \leq k \leq M$, $M = M(K)$, with step $\Delta := T/M$, such that, as $K \to \infty$, $\Delta \to 0$, and $K\Delta \to \infty$.

Hence,

$$\lambda_{ij}^*(t) = \sum_{k=1}^{M} \lambda_{ijk} \mathbf{1}_{(v_{k-1},v_k]}(t), \tag{7}$$

where $\mathbf{1}_{(v_{k-1},v_k]}(t)$ is equal to 1 if $t \in (v_{k-1}, v_k]$, and 0 otherwise. We get

$$l(K) = \sum_{i,j \in E} \sum_{k=1}^{M} (d_{ijk} \log \lambda_{ijk} - \lambda_{ijk} \nu_{ik}), \tag{8}$$

where $d_{ijk} = \sum_{r=1}^{K} \sum_{l=1}^{N^r(T)} \mathbf{1}_{\{J_{l-1}^r = i, J_l^r = j, X_l^r \in I_k\}}$ is the number of transitions from state i to state j for which the observed sojourn time in state i belongs to I_k, and ν_{ik} is the trace of the sojourn time in state i on the interval time I_k, given for $N(T) \geq 1$. The r.v. ν_{ik} can be represented by the sum of two r.v. as follows

$$\nu_{ik} := \nu_{ik}^1 + \nu_{ik}^2,$$

where ν_{ik}^1 is the trace of the sojourn time on the interval I_k, of the sojourn times in state i, and ν_{ik}^2 is the trace of the cumulated censored time T greater than v_k, in state i.

So, the maximum likelihood estimator $\widehat{\lambda}_{ijk}$ of λ_{ijk} is given by:

$$\widehat{\lambda}_{ijk} = \begin{cases} d_{ijk}/\nu_{ik} & \text{if} \quad \nu_{ik} > 0 \\ 0 & \text{otherwise.} \end{cases}$$

Thus, the estimator $\widehat{\lambda}_{ij}(t, K)$ of $\lambda_{ij}(t)$ is then given by

$$\widehat{\lambda}_{ij}(t, K) = \sum_{k=1}^{M} \widehat{\lambda}_{ijk} \mathbf{1}_{(v_{k-1},v_k]}(t). \tag{9}$$

Let us also define

$$\widehat{\Lambda}_i(t, K) = \sum_{j \in E} \int_0^t \widehat{\lambda}_{ij}(u, K) du, \quad \widehat{\Lambda}_{ik} = \widehat{\Lambda}_i(v_k, K) = \Delta \sum_{j \in E} \sum_{l=1}^{k} \widehat{\lambda}_{ijl}, \tag{10}$$

and

$$\nu_i^l(t) = \sum_{k=1}^{M} \nu_{ik}^l \mathbf{1}_{(v_{k-1},v_k]}(t), \quad l = 1, 2.$$

3 Maximum likelihood and empirical estimators of the semi-Markov kernel

Let us define estimators of the semi-Markov kernel by putting estimators (9) to (10), as follows

$$\widehat{Q}_{ij}(t, K) := \Delta \sum_{\{k: 0 \leq v_k \leq t\}} e^{-\widehat{\Lambda}_{ik}} \widehat{\lambda}_{ijk}.$$

Consider now the empirical kernel function defined by

$$\widetilde{Q}_{ij}(t,K) := \frac{1}{N_i}\sum_{r=1}^{K}\sum_{l=1}^{N^r}\mathbf{1}_{\{J_{l-1}^r=i,J_l^r=j,X_l^r\leq t\}}, \tag{11}$$

and the empirical density kernel given by:

$$\widetilde{q}_{ij}(t,K) = \frac{\widetilde{Q}_{ij}(v_k,K)-\widetilde{Q}_{ij}(v_{k-1},K)}{\varDelta}, \quad \text{if } t\in I_k.$$

Define also the estimator $\widetilde{H}_i(t,K)$ by

$$\widetilde{H}_i(t,K) := \sum_{j\in E}\widetilde{Q}_{ij}(t,K). \tag{12}$$

We define a function $G_i(\cdot,K)$, for $t\in I_k$, $1\leq k\leq M$, by

$$G_i(t,K) := \sum_{r=1}^{K}\Big\{\sum_{l=1}^{N^r}\frac{X_l^r-v_{k-1}}{N_i\varDelta}\mathbf{1}_{\{J_{l-1}^r=i,X_l^r\in I_k,X_l^r\geq t\}}+\frac{U_T^r-v_{k-1}}{N_i\varDelta}\mathbf{1}_{\{J_{N^r(T)}^r=i,U_T^r\in I_k\}}\Big\}.$$

Now, let us write estimator (9), as follows

$$\widehat{\lambda}_{ij}(t,K) = \frac{\widetilde{q}_{ij}(v_k,K)}{1-\{\widetilde{H}_i(v_k,K)-G_i(v_k,K)\}+h_i^r(t,K)}, \text{ if } t\in I_k,$$

where

$$h_i^r(t,K) := \frac{\nu_{ik}^2}{N_i\varDelta} = \frac{1}{N_i}\sum_{r=1}^{K}\mathbf{1}_{\{J_{N^r(T)}^r=i,U_T^r>v_k\}}.$$

In order to obtain a consistent estimator, we will neglect the term $h_i^r(t,K)$ from the denominator of estimator $\widehat{\lambda}_{ij}(t,K)$, and obtain a new modified estimator denoted by $\widehat{\lambda}_{ij}^0(t,K)$. That is,

$$\widehat{\lambda}_{ij}^0(t,K) = \frac{\widetilde{q}_{ij}(v_k,K)}{1-\{\widetilde{H}_i(v_k,K)-G_i(v_k,K)\}}, \quad \text{if } t\in I_k. \tag{13}$$

Denote the corresponding cumulative hazard rates estimator by $\widehat{\varLambda}_i^0(t)$, and $\widehat{\varLambda}_{ij}^0(t)$.

Lemma 1 *The estimator $\widehat{\lambda}_{ij}^0(t,K)$, is a consistent estimator of $\lambda_{ij}(t)$, as $K\to\infty$.*

Since $h_i^r(t,K)$ converges to a positive quantity, it is clear the estimator $\widehat{\lambda}_{ij}(t,K)$ is not consistent.

In the sequel of this paper, we will consider only the estimator $\widehat{\lambda}_{ij}^0(t,K)$. So, the MLE $\widehat{Q}_{ij}(t,K)$ in (11) is obtained by using this estimator. In the remaining of this section we will study the asymptotic properties of the semi-Markov kernel estimator given by (11).

Theorem 1 *The empirical estimator of the semi-Markov kernel is uniformly strongly consistent, in the sense that, as $K\to\infty$,*

$$\max_{i,j}\sup_{t\in[0,T]}\left|\widetilde{Q}_{ij}(t,K)-Q_{ij}(t)\right|\xrightarrow{a.s.}0.$$

We will prove now that the semi-Markov kernel estimator, obtained from modified PEXE of the hazard rate function $\widehat{\lambda}^0_{ij}(t,K)$, is asymptotically uniformly equivalent to the empirical estimator $\widetilde{Q}_{ij}(t,K)$.

Lemma 2 *Let i and j be any two fixed states. Then we have, for any $t \in [0,T]$,*

$$\widehat{Q}_{ij}(t,K) - \widetilde{Q}_{ij}(t,K) = O(K^{-1}), \quad \text{as } K \to \infty.$$

¿From the previous lemma, we conclude that the estimator of the semi-Markov kernel is asymptotically uniformly a.s. equivalent to the empirical estimator of the semi-Markov kernel for which we will prove the uniform strong consistency and derive a central limit theorem.

Corollary 1 *The estimator of the semi-Markov kernel $\widehat{Q}_{ij}(t,K)$ is uniformly strongly consistent, that is, when K tends to infinity,*

$$\max_{i,j} \sup_{t\in[0,T]} \left| \widehat{Q}_{ij}(t,K) - Q_{ij}(t) \right| \xrightarrow{a.s.} 0.$$

Theorem 2 *For any $i,j \in E$ and $t \in [0,T]$ fixed, $K^{1/2}[\widehat{Q}_{ij}(t,K) - Q_{ij}(t)]$ converges in distribution, as $K \to \infty$, to a zero mean normal random variable with variance $Q_{ij}(t)(1-Q_{ij}(t))[(\alpha\psi)(T)\mathbf{1}]$.*

4 The estimator of the reliability function and its asymptotic properties

After having outlined the problem of estimating the semi-Markov transition matrix, it is appropriate to give some concrete applications of these processes as models of evolution of the reliability function of some system.

Let the state space, E, be partitioned into two sets, $U = \{1,...,r\}$ the patient is in good health and $D = \{r+1,...,s\}$ the patient is ill due to some causes or the component is failed and under repair. Reliability models whose state space is partitioned in the above manner will be considered here. As indicated above, it is of interest to estimate the distribution function of the waiting time to hit down states (failure).

We focus on the estimation of the reliability function for a semi-Markov process which describes the stochastic evolution of system. The general definition of the reliability function in the case of semi-Markov processes is

$$R(t) = \mathbf{P}(Z_u \in U, \quad \forall\, u \leq t).$$

The reliability function $R(t)$ is given by:

$$R(t) = \sum_{i\in U} \alpha(i) R_i(t), \tag{14}$$

where $R_i(t)$ is the conditional reliability function, that the hitting time to D, starting from a state $i \in U$, is greater than the time t. It is easy to show, by a renewal argument, that $R_i(t)$ satisfies the following Markov renewal equation

$$R_i(t) - \sum_{i\in U} \int_0^t R_j(t-u) Q_{ij}(du) = 1 - H_i(t), \quad i \in U.$$

The solution of this MRE, see Section 2, together with (14), in matrix form, gives

$$R(t) = \alpha_0 (I - Q_0(t))^{(-1)} * (I - H_0(t))\mathbf{1}, \tag{15}$$

where $\mathbf{1} = (1, ..., 1)^\top$ is an r-dimensional column vector. Index 0 means restriction for matrices on $U \times U$, and for vectors on U.

We will give an estimator of the reliability function of semi-Markov processes and prove its uniform strong consistency and weak convergence properties as $K \to \infty$.

Let $\widehat{Q}$ be the modified MLE of PEXE type of the transition probability of the semi-Markov kernel Q. Then we propose the following estimator for the reliability function

$$\widehat{R}(t, K) = \widehat{\alpha}_0 (I - \widehat{Q}_0(t, K))^{(-1)} * (I - \widehat{H}_0(t, K))\mathbf{1}, \tag{16}$$

and we will prove now its uniform strong consistency and central limit theorems.

Theorem 3 *The estimator of the the reliability function of the semi-Markov process is uniformly strongly consistent in the sense that,*

$$\sup_{t \in [0,T]} \left| \widehat{R}(t, K) - R(t) \right| \xrightarrow{a.s.} 0, \quad K \to \infty.$$

Set

$$B_{ij}(t) := \sum_{n \in U} \sum_{k \in U} \alpha(n) B_{nijk} * (1 - H_k)(t).$$

Theorem 4 *For any fixed $t \in [0, T]$, the r.v. $K^{1/2}[\widehat{R}(t, K) - R(t)]$ converges in distribution to a zero mean normal random variable with variance*

$$\sigma_S^2(t) := \sum_{i \in U} \sum_{j \in U} \mu_{ii} \{[B_{ij} - (\alpha\psi)_i]^2 * Q_{ij}(t) - [(B_{ij} - (\alpha\psi)_i) * Q_{ij}(t)]^2\}.$$

5 Numerical Application

In this section we present a numerical example for a three state semi-Markov process for which we will consider $K = 50$ censored trajectories. The time interval is $[0, T]$, with $T = 1000$.

The conditional transition functions $F_{ij}(t)$ are the following $F_{12}(t)$, and $F_{31}(t)$ are exponential with parameters respectively 0.1 and 0.2, and $F_{21}(t)$, $F_{23}(t)$ are Weibull with parameters respectively $(0.3, 2)$, and $(0.1, 2)$ (scale and shape parameter). The other functions are identically 0.

The transition probabilities $P(2, 1)$ and $P(2, 3)$ are:

$$P(2, 1) = 1 - P(2, 3) = \int_0^\infty [1 - F_{23}(t)] dF_{23}(t).$$

The results obtained here are illustrated in figure 1. These results concern the reliability function.

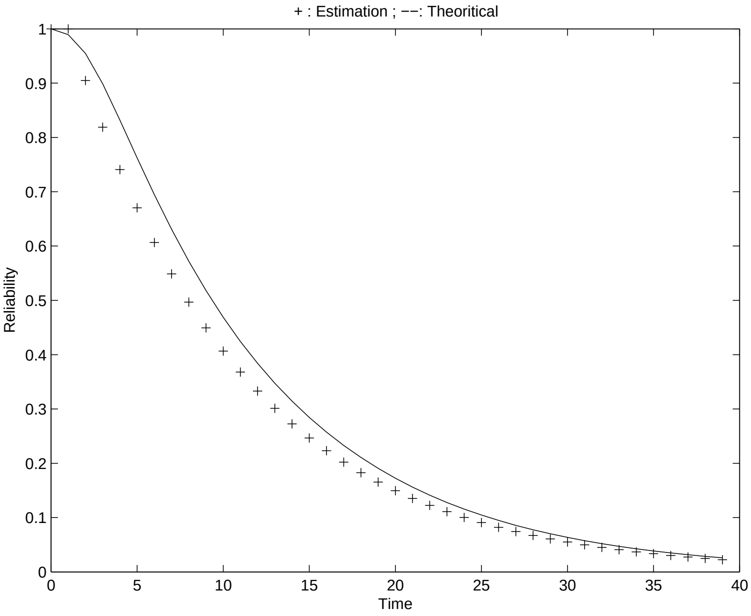


Fig. 1. Reliability estimation

References

[Alvarez, 2005]E.E.E. Alvarez (2005). Smothed nonparametric estimation in window censored semi-Markov processes, *J. Statist. Plann. Infer.*, 131, 209–229.

[Andersen *et al.*, 1993]P.K. Andersen, O. Borgan, R.D. Gill, N. Keiding (1993). *Statistical Models Based on Counting Processes*, Springer, N.Y.

[Dabrowska *et al.*, 1994]D. Dabrowska, G. Horowitz, M. Sun (1994). Cox regression in a Markov renewal model: an application to the analysis of bone marrow transplant data. *J. Amer. Statist. Ass.*, 89, 867-877.

[Gill, 1980]R.D. Gill (1980). Nonparametric estimation based on censored observations of Markov renewal process. *Z. Wahrsch. verw. Gebiete.* 53, 97–116.

[Greenwood and Wefelmeyer, 1996]P. E. Greenwood, W. Wefelmeyer (1996). Empirical estimators for semi-Markov processes. *Math. Methods Statist.* 5(3), 299–315.

[Lagakos *et al.*, 1978]S.W. Lagakos, C.J. Sommer, M. Zelen (1978). Semi-Markov models for partially censored data, *Biometrika*, 65(2), 311–317.

[Limnios, 2004]N. Limnios (2004). A functional central limit theorem for the empirical estimator of a semi-Markov kernel, *J. Nonparametric Statist.*, 16(1-2), pp 13-18.

[Limnios and Oprişan, 2001]N. Limnios, G. Oprişan (2001). *Semi-Markov Processes and Reliability*, Birkhäuser, Boston.

[Limnios and Ouhbi, 2003]N. Limnios, B. Ouhbi, "Empirical estimators of reliability and related functions for semi-Markov systems", In *Mathematical and*

Statistical Methods in Reliability, B. Lindqvist, K. Doksum (Eds.), World Scientific, 2003.

[Ouhbi and Limnios, 1996]B. Ouhbi, N. Limnios (1996). Non-parametric estimation for semi-Markov kernels with application to reliability analysis, *Appl. Stoch. Models Data Anal.*, 12, 209–220.

[Ouhbi and Limnios, 1999]B. Ouhbi, N. Limnios (1999). Non-parametric estimation for semi-Markov processes based on their hazard rate. *Statist. Infer. Stoch. Processes*, 2(2), 151–173.

[Ouhbi and Limnios, 2001]B. Ouhbi, N. Limnios (2001). The rate of occurrence of failures for semi-Markov processes and estimation. *Statist. Probab. Lett.*, 59(3), 245–255.

[Ouhbi and Limnios, 2003]B. Ouhbi, N. Limnios (2003). Nonparametric reliability estimation of semi-Markov processes. *J. Statist. Plann. Infer.*, 109(1/2), 155–165.

[Phelan, 1990]M.J.Phelan (1990). Estimating the transition probabilities from censored Markov renewal processes. *Statist. Probab. Letter.*, 10, pp 43–47.

[Pyke, 1961]R. Pyke (1961). Markov renewal processes: definitions and preliminary properties, *Ann. Math. Statist.*, 32, 1231–1241.

[Ruiz-Castro and Pérez-Ocon, 2004]J. E. Ruiz-Castro, R. Pérez-Ocon (2004). A semi-Markov model in biomedical studies, *Commun. Stat. - Theor. Methods*, 33(3).

[Voelkel and Cronwley, 1984]J.G. Voelkel, J. Cronwley (1984). Nonparametric inference for a class of semi-Markov processes with censored observations, *Ann. Statist.*, 12(1), pp 142–160.

Visual tracking and auxiliary discrete processes

Patrick Pérez[1] and Jaco Vermaak[2]

[1] IRISA/INRIA, Campus Universitaire de Beaulieu
35042 Rennes Cedex, France
(e-mail: perez@irisa.fr)
[2] Cambridge Univ. Eng. Dpt., Trumpington St.
Cambridge, CB2 1PZ, United Kingdom
(e-mail: jv211@eng.cam.ac.uk)

Abstract. A number of Bayesian tracking models involve auxiliary discrete variables beside the main hidden state of interest. These discrete variables usually follow a Markovian process and interact with the hidden state either via its evolution model or via the observation process, or both. Examples of such auxiliary variables include depth ordering for occlusion handling, switches between different state dynamics, exemplar indices, etc. We consider here a general model that encompasses all these situations, and show how Bayesian filtering can be rigorously conducted in this general setup. The resulting approach facilitates easy re-use of existing tracking algorithms designed in the absence of the auxiliary process. In particular we show how particle filters can be obtained based on sampling only in the original state space instead of sampling in the augmented space, as it is usually done. We finally demonstrate how this framework facilitates solutions to the critical problem of appearance and disappearance of targets, either upon scene entering and exiting, or due to temporary occlusions. This is illustrated in the context of color-based tracking with particle filters.
Keywords: Optimal Bayesian filter, Auxiliary discrete process, Particle filter, Visual tracking, Occlusion, Disappearance, Object detection.

1 Introduction and motivation

Visual tracking involves the detection and recursive localization of objects within video frames. In a number of visual trackers, the state of interest, e.g., size and location of the object, is associated with auxiliary discrete variables. Such variables show up for instance within the state evolution model, e.g., when different types of dynamics can occur (e.g., [North *et al.*, 2000]). More often, such auxiliary variables are introduced in the observation model. It is the case for appearance models based on a set of key views (e.g., [Toyama and Blake, 2001],[Wu *et al.*, 2003]) or silhouettes (e.g., [Gavrila, 2000] [Toyama and Blake, 2001]). Auxiliary variables are also used to handle partial or total occlusions (e.g., [Nguyen *et al.*, 2001]) or mutual occlusions when jointly tracking multiple objects (e.g., [MacCormick and Blake, 1999] [Wu *et al.*, 2003]). Finally, auxiliary variables can be used to assess the presence of tracked objects in the scene (e.g., [Vermaak *et al.*, 2002] [Isard and MacCormick, 2001]). When a Bayesian tracking approach is used with such augmented models, either specific filters are derived based on the detailed form of the model at hand or the optimal filter of the joint model is simply used. In the latter case, a practical implementation might be unnecessarily costly due to the increased dimension of the joint space. Sequential Monte Carlo approximations (SMC) in the joint space are for instance used in [Isard and MacCormick, 2001] [Toyama and Blake, 2001] [Vermaak *et al.*, 2002] [Wu *et al.*, 2003].

The first contribution of this paper is to propose a general and unified framework to easily derive the optimal Bayesian filter for the augmented model based on the one for a model with no (or frozen) auxiliary variables. In practice, this allows the re-use of existing tracking architectures, with a reasonable computational overhead in case the discrete auxiliary variable only takes a small number of values. This approach allows us in particular to introduce a generic SMC architecture that relies on sampling in the main state space only. This is exposed in Section 2.

The problem of appearing and disappearing objects, whether it is upon entering and exiting the scene, or upon getting occluded by another object, is critical in visual tracking. As we mentioned above, the different forms of this problem have already been addressed in the past based on auxiliary hidden processes. The second contribution of this paper is to re-visit these problems using our generic framework. The resulting filters are implemented using the generic SMC architecture proposed in Section 2. To handle occlusions, we introduce in Section 3 a binary visibility process that intervenes in the observation model. In this case, our generic approach allows us to derive a two-fold mixture filter that deal with temporary occlusions. In a similar fashion, we address the problem of “birth” and “death” of objects, which is crucial for multiple-object tracking, by introducing a binary existence process. This process impacts both the state evolution and the data model. The application of our approach leads in this case to a simple filter whose SMC approximation does not need to draw samples for the existence variable.

2 Tracking with an auxiliary process

2.1 Modeling assumptions

For visual tracking, we are interested in recursively estimating the object state $\mathbf{x}_t \in \mathbb{R}^{n_x}$, which specifies the position of the object in the image plane and, possibly, other parameters such as its size and orientation, based on a sequence of observations $\mathbf{y}^t = (\mathbf{y}_1 \cdots \mathbf{y}_t)$. We assume in addition that a discrete auxiliary variable a_t also has to be recursively inferred. This variable takes its values in a set of cardinality M that we will denote by $\{0 \cdots M-1\}$ for convenience.

The complete set of unknowns at time t is thus $\{\mathbf{x}_t, a_t\}$, for which we assume the following Markovian prior

$$p(\mathbf{x}_t, a_t|\mathbf{x}_{t-1}, a_{t-1}) = p(\mathbf{x}_t|\mathbf{x}_{t-1}, a_t, a_{t-1})p(a_t|a_{t-1}). \tag{1}$$

In other words, the state follows a Markov chain with its kernel parameterized by the current and previous values of the auxiliary variable, and the auxiliary process is a discrete Markov chain. Let $A = (\alpha_{ji})$ be its $M \times M$ transition matrix, with $\alpha_{ji} \doteq p(a_t = i|a_{t-1} = j)$. For brevity, we will also use the notation

$$p_{ji}(\mathbf{x}_t|\mathbf{x}_{t-1}) \doteq p(\mathbf{x}_t|\mathbf{x}_{t-1}, a_t = i, a_{t-1} = j). \tag{2}$$

As for the observation model, we assume in the normal way that the image data at successive instances are independent conditional on the hidden variables, i.e., $p(\mathbf{y}_t|\mathbf{x}_t, a_t, \mathbf{y}^{t-1}) = p(\mathbf{y}_t|\mathbf{x}_t, a_t)$. For notational convenience we will denote

$$p_i(\mathbf{y}_t|\mathbf{x}_t) \doteq p(\mathbf{y}_t|\mathbf{x}_t, a_t = i). \tag{3}$$

2.2 Bayesian filter

For tracking, we are interested in recursively estimating the joint filtering distribution

$$p(\mathbf{x}_t, a_t|\mathbf{y}^t) = p(\mathbf{x}_t|a_t, \mathbf{y}^t)p(a_t|\mathbf{y}^t), \tag{4}$$

from which the marginal filtering distribution can be deduced as

$$p(\mathbf{x}_t|\mathbf{y}^t) = \sum_i p(\mathbf{x}_t, a_t = i|\mathbf{y}^t) = \sum_i p_i(\mathbf{x}_t|\mathbf{y}^t)\xi_{i,t}, \tag{5}$$

where we used the notation

$$p_i(\mathbf{x}_t|\mathbf{y}^t) \doteq p(\mathbf{x}_t|a_t = i, \mathbf{y}^t) \text{ and } \xi_{i,t} \doteq p(a_t = i|\mathbf{y}^t). \tag{6}$$

Similar to our previous notation, we will now use the distribution subscript i to indicate conditioning with respect to the current auxiliary variable set to i, and the distribution subscript ji for conditioning on i and j being the current and previous values of the auxiliary variable.

We will first show how to compute the M conditional state posteriors $p_i(\mathbf{x}_t|\mathbf{y}^t)$. First note that

$$p_i(\mathbf{x}_t|\mathbf{y}^t) = \frac{p_i(\mathbf{x}_t, \mathbf{y}_t|\mathbf{y}^{t-1})}{p_i(\mathbf{y}_t|\mathbf{y}^{t-1})}. \tag{7}$$

The numerator can be expressed as

$$p_i(\mathbf{x}_t, \mathbf{y}_t|\mathbf{y}^{t-1}) = \sum_j p_{ji}(\mathbf{x}_t, \mathbf{y}_t|\mathbf{y}^{t-1})p(a_{t-1} = j|a_t = i, \mathbf{y}^{t-1}), \tag{8}$$

with

$$\begin{aligned} p_{ji}(\mathbf{x}_t, \mathbf{y}_t | \mathbf{y}^{t-1}) &= p_i(\mathbf{y}_t|\mathbf{x}_t) p_{ji}(\mathbf{x}_t|\mathbf{y}^{t-1}) \\ &= p_i(\mathbf{y}_t|\mathbf{x}_t) \int p_{ji}(\mathbf{x}_t|\mathbf{x}_{t-1}) p_j(\mathbf{x}_{t-1}|\mathbf{y}^{t-1}) d\mathbf{x}_{t-1}, \quad (9) \\ p(a_{t-1} = j | a_t = i, \mathbf{y}^{t-1}) &\doteq \tilde{\alpha}_{ji,t} \\ &\propto p(a_t = i | a_{t-1} = j, \mathbf{y}^{t-1}) p(a_{t-1} = j | \mathbf{y}^{t-1}). \quad (10) \end{aligned}$$

Based on the conditional independence structure of the model, one can show that the first term on the right hand side is independent of $\mathbf{y}^{t-1}$. We thus obtain, after normalization,

$$\tilde{\alpha}_{ji,t} = \frac{\alpha_{ji}\xi_{j,t-1}}{\sum_k \alpha_{ki}\xi_{k,t-1}}. \quad (11)$$

The predictive likelihood in the denominator of (7) is

$$p_i(\mathbf{y}_t|\mathbf{y}^{t-1}) = \sum_j \tilde{\alpha}_{ji,t} \int p_{ji}(\mathbf{x}_t, \mathbf{y}_t|\mathbf{y}^{t-1}) d\mathbf{x}_t. \quad (12)$$

The filtering distribution in (5) is then a mixture of the M conditional filtering distributions, i.e.,

$$p_i(\mathbf{x}_t|\mathbf{y}^t) = \frac{\sum_j \tilde{\alpha}_{ji,t} p_{ji}(\mathbf{x}_t, \mathbf{y}_t|\mathbf{y}^{t-1})}{p_i(\mathbf{y}_t|\mathbf{y}^{t-1})}, \quad (13)$$

each of which is obtained by combining M optimal Bayesian filters to compute (9) and (12).

We still need the marginal posterior of the auxiliary variable, $p(a_t|\mathbf{y}^t)$, to compute the weights $\xi_{i,t}$ in the mixture of (5). We have

$$\xi_{i,t} \propto p_i(\mathbf{y}_t|\mathbf{y}^{t-1}) \sum_j p(a_t = i | a_{t-1} = j, \mathbf{y}^{t-1}) \xi_{j,t-1}. \quad (14)$$

Since the first factor in the sum is independent of $\mathbf{y}^{t-1}$, we finally obtain, after normalization

$$\xi_{i,t} = \frac{p_i(\mathbf{y}_t|\mathbf{y}^{t-1}) \sum_j \alpha_{ji}\xi_{j,t-1}}{\sum_k p_k(\mathbf{y}_t|\mathbf{y}^{t-1}) \sum_j \alpha_{jk}\xi_{j,t-1}}. \quad (15)$$

Let us summarize the operations at time t for the generic algorithm:

- **Input**: $p_i(\mathbf{x}_{t-1}|\mathbf{y}^{t-1})$ and $(\xi_{i,t-1})$ for $i = 0 \cdots M-1$.

1. Compute $\tilde{\alpha}_{ji,t}$ as in (11), for $i = 0 \cdots M-1$.
2. Compute distributions $p_{ji}(\mathbf{x}_t, \mathbf{y}_t|\mathbf{y}^{t-1})$ as in (9), for $i, j = 0 \cdots M-1$.
3. Compute distributions $p_i(\mathbf{y}_t|\mathbf{y}^{t-1})$ as in (12), for $i = 0 \cdots M-1$.
4. Compute filtering distributions $p_i(\mathbf{x}_t|\mathbf{y}^t) =$ as in (13), for $i = 0 \cdots M-1$.

5. Compute posterior distribution $(\xi_{i,t})_{i=0\cdots M-1}$ of auxiliary variable as in (15).

- **Output**: distributions $p_i(\mathbf{x}_t|\mathbf{y}^t)$ and weights $\xi_{i,t}$.

At each time step, M^2 "elementary" filtering operations are required (step 2), one per possible occurrence of the pairing (a_t, a_{t-1}). In practice, not all M^2 values may be admissible, in which case the number of elementary filtering operations at each time step is reduced accordingly. As we will see, specificities of the model under consideration might also permit further computational savings.

The framework above is entirely general, both in terms of model ingredients (evolution and observation processes) and in terms of implementation. Regarding the latter, all existing techniques, whether exact or approximate, can be accommodated. If, for example, the filtering distributions $p_i(\mathbf{x}_t|\mathbf{y}^t)$ are to be represented by Gaussian mixtures, the mixtures components can be obtained by the Kalman filter for linear Gaussian models, and by the extended or unscented Kalman filters for non-linear and/or non-Gaussian models. For models of the latter kind it may sometimes be beneficial to adopt a particle representation, and use sequential importance sampling techniques to update the filtering distribution. This is especially true for the highly non-linear and multi-modal models used in visual tracking, hence the success of SMC techniques in the computer vision community. It is this type of implementation that we now consider.

2.3 SMC implementation

For a general SMC implementation, we will consider proposal distributions of the form $q_{ji}(\mathbf{x}_t|\mathbf{x}_{t-1}, \mathbf{y}_t) \doteq q(\mathbf{x}_t|\mathbf{x}_{t-1}, a_t = i, a_{t-1} = j, \mathbf{y}_t)$. Based on these proposals, different SMC architectures can be designed to approximate the generic algorithm of the previous section. We propose here an architecture that is based on systematic resampling. Assuming that each conditional posterior distribution $p_i(\mathbf{x}_{t-1}|\mathbf{y}^{t-1})$ at time $t-1$ is approximated by a set $(\mathbf{s}_{i,t-1}^{(n)})_{n=1\cdots N}$ of N equally weighted particles, we simply replace steps 2, 3 and 4 in the generic algorithm by:

2. For $j = 0 \cdots M-1$, for $i = 0 \cdots M-1$
   2a. Sample N particles $\tilde{\mathbf{s}}_{ji,t}^{(n)} \mathrm{sim} q_{ji}(\mathbf{x}_t|\mathbf{s}_{j,t-1}^{(n)}, \mathbf{y}_t)$.
   2b. Compute the *normalized* predictive weights

$$\pi_{ji,t}^{(n)} \propto \frac{p_{ji}(\tilde{\mathbf{s}}_{ji,t}^{(n)}|\mathbf{s}_{j,t-1}^{(n)})}{q_{ji}(\tilde{\mathbf{s}}_{ji,t}^{(n)}|\mathbf{s}_{j,t-1}^{(n)}), \mathbf{y}_t} \text{ with } \sum_n \pi_{ji,t}^{(n)} = 1. \tag{16}$$

3. Approximate the M predictive data likelihoods by

$$p_i(\mathbf{y}_t|\mathbf{y}^{t-1}) \approx \sum_j \sum_n w_{ji,t}^{(n)}, \tag{17}$$

where, for $i, j = 0 \cdots M-1$,

$$w_{ji,t}^{(n)} \doteq \tilde{\alpha}_{ji,t} p_i(\mathbf{y}_t | \tilde{\mathbf{s}}_{ji,t}^{(n)}) \pi_{ji,t}^{(n)}. \tag{18}$$

4. For $i = 0 \cdots M-1$, draw N particles $\mathbf{s}_{i,t}^{(n)}$ with replacement from the weighted set $(\tilde{\mathbf{s}}_{ji,t}^{(n)}, p_i(\mathbf{y}_t|\mathbf{y}^{t-1})^{-1} w_{ji,t}^{(n)})_{j,n}$ of $M \times N$ particles.

Steps 1 and 5 remain unchanged. At each instant t, posterior expectations can be approximated using the final particle sets. In particular,

$$\mathbb{E}[\mathbf{x}_t | a_t = i, \mathbf{y}^t] \approx \hat{\mathbf{x}}_{i,t} \doteq \frac{1}{N} \sum_n \mathbf{s}_{i,t}^{(n)}, \ \mathbb{E}[\mathbf{x}_t|\mathbf{y}^t] \approx \hat{\mathbf{x}}_t \doteq \sum_i \xi_{i,t} \hat{\mathbf{x}}_{i,t}. \tag{19}$$

If the proposal distribution does not depend on $a_t = i$, then step 2a can be performed M times instead of M^2 times, providing particles sets $(\tilde{\mathbf{s}}_{j,t}^{(n)})_n$ to be used in place of $(\tilde{\mathbf{s}}_{ji,t}^{(n)})_n$ in the remainder of the algorithm.

3 Appearance and disappearance

Most tracking algorithms assume the number of objects of interest to be constant in the sequence. However, in most cases objects of interest enter and exit the scene at arbitrary times. In addition, they can also disappear temporarily behind other occluding objects. In the latter case of occlusion, tracking should be continued blindly in the hope of locking back onto the objects when they re-appear. An object entering or exiting the scene should in contrast result in initiating or terminating tracking, respectively. In any case, these appearance and disappearance events, whether they are temporary or definitive, are themselves uncertain events. The associated concepts of "existence" and "visibility" should thus be treated jointly with the other unknowns within a probabilistic framework that can account for all the expected ambiguities. Exploiting the generic approach presented in the previous section, we propose to achieve this using two auxiliary binary processes. Although these two processes can be used jointly, we introduce them separately for the sake of clarity.

3.1 Visibility process

Explicit introduction of an occlusion process within the Bayesian tracking framework was proposed in [MacCormick and Blake, 1999] and [Wu *et al.*, 2003]. Both works, however, rely on specific modeling assumption (contour-based tracking in the former, luminance exemplars in the latter), and specific implementations (particle filter with partitioned importance sampling in the former vanilla bootstrap particle filter in the latter). In contrast, our approach relies on generic modeling assumptions and is independent of a specific implementation strategy, so that existing tracking architectures can be re-used. The occlusion modeling we propose can thus be used in conjunction with any Bayesian visual tracking technique, based for instance on the Kalman filter or one of its variants. In addition, using it within the SMC architecture of Section 2 allows restriction of the sampling to the object state space only.

Considering here only the case of complete occlusion, we introduce a binary visibility variable v_t that indicates whether the object is visible ($v_t = 1$) or not ($v_t = 0$) in the image at time t. The Markov chain prior on this binary variable is completely defined by the occlusion and desocclusion probabilities, α_{10} and α_{01}. The state evolution model is independent of the visibility variable, i.e.,

$$p_{ji}(\mathbf{x}_t|\mathbf{x}_{t-1}) = p(\mathbf{x}_t|\mathbf{x}_{t-1}). \tag{20}$$

Two data models,

$$p(\mathbf{y}_t|\mathbf{x}_t, v_t = 0) = p_0(\mathbf{y}_t) \text{ and } p(\mathbf{y}_t|\mathbf{x}_t, v_t = 1) = p_1(\mathbf{y}_t|\mathbf{x}_t), \tag{21}$$

will have to be specified, depending on whether the object of interest is visible in the image or not. In the former case, the likelihood is independent of the state value. Since our experiments are conducted in the context of color-based tracking we consider a simple observation model related to the more complex ones proposed in [Isard and MacCormick, 2001] and [Vermaak *et al.*, 2002]. Pixel-wise location independent background and foreground models, g_0 and g_1, respectively, are specified over the selected color space. Assuming conditional independence of color measures over a sub-grid S of pixels, we obtain

$$p_0(\mathbf{y}_t) = \prod_{s\in S} g_0(\mathbf{y}_{s,t}) \text{ and } p_1(\mathbf{y}_t|\mathbf{x}_t) = \prod_{s\in R(\mathbf{x}_t)} g_1(\mathbf{y}_{s,t}) \prod_{s\in \bar{R}(\mathbf{x}_t)} g_0(\mathbf{y}_{s,t}), \tag{22}$$

where $R(\mathbf{x}_t)$ is the image region associated with an object parameterized by the state $\mathbf{x}_t$, and $\mathbf{y}_{s,t}$ is the color at pixel s in frame t.

For this dynamic model, the SMC architecture of Section 2 can be simplified. Indeed, the independence of the state evolution with respect to the auxiliary variables allows step 2a to be performed only M times, and suggests the use of a unique proposal. A simple and classical choice is to take the state dynamics (20) as the proposal [Isard and Blake, 1996]. We will adopt this approach here, while bearing in mind that any data-based proposal, including the optimal one [Doucet *et al.*, 2000] in the rare cases that it is accessible, can be used in our generic framework.

Fig. 1 shows results obtained on a sequence where a walking person is successfully tracked despite a succession of severe and total occlusions caused by trees in the foreground. The tracking is initialized manually on the red top of the person. The initialization also provides the reference foreground model g_1, defined as a $5 \times 5 \times 5$ joint histogram in the RGB color space. The histogram for the reference background model g_0 is also obtained in the first frame based on the image complement of the initial selection. The unknown state $\mathbf{x}_t$ comprises the position in the image plane ($n_x = 2$) and its dynamics (20) is taken to be a random walk with independent Gaussian noise with variance 10^2 on each component. The parameters of the Markov chain on the visibility process are $\alpha_{01} = 0.8$ and $\alpha_{10} = 0.1$, and its initial distribution is given by $p(v_0 = 1) = 0.8$. We use $N = 200$ particles for the SMC implementation. The main quantities of interest are the marginal filtering distributions (5), which inform on the localization of the object of interest regardless of whether it is visible or not. We display the MC approximations of the state expectations $\hat{\mathbf{x}}_t$ relative to these distributions in Fig. 1. The algorithm also recursively estimates the marginal visibility posterior $p(v_t = 1|\mathbf{y}^t)$. The time evolution of this quantity for the pedestrian sequence is plotted in Fig. 2. It correctly drops to zero for each complete occlusion of the tracked person.

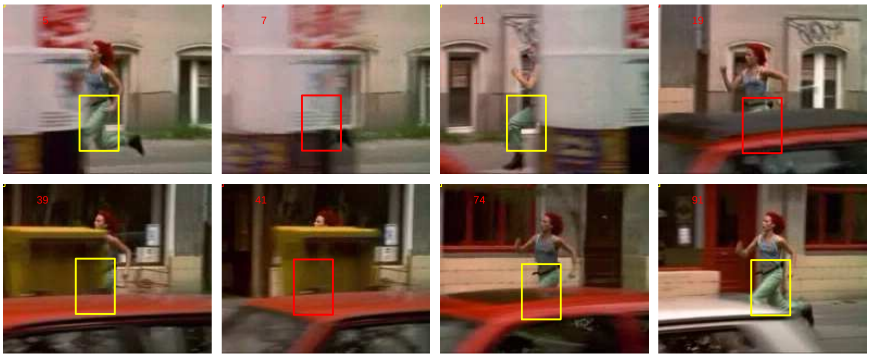


Fig. 1. Tracking under occlusions. The color-based tracker is initialized on the trousers of Lola (from movie "Run, Lola, run") who runs in the street. The rapid succession of partial, large or complete occlusions caused by cars, poles and mailbox is successfully handled thanks to the explicit modeling of visibility changes. In each of the displayed frames, the box corresponds to $\hat{\mathbf{x}}_t$ and its color is changed from yellow to red when $\xi_{1,t}$ drops below 0.5.

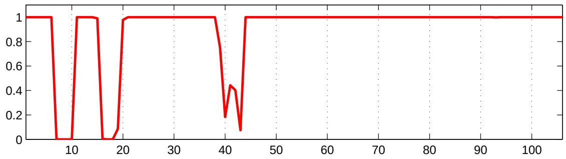


Fig. 2. Posterior visibility probability, $\xi_{1,t} = p(v_t = 1|\mathbf{y}^t)$, plotted against time for the example in Fig. 1. Occlusions and desocclusions make respectively the visibility probability drop, possibly down to zero, and increase back to unity.

3.2 Existence process

Using a Markovian binary variable to indicate presence in the scene is proposed in [Vermaak *et al.*, 2002] to determine in a probabilistic fashion the beginning and end of the track for a single object. We adopt the same model here. However, sequential Monte Carlo is the only inference mechanism considered in [Vermaak *et al.*, 2002], and it is conducted in the augmented state space. By comparison, our generic framework can be easily used with any Bayesian filtering technique and its SMC version implies sampling only in the object state space.

Following [Vermaak *et al.*, 2002], we introduce a binary existence variable e_t that indicates whether the object of interest is present ($e_t = 1$) or not ($e_t = 0$) in the scene at time t. The Markov chain prior on this binary variable is completely defined by the death and birth probabilities, α_{10} and α_{01}. Conditional on the existence variables the state dynamics is specified by

$$p_{00}(\mathbf{x}_t|\mathbf{x}_{t-1}) = p_{10}(\mathbf{x}_t|\mathbf{x}_{t-1}) = \delta_{\mathbf{u}}(\mathbf{x}_t) \tag{23}$$

$$p_{01}(\mathbf{x}_t|\mathbf{x}_{t-1}) = p_{\text{init}}(\mathbf{x}_t) \tag{24}$$

$$p_{11}(\mathbf{x}_t|\mathbf{x}_{t-1}) = p_{\text{dyn}}(\mathbf{x}_t|\mathbf{x}_{t-1}), \tag{25}$$

where $\mathbf{u}$ is the consuming state that corresponds to the object not existing, p_{init} is the initial state distribution, and p_{dyn} is the object dynamic model. From the data model point of view, the existence process is similar to the visibility process.

Due to the component (23) of the evolution model, non-existence $e_t = 0$ deterministically forces $\mathbf{x}_t$ into fictitious state $\mathbf{u}$. This is carried over in the posterior model, yielding

$$p_0(\mathbf{x}_t|\mathbf{y}^t) = \delta_{\mathbf{u}}(\mathbf{x}_t). \tag{26}$$

As a consequence, the algorithm only needs to recursively estimate the conditional filtering distribution for the case of the object existing, i.e., $p_1(\mathbf{x}_t|\mathbf{y}^t)$. Thus, within the SMC framework, only two proposal distributions, q_{01} and q_{11}, are required, instead of four. As in the previous section, we only consider the simple case where these distributions coincide with their counterparts in the evolution model.

In the following experiment, the observation model is defined as in the previous section. Yet again the state comprises the object location in the image plane, and in the state evolution model (24)-(25), p_{init} and p_{dyn} are respectively chosen as the uniform distribution over positions in the image plane and a random walk with independent Gaussian noise. The variance of the noise is 15^2 for each component for the car race sequence in Fig. 3. Also, the state distribution at time $t = 0$ coincides with p_{init}. Hence, contrary to the previous experiment, the tracker is not initialized manually at the beginning of the sequence (the reference foreground model is picked on an arbitrary red

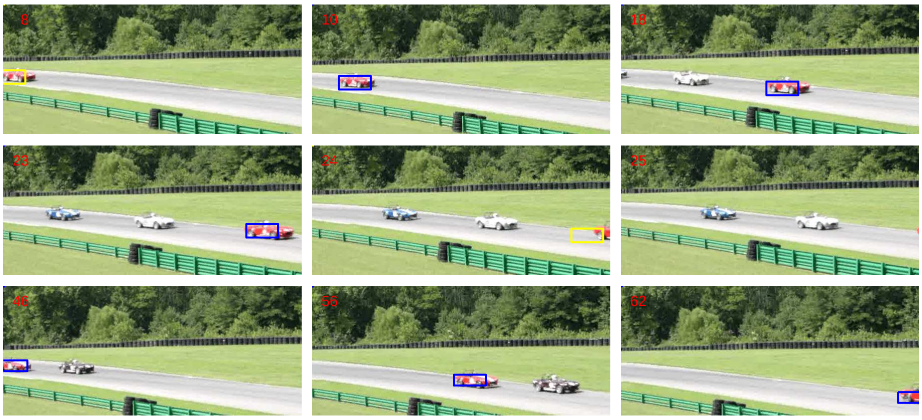


Fig. 3. Detection and tracking. A reference color model is initialized beforehand on one instance of a red car. The algorithm then successfully detects red cars that enter the scene, tracks them as long as they remain in view, and finally determines automatically when they disappear. In each of the displayed frames $\hat{\mathbf{x}}_{1,t}$ is displayed, provided that $\xi_{1,t}$ exceeds 0.2 (in blue if it is greater than 0.8 and in yellow otherwise).

car in a different part of the video). For this experiment, the death and birth probabilities are respectively set to $\alpha_{01} = 0.1$ and $\alpha_{10} = 0.1$, and the initial existence distribution is given by $p(e_0 = 1) = 0.1$. Finally, $N = 50$ particles were sufficient to detect the entrance and exit of red cars in the field of view and to track them while present in the scene. Entrance and exit events are clearly identified by the variations in the posterior existence probability $\xi_{1,t}$, as shown in Fig. 4. In this example, a single tracker successively locks on to different cars, each one appearing in the image after the previous one has been successfully detected and tracked until disappearance. In practice, distinction between different tracked objects would be necessary, especially if they are likely to be present simultaneously in the image. In this context, the information carried by the existence probabilities would facilitate the design of a mechanism that effectively initiates different trackers for each "detected" object and subsequently discards each tracker whose associated existence probability $\xi_{1,t}$ falls below a threshold.

4 Conclusion

In this paper we introduced a generic Bayesian filtering tool to perform tracking in the presence of a certain class of discrete auxiliary processes. The approach places no restriction on the ingredients of the evolution and observation models and on the selected type of filter (Kalman filter and its variants, particle filters). Hence the proposed framework allows re-use of existing architectures on a variety of tracking problems where the introduction

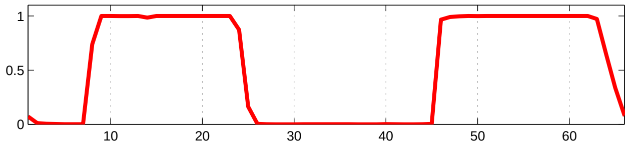


Fig. 4. Posterior existence probability, $\xi_{1,t} = p(e_t = 1|\mathbf{y}^t)$, against time for the example in Fig. 3. When the object of interest enters the scene the existence probability quickly ramps up to one, and falls back down to zero when it exits the field of view.

of auxiliary discrete variables is useful. We demonstrated in particular how the technique can be applied in visual tracking to handle occlusions and object appearance/disappearance via visibility and existence binary processes. Our generic frameworkwould now allow the combination of these two binary processes within a single tracking setup.

References

[Doucet *et al.*, 2000]A. Doucet, S. Godsill, and C. Andrieu. On sequential Monte Carlo sampling methods for Bayesian filtering. *Statistics and Computing*, 10(3):197–208, 2000.

[Gavrila, 2000]D. Gavrila. Pedestrian detection from a moving vehicle. In *Proc. Europ. Conf. Computer Vision*, Dublin, Ireland, June 2000.

[Isard and Blake, 1996]M. Isard and A. Blake. Contour tracking by stochastic propagation of conditional density. In *Proc. Europ. Conf. Computer Vision*, pages I:343–356, 1996.

[Isard and MacCormick, 2001]M. Isard and J. MacCormick. BraMBLe: a Bayesian multiple-blob tracker. In *Proc. Int. Conf. Computer Vision*, pages II: 34–41, Vancouver, Canada, July 2001.

[MacCormick and Blake, 1999]J. MacCormick and A. Blake. A probabilistic exclusion principle for tracking multiple objects. In *Proc. Int. Conf. Computer Vision*, pages 572–578, 1999.

[Nguyen *et al.*, 2001]H.T. Nguyen, M. Worring, and R. van den Boomgaard. Occlusion robust adaptive template tracking. In *Proc. Int. Conf. Computer Vision*, pages I: 678–683, Vancouver, Canada, July 2001.

[North *et al.*, 2000]B. North, A. Blake, M. Isard, and J. Rittscher. Learning and classification of complex dynamics. *IEEE Trans. Pattern Anal. Machine Intell.*, 22(9):1016–1034, 2000.

[Toyama and Blake, 2001]K. Toyama and A. Blake. Probabilistic tracking in a metric space. In *Proc. Int. Conf. Computer Vision*, pages II: 50–57, Vancouver, Canada, July 2001.

[Vermaak *et al.*, 2002]J. Vermaak, P. Pérez, M. Gangnet, and A. Blake. Towards improved observation models for visual tracking: selective adaptation. In *Proc. Europ. Conf. Computer Vision*, pages I: 645–660, Copenhagen, Denmark, May 2002.

[Wu *et al.*, 2003]Y. Wu, T. Yu, and G. Hua. Tracking appearances with occlusions. In *Proc. Conf. Comp. Vision Pattern Rec.*, Madison, Wisconsin, June 2003.

On Triplet Markov Chains

Wojciech Pieczynski and François Desbouvries

GET/INT/Département CITI and CNRS UMR 5157
9 rue Charles Fourier
91011 Evry, France
(e-mail: wojciech.pieczynski@int-evry.fr,
francois.desbouvries@int-evry.fr)

Abstract. The restoration of a hidden process X from an observed process Y is often performed in the framework of hidden Markov chains (HMC). HMC have been recently generalized to triplet Markov chains (TMC). In the TMC model one introduces a third random chain U and assumes that the triplet $T = (X, U, Y)$ is a Markov chain (MC). TMC generalize HMC but still enable the development of efficient Bayesian algorithms for restoring X from Y. This paper lists some recent results concerning TMC; in particular, we recall how TMC can be used to model hidden semi-Markov Chains or deal with non-stationary HMC.

Keywords: hidden Markov chains, hidden semi-Markov chains, pairwise Markov chains, triplet Markov chains, Bayesian segmentation, Kalman filtering and smoothing, iterative conditional estimation.

1 Introduction

An important problem in statistical data restoration consists in estimating a hidden random chain $X = \{X_i\}_{i=1}^n$ from an observed random chain $Y = \{Y_i\}_{i=1}^n$. Let X_i be discrete and Y_i continuous. Many Bayesian methods are available once the distribution of $Z = (X, Y)$ is simple enough. In particular, HMC with independent noise (HMC-IN), in which[1] $p(z) = p(x_1)p(x_2|x_1)\cdots p(x_n|x_{n-1})\ p(y_1|x_1)\cdots p(y_n|x_n)$ have been widely used and studied (see e.g. [Ephraim and Merhav, 2002] for a recent tutorial).

The pairwise Markov chains (PMC) model has been proposed recently [Pieczynski, 2003] and [Derrode and Pieczynski, 2004]. In a PMC one assumes that $Z = (X, Y)$ is an MC, i.e. that $p(z) = p(z_1)p(z_2|z_1)\cdots\ p(z_n|z_{n-1})$. Any HMC-IN is a PMC, but the converse is not true, because in a PMC X is no longer necessarily an MC; however, conditionally on Y, X remains an MC, and in turn this key computational property enables the development

[1] in this formula $p(z)$ denotes the probability density function (pdf) of Z w.r.t. $\kappa^n \otimes \mu^n$, $p(x_i)$ the pdf of X_i w.r.t. κ, and $p(y_i|x_i)$ the conditional pdf (w.r.t. μ) of Y_i given X_i, where κ denotes the counting measure and μ denotes the Lebesgue measure. Later on, other pdf or conditional pdf w.r.t. Lebesgue measure, counting measure, or product measures involving the Lebesgue and/or the counting measure(s) will also be considered; the true meaning of $p(.)$ or of $p(.|.)$ is easily deduced from the context.

of analogous Bayesian restoration algorithms [Lipster and Shiryaev, 2001, corollary 1 p. 72], [Pieczynski, 2003], [Pieczynski and Desbouvries, 2003], and [Desbouvries and Pieczynski, 2003b]. PMC have been further extended to TMC. In the TMC model one introduces a third chain $U = \{U_i\}_{i=1}^n$ (which can be physically meaningful or not) and assumes that the triplet $T = (X, U, Y)$ is an MC [Pieczynski *et al.*, 2002] and [Pieczynski, 2002]. TMC generalize some classical models in the sense that none of the chains X, U, Y, $V = (X, U)$, $Z = (X, Y)$ or (U, Y) needs to be an MC. The wider generality of PMC w.r.t. HMC and of TMC w.r.t. PMC can also be seen through the expression of $p(y|x)$. In an HMC-IN $p(y|x) = p(y_1|x_1) \cdots p(y_n|x_n)$, which is very simple, and undoubtedly too simple in some applications, including speech recognition [Wellekens, 1987] and [Ostendorf *et al.*, 1996]; in a PMC $p(y|x)$ is an MC, which is much richer; and in a TMC $p(y|x)$ is the marginal distribution of the MC $p(u, y|x)$, which is still much richer than an MC. In such applications as image processing, these increasingly complex models are likely to meet the growing need for a better modeling of the noise [Pérez, 2003].

Apart from this general discussion, the contribution of the TMC model (w.r.t. other possible extensions of the HMC-IN model) appears when describing how they encompass and extend some well known stochastic models. This is better appreciated at the local level, as we now see from a simple example. By definition, a TMC distribution is defined by $p(t_1)$ and by $p(t_{i+1}|t_i)$, which itself can be written by different expressions. In particular, the following factorizations will prove useful in the sequel :

$$p(t_{i+1}|t_i) = p(x_{i+1}|t_i)p(u_{i+1}|x_{i+1}, t_i)p(y_{i+1}|x_{i+1}, u_{i+1}, t_i) \tag{1}$$

$$= p(u_{i+1}|t_i)p(x_{i+1}|u_{i+1}, t_i)p(y_{i+1}|x_{i+1}, u_{i+1}, t_i). \tag{2}$$

The HMC-IN model is obtained from (1) if $p(x_{i+1}|t_i)$ reduces to $p(x_{i+1}|x_i)$, $p(u_{i+1}|x_{i+1}, t_i)$ to $\delta_{x_{i+1}}(u_{i+1})$ (with $\delta_{x_{i+1}}$ the Dirac mass, which simply means that $u_{i+1} = x_{i+1}$), and $p(y_{i+1}|x_{i+1}, u_{i+1}, t_i)$ to $p(y_{i+1}|x_{i+1})$. Other (non-trivial) examples will be given below.

The aim of this paper is to summarize some recent results (some of which are still under review) concerning the large family of TMC. In particular, we will see that the TMC model gathers some well known dynamical stochastic models (and thus provides a unifying framework for these models), as well as some new extensions of these models, and yet still enables the development of efficient hidden chain restoration and parameter estimation algorithms.

The rest of this paper is organized as follows. We will say that X (resp. U, Y) is discrete (resp. continuous) if each X_i (resp. U_i, Y_i) takes discrete (resp. continuous) values, and in this paper X and U can be either discrete or continuous (Y will be assumed to be continuous). So we have four possible situations, which are discussed in sections 2 to 5; as we will see, depending on the situation U admits a physical interpretation (see e.g. $\star 2$, item (iii), or $\star 3$, item (ii)) or not (see e.g. $\star 2$, item (i), or $\star 3$, item (i)). Finally section 6 is devoted to parameter estimation.

2 Discrete hidden chain with discrete auxiliary chain

Let X and U be discrete, with $X_i \in \Omega$ and $U_i \in \Lambda$. In this section we shall briefly recall why some classical Bayesian methods like Maximum Posterior Mode (MPM) can be used in TMC. Let $T = (X, U, Y)$ be an MC. The conditional law of $V = (X, U)$ given Y is then an MC, with initial pdf and transitions given by

$$p(v_1|y) = \frac{p(t_1)\beta_1(v_1)}{\sum_{v_1 \in \Omega \times \Lambda} p(t_1)\beta_1(v_1)}, \;\; p(v_{i+1}|v_i, y) = \frac{p(t_{i+1}|t_i)\beta_{i+1}(v_{i+1})}{\beta_i(v_i)}, \quad (3)$$

in which β_i can be computed via the classical backward recursions : $\beta_n(v_n) = 1$ and $\beta_i(v_i) = \sum_{v_{i+1} \in \Omega \times \Lambda} p(t_{i+1}|t_i)\beta_{i+1}(v_{i+1})$ for $1 \leq i \leq n-1$. Once $p(v_1|y)$ has been computed, the a posteriori marginals are computed recursively via $p(v_{i+1}|y) = \sum_{v_i \in \Omega \times \Lambda} p(v_i|y)p(v_{i+1}|v_i, y)$. Finally $p(x_i|y) = \sum_{u_i \in \Lambda} p(v_i|y)$, and thus the MPM estimate, which is defined by

$$[\hat{x}_{MPM}(y) = \{\hat{x}_i\}_{i=1}^n] \Longleftrightarrow [\text{for all } i, 1 \leq i \leq n, \hat{x}_i = \arg\max_{x_i} p(x_i|y)],$$

can be computed.

Let us now describe five particular applications of TMC in which this MPM restoration algorithm can be used.

(i) *Mixture approximation.* Assume that a given PMC (X, Y) is stationary, i.e. that $p(x_i, x_{i+1}, y_i, y_{i+1})$ does not depend on i. Then the distribution of (X, Y) is given by $p(x_1, x_2, y_1, y_2) = p(x_1, x_2)\ p(y_1, y_2|x_1, x_2)$. If $p(y_1, y_2|x_1, x_2)$ is not known exactly, one can approximate it by a mixture distribution (for instance a Gaussian one)

$$p(y_1, y_2|x_1, x_2) = \sum_{u_1, u_2 \in \Lambda \times \Lambda} p(u_1, u_2) p(y_1, y_2|x_1, x_2, u_1, u_2),$$

and in this case the model we implicitely deal with is actually a stationary TMC model, the distribution of which is defined by $p(t_1, t_2) = p(u_1, u_2)$ $p(x_1, x_2)\ p(y_1, y_2|x_1, x_2, u_1, u_2)$.

(ii) *"Switching" or "jumping" models".* One way to model non stationary hidden chains is to assume that for each i, $1 \leq i \leq n-1$, there are m possible transitions $p(x_{i+1}|x_i, u_i)$ with $u_i \in \Lambda = \{\lambda_j\}_{j=1}^m$. One usually considers that u_i is a realization of U_i, and $(U_1, \cdots, U_n)$ is an MC. If we directly assume that (X, U) is an MC then we obtain a more general model since U does not need to be an MC any longer. This model has been successfully applied in non stationary image segmentation [Lanchantin and Pieczynski, 2004a]. A further generalization consists in assuming that (X, U, Y) is a general TMC.

(iii) *Hidden semi-Markov chains (HSMC).* When X is an MC, the distribution of the sojourn duration in a given state is exponential, which is restrictive in some situations. In HSMC this distribution can be of any form; these models thus extend HMC, and yet still enable analogous processing, see e.g. [Yu and Kobayashi, 2003] [Moore and Savic, 2004] [Guédon, 2005]. Let q be a pdf on $\mathbb{N}^*$ modeling the probability distribution of the state duration, let $U_n \in \mathbb{N}^*$ be the time during which X_n remains in the same state, and let $\delta_{x_i}(.)$ be the Dirac mass on x_i. Then the semi-MC model can be written as

$$p(u_{i+1}|u_i) = \begin{cases} \delta_{u_i-1}(u_{i+1}) & \text{if } u_i > 1 \\ q(u_{i+1}) & \text{if } u_i = 1 \end{cases}; \tag{4}$$

$$p(x_{i+1}|x_i, u_i) = \begin{cases} \delta_{x_i}(x_{i+1}) & \text{if } u_i > 1 \\ p(x_{i+1}|x_i) & \text{if } u_i = 1 \end{cases}, \tag{5}$$

with $p(x_{i+1}|x_i) = 0$ for $x_{i+1} = x_i$. Consequently HSMC happen to be particular TMC (with auxiliary chain U), in which the three transition pdf in the r.h.s. of factorization (2) reduce respectively to $p(u_{i+1}|t_i) = p(u_{i+1}|u_i)$ given by (4), $p(x_{i+1}|u_{i+1}, t_i) = p(x_{i+1}|x_i, u_i)$ given by (5), and $p(y_{i+1}|x_{i+1}, u_{i+1}, t_i) = p(y_{i+1}|x_{i+1})$. Notice that the fact that HSMC are particular TMC enables to consider a lot of TMC models generalizing HSMC [Pieczynski, 2004].

(iv) *Non-stationary hidden chain X.* Let us consider the problem of unsupervised restoration using the classical HMC-IN $Z = (X, Y)$. The assumption that X is stationary cannot always be done, and yet this assumption is required when estimating the model parameters. However, the possible non stationarity of X can also be modeled by "mass functions", which can be seen as an extension of the probability distribution on discrete finite sets, and then the computation of the posterior distribution of X becomes a particular "Dempster-Shafer" fusion. Now, one can show that introducing mass functions is mathematically equivalent to considering some TMC, which in turn enables one to use different Bayesian algorithms. In particular, using TMC in unsupervised image segmentation enables to improve the results obtained with classical HMC [Lanchantin and Pieczynski, 2004b].

(v) *Vector auxiliary chain.* In a TMC $T = (X, U, Y)$ the chain U can be a vector one. For instance, it is possible to deal with non-stationary HSMC by introducing the pair $U = (W, S)$, in which W models the fact that an HSMC is a TMC, and S models the fact that the TMC (X, W, Y), which is seen as a PMC (V', Y) with $V' = (X, W)$, is not stationary.

3 Discrete hidden chain with continuous auxiliary chain

Let us now give two examples of TMC models with a discrete hidden chain and a continuous auxiliary chain; the first one, in which (U, Y) is Gaussian conditionally on X, enables to model complex noise distributions; while the second one, in which (U, Y) is not Gaussian conditionally on X, appears in radar signal or images modeling.

(i) Consider the following model : let $T = (X, U, Y)$ be an MC, X be an MC, and (U, Y) be Gaussian conditionnally on X. Since T is an MC, the conditional law of (U, Y) given X is an MC as well. However the conditional distribution of Y given X remains Gaussian but is no longer necessarily an MC (the proof of this result is an adaptation of the proof in [Pieczynski and Desbouvries, 2003] [Desbouvries and Pieczynski, 2003a]), so these simple assumptions can lead to "noise" models (i.e., $p(y|x)$) which are significantly more complex than those one usually deals with. Unfortunately, computing $p(x_i|y)$ exactly is not feasible and approximate methods are needed, as we now briefly explain. Let $x_i \in \Omega$. Since T is an MC, the distribution of (X, U) conditionally on Y is also an MC, the transitions of which can be computed by the "backward" recursion (with the difference that now U_i is continuous). As in section 2, let us classically set $\beta_n(v_n) = 1$ and

$$\beta_i(v_i) = \sum_{x_{i+1} \in \Omega} \int_{\mathbb{R}} p(t_{i+1}|t_i) \beta_{i+1}(v_{i+1}) du_{i+1} \text{ for } 1 \leq i \leq n-1. \quad (6)$$

Then $p(v_{i+1}|v_i, y) = \frac{p(t_{i+1}|t_i)\beta_{i+1}(v_{i+1})}{\beta_i(v_i)}$, so $p(v_{i+1}|v_i, y)$ can be computed if $\beta_i(v_i)$ can be computed. But we see from (6) that $\beta_i(v_i)$ is a rather rich mixture, containing, for k classes, k^{n-i} components.

(ii) *Speckle distribution in SAR images.* TMC with a discrete hidden chain and a continuous auxiliary chain are encountered for instance in radar signal or images, as we see from the following example. Let us consider a TMC $T = (X, U, Y)$ such that X is an MC, and $p(u, y|x) = \prod_{i=1}^{n} p(u_i, y_i|x_i)$. Let also $p(u_i, y_i|x_i) = p(u_i|x_i)p(y_i|u_i, x_i)$, in which $p(u_i|x_i)$ are Gamma distributions, and $p(y_i|u_i, x_i)$ are Gaussian distributions with mean $\mu(x_i)$ and variance $\sigma^2(u_i, x_i) = u_i\sigma^2(x_i)$. Then the distributions $p(y_i|x_i)$ are the so-called "K-distributions", and the chain U is the "speckle" process [Barnard and Weiner, 1996] [Delignon and Pieczynski, 2002] [Brunel and Pieczynski, 2005].

4 Continuous hidden chain with discrete auxiliary chain

In this section we assume that T is a TMC in which both X and Y are continuous, and U is discrete with $u_i \in \Lambda$. As in section 2, switching or jump-Markov models, i.e. models in which U is assumed to be an MC, and (X, Y) is an HMC-IN conditionally on U, are well known simple examples of such TMC; for such models the 3 factors in the r.h.s. of (2) reduce respectively to $p(u_{i+1}|t_i) = p(u_{i+1}|u_i)$, $p(x_{i+1}|u_{i+1}, t_i) = p(x_{i+1}|u_{i+1}, x_i)$, and $p(y_{i+1}|x_{i+1}, u_{i+1}, t_i) = p(y_{i+1}|x_{i+1}, u_{i+1})$.

Let us now consider the restoration problem. Although the physical meanings of the TMC models we deal with in this section are very different of those of section 3, the mathematical modeling and computational difficulties are indeed quite similar. Let us for instance consider the filtering problem, which consists in computing $p(x_i|y_{0:i})$. A recursive solution is given by

$$p(x_i|y_{0:i}) = \frac{\sum_{u_{i-1}\in\Lambda} \int p(x_i, u_i, y_i|x_{i-1}, u_{i-1}, y_{i-1})p(x_{i-1}, u_{i-1}|y_{0:i-1})dx_{i-1}}{p(y_i|y_{0:i-1})}$$

which, in general, cannot be computed in closed form. This computational problem is already encountered in the context of jump-Markov models. In particular, the linear Gaussian case has been studied for a long time, and as is well known the exact computation of the posterior filtered or smoothed estimates leads to a computational cost which grows exponentially with time (see e.g. [Tugnait, 1982] and the references therein). So approximate solutions have been proposed, see e.g. [Tugnait, 1982] [Kim, 1994] [Bar-Shalom and Li, 1995] [Doucet *et al.*, 2001]. Reformulating the jump-Markov model as a particular TMC does not help in solving the filtering problem; however, it can lead to interesting generalizations, to which the classical approximate methods designed for jump-Markov systems could be extended. For instance, in the TMC above U is a discrete MC and thus T can be viewed as a "hidden" MC. Such an HMC could then be extended to an HSMC, as specified in section 2, item (iii).

5 Continuous hidden chain with continuous auxiliary chain

TMC with continuous processes X, U and Y are used in some applications, including the extensions of the classical linear state-space system (7) to colored process and/or measurement noise. Let

$$\begin{cases} X_{n+1} = F_n X_n + G_n \eta_n \\ Y_n \quad\;\; = H_n X_n + J_n \xi_n \end{cases}, \tag{7}$$

in which η_n is the process noise and ξ is the measurement noise. F_n, G_n, H_n and J_n are known deterministic matrices, and processes $\eta = \{\eta_n\}_{n\in\mathbb{N}}$ and $\xi = \{\xi_n\}_{n\in\mathbb{N}}$ are assumed to be independent, jointly independent and independent of X_0. As a consequence, (X, Y) is an HMC-IN. The filtering problem consists in computing the posterior pdf $p(x_n|y_{0:n})$. From (7), $p(x_i|y_{0:i})$ can be computed recursively as

$$p(x_{i+1}|y_{0:i+1}) = \frac{p(y_{i+1}|x_{i+1}) \int p(x_{i+1}|x_i)p(x_i|y_{0:i})dx_i}{\int p(y_{i+1}|x_{i+1})[\int p(x_{i+1}|x_i)p(x_i|y_{0:i})dx_i]dx_{i+1}}. \tag{8}$$

If furthermore X_0 and (η_n, ξ_n) are Gaussian, then $p(x_n|y_{0:n})$ is also Gaussian and is thus described by its mean and covariance matrix. Propagating $p(x_n|y_{0:n})$ amounts to propagating these parameters, and (8) reduces to the celebrated Kalman filter [Kalman, 1960] see also [Ho and Lee, 1964] [Anderson and Moore, 1979] [Kailath *et al.*, 2000].

It happens that some classical extensions of model (7) are particular TMC. Consider for instance model (7), but in which we now assume that

$$\begin{bmatrix} \eta_{n+1} \\ \xi_{n+1} \end{bmatrix} = \underbrace{\begin{bmatrix} A_n^{\eta,\eta} & 0 \\ 0 & A_n^{\xi,\xi} \end{bmatrix}}_{A_n} \underbrace{\begin{bmatrix} \eta_n \\ \xi_n \end{bmatrix}}_{u_n} + \underbrace{\begin{bmatrix} \epsilon_n^\eta \\ \epsilon_n^\xi \end{bmatrix}}_{\epsilon_n}, \tag{9}$$

where $\epsilon^\eta = \{\epsilon_n^\eta\}_{n\in\mathbb{N}}$ (resp. $\epsilon^\xi = \{\epsilon_n^\xi\}_{n\in\mathbb{N}}$ is zero-mean, independent and independent of η_0 (resp. of ξ_0), and ϵ^η and ϵ^ξ are independent. Each one of the two processes $\eta = \{\eta_n\}_{n\in\mathbb{N}}$ and $\xi = \{\xi_n\}_{n\in\mathbb{N}}$ is thus an MC, and η is independent of ξ. Such a model has been introduced by Sorenson [Sorenson, 1966] (see also [Chui and Chen, 1999, ch. 5]). It is no longer an HMC (X is not an MC), but the whole model $T_n = (X_n, U_n, Y_{n-1})$ can be rewritten as

$$\underbrace{\begin{bmatrix} X_{n+1} \\ U_{n+1} \\ Y_n \end{bmatrix}}_{T_{n+1}} = \underbrace{\begin{bmatrix} F_n & \overline{G}_n & 0 \\ 0 & A_n & 0 \\ H_n & \overline{J}_n & 0 \end{bmatrix}}_{\mathcal{F}_n} \begin{bmatrix} X_n \\ U_n \\ Y_{n-1} \end{bmatrix} + \underbrace{\begin{bmatrix} 0 \\ \epsilon_n \\ 0 \end{bmatrix}}_{W_n} \tag{10}$$

(with $\overline{G}_n = [G_n, 0]$ and $\overline{J}_n = [0, J_n]$), and so $T = \{T_n\}$ is a TMC.

Model (10) is indeed a particular case of a linear TMC, defined by $T_{n+1} = \mathcal{F}_n T_n + W_n$, with $T_n = (X_n, U_n, Y_{n-1})$, and W_n independent and independent of T_0. $p(x_n|y_{0:n})$ is obtained by marginalizing $p(v_n|y_{0:n})$ which, in the Gaussian case, can be computed efficiently by a Kalman-like filtering algorithm [Desbouvries and Pieczynski, 2003a], [Ait-el-Fquih and Desbouvries, 2005b].

Kalman-like smoothing algorithms, extending to linear Gaussian TMC the two-filter and RTS smoothers, have also been derived [Ait-el-Fquih and Desbouvries, 2005a].

6 Parameter estimation

Let us finally mention that the model parameters can be estimated from the observed data Y, either by using the well-known "Expectation-Maximization" (EM) method [McLachlan and Krishnan, 1997] or the "Iterative conditional estimation" (ICE) method (some relationships between ICE and EM can be found in [Delmas, 1997]).

As an illustrative example, let us see how the model parameters can be estimated by ICE, which we first briefly recall. Parameter estimation according to the ICE principle can be performed once

(i) an estimator $\hat{\theta}(X, Y)$ of the parameters θ from the complete data (X, Y) is available; and
(ii) one can sample X according to $p(x|y)$.

Then ICE is described by the recursion $\theta^{q+1} = E(\hat{\theta}(X, Y)|Y = y, \theta^q)$, starting with some initial value θ^0. If for some components θ_j of θ this expectation cannot be computed, one samples $x^1, \cdots, x^l$ according to $p(x|y, \theta^q)$ and sets $\theta_j^{q+1} = \frac{1}{l}\sum_{i=1}^{l} \hat{\theta}_j(x^i, y)$.

Let us turn to parameter estimation in PMC and TMC. Let us first remark that the problem is identical in both cases, since a TMC $T = (X, U, Y)$ can be seen as a PMC (V, Y) with $V = (X, U)$. Let us as an illustrative example consider the case of a stationary PMC $Z = (X, Y) = (X_1, Y_1, \cdots X_n, Y_n)$ in which $p(z_i, z_{i+1})$ does not depend on i. So the distribution of Z is given by $p(z_1, z_2) = p(x_1, x_2)p(y_1, y_2|x_1, x_2)$. Assume that $X_i \in \Omega = \{\omega_1, \omega_2\}$ and that $p(y_1, y_2|x_1, x_2)$ are Gaussian. Then the model parameter consists of $\theta = (\alpha, \beta)$, where α gathers the four parameters $\alpha = \{\alpha_{i,j} = p(x_1 = \omega_i, x_2 = \omega_j)\}_{i,j=1}^2$, and $\beta = \{\beta_i\}_{i=1}^{20}$ the twenty parameters of the four Gaussian densities $\{p(y_1, y_2|x_1, x_2)\}_{x_1,x_2 \in \Omega\times\Omega}$ on $\mathbb{R}^2$.

Let us now apply ICE to this model. Let $\hat{\theta}(X, Y) = (\hat{\alpha}(X), \hat{\beta}(X, Y))$. $\hat{\alpha}(X)$ can be chosen as the classical frequency estimator, and $\hat{\beta}(X, Y)$ as the classical empirical means and variance-covariance matrices. Then $\alpha_i^{q+1} = E(\hat{\alpha}_i(X)|Y = y, \theta^q)$ can be computed, but $\beta_i^{q+1} = E(\hat{\beta}_i(X, Y)|Y = y, \theta^q)$ cannot. In practice, the interest of PMC over HMC-IN in unsupervised segmentation using the ICE principle has been proven by different experiments [Derrode and Pieczynski, 2004]. On the other hand, using copulas enables to extend ICE to the case where the exact nature of the noise distribution is not known (it can take different possible forms) [Brunel and Pieczynski, 2003].

References

[Ait-el-Fquih and Desbouvries, 2005a]B. Ait-el-Fquih and F. Desbouvries. Bayesian smoothing algorithms in pairwise and triplet Markov chains. In *submitted to the 2005 IEEE Workshop on Statistical Signal Processing*, Bordeaux, France, July 2005.

[Ait-el-Fquih and Desbouvries, 2005b]B. Ait-el-Fquih and F. Desbouvries. Kalman filtering for triplet Markov chains : Applications and extensions. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP 05)*, Philadelphia, USA, March 19-23 2005.

[Anderson and Moore, 1979]B. D. O. Anderson and J. B. Moore. *Optimal Filtering.* Prentice Hall, Englewood Cliffs, New Jersey, 1979.

[Bar-Shalom and Li, 1995]Y Bar-Shalom and X. R. Li. *Multitarget-multisensor tracking : principles and techniques.* YBS, 1995.

[Barnard and Weiner, 1996]T. J. Barnard and D. D. Weiner. Non-Gaussian clutter modeling with generalized spherically invariant random vectors. *IEEE Transactions on Signal Processing*, 44(10):2384–2390, 1996.

[Brunel and Pieczynski, 2003]N. Brunel and W. Pieczynski. Unsupervised signal restoration using copulas and pairwise Markov chains. In *Proceedings of the 2003 IEEE Workshop on Statistical Signal Processing*, St. Louis, MI, September 2003.

[Brunel and Pieczynski, 2005]N. Brunel and W. Pieczynski. Modeling temporal dependence of spherically invariant random vectors with triplet Markov chains. In *submitted to the 2005 IEEE Workshop on Statistical Signal Processing*, Bordeaux, France, July 2005.

[Chui and Chen, 1999]C.K. Chui and G. Chen. *Kalman Filtering with Real-Time Applications*. Berlin, DE: Springer, 1999.

[Delignon and Pieczynski, 2002]Y. Delignon and W. Pieczynski. Modeling non-Rayleigh speckle distribution in SAR images. *IEEE Transactions on Geoscience and Remote sensing*, 40(6):1430–1435, 2002.

[Delmas, 1997]J.-P. Delmas. An equivalence of the EM and ICE algorithm for exponential family. *IEEE Transactions on Signal Processing*, 45(10):2613–15, 1997.

[Derrode and Pieczynski, 2004]S. Derrode and W. Pieczynski. Signal and image segmentation using pairwise Markov chains. *IEEE Transactions on Signal Processing*, 52(9):2477–89, 2004.

[Desbouvries and Pieczynski, 2003a]F. Desbouvries and W. Pieczynski. Modèles de Markov triplet et filtrage de Kalman. *Comptes Rendus de l'Académie des Sciences - Mathématiques*, 336(8):667–670, 2003. in French.

[Desbouvries and Pieczynski, 2003b]F. Desbouvries and W. Pieczynski. Particle filtering in pairwise and triplet Markov chains. In *Proceedings of the IEEE - EURASIP Workshop on Nonlinear Signal and Image Processing (NSIP 2003)*, Grado-Gorizia, Italy, June 8-11 2003.

[Doucet *et al.*, 2001]A. Doucet, N. J. Gordon, and V. Krishnamurthy. Particle filters for state estimation of jump Markov linear systems. *IEEE Transactions on Signal Processing*, 49(3):613–24, March 2001.

[Ephraim and Merhav, 2002]Y. Ephraim and N. Merhav. Hidden Markov processes. *IEEE Transactions on Information Theory*, 48(6):1518–69, June 2002.

[Guédon, 2005]Y. Guédon. Hidden hybrid Markov/semi-Markov chains. *Computational Statistics and Data Analysis*, 2005. to appear.

[Ho and Lee, 1964]Y. C. Ho and R. C. K. Lee. A Bayesian approach to problems in stochastic estimation and control. *IEEE Transactions on Automatic Control*, 9:333–339, October 1964.

[Kailath *et al.*, 2000]T. Kailath, A. H. Sayed, and B. Hassibi. *Linear estimation*. Prentice Hall Information and System Sciences Series. Prentice Hall, Upper Saddle River, New Jersey, 2000.

[Kalman, 1960]R. E. Kalman. A new approach to linear filtering and prediction problems. *J. Basic Eng., Trans. ASME, Series D*, 82(1):35–45, 1960.

[Kim, 1994]C-J. Kim. Dynamic linear models with Markov switching. *J. of Econometrics*, 60:1–22, 1994.

[Lanchantin and Pieczynski, 2004a]P. Lanchantin and W. Pieczynski. Unsupervised non stationary image segmentation using triplet Markov chains. In *Advanced Concepts for Intelligent Vision Systems (ACVIS 04)*, Brussels, Belgium, August 31 - September 3 2004.

[Lanchantin and Pieczynski, 2004b]P. Lanchantin and W. Pieczynski. Unsupervised restoration of hidden non stationary Markov chain using evidential priors. *accepted for publication, IEEE Transactions on Signal Processing*, 2004.

[Lipster and Shiryaev, 2001]R. S. Lipster and A. N. Shiryaev. *Statistics of Random Processes, Vol. 2 : Applications*, chapter 13 : "Conditionally Gaussian Sequences : Filtering and Related Problems". Springer Verlag, Berlin, 2001.

[McLachlan and Krishnan, 1997]G. J. McLachlan and T. Krishnan. *EM Algorithm and Extensions.* Wiley, 1997.

[Moore and Savic, 2004]M. D. Moore and M. I. Savic. Speech reconstruction using a generalized HSMM (GHSMM). *Digital Signal Processing*, 14(1):37–53, 2004.

[Ostendorf *et al.*, 1996]M. Ostendorf, V.V. Digalakis, and O. A. Kimball. From HMMs to segment models: a unified view of stochastic modeling for speech recognition. *IEEE Transactions on Speech and Audio Processing*, 4(5):360–378, September 1996.

[Pérez, 2003]P. Pérez. *Modèles et algorithmes pour l'analyse probabiliste des images.* Université de rennes I. Habilitation à diriger les recherches, 2003. (in French).

[Pieczynski and Desbouvries, 2003]W. Pieczynski and F. Desbouvries. Kalman filtering using pairwise Gaussian models. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP 03)*, Hong-Kong, 2003.

[Pieczynski *et al.*, 2002]W. Pieczynski, C. Hulard, and T. Veit. Triplet Markov chains in hidden signal restoration. In *SPIE International Symposium on Remote Sensing*, Crete, Grece, September 22-27 2002.

[Pieczynski, 2002]W. Pieczynski. Chaînes de Markov triplet. *Comptes Rendus de l'Académie des Sciences - Mathématiques*, 335:275–278, 2002. in French.

[Pieczynski, 2003]W. Pieczynski. Pairwise Markov chains. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(5):634–39, May 2003.

[Pieczynski, 2004]W. Pieczynski. Chaînes semi-Markoviennes cachées et chaînes de Markov triplet. *submitted to : Comptes Rendus de l'Académie des Sciences - Mathématiques*, décembre 2004.

[Sorenson, 1966]H. W. Sorenson. Kalman filtering techniques. In C. T. Leondes, editor, *Advances in Control Systems Theory and Appl.*, volume 3, pages 219–92. Acad. Press, 1966.

[Tugnait, 1982]J. K. Tugnait. Adaptive estimation and identification for discrete systems with Markov jump parameters. *IEEE Transactions on Automatic Control*, 27(5):1054–65, October 1982.

[Wellekens, 1987]C. J. Wellekens. Explicit time correlation in hidden Markov models for speech recognition. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP 87)*, volume 12, pages 384–86, 1987.

[Yu and Kobayashi, 2003]S.-Z. Yu and H. Kobayashi. A hidden semi-Markov model with missing data and multiple observation sequences for mobility tracking. *Signal Processing*, 83(2):235–250, 2003.

Variations on Markovian Quadtree Model for Multiband Astronomical Image Analysis

Christophe J.-F. Collet and Farid Flitti

LSIIT UMR CNRS 7005, Université Strasbourg 1 (ULP)
Bd S. Brant, BP 10413, 67412 Illkirch, France
collet@lsiit.u-strasbg.fr , flitti@lsiit.u-strasbg.fr

Abstract. This paper is concerned with the analysis of multispectral observations, provided by space or ground telescopes. The large amount and the complexity of heterogeneous data to analyse lead us to develop new methods for segmentation tasks, which aim to be robust, fast and efficient. Some prior knowledge on the information to be extracted from the original image is available, and Bayesian statistical theory is known to be a convenient tool to take this *a priori* knowledge into consideration. In this paper, we investigate the use of the Bayesian inference on Markovian quadtrees for some reduction, fusion, segmentation or restoration problems of great importance in multiband astronomical imagery.

Keywords: Markovian quadtree, Bayesian inference, fusion, data reduction, copulas, astronomy.

1 Introduction

This paper deals with the unsupervised segmentation, reduction, fusion or restoration of multiband images. These different tasks are developed in an astronomical multispectral imagery framework, and validated on raw data cubes. The main goal of this presentation, consists in showing different processing chains describing the power, the efficiency and the fruitfulness of hierarchical Markovian modeling based on a quadtree topology. We will see that such modeling allows to deal with a large varieties of data : missing data, multiresolution data, multiband data, strongly noised data. In particular, we show how such approach is general and how this tool is able to face with a large number of various image processing tasks. The paper is organized as follows. The Markovian quadtree model is described in section 2. In section 3 a reduction methods on large data cube is coupled with quadtree modeling, in order to provide a single segmentation map avoiding thus the curse of dimensionality phenomenon. Then, in the fourth section, we propose to process the wavelet coefficients on the raw data cube, and feed a Markovian quadtree with the multiscale coefficients of the wavelet transform. Indeed, the quadtree topology exhibits a suitable structure to deal with multiscale coefficients : in this way, it becomes possible to use the different multi-scale segmentation maps obtained along a quadtree to restore and fused multiband images. Particularly, the problem of between-channels correlation modeling in the non-Gaussian case is briefly presented.

2 Markovian quadtree and segmentation tasks

Statistical Markovian approaches have proved to be fruitful to design robust and efficient images analysis methods. In the context of multispectral images, handling

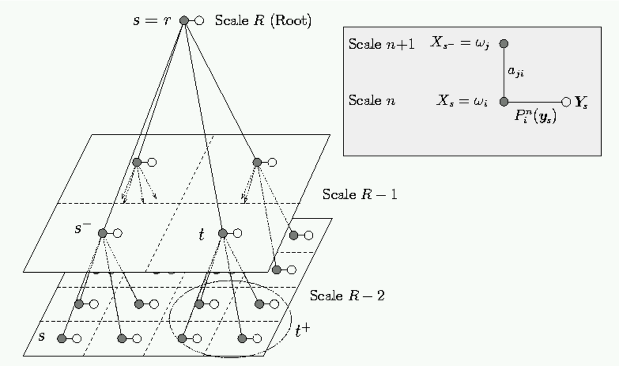


Fig. 1. Example of a dependency graph corresponding to a quadtree structure on a 16 × 16 lattice. Black circles represent labels and white circles represent multicomponent observations. Each node t has a unique parent t^- and four "children" t^+. a_{ij} stands for inter-scale probability of label transition, whereas $P_i^n(y_s)$ represents the likelihood to affect a label ω_i with observation y_s. Likelihood parameters and Markovian quadtree parameters (a_{ij} and root probabilities) can be estimated with, *e.g.,* an EM algorithm. The segmentation algorithm re-estimates iteratively the parameters of a given hidden in-scale-Markov model, to produce a new model which has a higher probability of generating the given observation sequence. This re-estimation procedure is continued until no more significant improvement in parameters can be obtained. The two-step computation of posterior marginals propagates available information all over the tree : on one hand, the bottom-up step spreads the influence of data to other levels up to the root, on the other hand the top-down step computes the posterior marginals taking into account this information. Thus, this proposed modeling scheme captures, over the quadtree, significant statistical dependencies and provides a robust scheme for segmentation.

correlated observed data requires a well-designed modeling framework. Resorting to a Bayesian scheme based on Markov models is indeed attractive when dealing with large amount of multispectral observations. Nevertheless, the well known Markov Field Models (MFM) lead to iterated optimization algorithms, not really well adapted [Geman and Geman, 1984, Graffigne *et al.*, 1995, Kato *et al.*, 1996] for many applications, even if some strategies to decrease the computing time have been proposed in the last decade (e.g., [Pérez *et al.*, 2000, Mignotte *et al.*, 2000]). This is due to the fact that most of Markov models are non-causal. As a consequence, inference must be conducted iteratively, which might turn prohibitively expensive. One way to circumvent this problem is to resort to a Markov model on a quadtree where in-scale causality[Laferté *et al.*, 2000, Provost *et al.*, 2003] permits non-iterative inference . A quadtree-based approach offers the well-known advantages of standard hierarchical techniques (improved robustness, ability to deal with multiresolution or missing data), while allowing for non-iterative inference as in the case of hidden Markov chains [Giordana and Pieczynski, 1997]. Let $G = (S, L)$ be a graph composed of a set S of nodes and a set L of edges. A tree is a connected graph with no cycle, where as a consequence, each node apart from the root r has a unique predecessor, its "parent", on the path to the root. A quadtree, as illustrated in Fig. 1, is a special case of tree where each node, apart from the terminal ones, the "leaves", has four "children". The set of nodes S can be partitioned into "scales", $S = S^0 \cup S^1 \ldots \cup S^R$, according to the path length from each node to the root. Thus, $S^R = \{r\}$, S^n involves 4^{R-n} sites, and S^0 is the finest scale formed by the leaves. We consider a labeling process X which assigns a class label X_s to each node of G : $X = \{X^n\}_{n=0}^{R}$ with $X^n = \{X_s, s \in S^n\}$ where X_s takes its values in the set $\Omega = \{\omega_1, ..., \omega_K\}$, of the K classes. A number of conditional independence properties are assumed. First, X is supposed to be Markovian in scale, *i.e.,*[1] $P(x^n|x^k, k > n) = P(x^n|x^{n+1})$. It is also assumed that the probabilities of inter-scale transitions can be factorized in the following way [Laferté *et al.*, 2000]:

$$P(x^n|x^{n+1}) = \prod_{s \in S^n} P(x_s|x_{s^-}), \tag{1}$$

where s^- designates the father of site s, as illustrated in Fig. 1. Finally, the likelihood of the multiband/multisensor observations $\boldsymbol{Y}$ conditionally to X is expressed as the following product (assuming conditional independence):

$$P(\boldsymbol{Y}{=}\boldsymbol{y}|x) = \prod_{n=0}^{R} P(\boldsymbol{y}^n|x^n) = \prod_{n=0}^{R} \prod_{s \in S^n} P(\boldsymbol{y}_s^n|x_s), \tag{2}$$

where $\forall s \in S^n$, $\forall n \in \{0, ..., R\}$, $P(\boldsymbol{y}_s|x_s = \omega_i) \overset{\triangle}{=} f_i(\boldsymbol{y}_s)$, captures the likelihood of the data y_s. Each site s of scale n can be associated with a label ω_i. If data are available at scale n, then the likelihood is expressed as $f_i^n(\boldsymbol{y}_s^n)$. Of course, if the data-driven terms do not follow a Gaussian law, the analytic expression of the multidimensional density $f_i^n(\boldsymbol{y}_s^n)$ is not always available. To overcome this difficulty, one may decorrelate bands via an adequate mapping, compute the multidimensional density of the decorrelated data as a simple product of the marginals and then obtain $f_i^n(\boldsymbol{y}_s^n)$ by Jacobian method [Provost *et al.*, 2003]. Another solution is to use copulas theory [Nelsen, 1998][Brunel *et al.*, 2005] (see Annexe). In section 3, we present a new way for multidimensional data-driven term computation, thanks to a regularized mixture of Probabilistic Principal Component.

Sometimes, the lack of observed data on some locations within the pictures leads to intricate segmentation problems but here, missing data can be easily inferred [Provost *et al.*, 2003]. In a general manner, we suppose the data available at different levels n, including the finest level ($n = 0$). On one hand, when no observation exists (for any given scale n), the likelihood $f_i^n(\boldsymbol{y}_s^n)$ is set to 1. On the other hand, if we have images of the same area at different levels of resolution, the quadtree structure can be still used and permits to properly consider all the available data. It is a way to conduct the segmentation while merging data. From these assumptions, it can be easily inferred that the joint distribution $P(x, \boldsymbol{y})$ can be factorized as follows :

$$P(x, \boldsymbol{y}) = P(x_r) \prod_{s \neq r} P(x_s|x_{s^-}) \prod_{n=0}^{R} \prod_{s \in S^n} P(\boldsymbol{y}_s|x_s). \tag{3}$$

One of the interests of this model lies in the possibility of computing exactly the posterior marginals $P(X_s|\boldsymbol{Y})$ and $P(X_s, X_s^-|\boldsymbol{Y})$ at each node s within two passes

[1] To simplify notation, we will denote the discrete probability $P(X = x)$ as $P(x)$.

in an unsupervised way [Delmas, 1997]. The segmentation label map $\hat{x}$ to be determined is finally given by:

$$\hat{x}_{s,s\in S^n} = \arg\max_{\omega_i \in \Omega} P(X_s = \omega_i | \boldsymbol{Y} = \boldsymbol{y}). \tag{4}$$

Equation (4) shows that we obtain a labeling of each pixel at each level of the quadtree, even if observations only lie on the finest level and even if there is missing data.

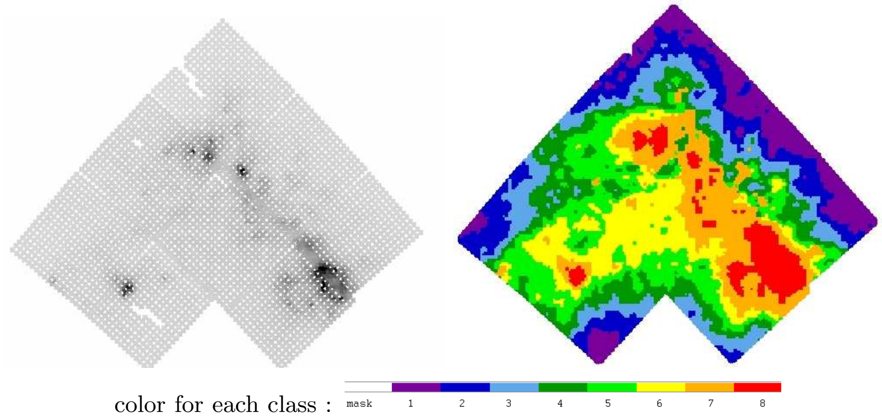


Fig. 2. On the left picture, composed of a mosaic of 9 observations, the missing pixel data, due to sampling adjustment problems, appear as a regular lattice of white dots. The Markovian quadtree allow to reconstruct a segmentation map without missing labels : the missing observations are labeled thanks to the Markovian *a priori* model.

3 Reduction/Segmentation on the Quadtree

Analysis of multicomponent data sets is a very hard task, due to the curse of dimensionality[Hughes, 1968]. Indeed, learning algorithms need a large diversity of observations to cover the behavior of the studied process. Especially, in the multidimensional case, the required number of samples grows quickly with the dimension, so that the process behavior becomes rapidly untractable in practice. This is the so-called Hughes phenomenon which corresponds to an important loss of accuracy in the process statistics estimation as dimensionality grows (more precisely the likelihood term in the quadtree). For example , for an observation size of $H \times W$ pixels by D spectral bands, one more channel observed adds $H \times W$ additional samples whereas the complexity deals with $\mathbb{R}^{D+1}$. To deal with this problem, one may carry out a space reduction step before classification [Landgrebe, 2003]. Fortunately, high dimensional observed data can often be described in a significantly smaller number of dimension than the original due to redundancy in data cube where neighboring bands are highly correlated. Many approaches were proposed to solve such analysis task. All seek a mapping on a reduced dimension space by maximizing a given criterion [Duda *et al.*, 2001]. More graceful solution consists on combining reduction and classification by associating a generative model to the observations within each class to compute the corresponding likelihood. Thus the observations are modeled as a mixture of such generative models [Tipping and Bishop, 1999, Lee *et al.*, 2000]. In this paper we propose to use a Markovian *a priori* associated with such generative models to regularize multidimensional pixel classification. In the sequel this approach will be illustrated using the Probabilistic Principal Component analysis (PPCA) generative model.

3.1 Probabilistic Principal Component analysis (PPCA)

The PPCA [Tipping and Bishop, 1999] is based on a latent variable model which lies each $D \times 1$ observed vector $\mathbf{y}$ to $q \times 1$ latent vector $\mathbf{t}$, $q < D$, as follows:

$$\mathbf{y} = A\mathbf{t} + \mu + \epsilon \tag{5}$$

where A is a $D \times q$ matrix, μ the observed data mean and ϵ is a random variable following an Gaussian $\mathcal{N}(0, \sigma^2 I)$ noise, I being the identity matrix. Given $\mathbf{t}$ and Eq. 5, the $\mathbf{y}$ probability distribution is :

$$P(\mathbf{y}/\mathbf{t}) = (2\pi\sigma^2)^{\frac{-D}{2}} exp\{-\frac{1}{2\sigma^2}\|\mathbf{y} - W\mathbf{t} - \mu\|^2\}. \tag{6}$$

Choosing Gaussian *prior* for $\mathbf{t}$, *i.e.*; $\mathcal{N}(0, I)$, the marginal distribution of $\mathbf{y}$ is

$$P(\mathbf{y}) = (2\pi)^{\frac{-D}{2}} |C|^{\frac{-1}{2}} exp\{\frac{-1}{2}(\mathbf{y} - \mu)^t C^{-1}(\mathbf{y} - \mu)\} \tag{7}$$

with $C = \sigma^2 I + AA^t$ a $D \times D$ matrix. Using the Bayes rule, the *a posteriori* probability of $\mathbf{t}$ is found to be [Tipping and Bishop, 1999] $\mathcal{N}(M^{-1}A^t(\mathbf{y} - \mu), \sigma^2 M^{-1})$ where $M = \sigma^2 I - A^t A$.

The maximization of the data log-likelihood $\mathcal{L} = \sum_{s \in S^0} \ln\{p(\mathbf{y}_s)\}$ gives the following parameter estimators :

$$\mu_{ML} = \frac{\sum_{s \in S^0} \mathbf{y}_s}{card(S^0)}; \;\; \sigma^2_{ML} = \frac{1}{D - q} \sum_{j=q+1}^{D} \lambda_j; \;\; A_{ML} = U_q(\Lambda_q - \sigma^2 I)^{\frac{1}{2}} R. \tag{8}$$

where λ_j are the eigenvalues of the data covariance matrix $\Sigma_x = \frac{1}{card(S^0)} \sum_{s \in S^0} (\mathbf{y}_s - \mu)(\mathbf{y}_s - \mu)^t$ given in descending order $(\lambda_1 \geq \cdots \geq \lambda_q)$, Λ_q is a diagonal matrix of the q largest eigenvalues, U_q the matrix of the corresponding eigenvectors, and R is an arbitrary orthogonal rotation matrix.

3.2 Regularized mixture of Probabilistic Principal Component analyzers

A mixture of PPCA (MPPCA) was proposed in [Tipping and Bishop, 1999] to model complex data structures as a combination of local PCA. For a K component MPPCA, the observations are partitioned in K clusters (*i.e;* classes) each one spanned by a local PPCA. Given this model, the distribution of the observations is $P(\mathbf{y}_s) = \sum_{i=1}^{K} \pi_i P(\mathbf{y}_s / x_s = \omega_i)$. Note that in this formulation the *prior* is the same for all $s \in S^0$ and thus, any information about the neighborhood is taken into account when classifying $\mathbf{y}_s$. We adapt this model by imposing a Markovian constraints via the quadtree modelling. The observation distribution become

$P(\mathbf{y}_s) = \sum_{i=1}^{K} P(x_s = \omega_i)P(\mathbf{y}_s/x_s = \omega_i)$, where X_s is drawn from a hierarchical Markovian process (Eq. 1) and

$$P(\mathbf{y}_s/x_s = \omega_i) = (2\pi)^{\frac{-D}{2}} |C_i|^{\frac{-1}{2}} exp\{\frac{-1}{2}(\mathbf{y} - \mu_\mathbf{i})^t C_i^{-1}(\mathbf{y} - \mu_\mathbf{i})\}. \qquad (9)$$

The matrix C_i is obtained in analog manner to Eqs. 7 and 8 by eigen-decomposition of the weighted covariance matrix $\Sigma_i = \frac{\sum_{s \in S^0} P(x_s=\omega_i/Y)(\mathbf{y}_s - \hat{\mu_i})(\mathbf{y}_s - \hat{\mu_i})^t}{\sum_{s \in S^0} P(x_s=\omega_i/Y)}$, where $\hat{\mu_i} = \frac{\sum_{s \in S^0} P(x_s=\omega_i/Y)\mathbf{y}_s}{\sum_{s \in S^0} P(x_s=\omega_i/Y)}$. The estimation of the *a priori* parameter remains the same as in the classical quadtree. To test our approach, we generate 3 sets of 3 images (2 classes (geometric shape and background) with Gaussian distribution (mean 120/120/128 and 136/136/128, standard deviation 16/16/16). Thus we obtain 9 images to segment. The obtained 4-classes segmentation map shows clearly the better behavior of our proposed approach towards MPPCA (cf. Fig. 3).

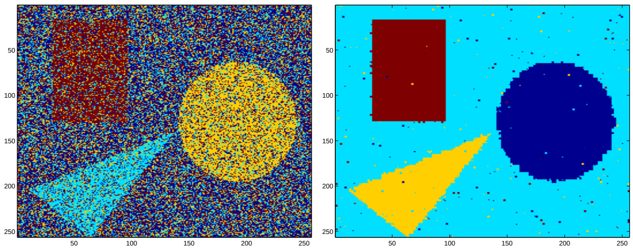


Fig. 3. Segmentation map obtained with the MPPCA on 9 images (left) remains noisy whereas the map obtained with the proposed technique (right) is well regularized.

4 Wavelet domain for restoration and fusion tasks

Fusion of multiband images is of great interest in astronomy, allowing to obtain an efficient summary of the whole multiband information in a single scene. Generally this task is more difficult for noisy observations. The wavelet domain is well adapted both for fusion [Zhang and Blum, 1999] and denoising [D.L.Donoho and Johnstone, 1994] tasks. Actually, wavelet coefficients measure local variations in the image and the sharper the discontinuity, the larger the coefficients. Intensity fluctuations corresponding to the noise, most of time considered as uncorrelated, are most important at the finest resolution and related wavelet coefficients decrease quickly as the scale increases. Real structures in the image will therefore lead to larger wavelet coefficient values at these coarsest resolutions. A threshold can be defined at each scale below which all the coefficients are discarded [D.L.Donoho and Johnstone, 1994]. Note that the result of such analysis depends strongly both on the wavelet used and on the thresholds chosen. Generally astronomical objects are diffuse and exhibit smooth edges so isotropic wavelet transforms are well adapted [Starck *et al.*, 1998]. We use the pyramidal algorithm with one wavelet which is an isotropic transform obtained by adapting the classical Laplacian pyramid [Starck *et al.*, 1998].

Few years ago [Crouse *et al.*, 1998], an efficient Markovian modeling of wavelets was introduced capturing interscale and spacial wavelet coefficient correlations. In this paper we use a more general Markovian framework modeling not only spatial and interscale dependencies as the existent models but also interband correlation for multiband image fusion and denoising. Moreover, the multidimensional likelihood may be efficiently modeled using the copulas theory [Nelsen, 1998] allowing us to use any kind of marginal densities with a given interband correlation. The

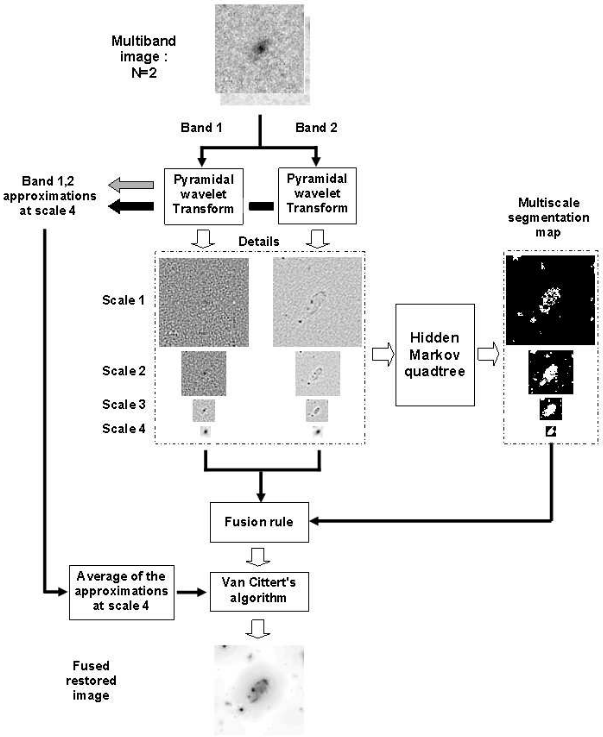

Multiband image : N=2
Band 1
Band 2
Band 1,2 approximations at scale 4
Pyramidal wavelet Transform
Pyramidal wavelet Transform
Details
Multiscale segmentation map
Scale 1
Scale 2
Scale 3
Scale 4
Hidden Markov quadtree
Fusion rule
Average of the approximations at scale 4
Van Cittert's algorithm
Fused restored image

Fig. 4. Fusion-restoration algorithm illustrated for a bi-band image. A pyramidal wavelet transform analyzes the two spectral bands (on the top). This leads to a multiresolution pyramid of wavelet coefficients for each band, up to scale 4. Then, all wavelet pyramids are combined to carry out two-class multiresolution Markovian segmentation map (on the right). This segmentation map masks small coefficients at different scales. The remainder coefficients are fused using an appropriate rule. The result with the average of coarsest approximations feed an iterative reconstruction procedure to give a unique fused restored image.

proposed approach is summarized in Fig.4. For a multiband image $\mathbf{Y}$ with D bands, a wavelet decomposition is carried out for each band b separately leading to a multiresolution pyramids $\mathcal{W}^b$, $b = 1, \cdots, D$. These D pyramids are combined in unique Multiband-Multiresolution Pyramid (MMP, *cf.* Fig. 4 and 5) $\mathcal{W}$ by considering details coefficients, $\mathcal{W}^1_{s \in S^j}, \cdots, \mathcal{W}^D_{s \in S^j}$, for space location s at scale j as a components of an unique vector $\mathcal{W}_j(s)$. The MMP is segmented in two-classes (*i.e.;* $\forall s \in S\ :\ x_s \in \{0,1\}$) using a vectorial hidden Markov quadtree (Fig.5) to separate significant wavelet coefficients from those associated with the noise. The selection relies now not only on the sole coefficient magnitude but also takes into account its neighbors : in space, in scale and with wavelength. This classification scheme produces a multiresolution binary mask highlighting significant wavelet coefficients and removing the others, corresponding to the noise contribution. The fusion of the *cleaned* wavelet coefficients is operated using the following rule :

$$\forall s \in S^n\ :\ W_s^{fused} = \frac{\sum_{i=1}^{D} \sigma_i^n x_s \mathcal{W}_s^i}{\sum_{i=1}^{D} \sigma_i^n x_s}, \tag{10}$$

σ_i^n being the standard deviation of the i^{th} marginal of the likelihood associated with class kept at scale n. The structure W^{fused} does not correspond to a smooth image since all non significant coefficients are put to zero before fusion. We seek instead a smooth solution $\hat{F}^{fused}$ which minimizes $\parallel (W^{fused} - O(\hat{F}^{fused})) \parallel$ where O is the wavelet transform operator. In practice we use the Van Cittert's algorithm [Starck *et al.*, 1998] to obtain the final restored-fused image (see Fig. 6).

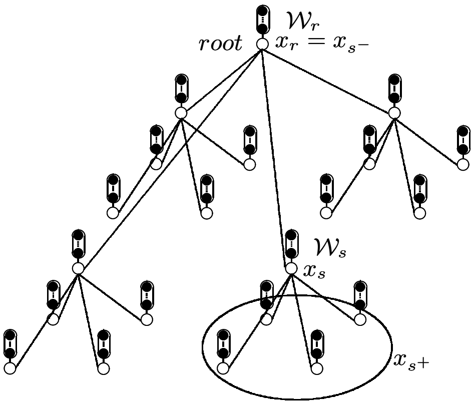


Fig. 5. Example of a dependency graph corresponding to a quadtree structure on a 4 × 4 lattice. White circles represent labels and black circles represent multiband observations $\mathcal{W}_s$, $s \in S$ in the wavelet domain.

Conclusion

This paper summarizes some variations around Markovian quadtree model, in order to show the efficiency of such a tool, to deal with unsupervised multiband image analysis, for e.g., reduction, segmentation, restoration, fusion tasks. Our motivations for using such a model are to provide fast computations and efficient structures to process multispectral and multiresolution large observations. Indeed, computer vision and astronomers communities need efficient tools to analyse and interpret large data cubes : ground or on-board telescopes provide larger amount of multispectral/multiresolution data cube every year, that have to be processed in an efficient way.

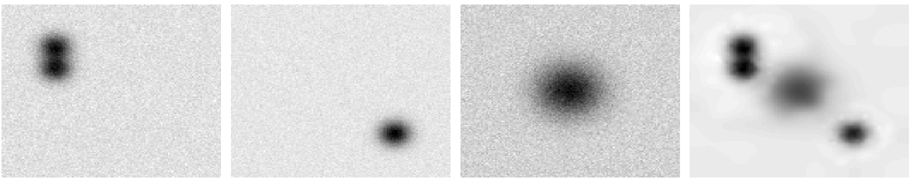


Fig. 6. Example of image fusion: from the left three simulated bands, the fusion result is on the right. All objects appearing in the three bands are present in fused image.

Annexe : Copulas for N-D likelihood computation

The basis of the copulas theory is Sklar's Theorem [Nelsen, 1998] which asserts the existence of a function C, called copula and defined on $[0,1]^N$, binding the joint cumulative distribution function $F(\mathbf{y}_s^1,\cdots,\mathbf{y}_s^N)$ to the marginal cumulative distribution functions $F^{[1]}(\mathbf{y}_s^1),\cdots,F^{[N]}(\mathbf{y}_s^N)$ as follows : $F(\mathbf{y}_s^1,\cdots,\mathbf{y}_s^N) = C(F^{[1]}(\mathbf{y}_s^1),\cdots,F^{[N]}(\mathbf{y}_s^N))$. If the marginals $F^{[1]},\cdots,F^{[N]}$ are continuous, then C is unique. Moreover, if C is differentiable it is possible to define a copula density as [Nelsen, 1998]:

$$\begin{aligned} f(\mathbf{y}_s^1,\cdots,\mathbf{y}_s^N) = f^{[1]}(\mathbf{y}_s^1) \times \cdots \times f^{[N]}(\mathbf{y}_s^N) \times \\ c(F^{[1]}(\mathbf{y}_s^1),\cdots,F^{[N]}(\mathbf{y}_s^N)) \end{aligned} \qquad (11)$$

where $f^{[j]}(\mathbf{y}_s^j)$ is the probability density function corresponding to $F^{[j]}(\mathbf{y}_s^j)$ and $c = \partial C/(\partial F^{[1]},\cdots,\partial F^{[N]})$ is the copula density. For multivariate Gaussian copula C_G, the copula density is given by [Nelsen, 1998]:

$$\forall\, \mathbf{t} = (t^1,\cdots,t^N)^T \in \mathbb{R}^N \ : c_G(\mathbf{t}) = |R|^{-\frac{1}{2}} \exp\left[-\frac{\tilde{\mathbf{t}}^T(R^{-1}-I)\,\tilde{\mathbf{t}}}{2}\right] \qquad (12)$$

where $\tilde{\mathbf{t}} = (\Phi^{-1}(t^1),\cdots,\Phi^{-1}(t^N))^T$ with $\Phi(.)$ the standard Gaussian cumulative distribution, R is the $N \times N$ correlation matrix of $\tilde{\mathbf{t}}$ and I the same size identity matrix. To model non-Gaussian multivariate densities, we use Eq. 11 with a Gaussian copula density (Eq. 12) and Generalized Gaussian marginal densities [Provost *et al.*, 2003] each one characterized by three parameters namely the mean, the standard deviation and the shape parameter. This modeling allows us to cover Upper-Gaussian (shape parameter < 2), Gaussian (shape parameter $= 2$) and Sub-Gaussian (shape parameter > 2) multidimensional densities. See [Nelsen, 1998] for more details on copulas theory.

References

[Brunel *et al.*, 2005]N. Brunel, W. Pieczynski, and S. Derrode. Copulas in vectorial hidden Markov chains for multicomponent image segmentation. In *ICASSP'05*, Philadelphia, USA, March 19-23 2005.

[Crouse *et al.*, 1998]M.S. Crouse, R.D. Nowak, and R.G. Baraniuk. Wavelet-based statistical signal processing using hidden markov models. *IEEE Trans. Image Processing*, 46(4), April 1998.

[Delmas, 1997]J.-P. Delmas. An equivalence of the EM and ICE algorithm for exponential family. *IEEE-T-SP*, 45(3):2613–2615, October 1997.

[D.L.Donoho and Johnstone, 1994]D.L.Donoho and I.M. Johnstone. Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, 81:425–455, September 1994.

[Duda *et al.*, 2001]R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern Classification.* John Wiley and Sons, 2001.

[Geman and Geman, 1984]S. Geman and D. Geman. Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, PAMI-6(6):721–741, Nov. 1984.

[Giordana and Pieczynski, 1997]N. Giordana and W. Pieczynski. Estimation of generalized multisensor hidden Markov chains and unsupervised image segmentation. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 19(5):465–475, 1997.

[Graffigne *et al.*, 1995]C. Graffigne, F. Heitz, P. Pérez, F. Prêteux, M. Sigelle, and J. Zerubia. Hierarchical Markov random field models applied to image analysis : a review. In *SPIE Neural Morphological and Stochastic Methods in Image and Signal Processing*, volume 2568, pages 2–17, San Diego, 10-11 July 1995.

[Hughes, 1968]G.F. Hughes. On the mean accuracy of statistical pattern recognizers. *IEEE Trans. Information Theory*, 14(1):55–63, 1968.

[Kato *et al.*, 1996]Z. Kato, M. Berthod, and J. Zérubia. A hierarchical Markov random field model and multitemperature annealing for parallel image classification. *Graphical Models and Image Processing*, 58(1):18–37, 1996.

[Laferté *et al.*, 2000]J.-M. Laferté, P. Pérez, and F. Heitz. Discrete markov image modeling and inference on the quad-tree. *IEEE-T-IP*, 9(3):390–404, March 2000.

[Landgrebe, 2003]D. Landgrebe. *Signal Theory Methods in Multispectral Remote Sensing.* John Wiley and Sons, 2003.

[Lee *et al.*, 2000]T.W. Lee, M.S. Lewicki, and T.J. Sejnowski. ICA mixture models for unsupervised classification of non-gaussian classes and automatic context switching in blind signal separation. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 22(10):1078–1089, October 2000.

[Mignotte *et al.*, 2000]M. Mignotte, C. Collet, P. Pérez, and P. Bouthemy. Sonar image segmentation using an unsupervised hierarchical MRF model. *IEEE Trans. on Image Processing*, 9(7):1–17, July 2000.

[Nelsen, 1998]R. B. Nelsen. *An introduction to copulas.* Lecture Notes in Statistics. Springer, New York, 1998.

[Pérez *et al.*, 2000]P. Pérez, A. Chardin, and J.-M. Laferté. Noniterative manipulation of discrete energy-based models for image analysis. *Pattern Recognition*, 33(4):573–586, April 2000.

[Provost *et al.*, 2003]J.N. Provost, C. Collet, P. Rostaing, P. Pérez, and P. Bouthemy. Hierarchical Markovian segmentation of multispectral images for the reconstruction of water depth maps. *Computer Vision and Image Understanding*, 93(2):155–174, December 2003.

[Starck *et al.*, 1998]J.-L. Starck, F. Murtagh, and A. Bijaoui. *Image Processing and Data Analysis: The Multiscale Approach.* Cambridge University Press, 1998.

[Tipping and Bishop, 1999]M. E. Tipping and C. Bishop. Mixtures of probabilistic principal component analysers. *Neural Computation*, (11):443–482, 1999.

[Zhang and Blum, 1999]Z. Zhang and R. S. Blum. A categorization of multiscale-decomposition-based image fusion schemes with a performance study for a digital camera application. *Proceedings of the IEEE*, 87(8):1315 – 1326, August 1999.

Estimation for partially observed semi-Markov processes via self-consistency equations

Odile Pons

INRA, Mathématiques et informatique appliquée, 78352 Jouy-en-Josas cedex, France
(e-mail: odile.pons@jouy.inra.fr)

Abstract. Nonparametric estimators of the survival function $S(t) = P(T \geq t)$ for a censored time variable T has been defined by several methods, in particular by integral self-consistency equations since Efron (1967) [Chang and Yang, 1987]. We establish explicit expressions of the estimators in an additive form and extend this approach to several cases: a left-truncated and right-censored variable, the left-censored or left-truncated sojourn times of a right-censored semi-Markov process.

Keywords: left-truncation, right-censoring, self-consistency, semi-Markov process.

1 Introduction

Semi-Markov processes are non-homogeneous models for the evolution of individuals or systems between several states or submitted to several kinds of damage. They may be applied to data in biomedicine, biology, demography and quality control. For instance, the comparison of two treatments in patients may involve not only the final event, death or recovery, but also their evolution between several health states or their quality of life during a disease. The transition times between the states are not always observed and their values may be missing due to several possible observation scheme. In some cases the estimation of the survival function has only solved by recursive algorithms. This paper presents the usual product-limit estimator of right-censored survival function as a sum and provide closed form expressions of a survival function under left and right censoring or truncation. The estimators are extended to estimate the distribution of sojour times of a semi-Markov process under similar censorship and truncation.

2 Estimation of right-censored and left-truncated variables

2.1 Right-censored variables

Let $(X_i, \delta_i)_{i \leq n}$ be a sample of real time variables and censoring indicators, $X_i = T_i \wedge C_i$ and $\delta_i = 1\{T_i \leq C_i\}$, where T and C have the distribution functions F and G, and survival functions S and $\bar{G}$. Let $N_n(t) = \sum_i \delta_i 1_{\{X_i \leq t\}}$

and $\widehat{S}_n$ satisfying the self-consistency equation

$$\widehat{S}_n(t) = n^{-1} \sum_{i=1}^{n} \{1_{\{X_i > t\}} + (1-\delta_i) 1_{\{X_i \le t\}} \frac{\widehat{S}_n(t)}{\widehat{S}_n(X_i)}\}, \tag{1}$$

then equation (1) uniquely defines an estimator of S if the censoring distribution is continuous,

$$\widehat{S}_n(t) = 1 - \int_0^t \frac{dN_n(s)}{n - \sum_{j=1}^n (1-\delta_j) 1_{\{X_j < s\}} \widehat{S}_n^{-1}(X_j)}$$

and $\widehat{S}_n(t) \equiv \widehat{\bar{\bar{F}}}_n(t)$, the Kaplan-Meier estimator.

2.2 Numerical example

Let $(X_{(1)} < X_{(2)} < \ldots < X_{(n)})$ be the ordered sample $(X_i)_{i \le n}$ and $\delta_{(i)}$ be the indicator related to $X_{(i)}$. The estimator $\widehat{S}_n$ is as a right-continuous decreasing step function with jumps at the uncensored observations, starting from $\widehat{S}_n(0) = 1$ and with

$$\widehat{S}_n(X_{(i)}) = \widehat{S}_n(X_{(i-1)}) - \frac{\delta_{(i)}}{n - \sum_{j=1}^n (1-\delta_j) 1_{\{X_j \le X_{(i-1)}\}} \widehat{S}_n^{-1}(X_j)}.$$

Consider a sample such that $(\delta_{(i)})_{i \le n} = (1,0,1,1,0,0,0,1,1,1)$, then the sequence $(\widehat{S}_n(X_{(i-1)}, \widehat{S}_n(X_{(i)}) - \widehat{S}_n(X_{(i-1)})_{i \le n}$ takes the values

$$((1, \frac{1}{10}), (\frac{9}{10}, 0), (\frac{9}{10}, \frac{9}{8} \times \frac{1}{10}), (\frac{9}{8} \times \frac{7}{10}, \frac{9}{8} \times \frac{1}{10}), (\frac{9}{8} \times \frac{6}{10}, 0), (\frac{9}{8} \times \frac{6}{10}, 0),$$
$$(\frac{9}{8} \times \frac{6}{10}, 0), (\frac{9}{8} \times \frac{6}{10}, \frac{9}{4 \times 10}), (\frac{9}{2 \times 10}, \frac{9}{4 \times 10}), (\frac{9}{4 \times 10}, \frac{9}{4 \times 10})).$$

The product-limit estimator of Kaplan-Meier is defined as

$$\widehat{\bar{\bar{F}}}_n(t) = \prod_{X_i \le t} \Big\{1 - \frac{\delta_i}{Y_n(X_i)}\Big\}$$

with $Y_n(t) = \sum_{i=1}^n 1_{\{X_i \ge t\}}$. For the above sample

$$(1 - N_n(X_{(i)})\, Y_n^{-1}(X_{(i)}))_{i \le n} = (\frac{9}{10}, 1, \frac{7}{8}, \frac{6}{7}, 1, 1, 1, \frac{2}{3}, \frac{1}{2}, 0),$$

$\widehat{\bar{\bar{F}}}_n$ is a step function with jumps at the $X_{(i)}$'s and the values $(\widehat{\bar{\bar{F}}}_n(X_{(i)}))_{i \le n}$ are

$$(\frac{9}{10}, \frac{9}{10}, \frac{9}{10} \times \frac{7}{8}, \frac{9}{10} \times \frac{6}{8}, \frac{9}{10} \times \frac{6}{8}, \frac{9}{10} \times \frac{6}{8}, \frac{9}{10} \times \frac{6}{8}, \frac{9}{10 \times 2}, \frac{9}{10 \times 4}, 0).$$

2.3 Left-truncated and right-censored variables

Under left-truncation and right-censoring, the variables X_i and δ_i for individual i are observed only if $X_i > U_i$, where T_i, C_i, U_i are independent with dfs $1 - S$, G and H respect. Let $Y_n(t) = \sum_{i=1}^n 1_{\{U_i < t \leq X_i\}}$. As in (1), self-consistency property for the estimator of $H(t)$ and $H(t)S(t) = P(U \leq t \leq T)$ may be written

$$\widehat{H}_n(t) = n^{-1}\{Y_n(t^+) + \sum_{i=1}^n 1_{\{U_i < X_i \leq t\}} \frac{\widehat{H}_n(t)}{\widehat{H}_n(X_i)}\}$$

$$\widehat{H}_n(t)\widehat{S}_n(t) = n^{-1}\{Y_n(t^+) + \sum_{i=1}^n (1-\delta_i) 1_{\{U_i < X_i \leq t\}} \frac{\widehat{H}_n(t)\widehat{S}_n(t)}{\widehat{H}_n(X_i)\widehat{S}_n(X_i)}\}, \quad (2)$$

Let $R_U(i)$ and $R_X(i)$ be the ranks of $U(i)$ and $X(i)$. A direct estimator of $H(t)$ as a right-continuous increasing step function with jumps at the observations U_i, $i = 1, \ldots, n$, and starting from $\widehat{H}_n(0) = 0$ is defined by

$$\widehat{H}_n(U_{(i+1)}) = \widehat{H}_n(U_{(i)}) + \frac{1_{\{U_{(i+1)} < X_{R_U(i+1)}\}}}{n - \sum_{j=1}^n 1_{\{X_j < U_j \leq U_{(i)}\}} \widehat{H}_n^{-1}(U_j)}. \quad (3)$$

Moreover, $\widehat{H}_n$ given by (3) is equal to the product-limit estimator of H ([Woodroof, 1985]),

$$\widehat{H}_n^{pl}(t) = \prod_{1 \leq i \leq n} \{1 - \frac{1_{\{Y_n(X_i) > 0\}} 1_{\{U_i < t \wedge X_i\}}}{Y_n(X_i)}\}.$$

By (2) an estimator $\widehat{S}_n$ is defined as a step function with jumps at the observed X_i, with $\widehat{S}_n(0) = 1$ and such that

$$(\widehat{H}_n\widehat{S}_n)(X_{(i)}) = (\widehat{H}_n\widehat{S}_n)(X_{(i-1)}) \\ - \frac{\delta_{(i)} 1_{\{U_{R_X(i)} < X_{(i)}\}}}{n - \sum_{j=1}^n (1-\delta_j) 1_{\{U_j < X_j \leq X_{(i-1)}\}} (\widehat{H}_n\widehat{S}_n)_n^{-1}(X_j)}.$$

3 Estimation of a semi-Markov process under right-censoring

We consider a n independent observations of a semi-Markov jump process in a finite state space $\{1, \ldots, m\}$. The i^{th} sample path of the process is defined by the sequence of the different sojourn states $J_i = (J_{i,k})_{k \geq 0}$ and by the sequence of the transition times $T_i = (T_{i,k})_{k \geq 0}$, with $T_{i,0} = 0$ and $T_{i,k}$ is the arrival time in state $J_{i,k}$, up to a random time t_i. For $k \geq 1$, $i = 1, \ldots, n$, the sojourn time $X_{i,k} = T_{i,k} - T_{i,k-1}$ in a transient state $J_{i,k-1}$ may therefore be right-censored by a random variable $C_{i,k}$ and the observations are $X_{i,k} \wedge C_{i,k}$ and the indicator $\delta_{i,k} = 1_{\{X_{i,k} \leq C_{i,k}\}}$. The variable $C_{i,k}$ is supposed independent of $(T_{i,1}, \cdots, T_{i,k-1})$ and $(J_{i,0}, \cdots, J_{i,k-2})$ but to depend only on $J_{i,k-1}$, $k \geq 1$, $i = 1, \ldots, n$. Let K_i be the random number of uncensored transitions of the process (J_i, T_i), $X_i^* = t_i - \sum_{k=1}^{K_i} X_{i,k}$ the last (censored) duration time; if $J_i^* = J_{i,K_i}$ is censored, $\delta_i^* = \delta_{i,K_i+1} = 0$, otherwise $X_i^* = 0$.

The model is defined by the transition functions from j to j', $F_{j'|j}(x) = P(X_{i,k} \leq x, J_{i,k} = j' | J_{i,k-1} = j)$ or, equivalently, by the transition probabilities $p_{j'|j} = P(J_{i,k} = j' | J_{i,k-1} = j)$ and the distributions of the sojourn times between two states j and j', $F_{|jj'}(x) = P(X_{i,k} \leq x | J_{i,k} = j', J_{i,k-1} = j)$. The distribution of a sojourn time in j is $F_j = \sum_{j'} F_{j'|j}$; the related survival functions are denoted $S_{|jj'}$ and S_j, and $S_{j'|j}(x) = p_{j'|j} - F_{j'|j}(x)$. The censoring variable of the sojourn times in state j has a distribution function G_j. These functions are all assumed to be continuous.

Let $N(j,n)$ be the total number of arrivals in state j, $Y^{nc}(x,j,j',n)$ the total number of sojourn times larger than x before a transition from j to j', $Y^{nc}(x,j,n)$ (resp. $Y^c(x,j,n)$) the total number of uncensored (resp. censored) sojourn times larger than x in j and $Y(x,j,n) = Y^{nc}(x,j,n) + Y^c(x,j,n)$. As in (1), the nonparametric maximum likelihood estimator $\widehat{S}_{n,j}$ of the survival function S_j in state j may be defined as a solution of the self-consistency equation

$$\widehat{S}_{n,j}(x) = \frac{1}{N(j,n)} \left\{ Y(x^+, j, n) + \sum_{i=1}^n (1 - \delta_i^*) 1_{\{J_i^* = j\}} 1_{\{X_i^* \leq x\}} \frac{\widehat{S}_{n,j}(x)}{\widehat{S}_{n,j}(X_i^*)} \right\}, \tag{4}$$

with $\widehat{S}_{n,j}(0) = 1$ and (4) determines the Kaplan-Meier estimator of S_j.

For the estimation of $S_{j'|j}$ and $S_{|jj'}$, we assume that the mean number of visits in j, $\pi_j^0 = n^{-1} EN(j,n)$, is finite and $\pi_j^0 p_{jj'} > 0$. Estimators $\widehat{S}_{n,j'|j}$ and $\widehat{S}_{n,|jj'}$ are unique solutions of

$$\begin{aligned} \widehat{S}_{n,j'|j}(x) = \frac{1}{N(j,n)} \Bigg\{ & Y^{nc}(x^+, j, j', n) + \sum_{i=1}^n (1 - \delta_i^*) 1_{\{J_i^* = j\}} 1_{\{X_i^* > x\}} \frac{\widehat{S}_{n,j'|j}(X_i^*)}{\widehat{S}_{n,j}(X_i^*)} \\ & + \sum_{i=1}^n (1 - \delta_i^*) 1_{\{J_i^* = j\}} 1_{\{X_i^* \leq x\}} \frac{\widehat{S}_{n,j'|j}(x)}{\widehat{S}_{n,j}(X_i^*)} \Bigg\}, \end{aligned} \tag{5}$$

$$\begin{aligned} \widehat{S}_{n,|jj'}(x) = \frac{1}{N(j,n)} \Bigg\{ & \frac{Y^{nc}(x^+, j, j', n)}{\widehat{p}_{n,jj'}} + \sum_{i=1}^n (1 - \delta_i^*) 1_{\{J_i^* = j\}} 1_{\{X_i^* > x\}} \frac{\widehat{S}_{n,|jj'}(X_i^*)}{\widehat{S}_{n,j}(X_i^*)} \\ & + \sum_{i=1}^n (1 - \delta_i^*) 1_{\{J_i^* = j\}} 1_{\{X_i^* \leq x\}} \frac{\widehat{S}_{n,|jj'}(x)}{\widehat{S}_{n,j}(X_i^*)} \Bigg\} \end{aligned} \tag{6}$$

where $\widehat{p}_{n,jj'} = \widehat{S}_{n,j'|j}(0)$. The estimators $\widehat{S}_{n,j'|j}$ and $\widehat{S}_{n,|jj'}$ solutions of equations (5) and (6) are defined from $\widehat{S}_{n,j'|j}(0) = \widehat{p}_{n,jj'}$ and $\widehat{S}_{n,|jj'}(0) = 1$. They are decreasing step functions with jumps at the observed durations before a transition from j to j' and their variations depend on the number of such transitions and on the number of censored durations in state j. The censored durations in j, before and after x, are dispatched onto all the observed duration times in j before a transition to another state according to weights depending on the previously calculated values of $\widehat{S}_{n,j'|j}$ (resp. $\widehat{S}_{n,|jj'}$) and $\widehat{S}_{n,j}$.

We denote $(X_{(1)} < X_{(2)} < \ldots)$ the ordered sample $((X_{i,1}, \ldots, X_{i,K_i}), X_i^*)_{i \leq n}$ and $\delta_{(l)}$ the indicator related to $X_{(l)}$,

$$\widehat{S}_{n,j'|j}(X_{(l-1)}) - \widehat{S}_{n,j'|j}(X_{(l)}) = \frac{Y^{nc}(X_{(l-1)}, j, j', n) - Y^{nc}(X_{(l)}, j, j', n)}{N(j,n) + \int_0^{X_{(l-1)}} \widehat{S}_{n,j}^{-1}(y)\, dY^c(y^+, j, n)}.$$

(5) defines the Kaplan-Meier estimator for $S_{j'|j}$ studied by [Gill, 1980] and $\widehat{S}_{n,|jj'}(x) = \widehat{p}_{n,jj'}^{-1} \widehat{S}_{n,j'|j}(x)$.

A self-consistency equation and a direct estimator of $p(j'|x,j) = P(J_{k,i} = j'|X_{k,i} \geq x, J_{k-1,i} = j)$

$$\widehat{p}_n(j'|x,j) = Y^{-1}(x^+, j, n)\{Y^{nc}(x^+, j, j', n) - \int_{x^+}^{\infty} \widehat{p}_n(j'|y,j)\, dY^c(y,j,n)\}. \tag{7}$$

Equation (7) defines an estimator of $p(j'|x,j)$ as a decreasing step function with jumps at the censored durations in j and at the uncensored durations related to transitions from j to j'. Starting from $\widehat{p}_n(j'|\infty, j) = 0$,

$$\widehat{p}_n(j'|X_{(l-1)}, j) = \widehat{p}_n(j'|X_{(l)}, j) + \frac{1 - \delta_{(l)} \widehat{p}_n(j'|X_{(l)}, j)}{Y^c(X_{(l)}, j, n)}.$$

4 Self-consistent estimation for observations by intervals

4.1 Doubly censored observations

For individual i, the k-th sojourn time $X_{i,k}$ of the process is observed on an interval $[U_{i,k}, C_{i,k}]$ with $U_{i,k} \leq C_{i,k}$ and $\cup_{0 \leq k \leq K_i} [T_{i,k} + U_{i,k}, T_{i,k} + C_{i,k}] \subset [0, t_i]$. The observations are $J_{i,k-1}$, $W_{i,k} = \max\{U_{i,k}, \min(X_{i,k}, C_{i,k})\}$, $\delta_{1,i,k} = 1_{\{X_{i,k} > U_{i,k}\}}$ and $\delta_{2,i,k} = 1_{\{X_{i,k} < C_{i,k}\}}$, and $X_{i,k}$ is observed only if $\delta_{1,i,k}\delta_{2,i,k} = 1$. We assume that the variables $U_{i,k}$ and $C_{i,k}$ are independent of $X_{i,k}$, with continuous d.f. H_j and G_j such that $\tau_{1,j} = \inf\{u; H_j(u) > 0\} = 0$ and $\tau_{2,j} = \sup\{u; S_j(u)\bar{G}_j(u) > 0\} = \infty$. For $x \geq \tau_{1,n,j}$, the notations of section 3 are modified as

$$
\begin{aligned}
Y^{nc}(x,j,j',n) &= \sum_{i=1}^{n}\sum_{k=1}^{K_i} \delta_{1,i,k}\delta_{2,i,k} 1_{\{J_{i,k}=j'\}} 1_{\{J_{i,k-1}=j\}} 1_{\{X_{i,k}\geq x\}},\\
Y^{c,1}(x,j,n) &= \sum_{i=1}^{n}\sum_{k=1}^{K_i} (1-\delta_{1,i,k}) 1_{\{J_{i,k-1}=j\}} 1_{\{U_{i,k}\leq x\}},\\
Y^{c,2}(x,j,n) &= \sum_{i=1}^{n}\sum_{k=1}^{K_i} (1-\delta_{2,i,k}) 1_{\{J_{i,k-1}=j\}} 1_{\{C_{i,k}\geq x\}},\\
\widehat{Y}^{c,1}(x,j,j',n) &= \int_0^x \frac{1-\widehat{S}_{n,j'|j}(y)}{1-\widehat{S}_{n,j}(y)}\, dY^{c,1}(y,j,n),\\
\widehat{Y}^{c,2}(x,j,j',n) &= -\int_x^\infty \frac{\widehat{S}_{n,j'|j}(y)}{\widehat{S}_{n,j}(y)}\, dY^{c,2}(y,j,n),
\end{aligned}
$$

$\widehat{Y}(x,j,j',n) = Y^{nc}(x,j,j',n) + \widehat{Y}^{c,1}(x,j,j',n) + \widehat{Y}^{c,2}(x,j,j',n)$ and $Y(x,j,n) = \sum_{j'} Y^{nc}(x,j,j',n) + Y^{c,1}(x,j,n) + Y^{c,2}(x,j,n)$.
The self-consistency equation for the estimator $\widehat{S}_{n,j'|j}$ of $S_{j'|j}$ is written as

$$
\begin{aligned}
N(j,n)\widehat{S}_{n,j'|j}(x) = \widehat{Y}(x^+,j,j',n) - \widehat{S}_{n,j'|j}(x)\int_0^x \frac{dY^{c,2}(j,n)}{\widehat{S}_{n,j}} \\
+\{1-\widehat{S}_{n,j'|j}(x)\}\int_x^\infty \frac{dY^{c,1}(j,n)}{1-\widehat{S}_{n,j}}
\end{aligned}
\tag{8}
$$

A sum over index j' gives an equation for $\widehat{S}_{n,j}$

$$
\begin{aligned}
\widehat{S}_{n,j}(x) = \frac{1}{N(j,n)}[Y(x^+,j,n) - \widehat{S}_{n,j}(x)\int_0^x \frac{dY^{c,2}(j,n)}{\widehat{S}_{n,j}} \\
+\{1-\widehat{S}_{n,j}(x)\}\int_x^\infty \frac{dY^{c,1}(j,n)}{1-\widehat{S}_{n,j}}],
\end{aligned}
$$

with $\widehat{S}_{n,j}(0) = 1$. This equation provides an algorithm for a decreasing estimator $\widehat{S}_{n,j}$ starting from $\widehat{S}_{n,j}(0) = 1$ and with jumps at the uncensored transitions times. Let $(W_{(1)} < W_{(2)} < \ldots)$ the ordered sample of the variables $W_{i,k}$, $k = 1,\ldots,K_i$, $i = 1,\ldots,n$ and let $\delta_{(l)}$, $\delta_{1,(l)}$ and $\delta_{2,(l)}$ the indicators related to $X_{(l)}$, then

$$
\begin{aligned}
\widehat{S}_{n,j}(W_{(l)}) &= \widehat{S}_{n,j}(W_{(l-1)}) - \frac{Y^{nc}(X_{(l-1)},j,n) - Y^{nc}(X_{(l)},j,n)}{d_{n,j,(l)}}, \text{ with}\\
d_{n,j,(l)} &= N(j,n) + \int_0^{W_{(l-1)}} \widehat{S}_{n,j}^{-1}(y)\, dY^{c,2}(y^+,j,n)\\
&\quad + \int_{W_{(l)}}^\infty (1-\widehat{S}_{n,j}(y))^{-1}\, dY^{c,1}(y^+,j,n).
\end{aligned}
$$

Since $U_{i,k} \leq C_{i,k}$, boundary constraints are $\widehat{H}_{n,j}(\infty) = 1 \geq \widehat{G}_{n,j}(\infty)$, $\widehat{G}_{n,j}(0) = 0 \leq \widehat{H}_{n,j}(0)$ and (8) uniquely defines estimators of $S_{j'|j}$, H_j, $\bar{G}_j$ and $p_{jj'}$,

$$\widehat{p}_{n,jj'} = \{N(j,n) + \int_0^\infty \frac{dY^{c,1}(j,n)}{1-\widehat{S}_{n,j}}\}^{-1}\{\widehat{Y}(0,j,j',n) + \int_0^\infty \frac{dY^{c,1}(j,n)}{1-\widehat{S}_{n,j}}\},$$

and, starting from $\widehat{S}_{n,j'|j}(0) = \widehat{p}_{n,jj'}$, $\widehat{S}_{n,j'|j}$ is a decreasing step-function with jumps at the uncensored transitions times,

$$\widehat{S}_{n,j'|j}(W_{(l)}) = \widehat{S}_{n,j'|j}(W_{(l-1)}) - \frac{Y^{nc}(X_{(l-1)},j,j',n) - Y^{nc}(X_{(l)},j,j',n)}{d_{n,j,(l)}}.$$

4.2 Left-truncated and right-censored observations

The k-th transition $X_{i,k}$ of the process for an individual i is now observed on an interval $[U_{i,k}, C_{i,k}]$, conditionally on $X_{i,k} \wedge C_{i,k} > U_{i,k}$. The variables $U_{i,k}$ and $C_{i,k}$ are only supposed to be independent and independent of $X_{i,k}$ but without $U_{i,k} < C_{i,k}$ and all the observations of the states and the duration times are missing for the transitions with $X_{i,k} \wedge C_{i,k} \leq U_{i,k}$. The nonparametric estimators of the survival functions are now defined from the counting processes

$$\begin{aligned}
Y^{nc,nt}(x,j,j',n) &= \sum_{i=1}^n \sum_{k=1}^{K_i} \delta_{i,k} 1_{\{J_{i,k}=j'\}} 1_{\{J_{i,k-1}=j\}} 1_{\{U_{i,k}<x\leq X_{i,k}\}},\\
Y^{c,nt}(x,j,n) &= \sum_{i=1}^n \sum_{k=1}^{K_i} (1-\delta_{i,k}) 1_{\{J_{i,k-1}=j\}} 1_{\{U_{i,k}<x\leq C_{i,k}\}},\\
N^{c,nt}(x,j,n) &= \sum_{i=1}^n \sum_{k=1}^{K_i} (1-\delta_{i,k}) 1_{\{J_{i,k-1}=j\}} 1_{\{U_{i,k}<C_{i,k}\leq x\}},\\
\widehat{Y}^{c,nt}(x,j,j',n) &= -\int_x^\infty \frac{\widehat{S}_{n,j'|j}(y)}{\widehat{S}_{n,j}(y)}\, dY^{c,nt}(y,j,n),\\
Y_n(x,j) &= \sum_{j'} Y^{nc,nt}(x,j,j',n) + Y^{c,nt}(x,j,n).
\end{aligned}$$

Self-consistency equations may be written for $\widehat{H}_{n,j}$, $\widehat{H}_{n,j}\widehat{S}_{n,j}$ and $\widehat{H}_{n,j}\widehat{S}_{n,j'|j}$,

$$\begin{aligned}
\widehat{H}_{n,j}(x) &= n^{-1}\{Y_n(x^+,j) + \sum_{i=1}^{n} 1_{\{U_{i,k}<X_{i,k}\le x\}} 1\{J_{i,k-1}=j\}\frac{\widehat{H}_{n,j}(x)}{\widehat{H}_{n,j}(X_{i,k})}\},\\
\widehat{H}_{n,j}(x)\widehat{S}_{n,j}(x) &= \frac{1}{N(j,n)}\{Y_n(x^+,j) + \widehat{H}_{n,j}(x)\widehat{S}_{n,j}(x)\int_0^x \frac{dN^{c,nt}(j,n)}{\widehat{S}_{n,j}\widehat{H}_{n,j}}\},\\
\widehat{H}_{n,j}(x)\widehat{S}_{n,j'|j}(x) &= \frac{1}{N(j,n)}\{Y^{nc,nt}(x,j,j',n) + \widehat{Y}^{c,nt}(x,j,j',n)\\
&\qquad + \widehat{H}_{n,j}(x)\widehat{S}_{n,j'|j}(x)\int_0^x \frac{dN^{c,nt}(j,n)}{\widehat{S}_{n,j}\widehat{H}_{n,j}}\}.
\end{aligned}$$

Let $(U_{(1)} < U_{(2)} < \ldots)$ and respectively $(X_{(1)} < X_{(2)} < \ldots)$ be the ordered sample of the variables $U_{i,k}$ and $X_{i,k}$, $k = 1,\ldots,K_i$, $i = 1,\ldots,n$ and let $\delta_{(l)}$ be the indicator related to $X_{(l)}$, $R_U(l)$ and $R_X(l)$ be the ranks of $U(l)$ and $X(l)$. The nonparametric estimator of $H_j(t) = \exp\{-\int_t^\infty H_j^{-1}\,dH_j\}$ may be defined as an increasing step function with jumps at the observations $U_{i,k}$ such that $J_{i,k-1} = j$, with $\widehat{H}_{n,j}(0) = 0$ and

$$\widehat{H}_{n,j}(U_{(l+1)}) = \widehat{H}_{n,j}(U_{(l)}) + \frac{1\{U_{(l+1)} \le X_{R_U(l+1)}\}1\{J_{R_U(l+1)} = j\}}{n - \sum_{l'} 1\{X_{l'} < U_{l'} \le U_{(l)}\}\widehat{H}^{-1}_{n,j}(U_{l'})}.$$

$\widehat{S}_{n,j}$ is deduced as a step function with jumps at the $X_{i,k}$ such that $J_{i,k-1} = j$, with $\widehat{S}_{n,j}(0) = 1$ and satisfying

$$\begin{aligned}
(\widehat{H}_{n,j}\widehat{S}_{n,j})(X_{(l)}) &= (\widehat{H}_{n,j}\widehat{S}_{n,j})(X_{(l-1)})\\
&- \frac{\delta_{(l)}1\{J_{R_X(l)} = j\}1\{U_{R_X(l)} < X_{(l)}\}}{n - \sum_{l'=1}^{n}(1-\delta_{l'})1\{U_{l'} < X_{l'} \le X_{(l-1)}\}(\widehat{H}_{n,j}\widehat{S}_{n,j})^{-1}(X_{l'})}.
\end{aligned}$$

An explicit expression of $\widehat{S}_{n,j'|j}$ is similar.

All the proposed estimators are all uniformly consistent on compact sets included in the support of the survival functions.

References

[Chang and Yang, 1987]N.M. Chang and G.L. Yang. Strong consistency of a nonparametric estimator of the survival function with doubly censored data. *Ann. Statist.*, pages 1536–1547, 1987.

[Gill, 1980]R. Gill. Nonparametric estimation based on censored observations of a markov renewal process. *Z. Wahrsch. verw. Gebiete*, pages 97–116, 1980.

[Woodroof, 1985]M. Woodroof. Estimating a distribution function with truncated data. *Ann. Statist.*, pages 163–177, 1985.

On the estimation of the entropy rate of finite Markov chains

Gabriela Ciuperca[1] and Valerie Girardin[2]

[1] Université LYON I, LaPCS, 50 Av. Tony-Garnier, 69366 Lyon cedex 07, France, gabriela.ciuperca@pop.univ-lyon1.fr
[2] LMNO, UMR6139, Campus II, Université de Caen, BP5186, 14032Caen, France, girardin@math.unicaen.fr

Abstract. We consider here ergodic homogeneous Markov chain with finite state spaces. We study an estimator of the entropy rate of the chain based on the maximum likelihood estimation of the transition matrix. We prove its asymptotic properties for estimation from one sample with long length or many independent samples with given length. This result has potential applications in all the real situations modeled by Markov chains, as detailed in the introduction.

Keywords: entropy rate, homogeneous Markov Chain, maximum likelihood estimation.

1 Introduction

Markov chains and entropy are linked since the introduction of entropy in probability theory by Shannon [24]. He defined the entropy of a distribution P taking values in a finite set, say $E = \{1, \ldots, s\}$, as $\mathbb{S}(P) = -\sum_{i=1}^{s} p_i \log p_i$, with the convention $0 \ln 0 = 0$.

For a discrete-time process $\mathbf{X} = (X_n)_{n \in \mathbb{N}}$, the entropy at time n is defined as the Shannon entropy of the n-dimensional marginal distribution of $\mathbf{X}$. Under suitable conditions, the entropy at time n divided by n converges. When the limit $\mathbb{H}(\mathbf{X})$ is finite, it is called the entropy rate of the process.

The entropy rate was first defined in [24] for an ergodic Markov chain with a finite state space E as the sum of the entropies of the transition probabilities $(p_{ij})_{j=1,\ldots,s}$ weighted by the probability of occurrence of each state according to the stationary distribution π of the chain, namely

$$\mathbb{H}(\mathbf{X}) = -\sum_{i=1}^{s} \pi_i \sum_{j=1}^{s} p_{ij} \log p_{ij}. \tag{1}$$

Shannon [24] proved the convergence of $n^{-1} \log \mathbb{P}(X_1 = i_1, \ldots, X_n = i_n)$ to $\mathbb{H}(\mathbf{X})$ in probability. McMillan [16] proved the convergence in mean for any stationary ergodic process with a finite state space. This constitutes the Shannon-McMillan theorem. The almost sure convergence proven by Breiman [4] is known as the Shannon-McMillan-Breiman theorem. Many extensions have been proven since (see [10] and the references therein), but the entropy rate has an explicit form only for Markov or semi-Markov processes (see [12]).

The entropy $\mathbb{S}(P)$ of a distribution P is widely used in all applications involving random variables; see [6], [8] and the references therein. The entropy rate $\mathbb{H}(\mathbf{Y})$ of an i.i.d. sequence with distribution P is the entropy $\mathbb{S}(P)$ of P. A whole statistical tool-box has been developed in this regard and applied to a wide range of applied domains. Having an explicit form for the entropy rate of a Markov chain allows one to use it similarly in all applications involving Markov modeling. For example, maximum entropy methods can be considered (see [9]).

It is well-known that in information theory, the entropy rate of a source measures its degree of complexity (see [6]), but the entropy rate is used in many other applied fields. In time series theory, the ApEnt coefficient describes the degree of hazard in a time series, and Pincus [20] proved that for a Markovian model, the ApEnt is equal to the entropy rate of the chain. In finance, Kelly [14] introduced entropy for gambling on horse races, and Breiman [5] for investments in general markets; Shannon-McMillan-Breiman theorem appears as an ergodic theorem for the maximum growth of compounded wealth when gambling on a sequence of random variables (see [6]), and the admissible self-financing strategy achieving the maximum entropy is a growth optimal strategy (see [15]).

When observations of the process are available, the need for estimating the entropy rate obviously arises.

Approximations of entropy can be obtained by numerical algorithms. The Ziv-Lempel algorithm allows one to get an approximation of the entropy of a binary process, whichever be its distribution. Plotkin & Wyner [21] derive an algorithmic estimator of the entropy rate for a queueing problem in telecommunication networks, for measuring the scattering and clustering of cells. Abundo et al. [1] compute numerical approximations of the entropy rate via the ApEnt to explain the degree of cooperativity of proteins in a Markov model with binomial transition distributions.

Basharin [3] introduced estimation of the entropy rate in the statistical theory of random processes by considering the maximum likelihood (ML) estimator $\widehat{p}_i = n^{-1} \sum_{k=1}^{n} \mathbb{1}_{(X_k=i)}$ and the plug-in estimator $\widetilde{H} = -\sum_{i=1}^{s} \widehat{p}_i \log \widehat{p}_i$ of $\mathbb{H}(\mathbf{Y})$, for an i.i.d. sequence $\mathbf{Y} = (Y_n)$ with distribution $P = (p_1, \dots, p_s)$ on a finite state space $E = \{1, \dots, s\}$. He proved that $\widetilde{H}$ is biased but strongly consistent and asymptotically normal. Misevichyus [18] considers an estimator of the entropy rate of an homogeneous stationary Markov chain with finite state space, based on the ML estimation of the transition probabilities.

For an estimation based on one sample of long length, problems may arise from the non-observation of some states, especially if s is large. Several procedures exist in order to avoid these problems.

Meeden [17] constructs an estimator of the transition matrix by a ML method modified by a Bayes procedure. He proves that this estimator is admissible when the loss function is the sum of individual squared error losses.

Another procedure consists in the series schemes (the number of observed states, their probabilities and the transition probabilities may vary with n). The main issue of these methods is the determination of the asymptotic distribution (possibly normal, but also Poisson, centered or non-centered chi-square, etc.) of the estimators thus obtained. For an i.i.d. sequence, Zubkov [27] gives conditions on the series scheme for the asymptotic normality of $\widetilde{H}$. Mukhamedkhanova [19] studies the class of asymptotic distributions of an estimator based on the ML estimation of the transition probabilities for a two-state stationary Markov chain.

Another approach consists in using several samples of finite length in which all the states are observed infinitely often; see [2], [13, Chapter V] or [23]. Moreover, practically, it may be simpler to observe many independent trajectories of the chain with short length rather than one long trajectory.

We study here ergodic homogeneous but non necessarily stationary Markov chains with finite state spaces. We study the estimator of the entropy rate for non-stationary chains and prove its asymptotic properties for an estimation based one sample in Section 3. We generalize it to an estimation based on several samples in Section 4. Some extension prospects are given in Section 5.

2 Notation and definitions

Let (X_n) be an homogeneous ergodic (that is irreducible and aperiodic) Markov chain with finite state space $E = \{1, \dots, s\}$ and stationary distribution $(\pi_i)_{i=1,\dots,s}$. Set, for $i, j = 1, \dots, s$,

$$\begin{aligned} p_i^{(n)} &= \mathbb{P}(X_n = i), \quad n \geq 0, \\ p_{ij} &= \mathbb{P}(X_n = j | X_{n-1} = i), \quad n \geq 1, \\ p_{(i,j)}^{(n)} &= p_{ij} p_i^{(n)} = \mathbb{P}(X_n = j, X_{n-1} = i), \quad n \geq 1, \end{aligned}$$

in which p_{ij} does not depend on n due to the homogeneity of the chain. We know from the ergodic theorem of Markov chains that $p_i^{(n)}$ converges to π_i when n tends to infinity (see, e.g., [11]).

We will also consider the bidimensional Markov chain (X_n, X_{n-1}), which is homogeneous and ergodic too, with transition probabilities

$$\mathbb{P}(X_{n+1} = l, X_n = k | X_n = j, X_{n-1} = i) = p_{ij} \delta_{jk}, \tag{2}$$

(where δ_{jk} denotes Kronecker's symbol). Its stationary distribution is given by $\pi_{(i,j)} = \pi_i p_{ij}$. Indeed, since π is the stationary distribution of $\mathbf{X}$, we have

$\sum_{i'=1}^{s} \pi_{i'} p_{i'i} = \pi_i$, or

$$\sum_{i'=1}^{s} \pi_{i'} p_{ij} p_{i'i} = \pi_i p_{ij}, \quad i,j = 1,\dots,s,$$

which is equivalent to

$$\sum_{i',j'=1}^{s} \pi_{(i',j')} p_{ij} \delta_{j'i} = \pi_{(i,j)} \quad i,j = 1,\dots,s.$$

Note that $p_{(i,j)}^{(n)}$ converges to $\pi_{(i,j)}$ when n tends to infinity.

The entropy rate of the chain $\mathbf{X}$, given in (1), can be written

$$\mathbb{H}(\mathbf{X}) = \sum_{i=1}^{s} \pi_i \log \pi_i - \sum_{i=1}^{s} \sum_{j=1}^{s} \pi_{(i,j)} \log \pi_{(i,j)}, \tag{3}$$

This decomposition will be the basis of the definition of the estimators of $\mathbb{H}(\mathbf{X})$ considered in the following.

3 Estimation from one sample with long length

Suppose we are given one observation of the chain, say $X = (X_0, \dots, X_n)$. Let us set for $i,j = 1,\dots,s$,

$$\mathbf{N}_n(i,j) = \sum_{m=1}^{n} \mathbb{1}_{\{X_{m-1}=i, X_m=j\}} \quad \text{and} \quad \mathbf{N}_n(i) = \sum_{m=1}^{n} \mathbf{N}_n(i,j).$$

It is well-known (see [2, Section 5] and the references therein, and also [23]) that the following estimators of the transition probabilities (p_{ij}),

$$\widehat{p}_{ij} = \frac{\mathbf{N}_n(i,j)}{\mathbf{N}_n(i)},$$

are their ML estimators. Clearly, the stationary distribution (π_i) is estimated by

$$\widehat{\pi}_i = \frac{\mathbf{N}_n(i)}{n}, \quad i,j = 1,\dots,s,$$

Note that when $\mathbf{N}_n(i) = 0$, it is necessary to set $\widehat{p}_{ij} = 0$ for all $j = 1,\dots,s$, and $\widehat{\pi}_i = 0$. When $\mathbf{N}_n(i) \neq 0$ and $\mathbf{N}_n(i,j) = 0$, we also have $\widehat{p}_{ij} = 0$ and suppose that $p_{ij} = 0$. Note that the scheme of estimation considered below in Section 4 constitutes a means of avoiding such problems of non-observation.

The asymptotic properties given in the following proposition derive from the law of large numbers and central limit theorem for Markov chains (see, e.g., [7]).

Proposition 1 *The estimators $\widehat{p}_{ij}$ and $\widehat{\pi}_i$ are strongly consistent and asymptotically normal, in mathematical words, when n tends to infinity,*

$$\begin{aligned} &\widehat{\pi}_i \xrightarrow{a.s.} \pi_i \ \ \textit{and} \qquad \sqrt{n}(\widehat{\pi}_i - \pi_i) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \pi_i(1-\pi_i)), \\ &\widehat{p}_{ij} \xrightarrow{a.s.} p_{ij} \ \textit{and} \ \sqrt{n}\pi_i(\widehat{p}_{ij} - p_{ij}) \xrightarrow{\mathcal{L}} \mathcal{N}(0, p_{ij}(1-p_{ij})). \end{aligned}$$

Replacing in (3) the probabilities by their estimators, we get the following estimator for the entropy rate,

$$\widehat{\mathbb{H}}_n = \sum_{i=1}^{s} \widehat{\pi}_i \log \widehat{\pi}_i - \sum_{i=1}^{s} \sum_{j=1}^{s} \widehat{\pi}_{(i,j)} \log \widehat{\pi}_{(i,j)},$$

where $\widehat{\pi}_{(i,j)} = \widehat{\pi}_i \widehat{p}_{ij} = n^{-1}\mathbf{N}_n(i,j)$.

Misevichyus [18] proved the following theorem in the particular case of a stationary chain (whose initial distribution is the stationary one). We give here a shorter proof which holds true for any chain.

Theorem 1 *Let $\mathbf{X}$ be an homogeneous ergodic Markov chain with a finite state space. Then the estimator $\widehat{\mathbb{H}}_n(K)$ of $\mathbb{H}(\mathbf{X})$ is*
1. strongly consistent;
2. asymptotically normal and unbiased when n tends to infinity.

Proof of Theorem 1

1. For proving that $\widehat{\mathbb{H}}_n$ converges almost surely to H when n tends to infinity, it is sufficient to apply [26, Theorem 1.10, p59].

2. Set

$$\widehat{\mathbb{H}}_1 = \sum_{i=1}^{s} \widehat{\pi}_i \log \widehat{\pi}_i \quad \text{and} \quad \widehat{\mathbb{H}}_2 = -\sum_{i=1}^{s} \sum_{j=1}^{s} \widehat{\pi}_{(i,j)} \log \widehat{\pi}_{(i,j)}.$$

Since by Proposition 1, $\widehat{\pi}_i$ converges almost surely to π_i when n tends to infinity, the Taylor's formula for $x \log x$ at π_i, for $\pi_i \neq 0$, implies that

$$\widehat{\mathbb{H}}_1 = H_1 + \sum_{i=1}^{s} (\log \pi_i + 1)(\widehat{\pi}_i - \pi_i) - \frac{1}{2} \sum_{i=1}^{s} \frac{(\widehat{\pi}_i - \pi_i)^2}{[\pi_i + \Theta_1(\widehat{\pi}_i - \pi_i)]^3},$$

for some $0 < \Theta_1 < 1$.

Clearly, $\mathbb{E}\,[\widehat{\pi}_i - \pi_i]$ converges to zero when n tends to infinity. We get from Proposition 1 that $\mathbb{E}\,[\widehat{\pi}_i - \pi_i]^2 = \mathrm{O}(n^{-1})$. Hence $\widehat{\mathbb{H}}_1$ is asymptotically unbiased.

By Proposition 1, $\sqrt{n}(\widehat{\pi}_i - \pi_i)$ converges in distribution to $\mathcal{N}(0, \pi_i(1-\pi_i))$ when n tends to infinity, hence the delta method (see, e.g., [25]) applies to prove that $\sqrt{n}(\widehat{\mathbb{H}}_1 - H_1)$ is asymptotically centered and normal.

Since $(\pi_{(i,j)})_{i,j=1,\dots,s}$ is the stationary distribution of the bidimensional chain given in (2), the same arguments hold for H_2, and the conclusion follows. □

4 Estimation based on several independent samples with fixed length

Suppose we are given K independent observations of the chain, say $X^{(k)} = (X_0^{(k)}, \ldots, X_n^{(k)})$, $k = 1, \ldots, K$, for a fixed integer n. Let us set

$$\begin{aligned} \mathbf{n}_K(i) &= \sum_{k=1}^{K} 1\!\!1_{\{X_0^{(k)}=i\}} = \mathrm{Card}\{X_0^{(k)} = i : \ k = 1, \ldots, K\}, \\ \mathbf{N}_{n,K}(i,j) &= \sum_{k=1}^{K} \sum_{m=1}^{n} 1\!\!1_{\{X_{m-1}^{(k)}=i, X_m^{(k)}=j\}} \\ \text{and} \quad \mathbf{N}_{n,K}(i) &= \sum_{j=1}^{s} \mathbf{N}_{n,K}(i,j). \end{aligned}$$

The following ML estimators of the transition probabilities (p_{ij}),

$$\widehat{p}_{ij}(n, K) = \frac{\mathbf{N}_{n,K}(i,j)}{\mathbf{N}_{n,K}(i)}, \quad i, j = 1, \ldots, s,$$

have been computed and studied in [2].

Suppose that when K tends to infinity, $\mathbf{n}_K(i)/K$ converges to a finite quantity, say η_i, for all $i = 1, \ldots, s$ (with $\eta_i > 0$ and $\sum_{i=1}^{s} \eta_i = 1$). Then, the ML estimators $\widehat{p}_{ij}(K)$ are strongly consistent and Anderson & Goodman [2] proved that

$$\sqrt{\mathbf{N}_{n,K}(i)} \ [\widehat{p}_{ij}(K) - p_{ij}] \xrightarrow{\mathcal{L}} \mathcal{N}(0, p_{ij}(1 - p_{ij})).$$

Note that for the above result to hold, the initial distribution of the chain $\mathbf{n}_K(i)$ can be supposed to be either non-random, with multinomial distribution $\mathcal{M}(K, (\eta_i)_{i=1,\ldots,s})$ or equal to the stationary distribution of the chain.

For estimating the stationary distribution from samples with finite length, it is easy to see that it is necessary for the chain to be stationary, with then

$$\widehat{\pi}_i(K) = \frac{\mathbf{n}_K(i)}{K}, \quad i = 1, \ldots, s.$$

Proposition 2 *Suppose that the chain is stationary and that K is such that $\mathbf{n}_K(i)/K$ converges to a finite quantity, say η_i, for all $i = 1, \ldots, s$, when K tends to infinity. Then, the estimators $\widehat{\pi}_i(K)$ and $\widehat{p}_{ij}(K)$ are strongly consistent and asymptotically normal, in mathematical words,*

$$\widehat{\pi}_i \xrightarrow{a.s.} \pi_i \text{ and } \sqrt{K}(\widehat{\pi}_i - \pi_i) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \pi_i(1 - \pi_i)) \tag{4}$$

$$\widehat{p}_{ij}(K) \xrightarrow{a.s.} p_{ij}, \text{ and } \sqrt{K}\widehat{\pi}_i(\widehat{p}_{ij}(K) - p_{ij}) \xrightarrow{\mathcal{L}} \mathcal{N}(0, p_{ij}(1 - p_{ij})). \tag{5}$$

Proof of Proposition 2 Since the K samples are supposed to be independent, (4) is a straightforward consequence of the strong law of large numbers and of the central limit theorem for i.i.d. sequences. Finally, (5) is proven in [2]. □

Setting $\widehat{\pi}_{(i,j)}(K) = \widehat{\pi}_i(K)\widehat{p}_{ij}(K) = n^{-1}\mathbf{N}_{n,K}(i,j)$ and replacing in (3) the probabilities by their estimators, we get the following estimator for the entropy rate,

$$\widehat{\mathbb{H}}(K) = \sum_{i=1}^{s} \widehat{\pi}_i(K) \log \widehat{\pi}_i(K) - \sum_{i=1}^{s}\sum_{j=1}^{s} \widehat{\pi}_{(i,j)}(K) \log \widehat{\pi}_{(i,j)}(K).$$

Theorem 2 *Let* $\mathbf{X}$ *be a stationary homogeneous ergodic Markov chain with a finite state space. Suppose that* $\mathbf{n}_K(i)/K$ *converges to a finite quantity, say* η_i, *for all* $i = 1, \ldots, s$, *when* K *tends to infinity. Then the estimator* $\widehat{\mathbb{H}}_n(K)$ *of* $\mathbb{H}(\mathbf{X})$ *is*
1. strongly consistent;
2. asymptotically normal and unbiased when K *tends to infinity.*

The proof follows the same lines as the proof of Theorem 1, with n replaced by K.

5 Conclusion

The above results have potential extensions in several directions. Extensions to a countable state space or to a general Borel state space can be considered. The parametric case, that is a Markov chain whose transition matrix depends continuously on a parameter with dimension less than s, would also be of interest for many applications; see for example [21] for a Bernoulli traffic source, [1] for a Markov chain with binomial transition probabilities modeling proteins interactions, or [6] for binary information source models.

References

1.Abundo, M., Accardi, L., Rosato, N. and Stella, L., Analyzing protein energy data by a stochastic model for cooperative interactions: comparison and characterization of cooperativity, *J. Math. Bio.* V44, pp341–359 (2002).

2.Anderson, T. W. and Goodman, L. A., Statistical inference about Markov Chains. *Ann. Math. Stat.* V 28, pp89–110 (1957).

3.Basharin, G.P., On a statistical estimation for the entropy of a sequence of independent random variables.*Theory Probab. Appl.* V4, pp333–36 (1959).

4.Breiman, L., The individual ergodic theorem of information theory. *Ann. Math. Stat.* V28, pp809–11 (1957) and V31, pp809–10 (1960).

5.Breiman, L., Optimal gambling system for favorable games in *Proc. 4th Berkeley Symp. Math. Stat. Prob. Berkeley, Ca: Univ. California Press* V1 pp65–78 (1960).

6.Cover, L., and Thomas, J., *Elements of information theory.* Wiley series in telecommunications, New-York (1991).

7.Dacunha-Castelle, D., and Duflo, M., *Probabilités et Statistiques. 2. Problèmes à temps mobile.* 2e édition, Masson, Paris (1994).

8.Föllmer, H., and Schied, A., *Stochastic Finance: An Introduction in Discrete Time.* Walter de Gruyter, Berlin (2002).

9.Girardin, V., Entropy maximization for Markov and semi-Markov processes. *Method. Comp. Appl. Probab.* V6, pp109–127 (2004).

10.Girardin, V., On the Different Extensions of the Ergodic Theorem of Information Theory, in: *Recent Advances in Applied Probability.* R. Baeza-Yates, J. Glaz, H. Gzyl, J. Hüsler and J. L. Palacios (Eds), Springer-Verlag (2005).

11.Girardin, V., and Limnios, N., *Probabilités en vue des applications*, Vuibert, Paris (2001).

12.Girardin, V. and Limnios, N., Entropy rate and maximum entropy methods for countable semi-Markov chains. *Commun. in Stat. : Theory and Methods* V33, pp609–622 (2004).

13.Gouriéroux, C., *Econométrie des variables quantitatives.* Economica, Paris (1984).

14.Kelly, J. L., A new interpretation for the information rate. *Bell Syst. Tech. J.* V35 pp917–26 (1956).

15.Li, P., and Yan, J., The growth optimal portfolio in discrete-time financial markets. *Adv. Math.* V31, pp537–42 (2002).

16.Mcmillan, M., The basic theorems of information theory. *Ann. Math. Stat.* V24, pp196–219 (1953).

17.Meeden, G., The admissibility of the maximum likelihood estimator for estimating a transition matrix. *Sankhya* V51, pp37–44 (1989).

18.Misevichyus, E. V., On the statistical estimation of the entropy of an homogeneous Markov chain. (in Russian) *Liet. Mat. Rink.* V6, pp393–95 (1966).

19.Mukhamedkhanova, R., The class of limit distributions for a statistical estimator of the entropy of Markov Chains with two states. *Soviet Math. Dokl.* V29, pp155–58 (1984).

20.Pincus, S. M., Approximate entropy as a measure of system complexity. *Proc. Natl. Acad. Sci. USA* V88, pp2297–301 (1994).

21.Plotkin, N. and Wyner, A., An entropy estimator algorithm and telecommunications applications. in *Maximum Entropy and Bayesian Methods.* Heidbreder, G.R. (Ed.), Kluwer Academic Publishers, pp35–50 (1996).

22.Sadek, A., and Limnios, N., Asymptotic properties for maximum likelihood estimators for reliability and failure rates of Markov chains. *Commun. Stat., Theory Methods* V31, pp1837–61 (2002).

23.Sadek, A., *Estimation des processus markoviens avec application en fiabilité.* Thèse Univ. Techn. Compiègne (2003).

24.Shannon, C., A mathematical theory of communication. *Bell Syst., Techn. J.* V27, pp379–423, 623-656 (1948).

25.van der Vaart, A.W., and Wellner, J.A., *Weak Convergence and Empirical Processes* Springer-Verlag, New-York (1996).

26.Shao, J., *Mathematical Statistics*, Sringer-Verlag, New York (2003).

27.Zubkov, A. M., Limit distribution for a statistical estimator of the entropy. *Theor. Probab. Appl.* V18, pp611–618 (1973).

Some reward paths in semi Markov models with stochastic selection of the transition probabilities

Aleka A. Papadopoulou[1] and George M. Tsaklidis[2]

[1] General Department of Applied Sciences
Technological Institute of Thessaloniki,
P.O. Box 14561, 54101 Thessaloniki, Greece
(e-mail: alepapa@uth.gr)

[2] Department of Mathematics
Aristotle University of Thessaloniki,
Thessaloniki 54124, Greece
(e-mail: tsaklidi@math.auth.gr)

Abstract. In the present, the reward paths in non homogeneous semi Markov systems in discrete time are examined with stochastic selection of the transition probabilities. First, the basic probability equations of the reward paths are derived in terms of the main parameters of the system and a general formula is given. Then the expected rewards for the one unit time intervals are presented in relation to the entrance probabilities.

Keywords: Stochastic selection, semi-Markov process, reward.

1 Introduction

The definition of the non homogeneous semi Markov process was provided in Iosifescu-Manu (1972) for the continuous time case, in Janssen & De Dominicis (1984) for the discrete case and in De Dominicis & Manca (1985). A general definition of rewards can be found in Limnios & Oprisan (2001) and the study of the asymptotic behaviour of semi Markov reward process in Reza-Soltani & Khorshidian (1998). Later on the non homogeneous semi Markov system in discrete time was examined in Vassiliou and Papadopoulou (1992), and the asymptotic behavior of the same model was studied in Papadopoulou and Vassiliou (1994). Important theoretical results and applications for semi Markov models can be found in the work of Cinlar (1969,1975,1975), Teugels (1976), Pyke and Schaufele (1964), Keilson (1969,1971), Mclean and Neuts (1967), Howard (1971), McClean (1980,1986), Janssen (1986) and in Janssen and Limnios (1999). Continuing this effort in the present, we study the behaviour of the rewards paid during an interval of time along the reward paths. We consider rewards to be discrete random variables depending on the state occupancies, transition probabilities which are stochastically selected for every time unit, and the time spent at the state we examine before and after the time of reference. In order to examine the characteristics of the reward paths, we derive a general formula expressing the rewards per time unit for every state of the system and every time point. The expected reward for every time unit and every state can be easily evaluated by means of this general formula. Moreover the basic function used, helps us to look into the relative reward structure in the course of time. A specific type of the holding time probability functions leads to a characteristic result for the rewards evolution.

2 The semi Markov reward model with stochastic selection of the transition matrix

We consider a population which is stratified into a set of states according to various characteristics and we denote by $\mathcal{S} = \{1, 2, \ldots, N\}$ the set of states assumed to be exclusive and exhaustive, so that each member of the system may be in one and only one state at any given time. Time t is considered to be a discrete parameter and the state of the system at any given time could be described by the vector $\mathbf{N}(t) = [N_1(t), N_2(t), \ldots, N_N(t)]'$ where $N_i(t)$ is the expected number of members of the system in the i-th state at time t. The expected number of members of the system at time t is denoted by $T(t)$ and $N_{N+1}(t)$ is the expected number of leavers during the time interval $(t-1, t]$. We assume that $T(t) = T$, i.e. the total number of leavers equals to the total number of recruits for every t and that the individual transitions between the states occur according to a non homogeneous semi Markov chain *(embedded non homogeneous semi Markov chain)*. In this respect we denote by $\mathbf{F}(t)_{t=0}^{\infty}$ the sequence of matrices, the (i, j)th element of which is the probability of a member of the system to make its next transition to state j, given that it entered state i at time t. Let also $\mathbf{p}_{N+1}(t)$ be the Nx1 vector whose i-th element is the probability of leaving the system from i, given that the entrance in state i occured at time t and $\mathbf{p}_o(t)$ the Nx1 vector the j-th element of which is the probability of entering the system in state j as a replacement of a member who entered his last state at time t. A member entering the system holds a particular membership which moves within the states with the members (see also Bartholomew (1982), Vassiliou and Papadopoulou (1992), Vassiliou *et al.* (1990)). Since the size of the system is constant, when a member decides to leave the system, the empty membership is taken by a new recruit who behaves like the former one. Denote by $\mathbf{P}(t)$ the matrix described by the relation

$$\mathbf{P}(t) = \mathbf{F}(t) + \mathbf{p}_{N+1}(t)\mathbf{p}'_o(t)$$

Obviously $\mathbf{P}(t)$ is a stochastic matrix with the (i, j)th element equal to the probability that a membership of the system which entered state i at time t makes its next transition to state j. For the present, we consider that the transition probability matrix $\mathbf{P}(t)$ is selected from a pool of matrices $\mathbf{L} = [\mathbf{P}_1(t), \mathbf{P}_2(t), \ldots, \mathbf{P}_v(t)]$ with corresponding probabilities

$c_1(t), c_2(t), \ldots, c_v(t)$. Thus, whenever a membership enters state i at time t, it selects state j for its next transition according to the probabilities $p_{ij}(t)$. However before the entrance into j, the membership 'holds' for a time in state i. Holding times for the memberships are described by the holding time mass function $h_{ij}(m)$ which equals to the probability, a membership which entered state i at its last transition holds m time units in i before its next transition, given that state j has been selected.

Let also $y_{ij}(t)$ be the reward that a membership earns at time t after entering state i for occupying state i during the interval $[t, t+1)$ when its successor state is j, and $b_{ij}(m)$ be the bonus reward that the membership earns for making a transition from state i to j, after holding time m time units in state i. Thus if a membership enters state i at time s and decides to make a transition to j after m time units in i, then the total reward that it earns equals to

$$\sum_{t=s}^{s+m-1} y_{ij}(t) + b_{ij}(m).$$

Now, denote:

$A_{ij;k}(t)$ =\{the reward that a membership earns during the time interval $[t, t+1)$ given that the membership entered state i at time 0, possesses state j at time t, and will undertake its next transition to state k\}.

Moreover, entering some state j implies stay at j at least one time unit. Also, denote

$e_{ij}(n,t)$ =prob\{that a membership which entered state i at time t will enter state j after n time units\},

with corresponding probability matrix $\mathbf{E}(n,t) = \{e_{ij}(n,t)\}$. It is apparent that $e_{ij}(n,t)$ depend on the transition probabilities $p_{ij}(t)$ (see also Papadopoulou (1997)) or equivalently on the transition probability matrices $\mathbf{P}(t)$ which are selected from the pool $\mathbf{L} = [\mathbf{P}_1(t), \mathbf{P}_2(t), \ldots, \mathbf{P}_v(t)]$ with probabilities $c_1(t), c_2(t), \ldots, c_v(t)$. Thus $e_{ij}(n,t)$ become (define) random variables and we are interested in the expected value of matrix $\mathbf{E}(n,t)$. From Papadopoulou (1997) we have that

$$\mathbf{E}(n,t) = \mathbf{P}(t)\diamond\mathbf{H}(n) + \textstyle\sum_{j=2}^{n}\{\mathbf{P}(t)\diamond\mathbf{H}(j-1)\{\mathbf{P}(t+j-1)\diamond\mathbf{H}(n-j+1)\} + \\ \textstyle\sum_{j=2}^{n}\sum_{k=1}^{j-2}\mathbf{S}_j(k,s,m_k)\{\mathbf{P}(t+j-1)\diamond\mathbf{H}(n-j+1)\},$$

for every $n \geq 1$ and $\mathbf{E}(0,t) = \mathbf{I}$, where:

$\mathbf{P}(t)\diamond\mathbf{H}(n)$ is the Hadamard product of the matrices $\mathbf{P}(t)$, $\mathbf{H}(n)$,

$$\mathbf{S}_j(k,s,m_k) = \textstyle\sum_{m_k=2}^{j-k}\sum_{m_{k-1}=1+m_k}^{j-k+1}\cdots\sum_{m_1=1+m_2}^{j-1}\prod_{r=-1}^{k-1}\{\mathbf{P}(t+m_{k-r}- \\ 1)\diamond\mathbf{H}(m_{k-r-1}-m_{k-r})\}.$$

where the i, r element of $\mathbf{S}_j(k,s,m_k)$ is the probability that a membership which entered state i at time s makes a transition to state r after $j-1$ time units and k intermediate transitions during the interval $(s, s+j-1)$. Thus,

it is easily seen that
$E[\mathbf{E}(n,t)] = E[\mathbf{P}(t)]\diamond\mathbf{H}(n)+\sum_{j=2}^{n}\{E[\mathbf{P}(t)]\diamond\mathbf{H}(j-1)\}\{E[\mathbf{P}(t+j-1)]\diamond\mathbf{H}(n-j+1)\} + \sum_{j=2}^{n}\sum_{k=1}^{j-2}\tilde{\mathbf{S}}_j(k,s,m_k)\{E[\mathbf{P}(t+j-1)]\diamond\mathbf{H}(n-j+1)\}$,
where:
$E[\mathbf{P}(t)] = \sum_{x=1}^{v} c_x(t)\mathbf{P}_x(t)$,
$\tilde{\mathbf{S}}_j(k,s,m_k,\beta) = \sum_{m_k=2}^{j-k}\sum_{m_{k-1}=1+m_k}^{j-k+1}\cdots\sum_{m_1=1+m_2}^{j-1}\prod_{r=-1}^{k-1} E[\mathbf{P}(t+m_{k-r}-1)]\diamond\mathbf{H}(m_{k-r-1}-m_{k-r})$.
There are three different ways for a membership starting from state i (at time $t=0$) to occupy state j at time t. The three different ways are exclusive and exhaustive, and are illustrated below:

time t	time $t+1$
new entrance in j	new entrance in k
age in j equal to $m_p \geq 0$	residual life in j equal to $m_f > 0$
age in j equal to $m_p > 0$	new entrance in k

Thus $A_{ij;k}(t)$ is a random variable taking the values: $y_{jk}(t)+b_{jk}(1)$, $y_{jk}(t)$, $y_{jk}(t)+b_{jk}(m_p+1)$. The corresponding probabilities can be easily evaluated, for example:

$P\{A_{ij;k}(t) = y_{jk}(t)+b_{jk}(1)\}$
=prob{a membership which entered state i at time 0 will enter state j at time t}·prob{a membership which entered state j at time t will take its next transition to state k}·prob{a membership which entered state j at its last transition holds one time unit in j before its next transition given that state k has been selected}
$= e_{ij}(0,t)p_{jk}(t)h_{jk}(1)$.
Similarly we have:

$P\{A_{ij;k}(t) = y_{jk}(t)\} =$
$= \sum_{m_p}\sum_{m_f} e_{ij}(0,t-m_p)p_{jk}(t-m_p)h_{jk}(m_p+m_f+1)$
where $m_p+m_f \leq M-1$, $m_f \neq 0$, $M \in N$,
$P\{A_{ij;k}(t) = y_{jk}(t)+b_{jk}(m_p+1)\} =$
$= e_{ij}(0,t-m_p)p_{jk}(t-m_p)h_{jk}(m_p+1)$.

The three cases given above can be summarized in the following general formula:

$P\{A_{ij;k}(t) =$
$= y_{jk}(t)+\delta_{(m_f-1)}b_{jk}(m_p+m_f+1)$
$= \sum_{m_p}\sum_{m_f} E[e_{ij}(0,t-m_p)]E[p_{jk}(t-m_p)]h_{jk}(m_p+m_f+1)$
where $\delta_{(m_f-1)}$ stands for the unit impulse , i.e. $\delta_{(n)} = \begin{cases} 1, \text{ if n=0} \\ 0, \text{ if n}\neq 0 \end{cases}$.

Now, let us number by $1, 2, ..., N_i(0)$ the memberships having started their motion from state i, and denote by $A_i^{(r)}(t)$ the reward of the r-th membership paid in the time interval $[t, t+1)$ and by $A_i(t)$ the rewards paid to all the $N_i(0)$ memberships having started their motion from state i.

Let us assume that the r-th membership having started its motion from state i, possesses at time t state j, having hold for m_p time units in j and having attained the next state k after m_f time units. Then we correspond to the r-th membership a $N \times N \times M \times M$ vector (M stands for the bound of m_p, m_f) having the value

$$A_{ij;k}^{(r)}(t; m_p, m_f){=}y_{jk}(t)\ +\ \delta_{(m_f-1)}b_{jk}(m_p+m_f)$$

in the position $(i-1)NM^2{+}(j-1)M^2{+}m_pM{+}m_f$ and 0 elsewhere.Then, the total reward paid in the interval $[t, t+1)$ for the memberships having started their motion from state i, is the r.v.

$$A_i(t) = \textstyle\sum_{r=1}^{N_i(0)} A_i^{(r)}(t).$$

Symbolize by $f_i^{(r)}(t)$ the probability generating function (p.g.f.) of $A_i^{(r)}(t)$ and by $F_i(t)$ the p.g.f. of $A_i(t)$. Since the r.v's. $A_i^{(r)}(t)$ are independent with common p.g.f. $f_i^{(r)}(t) = f_i(t)$, $r = 1, 2, .., N_i(0)$, then

$$F_i(t) = \textstyle\prod_{r=1}^{N_i(0)} f_i^{(r)}(t) = (f_i(t))^{N_i(0)}.$$

3 Conclusions

In the present paper we have derived, for the discrete time semi Markov system, formulas providing the probabilities of the rewards for one unit time interval by means of the entrance probabilities, the transition probabilities and the probabilities of the holding times. Then, relations for the evaluation of the total reward paid in one unit time interval to all the memberships of the system are given. The conclusions can be generalized for various reward paths of the memberships, and a reasonable perspective is treating the same questions for the continuous time case.

References

1.Bartholomew, D.J. (1982): *Stochastic models for social processes.* Wiley, Chichester.

2.Cinlar, E (1969): Markov renewal theory. *Adv. Appl. Prob.*, *1*, 123–187.

3.Cinlar, E (1975): Markov renewal theory: a survey. *Management Sci.*, *21*, 727–752.

4.Cinlar, E (1975): *Introduction to stochastic processes.*, Prentice-Hall, Englewood Cliffs, NJ.

5.De Dominicis, R. and R. Manca (1985): Some new results on the transient behaviour of semi Markov reward processes. *Methods of operations research and Computation*, *13*, 823–838.

6.Howard, R.A. (1971): *Dynamic Probabilistic systems*. Wiley, Chichester.

7.Iosifescu - Manu, A. (1972): Non homogeneous semi Markov processes. *Studiisi Cercetuari Matematice*, *24*, 529–533.

8.Janssen, J. (1986): *Semi-Markov models: Theory and Applications*. ed. J. Janssen, Plenum Press, New York.

9.Janssen, J. and R. De Dominics (1984): Finite non homogeneous semi Markov processes: Theoretical and computational aspects. *Insurance: Mathematics and Economics*, *3*, 157–165.

10.Janssen, J. and N. Limnios (1999): *Semi-Markov models and Applications*. J. Janssen and N. Limnios Eds, Kluwer Academic Publishers, Dordrecht.

11.Keilson, J (1969): On the matrix renewal function for Markov renewal processes. *Ann. Math. Statist.*, *40*, 1901–1907.

12.Keilson, J (1971): A process with chain dependent growth rate. Markov Part II: the ruin and ergodic problems. *Adv. Appl. Prob.*, *3*, 315–338.

13.Limnios, N. and G. Oprisan (2001): *Semi-Markov processes and reliability*. Birkhauser, Boston.

14.McClean, S.I. (1980): A semi Markovian model for a multigrade population. *J. Appl. Prob.*, *17*, 846–852.

15.McClean, S.I. (1986): Semi-Markov models for Manpower planning. In *Semi-Markov models: Theory and Applications*. Plenum Press, New York.

16.Mclean, R. A. and M. F. Neuts (1967): The integral of a step function defined on a semi Markov process. *Siam J. Appl. Math.*, *15*, 726–737.

17.Papadopoulou A.A. (1997): Counting transitions -Entrance probabilities in non homogeneous semi-Markov systems. *Applied Stoch. Models and Data Analysis*, *13*, 199–206.

18.Papadopoulou, A.A. & P.-C.G. Vassiliou (1994): Asymptotic behavior of non homogeneous semi-Markov systems. *Linear Algebra and Its Applications*, *210*, 153-198.

19.Pyke, R. and R. A. Schaufele (1964): Limit theorem for Markov renewal process. *Ann. Math. Statist.*, *55*, 1746–1764.

20.Reza Soltani, A. and K. Khorshidian (1998): Reward processes for semi-Markov processes: Asymptotic behaviour. *J. Appl. Prob.*, *35*, 833–842.

21.Teugels J.L. (1976): A bibliography on semi-Markov processes. *J. Comp. Appl. Math.*, *2*, 125–144.

22.Vassiliou, P.-C.G. and A.A. Papadopoulou (1992): Non homogeneous semi-Markov systems and maintainability of the state sizes. *J. Appl. Prob.*, *29*, 519–534.

23.Vassiliou, P.-C.G. , A. Georgiou and N. Tsantas (1990): Control of asymptotic variability in non homogeneous Markov systems. *J. Appl. Prob.*, *27*, 756–766.

Non-homogeneous Markov Mixture of Periodic Autoregressions for the Analysis of Air Pollution in the Lagoon of Venice

Roberta Paroli[1], Silvia Pistollato[2], Maria Rosa[2], and Luigi Spezia[3]

[1] Istituto di Statistica
Università Cattolica S.C., Milano, Italy
(e-mail: roberta.paroli@unicatt.it)
[2] Dipartimento ARPAV Provinciale di Venezia, Mestre, Italy
(e-mail: spistollato@arpa.veneto.it, mrosa@arpa.veneto.it)
[3] Dipartimento di Scienze Economiche e Metodi Quantitativi
Università degli Studi del Piemonte Orientale, Novara, Italy
(e-mail: luigi.spezia@eco.unipmn.it)

Abstract. Markov mixtures of autoregressions (MMAR) have been recently used to analyse the behaviour of non-linear and non-Gaussian time series. A special MMAR model with periodic components and a non-homogeneous hidden Markov chain is proposed here: the transition probabilities of the hidden chain are time-varying, because they depend, through logistic functions, on the dynamics of exogenous variables. We perform a complete Metropolis-within-Gibbs algorithm associated to the random permutation sampling for model choice and variable selection and to constrained permutation sampling for the estimation of the unknown parameters and the latent data. An environmental application is developed on the series of sulphur dioxide and meteorological variables recorded by an air pollution testing station in the lagoon of Venice.

Keywords: Time-varying transition probabilities, exogenous variables, Metropolis-within-Gibbs, random and constrained permutation sampling, sulphur dioxide.

1 Introduction

Non-linear and non-normal time series can be modelled by autoregressive processes assuming that different autoregressions, each one depending on a latent regime, alternate according to the regime switching, which is driven by an unobserved Markov chain. When the chain is supposed homogeneous these models are widely known as Markov switching autoregressive models, introduced in the econometric literature by [Hamilton, 1994] to study economic and financial time series. When the Markov chain is non-homogeneous we have that the transition probabilities are time-varying and depend on exogenous variables. The class of non-homogeneous hidden Markov models depending on deterministic exogenous variables has been proposed by [Diebolt *et al.*, 1994] in the classical framework.

In this paper we propose the Bayesian analysis of Markov mixtures of autoregressions (MMAR) models with a periodic component and a non-homogeneous Markov chain defined on a general number of states, whose transition probabilities depend on deterministic exogenous variables through a logistic function. We introduce Metropolis-within-Gibbs algorithms for the estimation of the unknown parameters and for the computation of the marginal likelihood, needed for model comparison. In both cases we consider the problem of label switching, which recently has become one of the most interesting topics in the Bayesian analysis of independent and Markov-dependent mixture models. We tackle label switching through constrained permutation sampling algorithm in the case of parameter estimation and through random permutation sampling in the case of marginal likelihood computation.

In the applications these models can be efficient tools to analyse environmental time series, whose main characteristics are: *i*) different unobserved levels of pollutant mean concentrations, depending on the weather conditions, *ii*) serially correlated data, *iii*) periodicities, *iv*) missing values, *v*) availability of meteorological covariates. So we will apply our methodology to analyse a three year series of hourly mean concentrations of sulphour dioxide recorded in the lagoon of Venice.

2 The non-homogeneous Markov mixtures of periodic autoregressions

The non-homogeneous Markov mixtures of periodic autoregressionsof order $(m;p)$ (NHMMAR$(m;p)$) are discrete-time stochastic processes $\{Y_t;X_t\}$, such that $\{X_t\}$ is an unobservable non-homogeneous discrete-time Markov chain with a finite number of states, m, while $\{Y_t\}$, given $\{X_t\}$, is an observed autoregressive process of order p with a periodic component and depending on exogenous variables with the conditional distribution of Y_t depending on $\{X_t\}$ only through the contemporary X_t.

Let $\{X_t\}$ be a discrete-time, first-order, non-homogeneous Markov chain on a finite state-space S_X with cardinality m $(S_X = \{1,\dots,m\})$. For any $t = 2,\dots,T$, $\mathit{\Gamma}_t = \left[\gamma^t_{i,j}\right]$ is the $(m \times m)$ transition matrix, where $\gamma^t_{i,j} = P(X_t = j \mid X_{t-1} = i)$, for any $i,j \in S_X$; the initial distribution is the vector $\delta = (\delta_1,\dots,\delta_m)'$, where $\delta_i = P(X_1 = i)$, for any $i \in S_X$; $x^T = (x_1,\dots,x_T)'$ is the sequence of the states of the Markov chain and, for any $t = 1,\dots,T$, x_t has values in S_X. At any time $t = 2,\dots,T$, the transition probabilities $\gamma^t_{i,j}$ can be obtained as logistic functions of the vector z_t of exogenous deterministic variables, i.e.

$$\text{logit}(\gamma^t_{i,j}) = \ln\left(\gamma^t_{i,j}\,/\gamma^t_{i,i}\right) = z'_t\alpha_{i,j} \qquad \text{for any } i,j \in S_X$$

$$\gamma^t_{i,j} = \left(\exp\left(z'_t\alpha_{i,j}\right)\right) \Big/ \Big(1 + \textstyle\sum_{j\neq i}\exp\left(z'_t\alpha_{i,j}\right)\Big) \quad \text{for any } i,j \in S_X$$

where $\alpha_{i,j}$ is an n-dimensional vector of parameters, $\alpha_{i,j} = (\alpha_{i,j,0}, \alpha_{i,j,1}, \ldots, \ \alpha_{i,j,n-1})'$, if $i \neq j$, and an n-dimensional vector of zeros, if $i = j$; z_t is an n-dimensional vector, $z_t = (1, z_{t,1}, \ldots, z_{t,n-1})'$, for any $t = 2, \ldots, T$. Instead of placing the first or the last entry of the transition matrix at the denominator of the logit as usual, we place there the diagonal entry because this statement allows us to perform constrained permutation sampling and random permutation sampling algorithms, as we shall see in Sections 3. Notice that when the last $n-1$ entries of z_t are equal to zero for any t, the Markov chain is homogeneous.

Hence, given the order-p dependence and the contemporary dependence conditions, the equation describing the NHMMAR model is

$$Y_{t(i)} = \mu_i + \sum_{\tau=1}^{p} \varphi_{\tau(i)} y_{t-\tau} + \sum_{j=1}^{q} \theta_{j(i)} w_{t,j} + \beta_{t(i)} + E_{t(i)}, \tag{1}$$

where $Y_{t(i)}$ denotes the generic variable Y_t when $X_t = i$, for any $1 \leq t \leq T$ and for any $i \in S_X$; the autoregressive coefficients $\varphi_{\tau(i)}$, for any $\tau = 1, \ldots, p$ and for any $i \in S_X$, depend on the current state i of the Markov chain; $w_{t,j}$, for any $1 \leq t \leq T$, are the observations of the j-th exogenous deterministic variable, for any $j = 1, \ldots, q$, that are elements of the matrix W of dimension $(T \times q)$, weighted by the coefficients $\theta_{j(i)}$, for any $j = 1, \ldots, q$ and for any $i \in S_X$, that depend on the current state of the Markov chain. The term $\beta_{t(i)}$ is the harmonic component of periodicity $2s$, depending on the current state i of the Markov chain

$$\beta_{t(i)} = \sum_{j=1}^{s^*} \left(\beta_{1,j(i)} \cos\left(\pi j t / s\right) + \beta_{2,j(i)} \sin\left(\pi j t / s\right) \right),$$

where s^* is the number of significant harmonics $(s^* \leq s)$. $E_{t(i)}$ denotes the Gaussian random variable E_t when $X_t = i$, with zero mean and precision λ_i $\left(E_{t(i)} \mathrm{sim} \mathcal{N}\left(0; \lambda_i\right)\right)$, for any $i \in S_X$, with the discrete process $\{E_t\}$, given $\{X_t\}$, satisfying the conditional independence and the contemporary dependence conditions.

By these statements the conditional distribution of any variables $Y_{t(i)}$, given state i, is normal,

$$Y_{t(i)} \mathrm{sim} \mathcal{N} \left(\mu_i + \sum_{\tau=1}^{p} \varphi_{\tau(i)} y_{t-\tau} + \sum_{j=1}^{q} \theta_{j(i)} w_{t,j} + \beta_{t(i)}; \lambda_i \right),$$

for any $t = 1, \ldots, T$ and for any $i \in S_X$, while the marginal distribution of any variable Y_t is a mixture of m normals, whose mixing distribution is a row of the transition matrix $\mathit{\Gamma}_t$,

$$Y_t \mathrm{sim} \sum_{i=1}^{m} \gamma_{x_{t-1},i} \mathcal{N} \left(\mu_i + \sum_{\tau=1}^{p} \varphi_{\tau(i)} y_{t-\tau} + \sum_{j=1}^{q} \theta_{j(i)} w_{t,j} + \beta_{t(i)}; \lambda_i \right),$$

for any t.

A sufficient condition for the stationarity of the process (1) is that all the m sub-processes generated by the m states of the chain are stationary, that is, for any $i \in S_X$, the roots of the auxiliary equations are all inside the unit circle. To automatically satisfy the constraint on any $\varphi_i = \left(\varphi_{1(i)}, \ldots, \varphi_{p(i)}\right)'$, we can reparametrize φ_i in terms of the partial autocorrelations $r_i = \left(r_{1(i)}, \ldots, r_{p(i)}\right)'$ of any sub-process, for any $i \in S_X$, according to [Jones, 1987]. Our inference will be based on the logarithmic transformation $R_{j(i)} = \ln\left(\frac{1+r_{j(i)}}{1-r_{j(i)}}\right)$, which maps any partial autocorrelation $r_{j(i)}$ from $(-1;1)$ to $\Re$, for any $j = 1, \ldots, p$ and any $i \in S_X$.

In the framework of the mixture models the problem of identifiability concerns the invariance of the mixture under permutation of the indices of the components. In model (1) we have m states and we have $m!$ ways to label them; so different models are interchangeable by permuting their labeling. This is often called the “label switching” problem and it can be overcome by placing some identifiability constraints on some parameters with a data-driven procedure based on random permutation sampling algorithm [Frühwirth-Schnatter, 2001]. In this paper we shall introduce the random permutation sampling and the constrained permutation sampling algorithms.

Furthermore to be able to estimate the state-dependent seasonal component we need to assume the same hidden state for all the s times of any sub-period.

The unknown parameters and latent data of the NHMMAR to be estimated are: α the matrix of the vectors $\alpha_{i,j}$; μ the vector of the signals; λ the vector of the precisions; R the matrix of the coefficients $R_{j(i)}$; θ the matrix of the coefficients $\theta_{j(i)}$; β the matrix of the state-dependent harmonic coefficients; x^T the sequence of the hidden states; y^* the vector of all the missing observations. For our Bayesian inference, we place independent multivariate normal priors on each entry of matrix α; independent normal priors on each entry of vector μ; independent gamma priors on each entry of vector λ; independent multivariate normal priors of dimension p on each entry of the vector R_i; independent multivariate normal priors of dimension q on each vector θ_i; independent multivariate normal priors of dimension $2s^*$ on each vector β_i.

Let $y^T = (y_1, \ldots, y_T)'$ be the sequence of the observations; the posterior distribution of the parameter vector $\psi = (\alpha, \mu, \lambda, R, \theta, \beta, x^T, y^*)$ is

$$\pi\left(\psi \mid y^T, y^0, Z, W, V, \delta\right) = f(\alpha, \mu, \lambda, R, \theta, \beta, x^T, y^* \mid y^T, y^0, Z, W, V, \delta) \propto$$
$$\propto f\left(y^T, y^* \mid \mu, \lambda, R, \theta, \beta, W, V, x^T, y^0\right) f\left(x^T \mid \alpha, Z, \delta\right) p(\alpha)p(\mu)p(\lambda)p(R)p(\beta)p(\theta),$$

where $y^0 = (y_{-p+1}, \ldots, y_0)'$ are the initial values fixed for the p-dependence condition; Z is the matrix of dimension $(T \times n)$ of z_t, the exogenous variables of the Markov chain; V is a $(T \times 2s^*)$ matrix whose generic element on the t-th row of the j-th odd column is $\cos(\pi jt/s)$, while the generic element on the t-th row of the j-th even column is $\sin(\pi jt/s)$, for any $j = 1, 2, \ldots, s^*$.

3 Bayesian analysis

Bayesian approach to inference of mixture models is based on MCMC methods. We introduce a Metropolis-within-Gibbs procedure for model choice, variable selection and for parameter estimation.

Model choice and variable selection can be performed by means of Bayes factors in which the marginal likelihoods of the competing models are computed according to [Chib, 1995] and [Chib and Jeliazkov, 2001] corrected by the random permutation sampling algorithm [Frühwirth-Schnatter, 2001]. For model choice we need to select the unknown cardinality of the state-space of the hidden Markov chain m and the autoregressive order p, while for variables selection we require to find the best subsets of explanatory variables Z and W among all the exogenous variables to be included in the final model. To encourage the moves between the $m!$ subspaces, we can use the random permutation sampling algorithm. So at the k-th iteration of the Metropolis-within-Gibbs algorithm we use to estimate the marginal likelihood, once $\psi^{(k)}$ has been drawn, we select randomly a permutation $(\rho(1), \ldots, \rho(m))'$ of the current labeling $(1, \ldots, m)'$ and then relabel the sequence of hidden states and the switching parameters.

We can estimate the unknown parameters of NHMMAR models via a Metropolis-within-Gibbs procedure, that we briefly discuss here.

To overcome label switching the Metropolis-within-Gibbs sampler is run on a subspace only, by placing some parameters in increasing or decreasing order. The identifiability constraint is chosen ex post after simulations by a data-driven procedure, based on random permutation sampling algorithm, so as to respect the geometry and the shape of the unconstrained posterior distribution; different identifiability constraints can be derived by different data sets. By plotting the couples of the outputs of the estimates, obtained via unconstrained Metropolis-within-Gibbs algorithm, performed associated with random permutation sampling, we can check if there are as many groups as the hidden states and if these groups can suggest special ordering in their labeling. Without loss of generality, and since for our data set the constraint is based on the precisions, we discuss our methodology assuming that the entries of λ must be in decreasing order ($\lambda_i > \lambda_j$, for $i < j$, $i, j \in S_X$), but the procedures can be easily adapted to any other type of constraint. If λ is not ordered, instead of rejecting the vector and going on sampling till an ordered vector is obtained, we adopt the constrained permutation sampling algorithm [Frühwirth-Schnatter, 2001]. At any k-th iteration of the MCMC sampler, after the generation of the sequence of the hidden states, we generate the vector of the precisions; so we have m couples $\left(i, \lambda_i^{(k)}\right)$. If the $\lambda_i^{(k)}$'s are unordered, we apply a permutation $\rho(\cdot)$ to order them; consequently also the corresponding i's must be permuted according to the permutation, $\{\rho(1), \ldots, \rho(m)\}$; then the permutation is extended to the sequence of states $x^{T(k)}$ just generated, and to the switching-parameters generated in the previous iteration, $\rho\left(\mu^{(k-1)}\right)$, $\rho\left(R^{(k-1)}\right)$, $\rho\left(\theta^{(k-1)}\right)$, $\rho\left(\beta^{(k-1)}\right)$, $\rho\left(\alpha^{(k-1)}\right)$; finally all the parameters and the missing observations are generated.

The iterative scheme of the Metropolis-within-Gibbs algorithm at the k-th iteration can be summarized as follows:
1) the sequence $x^{T(k)}$ of hidden states is generated by the forward filtering-backward sampling algorithm, [Carter and Kohn, 1994] and
[Frühwirth-Schnatter, 1994];
2) the parameters $\lambda_i^{(k)}$, for any $i \in S_X$, are generated independently from gamma distributions; the entries of the vector $\lambda^{(k)}$ must be in decreasing order to satisfy the identifiability constraint. If $\lambda^{(k)}$ is not ordered, we apply the constrained permutation sampling algorithm;
3) the parameters $\mu_i^{(k)}$, for any $i \in S_X$, are generated independently from normal distributions;
4) the parameters $R_{j(i)}^{(k)}$, for any $j = 1, \dots, p$ and any $i \in S_X$, are generated independently, by a Metropolis step, from the random walk $R_{j(i)}^{(k)} = R_{j(i)}^{(k-1)} + U_R$, where U_R is a Gaussian noise with zero mean and constant precision.
5) the parameters $\theta_i^{(k)}$, for any $i \in S_X$, are independently generated from normal distributions of dimension q;
6) the parameters $\beta_i^{(k)}$, for any $i \in S_X$, are independently generated from normal distributions of dimension $2s^*$;
7) the parameters $\alpha_{i,j}^{(k)}$, for any $i, j \in S_X$, with $i \neq j$, are generated independently, by a Metropolis step, from the random walk $\alpha_{i,j}^{(k)} = \alpha_{i,j}^{(k-1)} + U_A$, where U_A is a Gaussian noise with zero mean and constant precision matrix.
8) every missing observation y_t^* is generated from the conditional normal distribution.

Now, at the end of the k-th iteration of the MCMC sampler, the vector $\psi^{(k)}$ has been approximately simulated from $\pi(\psi \mid y^T, y^0)$, if k is large enough. We shall repeat these steps till we have an N-dimensional sample. This sample will be used to estimate each entry of ψ by means of posterior means, apart from the sequence of states, estimated thought posterior modes.

4 Application to air pollution in the lagoon of Venice

Air quality control includes the study of data sets recorded by air pollution testing stations. We are interested both in the analysis of the dynamics of the hourly mean concentrations of sulphur dioxide (SO2), in micrograms per cubic meter $\left(\mu g/m^3\right)$, recorded by an air pollution testing station in the lagoon of Venice (Italy), and in investigating its relationships with the daily meteorological variables. The series of the SO2 in the log scale from the 1st of January 2001 to the 31st of December 2003 (26280 observations) is plotted in Figure 1a and it can be noticed that some observations are missing. This happens either because sometimes the station must be stopped for automatic calibration or because of occasional mechanical failure, ordinary maintenance, or data quality inspections. Plotting the histogram of the values we can guess the presence of hidden states by noticing an asymmetric distribution.

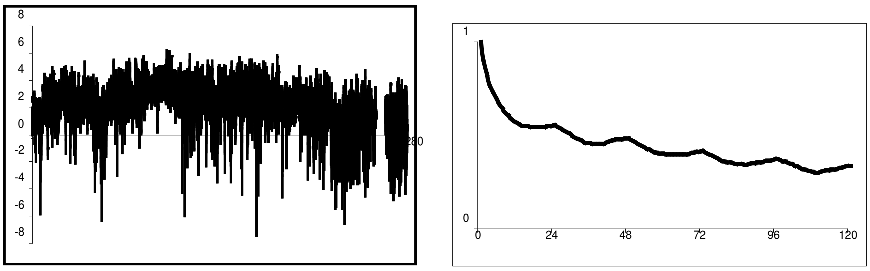


Fig. 1. (a) Series of the SO2 hourly log-concentrations; (b) 120 hours autocorrelations.

Just by looking at the series of observations we can notice a daily periodicity ($s = 24$) with 1 peak a-days ($s^* = 1$); the daily periodicity is confirmed by the correlogram of five days (Figure 1b). Atmospheric concentrations of the SO2 are influenced by many meteorological variables that are recorded together with the pollutant by the same station; we consider the following covariates: wind speed, temperature, atmospheric pressure, humidity, rainfall and solar radiation. Some of these variables will be included in the matrix W of the exogenous variables influencing the observed process and in the matrix Z of the covariates influencing the non-homogeneous Markov chain.

We develop our empirical analysis in three steps: *i*) model and variables selection, *ii*) constraint identification, *iii*) parameter estimation.

i) Model selection is performed for $m = 1, 2, 3, 4$ and $p = 0, 1, 2, 3, 4, 5, 6$ and the NHMMAR(3,1), i.e. a model with 3 hidden states and an autoregression component of order 1, is the best among all the competing models. Also variable selection is based on the values of the marginal likelihoods of all the models we analysed. The results show that temperature, humidity and wind are the variables to be included in the final model. They will be included both in the matrix W and in the matrix Z.

ii) In the second step of our analysis we have to select the identifiability constraint, which must respect the geometry and the shape of the unconstrained posterior distribution. Graphically analysing the outputs of the unconstrained NHMMAR(3;1) model, we chose the constraint on the precisions: $\lambda_1 > \lambda_2 > \lambda_3$ (Figures 2a) because the decreasing ordering is evident in the graph. Decreasing precisions is a reasonable constraint for these data, because when the low hidden state occurs, the variability of SO2 data depending on it is low and the concentrations of pollution are also low; by contrast when the high hidden state occurs, the variability of SO2 data depending on it is high and the concentrations of pollution are also high.

iii) Now we run constrained permutation sampling for the NHMMAR(3;1) model to estimate its parameters. The dynamics of the fitted values can be observed in Figure 2b: if we compare it with the dynamics of the actual data (in Figure 1a), we can see that these simulated values correctly follow the series according to the dynamics of the twenty-four hours. By this graph and by the values of the descriptive statistics we calculated to assess the fitting accuracy of the estimated model, we can argue that the fitting ability of the model is satisfactory.

Missing observations are simulated as extra latent variables; Figure 2c shows how simulated values fill the series according to the dynamics of the observed data. The dynamics of the hidden states, representing the three different levels of pollution occured during the analysed period, can be observed in Figure 2d, where we have depicted the sequence of the posterior modes of all generated states. State 3 underlies the observations with the highest level of pollution, while state 1 underlies those with the lowest level of pollution.

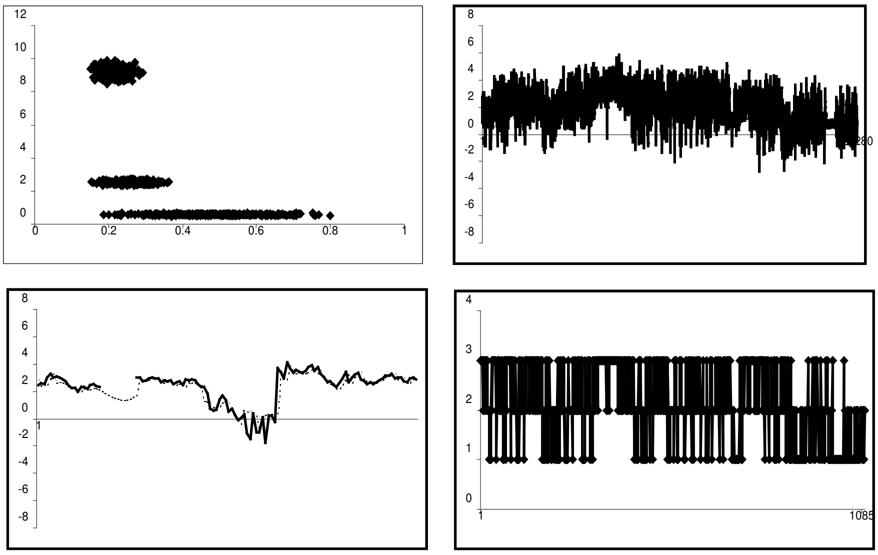


Fig. 2. (a) Couples of outputs of means and precisions of unconstrained algorithm with random permutations for m=3; (b) dynamics of the fitted values; (c) a subserie of actual (solid line) and fitted (dashes); (d) the sequence of the hidden states

5 Conclusions

We recurred to Bayesian non-homogeneous Markov mixtures of periodic autoregressions to analyse a time series about the hourly mean concentrations of sulphur dioxide, whose dynamics is characterized by cyclicity, non-normality and non-linearity. Model choice, exogenous variable selection and inference have been performed through Metropolis-within-Gibbs algorithms, considering the label switching problem, which has been efficiently tackled by permutation sampling.

References

[Carter and Kohn, 1994]C.K. Carter and R. Kohn. On gibbs sampling for state space models. *Biometrika*, pages 541–553, 1994.

[Chib and Jeliazkov, 2001]S. Chib and I. Jeliazkov. Marginal likelihoods from the metropolis-hastings output. *Journal of the American Statistical Association*, pages 270–281, 2001.

[Chib, 1995]S. Chib. Marginal likelihood from the gibbs output. *Journal of the American Statistical Association*, pages 1313–1321, 1995.

[Diebolt *et al.*, 1994]F.X. Diebolt, J.H. Lee, and G.C. Weinbach. Regime switching with time varying transition probabilities. In C.P. Hargreaves, editor, *Nonstationary Time Series Analysis and Cointegration*, pages 283–302, 1994.

[Frühwirth-Schnatter, 1994]S. Frühwirth-Schnatter. Data augmentation and dynamic linear models. *Journal of Time Series Analysis*, pages 183–202, 1994.

[Frühwirth-Schnatter, 2001]S. Frühwirth-Schnatter. Markov chain monte carlo estimation of classical and dynamic switching and mixture models. *Journal of the American Statistical Association*, pages 194–209, 2001.

[Hamilton, 1994]J.D. Hamilton. *Time Series Analysis*. Princeton University Press, Princeton, 1994.

[Jones, 1987]M.C. Jones. Randomly choosing parameters from the stationarity and invertible region of autoregressive-moving average models. *Applied Statistics*, pages 134–138, 1987.

On Occurrences of Words under Markovian Hypothesis

Ourania Chryssaphinou[1], Margarita Karaliopoulou[1], and Nikolaos Limnios[2]

[1] University of Athens
Department of Mathematics,
157 84, Athens,
Greece
(e-mail: ocrysaf@cc.uoa.gr, mkaraliop@math.uoa.gr)

[2] Universite de Technologie de Compiegne
Laboratoire de Mathematiques Appliquees Centre de Recherches de Royallieu
BP 20529,
60205 COMPIEGNE CEDEX, France
(e-mail: Nikolaos.Limnios@utc.fr)

Abstract. We consider a finite set of words $W = \{w_1, w_2, \ldots, w_\nu\}$ which are produced under the Markovian hypothesis. We study the distances between word occurrences and we give explicit formulae for the corresponding distributions in the case of having words of equal lengths.The obtained results can be applied to certain problems concerning DNA sequences, as well as, general sequential analysis.

Keywords: Word, Markov chain, distance between occurrences, semi-Markov, waiting time.

1 Preliminaries

Consider an alphabet $\Omega = \{\alpha_1, \ldots, \alpha_\ell\}$ with $\ell \geq 2$. We call word a finite sequence of elements of Ω. Let $W = \{w_1, \ldots, w_\nu\}$ a finite sets of words where $w_i = (\alpha_{i_1}, \ldots, \alpha_{i_{k_i}})$, $\alpha_{i_{n_i}} \in \Omega$, $n_i = 1, \ldots, k_i$ where k_i denotes the length of word w_i and let $k_i > 1$. We assume that the set of words is reduced. Let us consider a sequence of outcomes $\{J_n^*\}_{n\geq 1}$ generated by a Markov chain with state space Ω, and let $\mathbf{P} = (\wp(\alpha_i, \alpha_j))_{\alpha_i, \alpha_j \in \Omega}$, the transition probability matrix. We write $\mathbf{P}_l^n = (\wp^n(\alpha_i, \alpha_j))_{\alpha_i, \alpha_j \in \Omega}$, where

$$\wp^n(\alpha_i, \alpha_j) = P(J_{n+1}^* = \alpha_j | J_1^* = \alpha_i).$$

A word w_i occurs at time γ iff $J_{\gamma-k_i+1}^* = \alpha_{i_1}, \ldots, J_\gamma^* = \alpha_{i_{k_i}}$.

Definition 1 *Let W_λ a subset of W. We define*

$$U^* = \min\{\gamma \geq 1 : \ \textit{a word occurs at}\ \gamma\},$$

$$M_{W_\lambda}^* = \min\{\gamma \geq 1 : \ \textit{a word from the subset}\ W_\lambda \ \textit{occurs at}\ \gamma\},$$

and let y_0 be the first word which appears.

Clearly, the variable U^* indicates the waiting time (number of letters) for the first occurrence of any word from the set W, while the variable $M^*_{W_\lambda}$ indicates the waiting time (number of letters) for the first occurrence of any word from the subset W_λ. In section 2 we assume words of the same length and we obtain explicit formulae concerning the distributions of the above random variables. In section 3, under the same assumption we model the process of word occurrences via a semi-Markov model for which we derive the kernel as well as relative results. The under consideration random variables are of great interest for the study of biological sequences where the corresponding alphabet is $\Omega = \{A, C, G, T\}$.

Recurrent relations for variable $M^*_{\{w_i\}}$ are given by Blom and Thorburn (1982) in the i.i.d case, by Chryssaphinou and Papastavridis (1990) and Robin and Daudin (1999) in the Markov case.

2 Words of the same length and without a word at the beginning of the sequence

Let us examine the case where $k_i = k$, $\forall\, w_i \in W$. We construct a new Markov Chain $\{X^*_n\}_{n\geq 1}$ where

$$X^*_n = (J^*_n, \dots, J^*_{n+k-1}),\ n \geq 1, \tag{1}$$

with state space $\quad \Omega^k = \Omega \times \dots \times \Omega \quad$ and

$$u_i = (\alpha_1^{u_i}, \dots, \alpha_k^{u_i}),\ \forall\, i = 1, \dots, \ell^k \quad \text{and} \quad \alpha_n^{u_i} \in \Omega, \quad \forall\, n = 1, \dots, k,$$

The new transition matrix is

$$\tilde{P} = (\tilde{p}(u_i, u_j)),\ \ u_i,\ u_j \in \Omega^k, \tag{2}$$

where

$$\begin{aligned} \tilde{p}(u_i, u_j) &= \mathbb{P}(X^*_{n+1} = u_j | X^*_n = u_i) \\ &= \mathbb{P}(J^*_{n+1} = \alpha_1^{u_j}, \dots, J^*_{n+k} = \alpha_k^{u_j} \mid J^*_n = \alpha_1^{u_i}, \dots, J^*_{n+k-1} = \alpha_k^{u_i}) \\ &= I_{\{\alpha_2^{u_i} = \alpha_1^{u_j}, \dots, \alpha_k^{u_i} = \alpha_{k-1}^{u_j}\}} \wp(\alpha_{k-1}^{u_j}, \alpha_k^{u_j}). \end{aligned} \tag{3}$$

The initial distribution is

$$\begin{aligned} \mathbb{P}(X^*_1 = u_i) &= \mathbb{P}(J^*_1 = \alpha_1^{u_i}, \dots, J^*_k = \alpha_k^{u_i}) \\ &= \sigma(\alpha_1^{u_i}) \wp(\alpha_1^{u_i}, \alpha_2^{u_j}), \dots, \wp(\alpha_{k-1}^{u_i}, \alpha_k^{u_i}), \end{aligned} \tag{4}$$

where σ is the initial distribution of Markov chain J^*. We note

$$\tilde{P}_1 = (\mathbb{P}(X^*_1 = u_1), \dots, \mathbb{P}(X^*_1 = u_{\ell^k})). \tag{5}$$

Since $W \subseteq \Omega^k, \exists\, r_1, \ldots, r_\nu \in \{1, \ldots, \ell^k\} : w_1 = u_{r_1}, \ldots, w_\nu = u_{r_\nu}$. Let $B^c = \Omega^k \setminus B, \quad \forall\, B \subseteq \Omega^k$. We define the matrices

$$\tilde{P}_{B^cB^c}, \quad \tilde{P}_{B^cB}, \tag{6}$$

which are the restriction of the transition matrices $\tilde{P}$ in $B^c \times B^c$ and $B^c \times B$ respectively. Generally $\forall\, B_1, B_2 \subseteq \Omega^k$, let $\tilde{P}_{B_1B_2}$ the restriction of $\tilde{P}$ in $B_1 \times B_2$.

For $W_\lambda \subseteq W$, where $\mid W_\lambda \mid = \lambda$, we now define the nth-order transition matrix

$$\tilde{P}^n_{W^c_\lambda W^c_\lambda} = (\tilde{p}^n_{W^c_\lambda W^c_\lambda}(u_i, u_j)), \;\; u_i, \; u_j \in W^c_\lambda, \tag{7}$$

where $\tilde{p}^n_{W^c_\lambda W^c_\lambda}(u_i, u_j) = \mathbb{P}(X^*_{n+1} = u_j, X^*_n \in W^c_\lambda, \ldots, X^*_2 \in W^c_\lambda | X^*_1 = u_i)$, and $\mathbb{P}^1_{W^c_\lambda W^c_\lambda} = \mathbb{P}_{W^c_\lambda W^c_\lambda}, \quad \mathbb{P}^0_{W^c_\lambda W^c_\lambda} = \mathbb{I}_{\ell^k - \lambda}$. Finally, let us define

$$\tilde{P}_{W_\lambda} = (\mathbb{P}(X^*_1 = u_i)), \; u_i \in W_\lambda, \qquad \tilde{P}_{W^c_\lambda} = (\mathbb{P}(X^*_1 = u_i)), u_i \notin W_\lambda. \tag{8}$$

Now we are ready to present the following results.

Proposition 1 *With the above notation the distribution of the random variable $M^*_{W_\lambda}$ is given by*

$$\mathbb{P}(M^*_{W_\lambda} = n) = \begin{cases} 0, & n < k, \\ \tilde{P}_{W_\lambda} \boldsymbol{1}'_\lambda, & n = k, \\ [\tilde{P}_{W^c_\lambda}] \times [\tilde{P}^{n-k-1}_{W^c_\lambda W^c_\lambda}] \times [\tilde{P}_{W^c_\lambda W_\lambda}] \boldsymbol{1}'_\lambda, & n > k. \end{cases} \tag{9}$$

where $\boldsymbol{1}_\lambda = (1, \ldots, 1), \; (1 \times \lambda$ *matrix*)

Proof. It is $\mathbb{P}(M^*_{W_\lambda} = k) = \mathbb{P}(X^*_1 \in W_\lambda) = \tilde{P}_{W_\lambda} \mathbf{1}'_\lambda$. For $n > k$ we have

$$\begin{aligned} &\mathbb{P}(M^*_{W_\lambda} = n) = \mathbb{P}(X^*_{n-k+1} \in W_\lambda, X^*_{n-k} \in W^c_\lambda, \ldots, X^*_2 \in W^c_\lambda, X^*_1 \in W^c_\lambda) \\ &= \sum_{u_i \in W^c_\lambda} \mathbb{P}(X^*_{n-k+1} \in W_\lambda, X^*_{n-k} \in W^c_\lambda, \ldots, X^*_2 \in W^c_\lambda \mid X^*_1 = u_i) \\ &\qquad \mathbb{P}(X^*_1 = u_i) \\ &= \sum_{u_i \in W^c_\lambda} \sum_{u_j \in W_\lambda} \mathbb{P}(X^*_{n-k+1} = u_j \mid X^*_{n-k} \in W^c_\lambda, \ldots, X^*_2 \in W^c_\lambda, X^*_1 = u_i) \\ &\qquad \mathbb{P}(X^*_{n-k} \in W^c_\lambda, X^*_{n-k-1} \in W^c_\lambda \ldots, X^*_2 \in W^c_\lambda \mid X^*_1 = u_i) \mathbb{P}(X^*_1 = u_i) \\ &= \sum_{u_i \in W^c_\lambda} \sum_{u_j \in W_\lambda} \sum_{u_l \notin W_\lambda} \mathbb{P}(X^*_{n-k+1} = u_j \mid X^*_{n-k} = u_l) \\ &\qquad \mathbb{P}(X^*_{n-k} = u_l, X^*_{n-k-1} \in W^c_\lambda \ldots, X^*_2 \in W^c_\lambda \mid X^*_1 = u_i) \mathbb{P}(X^*_1 = u_i) \\ &= [\tilde{P}_{W^c_\lambda}] \times [\tilde{P}^{n-k-1}_{W^c_\lambda W^c_\lambda}] \times [\tilde{P}_{W^c_\lambda W_\lambda}] \mathbf{1}'_\lambda, \end{aligned}$$

which completes the proof.

Proposition 2 *For every $w_i \in W$ the following is valid*

$$\mathbb{P}(U^* = \gamma,\, y_0 = w_i) = \begin{cases} \tilde{P}_1 \; e'_{\ell^k;r_i}, & \gamma = k, \\ \tilde{P}_{W^c} \, [\tilde{P}_{W^c W^c}]^{\gamma-k-1} \, \tilde{P}_{W^c \Omega^k} \, e'_{\ell^k;r_i}, & \gamma > k, \end{cases} \tag{10}$$

where $e_{n;m} = (0, \ldots, \underbrace{1}_{m-position}, \ldots, 0)$ *($1 \times n$ matrix)*

Proof. The case of $\gamma = k$ is obvious, since $\mathbb{P}(U^* = k, y_0 = w_i) = \mathbb{P}(X_1^* = u_{r_i})$. For $\gamma > k$ we proceed as follows

$$\begin{aligned}
&\mathbb{P}(U^* = \gamma, y_0 = w_i) \\
&= \mathbb{P}(X^*_{\gamma-k+1} = u_{r_i}, X^*_{\gamma-k} \in W^c, \ldots, X^*_2 \in W^c, X^*_1 \in W^c) \\
&= \sum_{u_s \notin W_\lambda} \mathbb{P}(X^*_{\gamma-k+1} = u_{r_i}, X^*_{\gamma-k} \in W^c, \ldots, X^*_2 \in W^c \mid X^*_1 = u_s) \\
&\qquad \mathbb{P}(X^*_1 = u_s) \\
&= \sum_{u_s \notin W} \mathbb{P}(X^*_{\gamma-k+1} = u_{r_i} \;\mid\; X^*_{\gamma-k} \in W^c, \ldots, X^*_2 \in W^c, X^*_1 = u_s) \\
&\qquad \mathbb{P}(X^*_{\gamma-k} \in W^c, X^*_{\gamma-k-1} \in W^c \ldots, X^*_2 \in W^c \mid X^*_1 = u_s) \\
&\qquad \mathbb{P}(X^*_1 = u_s) \\
&= \sum_{u_s \notin W} \sum_{u_l \notin W} \mathbb{P}(X^*_{\gamma-k+1} = u_{r_i} \;\mid\; X^*_{\gamma-k} = u_l) \\
&\qquad \mathbb{P}(X^*_{\gamma-k} = u_l, X^*_{\gamma-k-1} \in W^c \ldots, X^*_2 \in W^c \;\mid\; X^*_1 = u_s) \\
&\qquad \mathbb{P}(X^*_1 = u_s) \\
&= \tilde{P}_{W^c} \; [\tilde{P}_{W^c W^c}]^{\gamma-k-1} \; \tilde{P}_{W^c \Omega^k} \; e'_{q^k;r_i} \,,
\end{aligned}$$

which ends the proof.

3 Words of the same length and with a word at the beginning of the sequence

We now consider $\{J_n\}$ where $J_n = J^*_{U^*+n}$, $\forall\, n \geq -U^* + 1$. We want to study the sequence $J_0, J_1, \ldots$ under the assumption that the word w_i has occurred with probability $\theta_i = \mathbb{P}(J_{-k_i+1} = \alpha_{i_1}, \ldots, J_0 = \alpha_{i_{k_i}})$, $i = 1, \ldots, \nu$. Without loss of generality we can take $\theta_i = \mathbb{P}(y_0 = w_i)$, $\quad i = 1, \ldots, \nu$.

The sequence $\{J_n, n \geq 0\}$ is a Markov chain with first order transition probabilities $\mathbb{P}(J_{n+1} = \alpha_j | J_n = \alpha_i) = \wp(\alpha_i, \alpha_j)$ and $\mathbb{P}(J_0 = \alpha_\zeta) = \sum_{i=1}^{\nu} I_{\{\alpha_{i_{k_i}} = \alpha_\zeta\}} \theta_i$.

In this case a word w_i occurs at time γ iff $J_{\gamma-k_i+1} = \alpha_{i_1}, \ldots, J_\gamma = \alpha_{i_{k_i}}$.

3.1 The Semi-Markov Model

In the case where words do not overlap, Biggins and Cannings (1987), introduced the idea of modelling the process of word occurrences via a semi-Markov model. Recently, Robin and Daudin (2001) generalized the idea considering the fact that words may overlap. Our aim is to determine the kernel of this new process. In order to present our results we need some more definitions and notations.

Definition 2 *Let us define the stochastic processes $\{U_n,\ n \geq 0\}, \{y_n,\ n \geq 0\}$, which describe the times of word occurrences and the corresponding words respectively, where $U_0 = 0$ and*

$$U_n = \min\{\gamma > U_{n-1} : a\ word\ from\, W\ occurs\ at\, \gamma\}, \quad n \geq 1, \tag{11}$$

$$y_n = w_i, \quad w_i \in W,\ i = 1, \ldots, \ell. \tag{12}$$

The process $\{(y_n, U_n), n \in \mathbb{N}\}$ is an homogenous Discrete time Markov Renewal Process(DTMRP) since

$$\begin{aligned}&\mathbb{P}(y_{n+1} = w_j, U_{n+1} - U_n = \gamma \,|\, y_0, \ldots, y_n = w_i, U_0, \ldots, U_n) = \\ &\mathbb{P}(y_{n+1} = w_j, U_{n+1} - U_n = \gamma \,|\, y_n = w_i) = \\ &\mathbb{P}(y_1 = w_j, U_1 = \gamma \,|\, y_0 = w_i) \ = \ q_{ij}(\gamma), \quad \forall n \geq 1\end{aligned}$$

Let us consider the following notation

- $\mathcal{M}_E$, the set of non negative matrices on $E \times E$.
- $\mathbf{I}_E \in \mathcal{M}_E$, the identity matrix, $0_{\overset{\circ}{A}} \in \mathcal{M}_E$, the null matrix.
- $\mathcal{M}_E(\mathbb{N})$, the set of matrix-valued functions: $\mathbb{N} \to \mathcal{M}_E$. If $A \in \mathcal{M}_E(\mathbb{N})$,we have $A = (A(\gamma) : \gamma \in \mathbb{N})$, where for fixed $\gamma \in \mathbb{N}, A(\gamma) = (A_{ij}(\gamma) : i, j \in E) \in \mathcal{M}_E$.

Then $q \in \mathcal{M}_E(\mathbb{N})$ $(E = \{1, \ldots \nu\})$ is the discrete time semi-Markov kernel relevant to the DTMRP $\{(y_n, U_n), n \in \mathbb{N}\}$. We have

$$q_{ij}^{(r)}(\gamma) = \mathbb{P}(y_r = w_j,\ U_r = \gamma \,|\, y_0 = w_i), \tag{13}$$

where $q^{(r)}$ is the r- fold convolution of q .Then we can define

$$\psi_{ij}(\gamma) = \sum_{r=0}^{\gamma} q_{ij}^{(r)}(\gamma), \quad w_i, w_j \in W,\ \gamma \in \mathbb{N}. \tag{14}$$

We can write

$$\psi_{ij}(\gamma) = q_{ij}(\gamma) + \sum_{s=1}^{\nu} \sum_{z=1}^{\gamma-1} \psi_{is}(z) q_{sj}(\gamma - z), \quad \text{for } \gamma \geq 1. \tag{15}$$

Definition 3 *For all $r \in \mathbb{N}^*, \forall w_i, w_j \in W$ let $M_{ij}^{(r)}$ be the number of letters of the r-th occurrence of w_j after w_i's occurrence.*

Definition 4 *For all $W_\lambda \subseteq W$ we define*

$$M_{iW_\lambda} = \min_{n \geq 1}\{U_n : y_n \in W_\lambda\} \text{ under the event } \{y_0 = w_i\}. \tag{16}$$

We will note M_{ij} for $M_{i\{w_j\}}$. Obviously $M_{ij} = M_{ij}^{(1)}$.

If $g_{ij}(\gamma) = \mathbb{P}(M_{ij} = \gamma)$ and $g_{ij}^{(r)}(\gamma) = \mathbb{P}(M_{ij}^{(r)} = \gamma)$, then

$$\psi_{ij}(\gamma) = \begin{cases} \sum_{r=0}^{\gamma} g_{jj}^{(r)}(\gamma), & i = j \\ \sum_{r=0}^{\gamma} g_{ij} * g_{jj}^{(r)}(\gamma), & i \neq j. \end{cases} \tag{17}$$

Definition 5 *We assume $k_i = k$ for all $w_i \in W$. We define*

$$X_n = (J_{n-k+1}, \ldots, J_n), \quad n \geq 0, \tag{18}$$

with

$$\mathbb{P}(X_0 = u_{r_i}) = \mathbb{P}(y_0 = w_i) = \mathbb{P}(J_{-k+1} = \alpha_{i_1}, \ldots, J_0 = \alpha_{i_k}),$$
$$\mathbb{P}(X_0 = u_i) = 0, \; \forall\, u_i \notin W.$$

Clearly, $\{X_n, n \geq 0\}$ is a Markov Chain with state space $\Omega^k = \{u_1, \ldots, u_{\ell^k}\}$, where $\forall\; i = 1, \ldots, \ell^k$, $u_i = (\alpha_1^{u_i}, \ldots, \alpha_k^{u_i})$, $\alpha_\zeta^{u_i} \in \Omega$, $\forall\, \zeta = 1, \ldots, k$ and $\tilde{P} = (\tilde{p}(u_i, u_j)) = (\mathbb{P}(X_{n+1} = u_j | X_n = u_i)), \forall\, u_i, u_j \in \Omega^k$.

Using the above definitions and notations we obtain the following results:

Proposition 3 *For every $w_i, w_j \in W$ we have:*

$$q_{ij}(\gamma) = \begin{cases} e_{\ell^k;r_i}\, \tilde{P}\, e'_{\ell^k;r_j}, & \gamma = 1 \\ e_{\ell^k;r_i}[\tilde{P}_{\Omega^k W^c}][\tilde{P}_{W^c W^c}]^{\gamma-2}[\tilde{P}_{W^c \Omega^k}]e'_{\ell^k;r_j}, & \gamma \geq 2 \end{cases} \tag{19}$$

Proof. It is

$$q_{ij}(1) = \mathbb{P}(X_1 = w_j | X_0 = w_i) = e_{\ell^k;r_i}\, \tilde{P}\, e'_{\ell^k;r_j}.$$

For $\gamma \geq 2$ we have

$$\begin{aligned}
&q_{ij}(\gamma) = \mathbb{P}(X_\gamma = w_j, X_1 \notin W, \ldots, X_{\gamma-1} \notin W \mid X_0 = w_i) \\
&= \mathbb{P}(X_\gamma = u_{r_j}, X_1 \notin W, \ldots, X_{\gamma-1} \notin W \mid X_0 = u_{r_i}) \\
&= \sum_{u_n, u_s \notin W} \\
&\mathbb{P}(X_\gamma = u_{r_j}, X_{\gamma-1} = u_n, X_{\gamma-2} \notin W, \ldots, X_2 \notin W, X_1 = u_s | X_0 = u_{r_i}) \\
&= \sum_{u_n, u_s \notin W} \\
&\mathbb{P}(X_\gamma = u_{r_j}, X_{\gamma-1} = u_n, X_{\gamma-2} \notin W, \ldots, X_2 \notin W, | X_1 = u_s, X_0 = u_{r_i}) \\
&\qquad \mathbb{P}(X_1 = u_s | X_0 = u_{r_i})
\end{aligned}$$

that is

$$
\begin{aligned}
&q_{ij}(\gamma)\\
&= \sum_{u_n, u_s \notin W} \mathbb{P}(X_\gamma = u_{r_j}, X_{\gamma-1} = u_n, X_{\gamma-2} \notin W, \dots, X_2 \notin W, |X_1 = u_s)\\
&\quad \mathbb{P}(X_1 = u_s | X_0 = u_{r_i})\\
&= \sum_{u_n, u_s \notin W} \mathbb{P}(X_\gamma = u_{r_j} | X_{\gamma-1} = u_n, X_{\gamma-2} \notin W, \dots, X_2 \notin W, X_1 = u_s)\\
&\ \mathbb{P}(X_{\gamma-1} = u_n, X_{\gamma-2} \notin W, \dots, X_2 \notin W, |X_1 = u_s) \mathbb{P}(X_1 = u_s | X_0 = u_{r_i})\\
&= \sum_{u_n, u_s \notin W} \mathbb{P}(X_\gamma = u_{r_j} | X_{\gamma-1} = u_n)\\
&\mathbb{P}(X_{\gamma-1} = u_n, X_{\gamma-2} \notin W, \dots, X_2 \notin W, |X_1 = u_s) \mathbb{P}(X_1 = u_s | X_0 = u_{r_i})\\
&= e_{\ell^k; r_i} [\tilde{P}_{\Omega^k W^c}][\tilde{P}_{W^c W^c}]^{\gamma-2} [\tilde{P}_{W^c \Omega^k}] e'_{\ell^k; r_j}\ .
\end{aligned}
$$

Proposition 4 *It is*

$$
\mathbb{P}(M_{iW_\lambda} = \gamma) = \begin{cases} e_{\ell^k; r_i}\ \tilde{P}_{\Omega^k W_\lambda}\ \boldsymbol{1}'_\lambda{}', & \gamma = 1 \\ e_{\ell^k; r_i}\ \tilde{P}_{\Omega^k W_\lambda^c} [\tilde{P}_{W_\lambda^c W_\lambda^c}]^{\gamma-2} \tilde{P}_{W_\lambda^c W_\lambda} \boldsymbol{1}'_\lambda, & \gamma \geq 2. \end{cases} \tag{20}
$$

Proof. The random variable M_{iW_λ} can be expressed as follows

$$
M_{iW_\lambda} = \min\{n \geq 1 : X_n \in W_\lambda\} \text{ over } \{X_0 = w_i\}. \tag{21}
$$

If $\gamma = 1$, then

$$
\mathbb{P}(M_{iW_\lambda} = 1) = \mathbb{P}(X_1 \in W_\lambda \ \mid\ X_0 = w_i) = e_{\ell^k; r_i}\ \tilde{P}_{\Omega^k W_\lambda}\ \mathbf{1}'_\lambda.
$$

If $\gamma \geq 2$, then

$$
\begin{aligned}
&\mathbb{P}(M_{iW_\lambda} = \gamma) = \mathbb{P}(X_\gamma \in W_\lambda, X_{\gamma-1} \notin W_\lambda, \dots, X_1 \notin W_\lambda \ \mid\ X_0 = w_i)\\
&= \sum_{u_r \in W_\lambda} \mathbb{P}(X_\gamma = u_r, X_{\gamma-1} \notin W_\lambda, \dots, X_1 \notin W_\lambda \ \mid\ X_0 = w_i)\\
&= \sum_{u_r \in W_\lambda} \sum_{u_s, u_n \notin W_\lambda,} \mathbb{P}(X_\gamma = u_r, X_{\gamma-1} = u_s, \dots, X_1 = u_n \mid X_0 = w_i)\\
&= \ \dots\ = \sum_{u_r \in W_\lambda} \sum_{u_s \notin W_\lambda} \sum_{u_n \notin W_\lambda} \mathbb{P}(X_\gamma = u_r \ \mid\ X_{\gamma-1} = u_s)\\
&\quad \mathbb{P}(X_{\gamma-1} = u_s, X_{\gamma-2} \notin W_\lambda, \dots, X_2 \notin W_\lambda \ \mid\ X_1 = u_n)\\
&\quad \mathbb{P}(X_1 = u_n \mid X_0 = w_i)\\
&= e_{\ell^k; r_i}\ \tilde{P}_{\Omega^k W_\lambda^c}\ [\tilde{P}_{W_\lambda^c W_\lambda^c}]^{\gamma-2} \tilde{P}_{W_\lambda^c W_\lambda}\ \mathbf{1}'_\lambda.
\end{aligned}
$$

References

[Barbu *et al.*, 2004]Barbu, V., Boussemart, M., Limnios, N.,(2004), Discrete time semi-Markov processes for reliability and survival analysis, *Communications in Statistics - Theory and Methods*, 33(11).

[Biggins and Cannings, 1987]Biggins J D. , Cannings C.,(1987),Markov renewal processes, counters and repeated sequences in markov chains. *Advanced Applied probability trust 19*, 521-545.

[Blom and Thorburn, 1982]Blom and Thorburn,(1982) How many random digits are required until given sequences are obtained, *Applied probability trust.*

[Chryssaphinou and Papastavridis, 1990]Chryssaphinou. O., Papastavridis S., (1990), The occurrence of sequence patterns in repeated dependent experiments,*Th. Probab. Appl. 35* , 145-152.

[Feller, 1968]Feller W.,(1968) An introduction to Probability Theory and its Applications, Wiley, New York.

[Guibas and Odlyzko, 1981]Guibas and Odlyzko ,(1981), String Overlaps, pattern matching and nontransitive games, *Journal of combinatorial Theory.*

[Pittenger, 1987]Pittenger, (1987) Hitting times of sequences , *Stochastic processes and their applications* 24, 225-240 North Holland.

[Robin and Daudin, 1999]Robin S., Daudin J.J., (1999), Exact distribution of word occurrences in a random sequence of letters, *Journal of applied probability* , Vol 36, 179-193.

[Robin and Daudin, 2001]Robin S., Daudin J.J., (2001) Exact distributions of the distances between any occurrence of a set of words, *Annals of the Institute of Statistical Mathematics* , Vol 36 (4) 895-905.

Markovian auto-models with mixed states

Cécile Hardouin[1] and Jian-Feng Yao[2]

[1] SAMOS - Université Paris 1
90 rue de Tolbiac,
75634 Paris Cedex 13, France
(e-mail: hardouin@univ-paris1.fr)

[2] IRMAR - Université de Rennes 1
Campus de Beaulieu,
35042 Rennes Cedex, France
(e-mail: Jian-Feng.Yao@univ-rennes1.fr)

Abstract. We present a new class of Markovian auto-models with a mixed state space $E = \{0\} \cup]0; +\infty[$ involving both discrete and continuous states. We first introduce an extension of the Besag's auto-models to the multivariate case ; then we define the specific Markovian random field defined on a lattice S, whose components are valued in E with conditional distribution belonging to an exponential family. We study two particular examples, based on the use of the exponential distribution and the Gaussian positive distribution, and look for the admissibility conditions for such models. Last, we present briefly some experimental results obtained for the analysis of motion measurements of video sequences.

Keywords: Auto-models, Mixed states.

1 Besag auto-models : multivariate extension

We consider a set of sites $S = \{1, ..n\}$, a measurable space $(E, \mathcal{E})$ (usually a subset of $\mathbb{R}^d$) equipped with the measure ν. The product space is $(\Omega, \mathcal{O}) = (E^S, \mathcal{E}^{\otimes S})$ with the product measure $\nu^S = \nu^{\otimes S}$. A random field is a probability measure μ over $(\Omega, \mathcal{O})$; we assume that μ admits a probability density f everywhere positive w.r.t ν^S.

The set of sites S is equipped with a graph structure $\mathcal{G}$, symmetrical and reflexive called the neighborhood graph; $\langle i, j \rangle$ denotes that i and j are neighbors, for $i \neq j$. A non empty set $C \subseteq S$ is a clique if C is a single point or if all elements of C are pairwise neighbors.

The field is Markovian if all the conditional distributions on the outer configurations depend on the configurations on the neighborhoods.

Let us note $\mathbf{0}$ a reference layout of Ω ($\mathbf{0}$ is 0 when $E = \mathbb{N}$, $\mathbb{R}$ or $\mathbb{R}_+$), then we can write

$$\mu(dx) = f(x)\nu^S(dx), \qquad f(x) = f(\mathbf{0}) \exp U(x)$$

with $U(\mathbf{0}) = 0$. Moreover, the energy U is the sum of potentials ϕ_A, $A \in \mathcal{C}$ the set of cliques.

According to Besag's definition ([Besag, 1974]), a real-valued field X is an auto-model if its distribution μ can be written as

$$U(x) = \sum_{i \in S} \phi_i(x_i) + \sum_{\{i,j\}} \beta_{ij} x_i x_j \ ,$$

with $\beta_{ij} = \beta_{ji}$. Thus an auto-model is a Markovian field with cliques of at most two points and linear pairwise interaction potentials.

We denote the conditional law on a site i by $f_i(x_i|.)$. The following result characterizes Besag's auto-models in the $d-$dimensional case.

Theorem 1 *We assume that for each site i, the conditional density belongs to a multi-parameter exponential family:*

$$\ln f_i(x_i|.) = \langle A_i(.), B_i(x_i) \rangle + C_i(x_i) + D_i(.) \ , \ A_i \in \mathbb{R}^d \ , \ B_i(x_i) \in \mathbb{R}^d \ . \quad (1)$$

with $B_i(0) = C_i(0) = 0$ for $0 \in E$. And that the family of sufficient statistics $\{B_i(x_i)\}$ is regular in the sense that

$$\textit{for all } i \in S, \quad Span\{B_i(x_i),\ x_i \in E\} = \mathbb{R}^d \ .$$

Then there exist for all $i, j \in S,\ i \neq j$, a family of vectors $\alpha_i \in \mathbb{R}^d$ and a family of $d \times d$ matrices satisfying $\beta_{ij} = \beta_{ji}^t$ such that

$$A_i(.) = \alpha_i + \sum_{j \neq i} \beta_{ij} B_j(x_j) \ . \quad (2)$$

Consequently the set of potentials is given by

$$\phi_i(x_i) = \langle \alpha_i, B_i(x_i) \rangle + C_i(x_i) \ , \quad (3)$$

and

$$\phi_{ij}(x_i, x_j) = \phi_{ij}(x_i, x_j) = B_i^t(x_i) \beta_{ij} B_j(x_j) \ . \quad (4)$$

See [Hardouin and Yao, 2004] for the proof.

Conversely, a Gibbs distribution with potentials (3) and (4) has conditional distributions given by (1) and (2) as soon as the energy U is admissible, i.e. $\int_\Omega \exp U(x) \nu^S(dx) < \infty$.

2 Random variable with mixed states

2.1 Distribution of mixed exponential family $\mathcal{L}(p, \xi)$:

We consider X which takes values in $E = \{0\} \cup]0, +\infty[$, equipped with the measure

$$\nu(dx) = \delta(dx) + \lambda(dx) \quad (5)$$

where δ is the Dirac measure at 0, and λ is the Lebesgue measure on $\mathcal{B}(]0, +\infty[)$.

We define the random variable X with mixed exponential family distribution on E. Let $p \in]0,1[$; then $X = 0$ with probability p, and with probability $1-p$, $X > 0$ follows a distribution which belongs to an exponential family, with the probability density :

$$g_\xi(x) = G(\xi)\ \exp\langle \xi, T(x)\rangle\ ,\ x > 0$$

where T is defined such as $T(0) = 0$. The probability density of X on E is (w.r.t. ν) :

$$\begin{aligned} f_\theta(x) &= p\delta(x) + (1-p)g_\xi(x) \\ &= p\ \exp\left\{(1-\delta(x))\ln\frac{(1-p)G(\xi)}{p} + \langle \xi, T(x)\rangle\right\} \\ &= Z^{-1}(\theta)\ \exp\langle \theta, B(x)\rangle \end{aligned}$$

where $\theta = (\theta_1, \theta_2)^t = (\ln\frac{(1-p)G(\xi)}{p}, \xi)^t$ and $B = (\delta^*, T^t)^t$ where we set $\delta^* = 1 - \delta$ in order to have $B(0) = 0$.

We denote this mixed distribution by $\mathcal{L}(p, \xi)$. Let us precise two particular cases of further use.

Mixed exponential distribution $\mathcal{E}(p, \lambda)$:

Let $g_\xi(x) = \lambda\exp\{-\lambda x\}\ ,\ x > 0$. Then

$$f(x) = p\exp\{\delta^*(x)\ln\frac{(1-p)\lambda}{p} - \lambda x\} = Z^{-1}(\theta)\ \exp\langle\theta, B(x)\rangle$$

Here $\theta = (\theta_1, \theta_2)^t = (\ln\frac{(1-p)\lambda}{p}, \lambda)^t$ and the sufficient statistics is $B(x) = (\delta^*(x), -x)^t$. Conversely we have $\lambda = \theta_2$ and $p = \frac{\theta_2}{\theta_2 + \exp\theta_1}$.

Mixed positive Gaussian distribution $G(p, \sigma^2)$

With probability $1-p$, X=$|Z|$ where Zsim$N(0, \sigma^2)$. The probability density of X is given by $f(x) = Z^{-1}(\theta)\ \exp\langle\theta, B(x)\rangle$ with $\theta = (\theta_1, \theta_2)^t = (\ln\frac{2(1-p)}{p\sigma\sqrt{2\pi}}, \frac{1}{2\sigma^2})^t$ and $B(x) = (\delta^*(x), -x^2)^t$. We get also $\sigma^2 = \frac{1}{2\theta_2}$ and $p = \frac{2}{2+\sqrt{2\pi\theta_2}\exp\theta_1}$.

3 Markovian auto-models with mixed states

We now consider a random field X on $S = \{1, 2, \cdots, n\}$, $X = (X_1, X_2, \cdots, X_n)$, in $F = E^S = (\{0\}\cup]0, +\infty[)^S$.

We assume that the family of the conditional distributions $f_i(x_i|.)$ belongs to the family of mixed distributions $\mathcal{L}(p_i(.), \xi_i(.))$ described previously. In other words, we can write (**??**) with

$$\ln f_i(x_i|.) = \mathcal{L}(p_i(.), \xi_i(.)) = \langle A_i(.), B_i(x_i)\rangle + C(x_i) + D_i(.)\ ,$$

with $B_i(x_i) = (\delta^*(x_i), T_i^t(x_i))^t$. Theorem 1 ensures that there exists vectors $\alpha_i \in \mathbb{R}^2$ and $2\times 2-$matrices β_{ij} verifying $\beta_{ij} = \beta_{ji}^t$ such that

$A_i(.) = \theta_i(.) = \alpha_i + \sum_{j\neq i}\beta_{ij}B_j(x_j)$
and the potentials of the joint energy are given by (3) and (4).

Let us specify the resulting auto-models when we take for the density g_ξ of the positive component in each site first the exponential distribution and next the positive Gaussian distribution. For each example, we give conditions ensuring the admissibility of the models; then we specify them to the four nearest neighbors system, with or without isotropy. We further use the resulting models in two different contexts: we look for a "good" suitable set of parameters of the auto-exponential model in the rainfall framework, and apply the positive Gaussian auto-model to motion measurements of video sequences.

3.1 Mixed auto-exponential models

We suppose that the conditional distributions are in the family of mixed exponential distributions $\mathcal{E}(p_i(.),\lambda_i(.))$. Then, there exist $\alpha_i = (a_i, b_i)^t$, $\beta_{ij} = \begin{pmatrix} c_{ij} & d^*_{ij} \\ d_{ij} & e_{ij} \end{pmatrix}$ verifying $c_{ij} = c_{ji}$, $e_{ij} = e_{ji}$ and $d_{ij} = d^*_{ji}$ such that we can write the global energy as:

$$U(x) = \sum_{i\in S}\alpha_i^t B(x_i) + \sum_{(i,j):\langle i,j\rangle} B^t(x_i)\beta_{ij}B(x_j) \tag{6}$$

$$\begin{aligned} U(x) = \textstyle\sum_{i\in S} a_i\delta^*(x_i) - \sum_{i\in S} b_i x_i &+ \textstyle\sum_{\langle i,j\rangle} c_{ij}\delta^*(x_i)\delta^*(x_j) \\ &- \textstyle\sum_{(i,j):\langle i,j\rangle} d_{ij}x_i\delta^*(x_j) + \sum_{\langle i,j\rangle} e_{ij}x_ix_j \end{aligned} \tag{7}$$

We note that potential $\phi(x_i, x_j) = x_i\delta(x_j)$ is not symmetric in (x_i, x_j).

Proposition 1 *We assume that U satisfies the following condition* **(A)** *:*

$$(\mathbf{A}) : \begin{cases} \forall i \in S, \forall A \subset \partial i, \quad b_i + \sum_{j\in A} d_{ij} > 0 \\ \forall i, j \in S, \quad e_{ij} \leq 0 \end{cases} \tag{8}$$

Then the energy U is admissible.

Proof: see [Hardouin and Yao, 2004].

Under condition **(A)**, the model with density defined by $f(x) = Z^{-1}\exp U(x)$ where U satisfies (6) or (7) is called mixed exponential auto-model.

Conditional distributions:
By construction, for each i, $f_i(x_i|.)$ sim $\mathcal{E}(p_i(.),\lambda_i(.))$.
$f(x_i|\,x^i) = Z^{-1}(\theta, x^i)\exp\{\theta_1(x^i)\delta^*(x_i) - \theta_2(x^i)x_i\}$, where

$$\theta_1(x^i) = a_i + \sum_{j:\langle i,j\rangle} \{c_{ij}\delta^*(x_j) - d^*_{ij}x_j\} \text{ and } \theta_2(x^i) = b_i + \sum_{j:\langle i,j\rangle} \{d_{ij}\delta^*(x_j) - e_{ij}x_j\}$$

Example 1 : Mixed exponential auto-model with the 4 nearest neighbors.

We consider $S = [1, M] \times [1, N]$, and suppose that the energy is isotropic. Then we can write the energy depending on 5 parameters $\theta = (a, b, c, d, e)$:

$$U(x) = \sum_{i\in S}(a\delta^*(x_i) - bx_i) + \sum_{\langle i,j\rangle}\{c\delta^*(x_i)\delta^*(x_j) + ex_ix_j\} - d\sum_{(i,j):\langle i,j\rangle} x_i\delta^*(x_j)$$

$(\mathbf{A})$: $b > 0$, $b + 4d > 0$ and e ≤ 0

Conditional distribution is defined by :

$$f(x_i|x^i) = Z^{-1}(\theta, x^i)\exp U_i(x_i|x^i) \text{ , where } U_i(x_i|x^i) = \theta_1(x^i)\delta^*(x_i) - \theta_2(x^i)x_i$$

$$\text{with : } \begin{cases} \theta_1(x^i) = a + c(4 - v_i(0)) - dv_i(+) \\ \theta_2(x^i) = b + d(4 - v_i(0)) - ev_i(+) \\ v_i(0) = \sum_{j:\langle i,j\rangle}\delta(x_j) \text{ and } v_i(+) = \sum_{j:\langle i,j\rangle} x_j \end{cases}$$

Particularly, $\left(X_i \mid x^i,\, X_i > 0\right) \mathrm{sim}\mathcal{E}xp(\theta_2(x^i))$ and $P\left(X_i = 0 \mid x^i\right) = \frac{\theta_2(x^i)\exp\{-\theta_1(x^i)\}}{1+\theta_2(x^i)\exp\{-\theta_1(x^i)\}}$.

Application: We now assume that the context is rainfall data. We note $x_i = 0$ when it does not rain at the site i, and $x_i > 0$ otherwise. The model should satisfy conditions such that the rain increases with $v_i(+)$, and is decreasing w.r.t $v_i(0)$, where $v_i(+)$ and $v_i(0)$ are the cumulated height of rainfall on the neighbor sites and the number of neighbor sites where it does not rain. This implies the following constraints on the parameters:

$a \in \mathbb{R}$, $c > 0$, $d \leq 0$, $b > -4d$, $e = 0$. We remark here that $e = 0$; we then propose other models involving $e \neq 0$, which induces cooperation. One solution is to consider a censored or a truncated exponential distribution on the positive component, i.e the state space is $E = \{0\}\cup]0, K]$ where K is a fix positive constant. This model is then admissible without any condition on the parameters and therefore permits to introduce cooperation between neighbor sites, via parameter $e \neq 0$. Another solution which we propose in the following example is to apply the mixed auto-model feature.

Example 2 : Double mixed exponential auto-model:

$E = \{0\}\cup]0, K[\cup\{K\}$. We are in the context of a 3-dimensional variable: let p, q $\in]0, 1[$; we set $X = 0$ with probability p, $X = K$ with probability q, and X $\in]0, K[$ with probability $1 - p - q$, according to an exponential distribution on this interval. Again, Theorem 1 ensures the model is well defined; moreover, the model is admissible and allows cooperation between neighbor sites.

3.2 Gaussian positive auto-model

We now suppose that the conditional distributions belong to the family of positive Gaussian mixed-state distribution $G(p_i(.), \sigma_i^2(.))$ given above. Then, there exist a family of vectors $\alpha_i = (a_i, b_i)^t$, and matrices $\beta_{ij} = \begin{pmatrix} c_{ij} & d^*_{ij} \\ d_{ij} & e_{ij} \end{pmatrix}$ verifying $c_{ij} = c_{ji}$, $e_{ij} = e_{ji}$ and $d_{ij} = d^*_{ji}$ such that we can write the global energy as:

$$U(x) = \sum_{i\in S} a_i \delta(x_i) - \sum_{i\in S} b_i x_i^2 + \sum_{\langle i,j\rangle} c_{ij}\delta(x_i)\delta(x_j) \\ - \sum_{(i,j):\langle i,j\rangle} d_{ij} x_i^2 \delta(x_j) + \sum_{\langle i,j\rangle} e_{ij} x_i^2 x_j^2 \tag{9}$$

Let us describe in more details the local distributions. By construction, in each site i, the conditional distribution is $G(p_i(.), \sigma_i^2(.))$ with parameters
$\theta_{i,1}(.) = a_i + \sum_{j\neq i}[c_{ij}\delta(x_j) - d^*_{ij} x_j^2]$
$\theta_{i,2}(.) = b_i + \sum_{j\neq i}[d_{ij}\delta(x_j) - e_{ij} x_j^2]$

Particularly, $\theta_{i,2}(.) = \frac{1}{2\sigma_i^2(.)}$ et $p_i(.) = \frac{2\exp\theta_{i,1}(.)}{\sqrt{\pi/\theta_{i,2}(.)}+2\exp\theta_{i,1}(.)}$. It follows that necessarily for all i and its possible neighboring configuration $(.) = (x_j,\ j \neq i)$, the variance parameter of the Gaussian component must be positive i.e. $\frac{1}{2\sigma_i^2(.)} > 0$.

Proposition 2 *We assume that U satisfies the following condition* **(B)** *:*

$$(\mathbf{B}) : \begin{cases} \forall i \in S, \forall A \subset S \setminus i,\ b_i + \sum_{j\in A} d_{ij} > 0 \\ \forall i, j \in S,\ e_{ij} \leq 0 \end{cases} \tag{10}$$

Then the energy U is admissible. Consequently, the associated positive Gaussian auto-model is well defined.

See [Bouthemy et al., 2004] for the proof.

Let us now describe the particular model using the four nearest neighbors system ; we denote here by $\{i \pm (1,0),\ i \pm (0,1)\}$ the four neighbors of i ; furthermore, we assume that the field is homogeneous in space, i.e. the parameters are the same for all sites. Moreover, we will allow possible anisotropy between the horizontal and vertical directions. Under all these considerations and by the previous results, there exist a vector $\alpha = (a, b)$ and two 2×2 matrices

$$\beta^{(k)} = \begin{pmatrix} c_k & d^*_k \\ d_k & e_k \end{pmatrix} \qquad , \quad k = 1,\, 2$$

such that $\forall i,\ \ \alpha_i = \alpha$, $\forall\{i, j\},\ \ \beta_{ij} = 0$ unless i and j are neighbors where

$$\beta_{ij} = \beta^{(1)} \text{ for } j = i \pm (1,0), \qquad \beta_{ij} = \beta^{(2)} \quad \text{for } j = i \pm (0,1)$$

We need further to set parameters d^*_1, d^*_2, e_1, e_2 to zero, since otherwise we get a repulsive field with neighbor sites in competition which is not suited to the homogeneous motion textures we intend to analyze below.The model has then 6 parameters $(a,\ b,\ c_1,\ c_2,\ d_1, d_2)$.

Now we come for an application to video sequences. Temporal textures (or dynamic textures) designate video contents involving natural (almost stationary) dynamic phenomena such as rivers, sea waves, moving foliage, etc. Mixed state auto-models allow us to specify non linear models, to take into account the spatial context and to introduce both symbolic information (no motion) and continuous motion values, which is of great interest to handle dynamic pictures; we do not model the time-varying intensity function but the motion measurements themselves.

In order to evaluate the performance of the proposed modeling, we examine if the introduced auto-models can realize two fundamental characteristics of a homogeneous texture, namely spatial isotropy and spatial stationarity. For the positive Gaussian auto-models used here, isotropy occurs if (and only if) $c_1 = c_2$ and $d_1 = d_2$. The admissibility condition given in the former result is then reduced to the unique simple condition $b > 0$.

In each experiment, we estimate the parameters by the usual pseudo-likelihood method; this method has good consistency properties for classical one-parameter auto-model and we conjecture that it is still the case for the multi-parameters auto-models considered here. The full description and discussion of the empirical results can be found in [Bouthemy et al., 2004].

The first experiment is to consider motion from trees, which is believed to be spatially isotropic, and close-up shots of a moving escalator, which is clearly anisotropic (vertical motion). In the first case, we fit both the 6-parameter (a, b, c_1, c_2, d_1, d_2) anisotropic (positive Gaussian) auto-model and the 4-parameter (a, b, c, d) isotropic one. The obtained estimates of c_1 and c_2 in one hand, and of d_1 and d_2 in another hand are almost identical, and are moreover very close to the estimated values obtained for c and d in the istropic feature. While for the moving escalator, we get significant differences between c_1 and c_2 as well as between d_1 and d_2.

The second experiment was conducted to analyze spatial sationarity. For a given texture, we divide the motion map into 12 blocks of the same size and fit an anisotropic positive gaussian auto-model to each block. This has been applied to sea-waves images and to a river motion texture. The obtained results show that the 12 sets of the estimated parameters for the sea waves texture are nearly the same, reflecting the expected spatial stationarity; while they are significantly different for the river, which confirms the assumption of non spatial stationarity for this kind of motion texture.

4 Conclusion

We have introduced a new class of random field models, namely mixed state auto-models. This approach is made possible by extending Besag's one parameter auto-models to the multi-parameter case. We provide a construction of these models and show via the given examples how useful and promising these mixed state auto-models can be; we point out for instance their performance to realize some fundamental characteristics of an homogeneous dynamic motion texture. We are currently developing other applications of these new auto-models, namely fitting pluviometric measures; there are many other possible applications in various domains, as soon as the data involves both discrete and continuous components.

There are still several questions which need further investigations; first, the convergence of the pseudo-likelihood has to be established; also, some efficient Monte Carlo simulation algorithms have to be designed for these mixed auto-models.

References

[Allcroft and Glasbey, 2003]D.J. Allcroft and C. Glasbey. A latent gaussian Markov random field model for spatio-temporal rainfall disagregation. *Applied Statistics 52*, pages 487-498, 2003.

[Arnold et al., 1999]B.C. Arnold, E. Castillo, JM Sarabia. *Conditional specification of statistical models*, Springer verlag, New York,1999.

[Besag, 1974]J. Besag. Spatial interactions and the statistical analysis of lattice systems. *JRSS B*, 148**,** pages 1-36, 1974.

[Bouthemy et al., 2004]P. Bouthemy, C. Hardouin, G. Piriou, J. Yao. Auto-models with mixed states and analysis of motion textures. *Preprint IRISA n° 1682,* 2005.

[Guyon, 1995]X. Guyon. *Random fields on a network: Modeling, Statistics and applications.* Springer-Verlag, New York, 1995.

[Guyon and Hardouin, 2002]X. Guyon, C. Hardouin. Auto modèles markoviens à états mixtes, *Annales des journées de statistique SFDS, Bruxelles, 2002.*

[Hardouin and Yao, 2004]C. Hardouin and J. Yao. Markovian auto-models with mixed states. *Technical report, IRMAR, Université de Rennes 1*, 2004.

[Whittle, 1963]P. Whittle. Stochastic processes in several dimensions. *Bull. Inst. Statist. Inst. 40,* pages 974-994, 1963.

A new detection scheme: the sequential Markov detector

Didier Billon

Thales Underwater Systems
Brest, France
(e-mail: didier.billon@fr.thalesgroup.com)

1 Introduction

Automatic detection and tracking (ADT), as performed in radar and sonar, process data $y(t, \omega)$ depending on time variable t and observation variable ω, frequency or direction for instance. ADT decides whether signal from an object to be detected is present or not at any time t and any location ω: this is detection. Once such a positive decision has been taken at some time t_0, ADT has to estimate the location $\omega(t)$ of the signal at further times $t \geq t_0$: this is tracking. The detection step is also named track initiation. Track initiation and tracking are both time association processing. But, because tracking performs only on the data in the vicinity of the tracks, while track initiation has to perform on the whole data domain, tracking may use more computationally intensive algorithms, especially algorithms based on a state model that maps a space of states $\{x\}$ for the detected object to the data domain $\{\omega\}$. Nevertheless this situation may be paradoxical with respect to the fact that deciding that some object is present is at least as important in some applications as once this decision is taken, estimating the state of the object along time. Indeed the detection performance, expressed in terms of detection probability and false alarm probability, is fully achieved by the track initiation step. The continuous increase of the real time computation power let us to envisage the application of the same kind of principle for track initiation like for tracking. In this paper we propose such a new algorithm, a sequential detector based on a hidden Markov model (HMM), that we named the sequential Markov detector (SMD).

The track initiation in most existing radars and sonars rely on a same principle: the P out of N detection. It performs on events detected from single data elements that exceed a detection threshold r_1. The integer N is the duration of the detection test window along the discrete time axis. A signal is detected when there are events at P times at least within the window. This criterion may be refined with the supplementary condition that the mean value of the data corresponding to the highest P events shall be larger than a second detection threshold $r_2 > r_1$. The false alarm probability depends on r_1 and r_2 and on the extent $\delta\omega$ of the test window in the data domain. Increasing $\delta\omega$ results in increasing both the false alarm probability and the capability for accommodating a signal drift along time in the observation domain. In practice, $\delta\omega$ is generally set up at most equal to the resolution of the sensor and the time duration N is set up small enough so that the signal drift cannot exceed this extent. Then N may have to be limited to a few units, a constraint that prevents from taking the full benefit for detection from long signal duration. Whatever be the P out of N variant used in practice, N is most often smaller than 10. This may be also a limit related to the duration of the shortest signal to be detected.

In the "track-before-detect" (TBD) approach for ADT, tentative tracks are formed before being validated. Because the data are integrated on a longer time duration according to some model for the dynamics of the object to be detected, this approach is better suited to low signal-to-noise ratio condition. This processing is most often done on data blocks of fixed duration. In [Tonissen and Evans, 1996], the data are integrated along candidate paths by means of a dynamic programming algorithm. In [Barrett and Holdsworth, 1993], HMM is used for likelihood ratio testing. Sequential detection, which allows for taking a decision about presence or absence of a signal embedded in noise, from a variable number of data frames, is proposed for a constant velocity target model in [Blostein and Richardson, 1994], but the test is truncated and the data still are structured in blocks of fixed duration.

We introduce in this paper a new track initiation scheme combining HMM and sequential detection, named the sequential Markov detector (SMD). HMM allows for testing any path $(x(t))_{t_0 \leq t \leq t_0+\Delta t}$ in the state space from an exact expression of the joint likelihood ratio of this path and the data sequence $(y(t, \omega_x(t)))_{t_0 \leq t \leq t_0+\Delta t}$ along the corresponding path $(\omega(t))_{t_0 \leq t \leq t_0+\Delta t}$ in the data domain. Like the sequential probability ratio test (SPRT) [Marano *et al.*, 2005], SMD does not require fixing Δt. But there is no fixed fail threshold in SMD, which involves a factor exponentially decreasing as a function of Δt controlling an automatic reset process.

We review the principle of HMM detection in part 2 and introduce the sequential Markov detector in part 3. Application of SMD for detection of spectral lines in the time-frequency domain is presented in part 4. For this application, we compare SMD to P out of N by means of a Monte-Carlo simulation in part 5.

2 HMM detection

We assume that the time behaviour of the object to be detected from its signal embedded in background noise is a Markov process taking its values in a finite state space $\{x_1 \dots x_N\}$. Then the a priori probability for any path $X(t_0, \Delta t) = (t_0, \Delta t)_{t=t_0 \dots t_0+\Delta t}$ being the path of the object equals the product of the initial state probability and the probabilities of the transitions between the successive states $x(t-1)$ and $x(t)$ for $t_0+1 \leq t \leq t_0+\Delta t$. The N initial state probabilities $P_n = \mathcal{P}[x(0) = x_n]$ and the N^2 transition probabilities $P_{n,m} = \mathcal{P}[x(t) = x_n | x(t-1) = x_m]$ are known parameters of our Markov model. We assume that the data serie $Y_X(t_0, \mathit{\Delta} t) = (y(t, \omega_x(t)))_{t=t_0 \ldots t_0+\mathit{\Delta} t}$ can be modelled as an independent random process with probability densities p_0 for the background noise and p_1 for the mix of noise and signal:

$$\begin{aligned} p_0(y(t, \omega_x(t))) &\equiv \mathcal{P}[y(t, \omega_x(t)) | H_0] \\ p_1(y(t, \omega_x(t))) &\equiv \mathcal{P}[y(t, \omega_x(t)) | x(t)] \end{aligned}$$

where H_0 is the hypothesis that there is no object. Then the joint probability of $X(t_0, \mathit{\Delta} t)$ and $Y_X(t_0, \mathit{\Delta} t)$ can be written as

$$\begin{aligned} \mathcal{P}[X(t_0, \mathit{\Delta} t), Y_X(t_0, \mathit{\Delta} t)] = &\ \mathcal{P}[y(t_0 + \mathit{\Delta} t, \omega_x(t_0 + \mathit{\Delta} t)) | X(t_0, \mathit{\Delta} t)] \\ &\ \mathcal{P}[Y_X(t_0, \mathit{\Delta} t - 1) | X(t_0, \mathit{\Delta} t)] \\ &\ \mathcal{P}[x(t_0 + \mathit{\Delta} t) | x(t_0 + \mathit{\Delta} t - 1)] \\ &\ \mathcal{P}[X(t_0, \mathit{\Delta} t - 1)] \end{aligned}$$

The last condition for our model being an HMM [Rabiner and Juang, 1986] is that the information from the state process about the data at some time comes from the state at this time. Then the previous equation takes the following recursive form:

$$\begin{aligned} \mathcal{P}[X(t_0, \mathit{\Delta} t), Y_X(t_0, \mathit{\Delta} t)] = &\ p_1(y(t_0 + \mathit{\Delta} t, \omega_x(t_0 + \mathit{\Delta} t))) \\ &\ \mathcal{P}[x(t_0 + \mathit{\Delta} t) | x(t_0 + \mathit{\Delta} t - 1)] \\ &\ \mathcal{P}[X(t_0, \mathit{\Delta} t - 1), Y_X(t_0, \mathit{\Delta} t - 1)] \end{aligned}$$

Since we have $\mathcal{P}[Y_X(t_0, \mathit{\Delta} t) | H_0] = \prod_{t=t_0}^{t=t_0+\mathit{\Delta} t} p_0(y(t, \omega_x(t)))$, we get also such a recursive form for the likelihood ratio $\mathit{\Lambda}_{X,Y}(t_0, \mathit{\Delta} t)$ of $(X(t_0, \mathit{\Delta} t), Y_X(t_0, \mathit{\Delta} t))$:

$$\mathit{\Lambda}_{X,Y}(t_0, 0) = \frac{p_1(y(t_0, \omega_x(t_0)))}{p_0(y(t_0, \omega_x(t_0)))} \mathcal{P}[x(t_0)]$$

$$\mathit{\Lambda}_{X,Y}(t_0, \mathit{\Delta} t) = \frac{p_1(y(t_0 + \mathit{\Delta} t, \omega_x(t_0 + \mathit{\Delta} t)))}{p_0(y(t_0 + \mathit{\Delta} t, \omega_x(t_0 + \mathit{\Delta} t)))} \mathcal{P}[x(t_0 + \mathit{\Delta} t) | x(t_0 + \mathit{\Delta} t - 1)] \mathit{\Lambda}_{X,Y}(t_0, \mathit{\Delta} t - 1)$$

For each state x_n, let us consider the maximum value of $\mathit{\Lambda}_{X,Y}(t_0, \mathit{\Delta} t)$ for all paths $X(t_0, \mathit{\Delta} t)$ ending at x_n:

$$\mathit{\Lambda}(x_n, t_0, \mathit{\Delta} t) \equiv \max \{ \mathit{\Lambda}_{X,Y}(t_0, \mathit{\Delta} t) | x(t_0 + \mathit{\Delta} t) = x_n \}$$

It can be computed recursively by means of the Viterbi algorithm:

$$\begin{aligned} \mathit{\Lambda}(x_n, t_0, 0) &= \frac{p_1(y(t_0, \omega_n))}{p_0(y(t_0, \omega_n))} P_n \\ \mathit{\Lambda}(x_n, t_0, \mathit{\Delta} t) &= \frac{p_1(y(t_0 + \mathit{\Delta} t, \omega_n))}{p_0(y(t_0 + \mathit{\Delta} t, \omega_n))} \max_{1 \le m \le N} \{ P_{n,m} \cdot \mathit{\Lambda}(x_m, t_0, \mathit{\Delta} t - 1) \} \end{aligned}$$

ω_n being the location in the data domain corresponding to the state x_n.

By comparing to a threshold the values of $\Lambda(x_n, t_0, \Delta t)$ for $1 \leq n \leq N$, we perform a detection test that, among all tests operating on the same time window, maximises the detection probability for a fixed false alarm probability determined by the detection threshold value. This holds for any signal starting before t_0 and ending after $t_0 + \Delta t$. For some given signal, the best performance is achieved when the processing time window equals the time interval when the signal is present. In practice, this interval is often unknown. A way for handling this problem is to perfom the processing for all possible values of t_0 and Δt. In practice, a trade-off has to be found between the computation cost and the detection performance by taking $(t_0, \Delta t)$ from some reduced subset into the set of the possible values.

3 Sequential Markov detector

SQPRT [Marano *et al.*, 2005] is a detection test that does not require fixing a priori Δt. It computes recursively the likelihood ratio of an i.i.d. data time serie and compares its current value to a downer threshold, the fail threshold, and an upper one, the detection threshold. If the likelihood ratio value is smaller than the fail threshold, H_0 is decided. If it stands between both thresholds, the likelihood ratio is multiplied by the likelihood ratio of the next data element and a new test is performed. If it is higher than the detection threshold, the signal presence hypothesis H_1 is decided. The false alarm probability $P_\mathrm{f} = \mathcal{P}[H_1 decided | H_0]$ relates mainly on the detection threshold value, approximately equal to $P_\mathrm{d}/P_\mathrm{f}$, P_d being the desired detection probability $\mathcal{P}[H_1 decided | H_1]$. The detection probability relates mainly on the fail threshold, approximately equal to $(1 - P_\mathrm{d})/(1 - P_\mathrm{f})$, close to $1 - P_\mathrm{d}$ if $P_\mathrm{f} \ll 1$.

We look now at how SQPRT could be applied to the maximum likelihood ratio $\Lambda(x_n, t_0, \Delta t)$ defined in section 2. Testing $\Lambda(x_n, t_0, \Delta t)$ for detection is equivalent to testing all the likelihood ratio values of the paths $X(t_0, \Delta t)$ ending at state x_n. The number of these paths is growing exponentially as a function of Δt because of the number of state transitions allowed at each time step. So does the false alarm probability of the test performed on the maximum value $\Lambda(x_n, t_0, \Delta t)$. For making this probability independent on Δt, we should decrease by an inverse factor the SQPRT false alarm probability P_f, so increase inversely the detection threshold value $P_\mathrm{d}/P_\mathrm{f}$. Equivalently the threshold value may be kept constant and the likelihood ratio multiplied at each update step by a constant factor K smaller than 1.

In the standard SQPRT, H_0 is definitely decided and the test is ended when the test value goes below the fail threshold approximately equal to $(1 - P_\mathrm{d})/(1 - P_\mathrm{f})$. This is because of the assumption that either H_0 holds for the whole data serie or H_1 does. If the signal may be present only during some time interval within the time interval of the data, the test must be reset once it failed in order to cope with the possibility that the past data might be noise only and that signal might start at some further time. Then a rather logical reset process would be to disregard the past data if, by doing so, the current test value, and consequently the further ones because of the recursive computation, are increased. Such reset process aims to prevent the signal detection from being jeopardized by the noise data before the starting time of the signal.

From the above principles, we can now introduce our new test, the Sequential Markov Detector. Its test value $\Lambda_n(t)$ at each time t and at each point of an HMM state space has the following recursive definition:

$$\begin{aligned}\Lambda_n(0) &= \frac{p_1(y(0,\omega_n))}{p_0(y(0,\omega_n))}P_n \\ \Lambda_n(t) &= \frac{p_1(y(t,\omega_n))}{p_0(y(t,\omega_n))}\max\{K\max_{1\le m\le N}\{P_{n,m}\cdot\Lambda_m(t-1)\},P_n\}\end{aligned}$$

where K is a constant smaller than 1. There is the following relation between $\Lambda_n(t)$ and the maximum likelihood ratio $\Lambda(x_n,t_0,\Delta t)$ presented in section 2:

$$\Lambda_n(t) = K^{t-\tau_n(t)}\Lambda(x_n,\tau_n(t),t-\tau_n(t))$$

where $\tau_n(t)$ is the latest time anterior or equal to t when Λ_n was reset. As discussed above, this relation is the one intended for making the false alarm probability independent on Δt and the $\Lambda_n(t)$ reset condition is that the test value from the current data only $\Lambda(x_n,t,0)$ is larger than the test value taking the past data into account $K^{t-\tau_n(t-1)}\Lambda(x_n,\tau_n(t-1),t-\tau_n(t-1))$. K is set up so that the following relation holds:

$$\mathcal{P}[K\max_{1\le m\le N}\{P_{n,m}\cdot\Lambda_m(t-1)\} > P_n|H_0] = \frac{1}{2}$$

expressing the fact that the probabilities for a path being continued or reset are equal when there is no signal.

$\Lambda_n(t)$ can be computed according to the expression of its above recursive definition. This computation is quite similar to the Viterbi algorithm, except for the reset process. In the next paragraph, we show how to apply it to detection of spectral lines in the time-frequency plane.

4 Application to detection of spectral lines in time-frequency data

The observed data $y(t,f) = |S(t,f)|^2$ are the square magnitude of the output of a time sliding Fourier transform performed on a scalar signal. Hence the observation variable noted previously ω is the frequency, noted from now f.

We define the state as being the pair composed by the frequency of the signal to be detected and its time derivative, which we call the slope:

$$x(t) = (f_x(t), \dot{f}_x(t))$$

We assume that the complex noise component of $S(t, f)$ is zero-mean gaussian with unit variance and that the signal amplitude is constant with signal-to-noise power ratio r_0. Then p_0 and p_1 are homothetic to centered and uncentered χ^2 laws with 2 degrees of freedom:

$$\begin{aligned} p_0(y) &= \exp(-y) \\ p_1(y) &= \exp(-y - r_0) \cdot I_0(2\sqrt{r_0 y}) \end{aligned}$$

with $I_0(z) = \frac{1}{\pi} \int_0^\pi e^{z\cos(\theta)} d\theta$.

Let us be T_{FT} the length of the sliding time window of the Fourier transform and k_t and k_f the coefficients such that the time step of the data $y(t, f)$ equals T_{FT}/k_t and the frequency step of the states equals T_{FT}^{-1}/k_f. Then the slope step is taken equal to $(k_t/k_f)T_{\mathrm{FT}}^{-2}$, the ratio of the frequency step to the time step. So the state space is a finite grid in the real plane with mesh $(T_{\mathrm{FT}}^{-1}/k_f, (k_t/k_f)T_{\mathrm{FT}}^{-2})$. Within this grid, we define the transition probabilities $P_{n,m}$ as following:

$$\begin{aligned} &\text{if} \quad \frac{-1}{2k_f T_{\mathrm{FT}}} \le f_n - f_m - \frac{\dot{f}_n + \dot{f}_m}{2} \frac{T_{\mathrm{FT}}}{k_t} < \frac{1}{2k_f T_{\mathrm{FT}}} \\ &\text{then} \quad P_{n,m} = h\left(\frac{k_f T_{\mathrm{FT}}^2}{k_t} |\dot{f}_n - \dot{f}_m|\right) \\ &\text{else} \quad P_{n,m} = 0 \end{aligned}$$

where h may be any decreasing function such that

$$h(0) + 2\sum_{i=1}^{\infty} h(i) = 1$$

The above relations means that the probability of the transition from the state $(f_m, \dot{f}_m)$ to the state $(f_n, \dot{f}_n)$ is non zero if and only if the frequency change $f_n - f_m$ equals the mean slope value $(\dot{f}_n + \dot{f}_m)/2$ multiplied by the time step T_{FT}/k_t within an error less than half the frequency quantization step T_{FT}^{-1}/k_f. Then the transition probability is a decreasing function of the absolute value of the slope change $\dot{f}_n - \dot{f}_m$. For a given state $(f_m, \dot{f}_m)$ and a given slope change $\dot{f}_n - \dot{f}_m$, there is only one frequency f_n which fulfils the first relation. Then the last relation is equivalent to the condition $\sum_n P_{n,m} = 1$, which expresses the fact that the sum of the probabilities, conditional with respect to some state m, of all its possible successors n, equals 1.

In practice the setting of the function h that determines the transition probabilities may be rather arbitrary because a statistical model for the frequency fluctuation of the signals to be detected is seldom available. The broader is the peak of h at 0, the better is the processing capability to cope with fast fluctuation of the frequency slope, but the lower is the performance achieved on constant frequency slope signals, especially stable frequency signals. The performance decrease on stable frequency signals when the model is changed from a setting suited to them to a setting suited to fluctuating frequency slope signals is illustrated by results shown in the next paragraph.

5 Performance evaluation

We compared the performances of SMD and P out of N detector by means of a Monte-Carlo simulation for detection of spectral lines in magnitude-square FFT data as described in the previous paragraph. The data and the states have the same time and frequency steps with $k_t = 2$ and $k_f = 4$. Note that since k_t is larger than 1, the assumption of time independent data is not valid. This deviation with respect to the HMM theoretical frame, rather usual because the data sampling frequency is above the Shannon bound in many applications, is not expected to have a significant impact on the performance.

We tested two SMD settings having both their signal-to-noise ratio parameter r_0 equal to 0.5 and uniform probability law for the initial states: $P_n = 1/N$ for any n. In the first setting, the slope set is $\{0\}$. Then our model is equivalent to the one where the states are the frequencies and $P_{n,m}$ equals 1 if $n = m$ and 0 otherwise ($h(0) = 1$ and $h(i) = 0$ for $i > 0$). This setting is suited to detection of stable frequency signals. In the second setting, the set of the values for the normalised slope $k_f \dot{f}_n T_{\mathrm{FT}}^2 / k_t$ is $\{-2, -1, 0, 1, 2\}$ and the slope change $k_f(\dot{f}_n - \dot{f}_m)T_{\mathrm{FT}}^2 / k_t$ takes its value in $\{-1, 0, 1\}$ with an uniform probability law: $h(0) = h(1) = 1/3$ and $h(i) = 0$ for $i > 1$. The SMD output $\Lambda_n(t)$ computed in the state space is projected to the frequency axis according to the relation

$$D(t, f) = \max\{\Lambda_n(t) | f_n = f\}$$

The P out of N detector window covers four frequency channels; so its bandwidth equals the frequency resolution T_{FT}^{-1} of the Fourier transform. The threshold r_1 equals 2. The detector output $D(t, f)$ is computed by placing the $(N, 4)$ window so that one of its points, arbitrarily fixed, is located at the point (t, f) of the time-frequency data. Then $D(t, f)$ equals the mean of the P highest events when the P out of N condition is fulfilled and it equals 0 otherwise. Two P out of N detectors were tested: $(P, N) = (3, 4)$ and $(P, N) = (6, 8)$.

Each value of the detection probability estimate $\hat{P}_{\mathrm{d}}$ in tables 1 to 3 is the mean of the results of 3 independent Monte-Carlo runs, each run involving two files of 2000×1000 time-frequency complex data. One file consists in noise only samples $S_0(t, f)$. The data $S_1(t, f)$ of the second file are samples of the sum of the noise $S_0(t, f)$ and $I = 30$ test signals having the same detection features (signal-to-noise ratio, time duration, frequency fluctuation). Hence 90 test signals were used for each value of the estimate $\hat{P}_{\rm d}$.

The I test signals in the data $S_1(t,f)$ are located in I non-overlapping time-frequency blocks $[t_{0,\min,i}, t_{0,\max,i}] \times [f_{\min,i}, f_{\max,i}]$ having the same size $\Delta t \times \Delta f$. The computation of $\hat{P}_{\rm d}$ involves the processing outputs $D_0(t,f)$ and $D_1(t,f)$ from the noise only and signal plus noise input data $S_0(t,f)$ and $S_1(t,f)$. For one Monte-Carlo run, it has the expression:

$$\hat{P}_{\rm d} = \frac{1}{I^2}\sum_{i=1}^{I}\sum_{j=1}^{I}\left|\begin{array}{l}\max\{D_1(t,f_i(t))|t_{1,\min,i} \le t \le t_{1,\max,i}\} \qquad > \\ \max\{D_0(t,f_i(t))|t_{0,\min,j} \le t \le t_{0,\max,j}, f_{\min,j} \le f \le f_{\max,j}\}\end{array}\right|$$

where $|true|$ equals 1, $|false|$ equals 0, $[t_{1,\min,i}, t_{1,\max,i}]$ is the time interval of the ith signal included in $[t_{0,\min,i}, t_{0,\max,i}]$, and $f_i(t)$ is the signal frequency determined by the relation

$$f_i(t) = \arg\max_{\rm f}\left\{|S_1(t,f) - S_0(t,f)|\Big|f_{\min,i} \le f \le f_{\max,i}\right\}$$

The I detection threshold values are the maximum output values on the I noise-only data blocks. Hence they relate to a mean number of false alarms per data block equal to 1. We define the false alarm probability $P_{\rm f,in}$ as the ratio of the output false alarm rate to the input rate of independent data. Each data block containing $\Delta t \times \Delta f$ independent data elements, we have:

$$P_{\rm f,in} = \frac{1}{\Delta t \times \Delta f}$$

Each data block consists in 300 time lines of 200 adjacent frequency channels. So we have $P_{\rm f,in} = 1/(300 \times T_{\rm FT}/k_t)/(200 \times T_{\rm FT}^{-1}/k_f) = 1.3 \times 10^{-4}$.

Defining the false alarm probability with respect to the input data allows a fair comparison between detectors having different rates of independent decisions. In order to validate the above method, the detection probability of the integrator of time constant equal to the signal duration $T_{\rm sig}$ was estimated with the above method in the cases of stable frequency signals with $T_{\rm sig} = 100 \times T_{\rm FT}$ (Table 1) and $T_{\rm sig} = 10 \times T_{\rm FT}$ (Table 2). The well-known ROC curves for the Rice case give the theoretical value of the signal-to-noise ratio r corresponding to the measured detection performance $(\hat{P}_{\rm d}, P_{\rm f,out})$, $P_{\rm f,out} \approx (T_{\rm sig}/T_{\rm FT}) \times P_{\rm f,in}$ being the false alarm probability at the detector output as considered in these curves. This theoretical value is given in parentheses below the $\hat{P}_{\rm d}$ value in tables 1 and 2. The difference between both signal-to-noise ratio values was always found smaller than 1 dB. This difference may be caused not only by the estimation error on $\hat{P}_{\rm d}$ but also likely by the fact that in our test we have a random detection threshold and a fixed false alarm rate, while the ROC curves hold for a deterministic threshold and a random false alarm rate.

From the results in tables 1 and 2, the cost of not knowing the signal duration in stable state SMD with respect to the performance achieved by

$10\log_{10}(r)$	-5	-4	-3	-2	-1	0	1	2	3
integrator $100 \times T_{\mathrm{FT}}$ (SNR(dB) for $P_{\mathrm{f,out}} = 0.013$)	0.53 (-5.9)	0.83 (-4.5)	0.93 (-3.6)	0.98 (-2.8)					
SMD – stable states	0.30	0.55	0.74	0.90	0.96				
SMD – 5 slopes $h(0) = h(1) = 1/3$		0.30	0.46	0.60	0.79	0.97			
3 out of 4					0.31	0.40	0.50	0.67	0.90

Table 1. $\hat{P}_{\mathrm{d}}$ for $P_{\mathrm{f,in}} = 1.3 \times 10^{-4}$ – Stable frequency – $T_{\mathrm{sig}} = 100 \times T_{\mathrm{FT}}$.

$10\log_{10}(r)$	2	3	4	5	6	7
integrator $100 \times T_{\mathrm{FT}}$ (SNR(dB) for $P_{\mathrm{f,out}} = 0.013$)	0.54 (1.2)	0.88 (3.0)	0.92 (3.2)			
SMD – stable states	0.47	0.84	0.90	0.99		
SMD – 5 slopes $h(0) = h(1) = 1/3$		0.49	0.63	0.89	0.95	0.99
3 out of 4			0.33	0.66	0.79	0.95

Table 2. $\hat{P}_{\mathrm{d}}$ for $P_{\mathrm{f,in}} = 1.3 \times 10^{-4}$ – Stable frequency – $T_{\mathrm{sig}} = 10 \times T_{\mathrm{FT}}$.

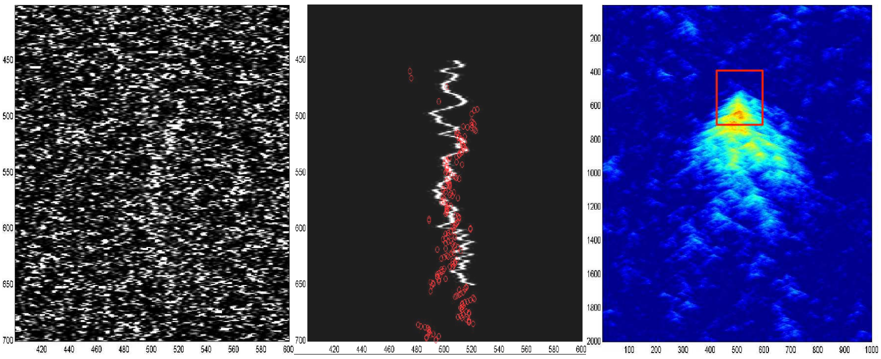
the time integrator with time constant equal to the signal duration appears being close to 1 dB when the signal contains 100 independent samples and smaller than 1 dB when it contains 10 independent samples. The ability to perform also on fluctuating frequency slope signal with the second SMD setting is provided with an additional cost standing between 1 dB and 2 dB in detection of stable frequency signals. Then the gain with respect to 3 out of 4 detection stands between 3 and 4 dB for 100 independent sample signal and between 1 and 2 dB for 10 independent sample signal.

Performances on fluctuating frequency signals are presented in Table 3. The frequency fluctuation is gaussian with standard deviation σ_f taking values 0, T_{FT}^{-1}, $2T_{\mathrm{FT}}^{-1}$ and $3T_{\mathrm{FT}}^{-1}$. The time length of the fluctuation correlation equals $15 \times T_{\mathrm{FT}}$. As expected, the performance from the integrator with time constant equal to signal duration is much sensitive to signal frequency fluctuation. 6 out of 8 detector performs better than 3 out of 4 detector only when σ_f is smaller than $2T_{\mathrm{FT}}^{-1}$. This illustrates the fact that the N parameter of the P out of N detector is limited by the expected drift of the signal to be detected. Anyway SMD still performs significantly better than P out of N in all test cases.

An example of input and output data is displayed on Figure 1 where, for illustration clarity, only one signal is embedded in noise, the data format being the same than the one described above. The signal features are $10\log_{10}(r) = -1$ dB, $T_{\mathrm{sig}} = 100 \times T_{\mathrm{FT}}$ and $\sigma_f = 3T_{\mathrm{FT}}^{-1}$. In practice, the SMD output would be reset in the vicinity of a detection and a specific tracking process should be started for maintenance and termination testing of the newly validated track. This further tracking process, not performed in this work

$\sigma_f \times T_{\mathrm{FT}}$	0	1	2	3
integrator $100 \times T_{\mathrm{FT}}$	1.00	0.58	0.21	0.14
SMD – 5 slopes $h(0) = h(1) = 1/3$	0.79	0.75	0.61	0.53
3 out of 4	0.31	0.28	0.23	0.17
6 out of 8	0.48	0.34	0.20	0.10

Table 3. $\hat{P}_{\mathrm{d}}$ for $P_{\mathrm{f,in}} = 1.3 \times 10^{-4}$ – Fluctuating frequency – $T_{\mathrm{sig}} = 100 \times T_{\mathrm{FT}}$ – $10 \log_{10}(r) = -1\mathrm{dB}$.



Signal+ noise input $|S_1(t,f)|^2$ Zoom 300 times × 200 frequencies

Track: amplified signal $|S_1(t,f) - S_0(t,f)|^2$ Red circles: frequency bin in $[1, 1000]$ of maximum SMD output $D_1(t,f)$ Zoom 300 times × 200 frequencies

SMD signal + noise output $D_1(t,f)$ Full time-frequency range 2000 times × 1000 frequencies Zoom window in red

Fig. 1. Example of input and output data.

entirely devoted to the track initiation problem, would avoid the spreading of the SMD output peaks seen on Figure 1.

6 Conclusion

We presented a new track initiation method named the Sequential Markov Detector. It is a "track-before-detect" processing which combines HMM tracking and sequential detection. The detection test value is the a posteriori likelihood ratio weighted by a factor exponentially decreasing as a function of the time duration of the tested path in the state space. It is reset when taking into account only the current data provides a larger value.

This new detector was shown to perform significantly better than the usual P out of N detector for spectral line detection from time-frequency data. The margin for further performance improvement from the same kind of data and the same a priori information about the signal is likely small since the detection loss on a stable spectral line with respect to the constant frequency integrator matched to the signal duration is at most 3 dB for signal bandwidth-time product at most equal to 100, while at least a part of this loss is the unavoidable cost for SMD ability to perform on unknown-duration unstable-frequency signal. Further research should rather to look at how exploiting richer data, for instance complex spectral data instead of magnitude data, or more accurate a priori information within the state model.

References

[Tonissen and Evans, 1996]S.M. Tonissen and R. J. Evans, "Performance of Dynamic Programming Techniques for track-before-detect", *IEEE Tr. AES*, vol. 32, no. 4, pp. 1440–1451, October 1996.

[Barrett and Holdsworth, 1993]R. F. Barrett and D. A. Holdsworth, "Frequency Tracking Using Hidden Markov Models with Amplitude and Phase Information" *IEEE Tr. AES*, vol. 41, no. 10, pp. 2965–2976, October 1993.

[Blostein and Richardson, 1994]S. D. Blostein and H. S. Richardson, "A Sequential Detection Approach to Target Tracking", *IEEE Tr. AES*, vol. 30, no. 1, pp. 197–212, January 1994.

[Marano *et al.*, 2005]S. Marano, V. Matta and P. Willett, "Sequential Detection of Almost-Harmonic Signals", *IEEE Tr. SP*, vol. 51, no. 2, pp. 395–406, February 2003.

[Rabiner and Juang, 1986]L.R. Rabiner and B.H. Juang, "An Introduction to Hidden Markov Models", *IEEE ASSP Magazine*, pp. 4–16, January 1986.

Markov Random Fields for Recognizing Textures modeled by Feature Vectors

Juliette Blanchet, Florence Forbes, and Cordelia Schmid

Inria Rhône-Alpes, 655 avenue de l'Europe,
Montbonnot, 38334 Saint Ismier Cedex, France
(e-mail: florence.forbes@inrialpes.fr, juliette.blanchet@inrialpes.fr,
cordelia.schmid@inrialpes.fr)

Abstract. This paper decribes a new probabilistic framework for recognizing textures in images. Images are described by local affine-invariant descriptors and by spatial relationships between these descriptors. We propose to introduce the use of statistical parametric models of the dependence between descriptors. Hidden Markov Models (HMM) are investigated for such a task using recent estimation procedures based on the mean field principle to perform the non trivial parameter estimation they require. Preliminary experiments obtained with 140 images of seven different natural textures show promising results.

Keywords: Hidden Markov Models, Mean Field approximation, Statistical learning, Texture recognition.

1 Introduction

Image descriptors is a key notion in computer vision. Descriptors are local characteristics whose geometric organization can be very informative when carrying out pattern recognition tasks. The most important characteristics for efficient image descriptors are good discrimination, locality (for resistance to occlusions), and sufficient invariance to various image transformations. Local descriptors that meet these requirements exist, but incorporating information about the relative spatial organization of such descriptors is still an open issue. It is not yet clear which organizational models will prove to be the most useful, and many statistical issues relating to the estimation and selection of such models remain to be resolved. In this paper, we propose organizational models based on Markov Random Fields and we focus on a texture recognition task as a first investigation of these models. We specify how to select and estimate such models from the data.

The approach we consider for texture recognition is the use of affine-invariant region detectors. Such representations have several advantages but they do not account for the way detected regions are organized within the image. An attempt to include neighborhood statistics was described in [Lazebnik *et al.*, 2003a]. This was done by adding in the recognition stage, a relaxation step [Rosenfel *et al.*, 1976] to refine texture membership probabilities but was not using an explicit organizational model for the data in the learning stage. Our claim is that there is some gain in assuming that the feature vectors are dependent statistical variables and consequently in using parametric statistical models to account for this dependencies explicitly. We show that recognition can be improved by using Hidden Markov Models (HMM) as organizational models when learning the texture classes. Estimating the parameters of such models in this context is not trivial. We use recent estimation procedures (EM-like algorithms) based on the Expectation-Maximization (EM) algorithm and on the mean field principle of statistical physics [Chandler, 1987].

2 Hidden Markov Models for textures

For the feature extraction stage, we follow the texture representation method described in [Lazebnik *et al.*, 2003a] for its advantages over methods proposed in recent literature. It is based on an interest point detector that leads to a sparse representation selecting the most perceptually salient regions in an image and on a *shape selection* process that provides affine invariance. Informally (see [Lindeberg and Garding, 1997] for details), regions are represented by ellipses of various volume and shape and centered at various locations (points found by the detector). The neighborhood of a region represented by a given ellipse can then be naturally computed by adding a constant amount (15 pixels in our implementation) to the major and minor axes and to let the neighborhood consists of all points that fall inside this enlarged ellipse. We can then think of an image as a graph with edges emanating from the center of each region to other centers within its neighborhood. To each detected region is then associated a feature vector (descriptor). The descriptors we use are intensity domain *spin images* [Lazebnik *et al.*, 2003b] rescaled to have a constant norm and flattened into 80-dimensional feature vectors. The basic assumption is that descriptors are random variables with a specific probability distribution in each texture class. In [Lazebnik *et al.*, 2003a], the distribution of descriptors in each texture class is modeled as a Gaussian mixture model where each component corresponds to a sub-class. This is assuming that the descriptors are independent variables although it naturally exists strong neighborhood relationships between feature vectors within the same image. To take that into account, we propose to improve on the Gaussian mixture model by assuming that for each image from a single texture, the distribution of descriptors is that of a Hidden Markov Model (HMM) with K components and appropriate parametrization to be specified below.

Let $x_1, \ldots, x_n$ denote the n descriptors (80-dimensional vectors) extracted at locations denoted by $\{1, \ldots, n\}$ from an image. Let m denotes the texture class of this image. For $i = 1, \ldots, n$, we model the probability of observing descriptor x_i when the image is from texture m as

$$P(x_i|\Psi_m) \;=\; \sum_{k=1}^{K} P(Z_i = c_{mk}|\beta_m)\; f(x_i|\theta_{mk}),$$

where $f(x_i|\theta_{mk})$ denotes the multivariate Gaussian distribution with parameters θ_{mk} namely the mean μ_{mk} and covariance matrix Σ_{mk}. Notation Z_i denotes the random variable representing the sub-class of descriptor x_i. It can take values in $\{c_{mk}, k = 1 \ldots K\}$ denoting the K possible sub-classes for texture m. Note that for simplicity we assume K being the same for each texture but this can be generalized (see section 5). Notation β_m denotes additional parameters defining the distribution of the Z_i's and Ψ_m denotes the whole model parameters *i.e.* $\Psi_m = (\theta_{mk}, \beta_m, k = 1 \ldots K)$. Our approach differs from [Lazebnik *et al.*, 2003a] in that our aim is to account for spatially dependent descriptors. More specifically, the dependencies between neighboring descriptors are modeled by further assuming that the joint distribution of $Z_1, \ldots, Z_n$ is a discrete Markov Random Field on the graph defined above. Denoting $z = (z_1, \ldots, z_n)$ specified values of the Z_i's, we define

$$P(z|\beta_m) = W(\beta_m)^{-1} \, \exp(-H(z, \beta_m)),$$

where $W(\beta_m)$ is a normalizing constant and H is a function assumed to be of the following form (we restrict to pair-wise interactions),

$$H(z, \beta_m) = \sum_{i=1}^{n} V_i(z_i, \beta_m) + \sum_{\substack{i,j \\ i \mathrm{sim} j}} V_{ij}(z_i, z_j, \beta_m),$$

where the V_i's and V_{ij}'s are respectively referred to as singleton and pair-wise potentials. We write $i\mathrm{sim}j$ when locations i and j are neighbors on the graph, so that the second sum above is over neighboring locations. The spatial parameters β_m consist of two sets $\beta_m = (\alpha_m, \mathbb{B}_m)$ where α_m and $\mathbb{B}_m$ are defined as follows. We consider pair-wise potentials V_{ij} that only depend on z_i and z_j (not on i and j). Since the z_i's can only take a finite number of values, we can define a $K \times K$ matrix $\mathbb{B}_m = (b_m(k, l))_{1 \leq k,l \leq K}$ and write without lost of generality

$$V_{ij}(z_i, z_j, \beta_m) = -b_m(k, l) \text{ if } z_i = c_{mk} \text{ and } z_j = c_{ml}.$$

Similarly we consider singleton potentials V_i that only depend on z_i so that denoting by α_m a $K-$dimensional vector, we can write

$$V_i(z_i, \beta_m) = -\alpha_m(k) \text{ if } z_i = c_{mk},$$

where $\alpha_m(k)$ is the k^{th} component of α_m. This vector α_m acts as weights for the different values of z_i. When α_m is zero, no sub-class is favored, *i.e.* at a given location i, if no information on the neighboring locations is available, then all sub-classes appear with the same probability at location i. When $\mathbb{B}_m$ is zero, there is no interaction between the locations and the Z_i's are independent. When $\mathbb{B}_m$ is zero, β_m reduces to α_m and it comes that for $i = 1, \ldots, n$ and $k = 1, \ldots, K$,

$$P(Z_i = c_{mk}|\alpha_m) = \frac{\exp(\alpha_m(k))}{\sum_{l=1}^{K} \exp(\alpha_m(l))},$$

which clearly shows that α_m acts as weights for the different possible values of z_i. Conversely, when α_m is zero and $\mathbb{B}_m = \beta \times I$ where β is a scalar, the spatial parameters β_m reduce to a single scalar interaction parameter β

and we get the Potts model traditionnaly used for image segmentation. Note that this model is not necessarily appropriate for textures since it tends to favor neighbors that are in the same sub-class. In practice we observed in our experiments that when learning texture classes, $\mathbb{B}_m$ could be far from $\beta \times I$. Texture m is then represented by an HMM defined by parameters Ψ_m being $\Psi_m = (\mu_{mk}, \Sigma_{mk}, \alpha_m(k), \mathbb{B}_m, k = 1, \ldots, K)$.

3 Learning the descriptors distribution and organization

In a supervised framework, we first learn the distribution for each texture class based on a training data set. Our learning step is based on an EM-like algorithm and this framework allows to incorporate unsegmented multi-texture images. However, we refer to the work of [Nigam *et al.*, 2000] and [Lazebnik *et al.*, 2003a] for more details on how to implement this generalization.

In this presentation the training data consists then of single-texture images from each texture class $m = 1, \ldots, M$. Each texture class is learned successively. Using all the feature vectors and neighborhood relationships extracted from the images belonging to class m, we estimate an HMM as described in section 2. The EM algorithm is a commonly used algorithm for parameters estimation in problems with hidden data (here the sub-class assignments). For Hidden Markov Random Fields, due to the dependence structure, the exact EM is not tractable and approximations are required to make the algorithm tractable. In this paper, we use some of the approximations based on the mean field principle presented in [Celeux *et al.*, 2003]. This allows to take the Markovian structure into account while preserving the good features of EM. The procedures in [Celeux *et al.*, 2003] are based on mean field approximation. More specifically, we used the so-called *simulated field* algorithm for it shows better performance in some segmentation tasks (see [Celeux *et al.*, 2003]). Note that in practice, we had to extend these algorithms to incorporate the estimation of matrix $\mathbb{B}_m$ and to include irregular neighborhood structure coming from descriptors locations and not from regular pixel grids like in [Celeux *et al.*, 2003].

Briefly, these algorithms can be presented as follow. They are based on the EM algorithm which is an iterative algorithm aiming at maximizing the log-likelihood (for the observed variables x) of the model under consideration by maximizing at each iteration the expectation of the complete log-likelihood (for the observed and hidden variables x and z) knowing the data and a current estimate of the model parameters. When the model is an Hidden Markov Model with parameters Ψ_m, there are two difficulties in evaluating this expectation. Both the normalizing constant $W(\beta_m)$ and the conditional probabilities $P(z_i \mid x, \Psi_m)$ and $P(z_i, z_j, j \in N(i) \mid x, \Psi_m)$ cannot be computed exactly ($N(i)$ denotes the neighbors of i). Informally, the mean field approach consists in approximating the intractable probabilities by neglecting fluctuations from the mean in the neighborhood of each location i. More generally, we talk about mean field-like approximations when the value at location i does not depend on the value at other locations which are all set to constants (not necessarily to the means) independently of the value at location i. These constant values denoted by $\tilde{z}_1, \ldots, \tilde{z}_n$ are not arbitrary but satisfy some appropriate consistency conditions (see [Celeux *et al.*, 2003]). It follows that $P(z_i \mid x, \Psi_m)$ is approximated by $P(z_i \mid x, \tilde{z}_j, j \in N(i), \Psi_m)$ and $P(z_i, z_j, j \in N(i) \mid x, \Psi_m)$ by $P(z_i \mid x, \tilde{z}_l, l \in N(i), \Psi_m)\ P(z_j \mid x, \tilde{z}_l, l \in N(j), \Psi_m)$. Using such approximations leads to algorithms which in their general form consist in repeating two steps. At iteration q,

(1) Create from the data x and some current parameter estimates $\Psi^{(q-1)}$ a configuration $\tilde{z}_1^{(q)}, \ldots \tilde{z}_n^{(q)}$, *i.e.* values for the Z_i's. Replace the Markov distribution $P(z|\beta_m)$ by the factorized distribution $\prod\limits_{i=1}^{n} P(z_i|\tilde{z}_j^{(q)}, j \in N(i), \Psi_m)$. It follows that the joint distribution $P(x, z|\Psi_m)$ can also be approximated by a factorized distribution and the two problems encountered when considering the EM algorithm with the exact joint distribution disappear. The second step is therefore,

(2) Apply the EM algorithm for this factorized model with starting values $\Psi^{(q-1)}$, to get updated estimates $\Psi^{(q)}$ of the parameters.

In particular the *mean field* algorithm consists in using mean values for the $\tilde{z}_i{}^{(q)}$'s while the *simulated field* algorithm consists in obtaining $\tilde{z}_i{}^{(q)}$'s by simulation. In practice, at step **(2)**, performing one EM iteration is usually enough. In this case the *mean field* algorithm is the algorithm in [Zhang, 1992]. In Section 5, results are reported for the simulated field algorithm. Results for the mean field algorithm were at best equivalent. Then, for each texture, the HMM estimation provides us with estimations for the means and covariance matrices of the K Gaussian distributions, namely μ_{mk} and Σ_{mk} for $k = 1, \ldots K$, but also for the hidden field parameters, matrix $\mathbb{B}_m$ and vector α_m. This set of parameters is then associated to the texture class and used to classify regions in test images in one of the learned textures as specified in the next section.

For comparison we also consider a different way to learn texture that do not use the HMM formalism. We used a penalized EM algorithm for spatial data called NEM for Neighborhood EM [Ambroise *et al.*, 1997]. It provides a way to add spatial information when dealing with data represented as independent mixture models. It leads to a simple procedure but is not as flexible as the HMM approach which includes spatial information directly in the model. NEM can be seen as intermediate between the use of independent mixture models as in [Lazebnik *et al.*, 2003a] and our approach. To use it in our experiments we had to generalize its Potts-like penalization to a penalization term appropriate for textures. We used a matrix $\mathbb{B}$ as in Section 2.

A set of parameters is then associated to each texture class and used to classify regions in test images in one of the learned textures as specified in the next section.

4 Classification and retrieval

Images in the test set are not labeled and may contain several texture classes. Our aim is first to classify each region individually in one of the M texture classes under consideration. Then, each region can possibly be in one of $M \times K$ sub-classes. To identify these sub-classes, the model for the descriptor distribution has to incorporate the information learned from each texture in the learning stage. To do so, at recognition time, the descriptors distribution is assumed to be that of a Gaussian HMM as presented in Section 2 but with a discrete hidden field taking values in $\{c_{mk}, m = 1, \dots, M, k = 1, \dots, K\}$ *i.e.* with $M \times K$ components instead of K in the learning stage. In addition, the parameters of this HMM are given: for $m = 1, \dots, M$ and $k = 1, \dots, K$, the conditional distributions $f(x_i|\theta_{mk})$ are assumed to be Gaussian with means and covariance matrices learned at learning time. As regards, the hidden field, the pair-wise potentials are defined through a square matrix of size $M \times K$ denoted by $\mathbb{B}$ and constructed from the learned $\mathbb{B}_m$ matrices as follows: we first construct a bloc diagonal matrix using the learned $\mathbb{B}_m$ as blocs. The other terms correspond to pairs of sub-classes belonging to different classes. When only single-texture images are used in the learning stage, these terms are not available. As mentionned in [Lazebnik *et al.*, 2003a] even when multi-texture images are used for learning, the estimations for such terms are not reliable due to the fact that only a few such pairs are present in the training data. Unless the number of texture classes is very small, it is quite difficult to create a training set that would include samples of every possible boundary. In practice the missing values in $\mathbb{B}$ are set to a constant value chosen as a "smootheness constraint". The potentials on singletons, which are related to the proportions of the different sub-classes as mentioned in Section 2 are fixed to the values learned for each texture. Then the EM-like algorithm of Section 3 can be used with all parameters fixed to estimate the membership probability for each of the $M \times K$ sub-classes. The algorithm can be seen as iterations refining initial membership probabilities by taking into account the learned HMM's. This is not possible with standard EM for Gaussian mixtures since without spatial information, when all parameters are fixed, the algorithm reduces to a single iteration.

Membership probabilities are then also obtained for each texture class. For each region located at i, we get $P(Z_i = c_{mk}|x_i)$ for $m = 1, \dots M$ and $k = 1, \dots K$ and $P(Y_i = m|x_i)$ if Y_i denotes the unknown texture class. We have $P(Y_i = m|x_i) = \sum_{k=1}^{K} P(Z_i = c_{mk}|x_i)$. Determining the texture class of the region located at i consists then in assigning it to the class m that maximizes $P(Y_i = m|x_i)$. At the image level, a global score can be defined for each texture class. For instance, the score for class m can be computed by summing over all n regions found in the image, *i.e.* $\sum_{i=1}^{n} P(Y_i = m|x_i)$, and the image assigned to the class with the highest score.

Note that in a previous study, the HMM in the test stage was only partly defined. All parameters were fixed as above except the potentials on singletons which were estimated using the EM-like algorithm as in Section 3. This required much more computation and did not lead to better recognition rates in our experiments, except for some rare cases. However this possibility would worth further investigation.

5 Experimental Results

Preliminary experiments are made on a data set containing seven different textures (Figure 1). The data set is partitionned into a training and a test set containing 10 single texture images each. For simplicity, we set $K = 10$ for each texture. In some preliminary study we selected varying K using the Bayesian Information Criterion (BIC) of Schwarz [Forbes and Peyrard, 2003] but we did not observe significantly better recognition results. For the Gaussian distributions we restrict to diagonal covariance models. For each texture class m, using BIC we select among these models, the ones with $\Sigma_{mk} = \sigma_m^2 I$ for all $k = 1, \ldots K$. Table 1 shows classification results for individual regions that is the fraction of all individual regions in the test images that were correctly classified. The "Max likelihood" column refers to the method that consists in assuming that all texture class has the same probability to occur in the test image independently of the image. A region is then classified as belonging to the texture class with the best mixture likelihood (learned parameters). The "Relaxation" column refers to the method used in [Lazebnik *et al.*, 2003a]. The procedure uses as initial probabilities the ones that can be computed from the learned mixture models. These probabilities are then modified, through a relaxation step [Rosenfel *et al.*, 1976], using some additional spatial information deduced from the learning stage using co-occurence statistics. The results in Table 1 show that the rates improve significantly on the Maximum Likelihood rates for textures 1 to 5 but much less for textures 6 and 7. This points out one drawback of Relaxation which is sensitive to the quality of the initial probability estimates. The following columns refer to methods investigated in this paper. When all parameters are fixed, as this is the case in the test stage, NEM iterations can be reduced to update equations for the membership probabilities. These equations can be compared to Relaxation equations which similarly consist in updating membership probabilities. However, a main difference is that NEM is originally made for mixture models and therefore the mixture model is taken into account at each iteration. In the Relaxation algorithm, no model assumption is made and iterations are independent of the model used for the data. In a context where learning is made by assuming mixture models, using NEM seems then more consistent and appropriate. Table 1 shows better rates for NEM when compared to Relaxation. The method using HMM's is the only one where the descriptors are modeled as statistically dependent variables. It provides a way to analyse and control theses dependencies through a number of parameters. The "simulated Field" columns refer to our HMM model. When all parameters are fixed, the Simulated Field algorithm also reduces to update equations comparable to Relaxation but with the advantage of including the Markov model explicitly. The rates increase when compared to Relaxation. When comparing to NEM, rates increase for textures 5 to 7 and decrease for textures 1 to 4 but on average the Simulated Field algorithm performs better. As a global comment, one can observe that all methods have more trouble in recognizing textures 6 and 7. The corresponding data sets both contain images with very strong luminosity changes and some fuzzy images suggesting that the descriptors and/or the neighborhood structure we used may not be invariant enough. These preliminary experiments show however that there is significant gain in incorporating spatial relationships between descriptors. It appears that there is some gain in doing that using statistical parametric models, such as mixture models (NEM) or their extension HMM's (Simulated Field Algorithm), in the learning stage as well as in the test stage.

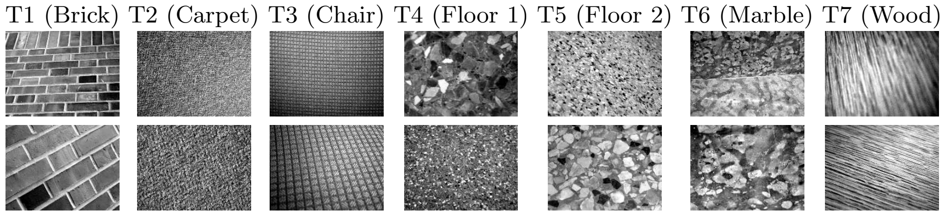


Fig. 1. Samples of the texture classes used in the experiments.

Class	T1	T2	T3	T4	T5	T6	T7
Max. Likelihood	48	77	52	56	50	17	30
Relaxation	78	96	72	86	80	19	42
NEM	82	98	78	88	80	20	43
Simulated Field	81	97	77	80	86	26	46

Table 1. Classification rates in % for individual regions of single-texture images.

6 Conclusions

We based our work on recent techniques for image description going further regular grid of pixels to sets of irregularly spaced feature vectors. Our aim was to show that statistical parametric models could be introduced to account for spatial or geometric relationships between feature vectors. We show that Hidden Markov Models were natural candidates and focused on a texture recognition task as an illustration. For such a task Markov Models have been used to model grey-level values on regular pixel grids but their introduction in the context of feature vectors at irregular locations is new. In this context, they provide parametric models where the parameters have a natural interpretation. Some of them (the α_{mk}'s) can be related to texture proportions while others (matrix $\mathbb{B}$) to pair-wise interactions (see Section 2). In our method, parameters can be estimated or tuned, for instance, to incorporate a priori knowledge regarding texture proportions or strenght of interactions. Other methods such as Relaxation are much less readable in that sense.

Preliminary results are promising and illustrate a general methodology. It provides a statistical formalism to be investigated in other contexts. Future work would be to study its application for object recognition or more complex classes recognition. Before that, more specific analysis would be necessary as regards the choice of the neighborhood structure. In particular, the use of stronger geometric neighborhood relationships that take into account affine shape while preserving the maximum amount of invariance would worth additional investigation. Also the methodology presented here for feature vectors derived from interest points and spin images, could be investigated with other image description techniques.

References

[Ambroise *et al.*, 1997]C. Ambroise, V. Mo Dang, and G. Govaert. Clustering of spatial data by the EM algorithm. In Kluwer Academic Publishers Dordrencht, editor, *geoENV I- Geostatistics for Environmental Applications,Quantitative Geology and Geostatistics*, volume 9, pages 493–504, 1997.

[Celeux *et al.*, 2003]G Celeux, F. Forbes, and N. Peyrard. EM procedures using mean field-like approximations for Markov model-based image segmentation. *Pattern Recognition*, 36(1):131–144, 2003.

[Chandler, 1987]D. Chandler. *Introduction to Modern Statistical Mechanics.* Oxford University Press, 1987.

[Forbes and Peyrard, 2003]F. Forbes and N. Peyrard. Hidden markov random field model selection criteria based on mean field-like approximations. *IEEE trans. PAMI*, 25(8), 2003.

[Lazebnik *et al.*, 2003a]S. Lazebnik, C. Schmid, and J. Ponce. Affine-invariant local descriptors and neighborhood statistics for texture recognition. In *Proc. ICCV*, 2003.

[Lazebnik *et al.*, 2003b]S. Lazebnik, C. Schmid, and J. Ponce. Sparse texture representation using affine-invariant regions. In *Proc. CVPR*, 2003.

[Lindeberg and Garding, 1997]T. Lindeberg and J. Garding. Shape-adapted smoothing in estimation of 3-d depth cues from affine distorsions of local 2-d brightness structure. *Image and Vision Computing*, 15:415–434, 1997.

[Nigam *et al.*, 2000]K. Nigam, A. McCallum, S. Thrun, and T. Mitchell. Text classification from labeled and unlabeled documents using EM. *Machine Learning*, 39(2/3):103–134, 2000.

[Rosenfel *et al.*, 1976]A. Rosenfel, R. Hummel, and S. Zucker. Scene labeling by relaxation operations. *IEEE Trans. Systems, Man, and Cybernetics*, 6(6):420–433, 1976.

[Zhang, 1992]J. Zhang. The mean field theory in em procedures for markov random fields. *IEEE Trans. Signal Proc.*, 40(10):2570–2583, 1992.

Part XII

Queues and transportation

A multicriteria decision method to evaluate local transport service

Laura Grassini and Alessandro Viviani

Statistics Department
University of Florence
I-50134 Firenze, Italy
(e-mail: grassini@ds.unifi.it, viviani@ds.unifi.it)

Abstract. This paper deals with the use of multiple criteria decision methods to evaluate a number of bus routes operating in the territory of Florence, on the basis of a set of variables describing the effectiveness level of the service.

Keywords: Multiple criteria decision, Promethee methods, transport service.

1 Introduction

In Italy, legislative decree n. 422/1997 vested regional governments with the responsibility for programming and financing expenditure decisions. The Authority also expressed its preference for a less frequent recourse to the use of public franchise for local transport services in favour of a system of licenses or permits.

Under this law, a partial liberalization of the local public transport market and the renewing of fleet occurred. A number of new companies were born to manage lines pertaining to railway and road transportation.

According to these regulations, the vested organization (Region government) must be provided of a support tool in defining the organizational architecture of the local transport system. Moreover, the regulation of competition allows for achieving greater overall system efficiency, by offering an integrated service (tariff integration included) where multimodality can help in optimizing the use of the system. Hence, at first, it is necessary to set the commercial value of the running programme which is able to satisfy transport demand, on the basis of actual operative conditions.

According to the legislative decree n. 422/1997, in several phase concerning both planning and management of local public transport, it becomes necessary to make evaluations on only a part of the programmed service like the market value of a single transit line, for example, a bus route. In this respect, the paper presents the results of a statistical analysis aimed to provide a performance evaluation of the bus routes operating in the urban and suburban area of Florence.

At the first step, we considered a set of variables describing the level of effectiveness of the services. From the use of the multicriteria decision methods PROMETHEE II (**P**reference **R**anking **O**rganization **METH**od for

Enrichment **E**valuations), it is possible to derive an ordinal indicator for the bus routes. Outranking methods like PROMETHEE are decision support systems but they have also been used to rank alternatives in other kinds of problems. For example, they were used to evaluate the importance of a number of service attributes for the measurement of customer satisfaction [Franceschini and Rossetto, 1997]. In addition, PROMETHEE II is relatively simple in the involvement of criteria importance (weights) and in the computational procedure.

At the second step, we will briefly discuss the possibility to use the ordinal indicator provided by the PROMETHEE with a productivity index (Km./costs) to obtain an overall performance measure of a bus route.

The paper is structured as follows. In the next paragraph, we briefly describe the PROMETHEE approach, an outranking method for multicriteria decision problems. Finally, in the last two paragraphs, the empirical analysis is presented and discussed.

2 PROMETHEE decision methods

Some of the widely developed methods in the field of decision theory include utility theory, outranking methods and the Analytical Hierarchy Process [Gupta and Berger, 1994], [Roy, 1990]. Within these schools of thought there are many alternative approaches which correspond to different classes of problems, or different solution requirements. It is difficult to see how any one of these theories might become the best one, as each has its own advantages and disadvantages.

In this paper, we consider the outranking methods. These methods split the alternatives according to an *A is at least as good as B* hypothesis, and then explore the concordance and discordance using a decision algorithm.

A well known outranking method, that is also very intuitive and easy to use, is PROMETHEE, originally developed by Brans and Vincke [Brans and Vincke, 1985]. PROMETHEE allows a direct use of the data in a simple multicriteria table. Instead of having to perform a large number of comparisons, the decision-maker has to define his own scales of measure (without limitation), to indicate his priorities and his preferences for every criterion (by focusing on value, without having to worry about the method of calculation).

Let us consider two potential alternative A and B, and one evaluation criteria $f(.)$. Each single evaluation is expressed by $f(A), f(B)$ and gives a real number. This criterion may have to be minimized or maximized.

In order to rank the two alternatives, PROMETHEE requests additional information. For the criterion, a specific preference function must be defined. For example, we assume that the preference function *P(A,B)* is such that:

$$P(A,B) = \begin{cases} 0 & if \;\; f(A) \leq f(B) \\ p(f(A) - f(B)) & if \;\; f(A) > f(B) \end{cases} \qquad (1)$$

where $P(A, B)$ depends on the difference $f(A) - f(B)$. $p(.)$ is a function such that: if it is zero, A and B are indifferent choices; if it is close to zero, there is a weak preference for A; if it is close to 1, there is a strong preference for A; if the preference function is 1, there is a strict preference for A.

A wide used shape for a preference function is the linear form like, for example:

$$p(x) = \begin{cases} 1 & if \quad x > m \\ x/m & if \quad x \leq m \end{cases} \tag{2}$$

and $x = (f(A) - f(B)) \geq 0, m > 0$.

According with (2), the decision maker progressively prefers A over B for increasing differences $f(A) - f(B)$. The intensity of the preference progressively grows; when $x > m$ there is strict preference for A.

If there are k criteria and therefore k preference functions $p_i(A, B)$, $i = 1, ..., k$, different weights can be attached to different decision criteria. Such weights represent the importance of the different criteria in decision making.

These weights are used to derive the *outranking* index $\pi(A, B)$ of A over B, which is:

$$\pi(A, B) = \frac{\sum_{i=1}^{k} w_i p_i(A, B)}{\sum_{i=1}^{k} w_i} \tag{3}$$

This index provides a measure of the preference for A on B over all the criteria. As $0 \leq p_i(A, B) \leq 1$, expression (3) will assume values between 0 and 1.

In the case of n alternatives, PROMETHEE method calculates positive and negative preference flows for each alternative. The positive flow of A expresses how much the alternative A is dominating the others; the negative flow expresses how much it is dominated by the other ones. Positive and negative flows for the alternative A are expressed by the following formulas:

$$\phi^+(A) = \sum_b \pi(A, b) \tag{4}$$

$$\phi^-(A) = \sum_b \pi(b, A) \tag{5}$$

where the summation is over b, that is over the alternatives different from A. $\phi^+(A)$ expresses how much A outranks the other alternatives; $\phi^-(A)$ expresses how much the other alternatives outrank A.

The version labelled PROMETHEE II provides a complete ranking of the alternatives on the basis of the net flow:

$$\phi(A) = \phi^+(A) - \phi^-(A) \tag{6}$$

Therefore, we have:

- A outranks B iff $\phi(A) > \phi(B)$
- A and B are indifferent alternatives iff $\phi(A) = \phi(B)$

3 The case study: public transport in Florence

In this paragraph we describe the data used for the analysis and the main features of the Florence public transport system with special reference to ATAF (Azienda Trasporti Area Fiorentina) that is the main service provider.

Public transport in Florence in almost exclusively based on a system of bus routes.

ATAF operates with about 450 buses producing more than 18 million kilometers a year over a total route length of 450 km and serving a population of more than 580,000 units (inhabitants in the Municipality of Florence and other municipalities). Since 2001, some of the original ATAF suburban bus routes have been transferred to a new company (LI.NEA).

The percentage of regular users of ATAF service is about 40% of the served population. Of these, 67.2% are women and only 39.1% are occupied. The total population of bus users is characterized by a large presence of not-occupied individuals. For most of these, bus is the only transportation mean to move within the Florentine territory.

In 2000, a form of ticket integration was introduced for several of transportation providers (ATAF and other bus services, railways). Anyway, the use of train or non ATAF providers within the territory around Florence is rare. There is not an actual intermodal transport as the various transport modes are not efficiently integrated to provide a user-friendly service.

Data for the empirical analysis are derived from three sources.

- *ATAF database.* It provides the most important data related with the structure of the organization, the planned routes and terminals, the network system.
- *ATAF customer satisfaction survey.* It is a yearly CATI survey on the total served population (i.e. the inhabitants of the Municipality of Florence and of the other Municipalities served by ATAF), carried out to monitor mobility behavior. This data source provides information about the importance of some items describing the effectiveness of the service (i.e. the weights for the PROMETHEE analysis).
- *Interview of ATAF management staff.* This source provides information about the weights for the PROMETHEE analysis, from the managers' point of view.

4 The case study: results of the empirical analysis

In this section we describe variables, criteria and preference functions used for ranking a number of bus routes operating in the territory of Florence. We considered 16 bus routes, that resulted the most used (in 2002) from the ATAF customer survey. Moreover, in our analysis, we considered a total of 9 criteria, that cover some features of the transport service. Table 1 describes the data recorded for each route and the related optimality direction. The values related to each of the 16 bus transit lines are derived from internal agency data.

Criteria	Description	Users weights	ATAF weights
C1 min	Network length/N. stops	0.100	0.101
C2 max	N. bus shelters/N. stops	0.100	0.087
C3 max	N. stops with schedule information /N. stops	0.075	0.130
C4 max	N. of produced runs/N. planned runs	0.050	0.072
C5 max	Speed (Km/h)	0.125	0.116
C6 max	N. served municipalities/N. municip.s in the network	0.100	0.087
C7 max	Pollution limitations 0/3)	0.100	0.130
C8 max	Importance for tourism (ordinal 0/3)	0.100	0.130
C9 max	Number of passengers	0.250	0.145

Table 1. Criteria for the decision and related scaled weights

Table 1 shows the weight system adopted. Specifically, we considered two types of weights.

- ATAF weights: they are obtained through an interview to ATAF managers.
- Users' weights: for C1-C8, they are derived from the customer satisfaction survey described above; for C9, ATAF weight is attributed also to users.

We computed the mean of the evaluations (attributed by the respondents on a 10 points scale) about the importance of a number of services characteristics [Zeithaml *et al.*, 1990]. The weights have been proportionally scaled to sum up 1.

The PROMETHEE method also requires the specification of a preference function. In this application we adopted the linear form of the following type:

$$p(x) = \begin{cases} 0 & if \quad x \leq 0 \\ x/R & if \quad x > 0 \end{cases} \tag{7}$$

where R is the variation range of the criterium variable and x is expressed according to the maximization or minimization orientation of the criterion. Note that x/R gives a standardized value as requested by the function $p(x)$ in formula (2).

Before applying PROMETHEE method,we have conducted a principal component analysis (PCA) on the variables involved in the decision problem (Table 1). This analysis is useful to investigate the presence of any conflicting character of the criteria [Brans and Mareschal*et al.*, 1994]. To facilitate the interpretation of PCA results, the sign of C1 (which is *'min'* oriented) has been changed to negative.

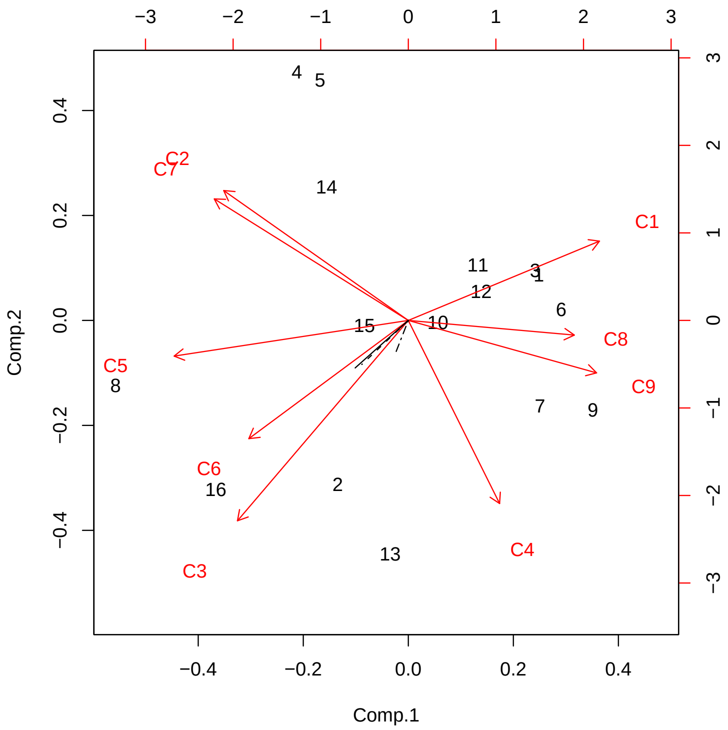


Fig. 1. Biplot from the PCA on criteria data (correlation matrix)

Figure 1 provides an approximate representation of the information related to this problem, because only 60% of the variance is reproduced by the first two principal components [Brans and Mareschal*et al.*, 1994]. We can see that some criteria (lines with arrows)are oriented in the opposite direction. That is the case, for example, of: C1 *vs* C6 and C3, C2 *vs* C7, C5 *vs* C8 and C9. Some cases are easy to be understood. The opposition of C5 against C8 and C9 is determined by the fact that bus routes serving the center of Florence are characterized by a strong importance for tourism (C8), are generally crowded (C9) and travel at a lower speed (C5). Viceversa occurs for buses travelling in suburban areas. The numbers in the figure label the 16 bus routes.

In a situation like the one represented in the biplot, the results of a multicriteria decision method could be quite sensitive to the weight system. Figure 1 shows also the projection of weights (dot-dash line: users weights, dashed line: equal weights, solid line: ATAF weights; the last two are partially overlayed). In the case of a decision problem, one should look at the alternatives located in the direction of the weights [Brans and Mareschal*et al.*, 1994].

In the case investigated in the paper, the projection of weights can give an approximate idea (because only 60% of variance is absorbed by the two components) of the compromise resulting in the ranking process and can allow a comparison among different systems of weights. We can see, for example, that Ataf and users weights are oriented almost in the same direction.

Table 2 contains the rank of the 16 bus routes, obtained through the application of PROMETHEE II method with, respectively: equal weights, users' weights and ATAF weights.

Table 2 contains also a label indicating the type of bus transit line: R means *radial* route, that links the center of the town with suburban sites; L indicates *longitudinal* suburban bus route, that links opposite suburban sites (for example, West-East, North-South, etc.). The radial routes are placed, on the average, at better rank position than longitudinal routes (mean rank larger than 9 *vs.* 7 of the longitudinal transit lines).

Though the presence of some conflicting character among the 9 criteria, the ranking obtained through different weights are similar. In fact, the rank correlation coefficient is 0.85, 0.88, 0.91 respectively for equal weights *vs* users' weights rankings, equal weights *vs* ATAF weights rankings, ATAF weights *vs* users' weights rankings.

Bus		Ranks			PCA	
lines	Type	Equal weights	Users weights	ATAF weights	Scores	Rank
1	L	10	9	10	0.5792	11
2	L	9	12	9	0,1447	9
3	L	13	13	13	1,2099	13
4	L	11	11	11	0,4738	10
5	R	12	10	12	-0,1094	8
6	L	14	14	14	1,6911	15
7	L	5	2	5	-0,4845	6
8	R	4	6	4	-1,8037	2
9	L	7	5	6	-0,6073	5
10	L	15	16	16	1,4800	14
11	R	16	15	15	1,9793	16
12	L	3	1	3	-1,3787	4
13	R	8	7	7	-0,4356	7
14	R	2	3	2	-1,5765	3
15	R	6	8	8	1,0921	12
16	R	1	4	1	-2,2546	1

L: longitudinal R: radial

Table 2. Results of PROMETHEE II method and subsequent PCA

In order to estimate the market value of a bus route, the effectiveness measure obtained through the PROMETHEE method is not sufficient because also economic features must be considered. In this respect, we have investigated the relationship between PROMETHEE ranks and the productivity indicator Km/costs, that is available for each bus route. If we consider PROMETHEE ranks as a quantitative variable and by using a negative sign for the variable km/costs (so that it is oriented in the same direction of PROMETHEE ranks), the correlation is 0.716. In this case, a scalar performance measure could be obtained through PCA. The first component absorbs more than 90% of variance and can summarize the effectiveness and productivity indicators. Table 2 shows the scores and the related ranks obtained from PCA.

5 Concluding remarks

The customer and user oriented approach requires the monitoring and measure of service's effectiveness. In this paper, effectiveness of bus routes operating in the territory of Florence is based on information derived from a customer satisfaction survey and internal agency data. A multicriteria decision approach (the PROMETHEE outranking method) has been used to derive a rank ordering of the different bus routes.

This ranking, together with a measure of productivity, has been used to provide a measure of overall performance for the bus routes. Of course, the use of PCA is only a compromise solution.

The empirical analysis here carried out shows a possible use of customer survey data and internal data in order to estimate the market value of the service. In particular, the PROMETHEE method could be a way to synthesize indicators of different nature and importance.

References

[Brans and Mareschal*et al.*, 1994]Brans J.P., B. Mareschal (1994), "PROMCALC & GAIA: A new decision support system for multicriteria decision aid", *Decision Support Systems*, 12, 297-310.

[Brans and Vincke, 1985]Brans J.P., P. Vincke (1985), "A preference ranking organization method: the Promethee method for multiple criteria decision making", *Management Science*, 31.6, 647-656.

[Franceschini and Rossetto, 1997]Franceschini F., S. Rossetto (1997), "La valutazione e il controllo in linea della qualita' dei servizi", *De Qualitate*, 1, 43-57.

[Gupta and Berger, 1994]Gupta S.S., J.O. Berger (1994), *Statistical decision theory and related topics*, Springer Verlag.

[Mareschal, 1988]Mareschal B. (1988), *Weight stability interval in multicriterion decision making*, the MIT Press.

[Roy, 1990]Roy, B. (1990), "Decision Aid and Decision Making", *European Journal of Operaitonal Research*, 45, 324-331.

[Zeithaml *et al.*, 1990]Zeithaml V.A., A.Parasuraman, L.L.Berry (1990), *Delivering Quality Service - Balancing Customer Perceptions and Expectations*, The Free Press.

Queues with server vacations in urban traffic control

Maria de Lurdes Simões[1], Paula Milheiro Oliveira[1], and Américo Pires da Costa[2]

[1] Faculdade de Engenharia da Universidade do Porto
DEC – CEC
Rua Dr. Roberto Frias, s/n P-4200-465 PORTO, Portugal
(e-mail: maires@fe.up.pt, poliv@fe.up.pt) Partially supported by PRODEP III-5.3

[2] Faculdade de Engenharia da Universidade do Porto
DEC – CITTA
Rua Dr. Roberto Frias, s/n P-4200-465 PORTO, Portugal
(e-mail: aapc@fe.up.pt)

Abstract. A queuing system resulting from a semaphorized intersection regulated by semi-actuated control in a network urban traffic is considered. Modelization of the queue length and of the delay of vehicles is crucial in the study of the performance of intersections equipped with traffic signals. In these systems, the server (green signal) is desactivated (red signal) during a random period of time. Due to this particularity, models for classic queues such as $M/M/1$, $M/G/1$ and $G/M/1$ are not appropriate. In the urban traffic literature, the frequent desactivation of the server as well as the variation of the service period are not well formulated. In the present work a $M/G/1$ queue where the server occasionally takes vacations and the service discipline is a non-gated time-limited policy is analyzed. The present analysis follows [Leung and Eisenberg, 1991] who consider an application of these models in telecommunications. Their implementation, given its complexity, is made possible by using Laguerre functions when looking for an approximate solution of the differential equations involved. One concludes that the mean delays of vehicles given by this model are slightly smaller than those obtained by simulation procedures, but they are able to give us a good approximation for larger flows, which is of interest for traffic engineers, since, in that case, the approximations one can find in the traffic literature are known not to be adequate.

Keywords: Queues, Server vacations, Traffic models.

1 Introduction

Waiting systems that admit interruptions of service often appear when the server uses idle periods of time of one queue or one task to serve clients in another queue or to perform another task. What matters is that, for these idle periods, the server is not available nor operational for new arrivals to the system (see *e.g.* [Doshi, 1986] for an interesting briefing on the subject). Among other applications these waiting systems appear in the literature as models for computer networks and telecommunications, production and quality control.

Models with interruptions of the server have been analyzed for different waiting systems, as the M/GI/1 or the GI/GI/1 queues with a single server, no restrictions existing on the arrivals process or the service time distribution, as long as the stationarity is maintained. In what regards the pause of the server, the model may fall into different classes, depending on the situations that trigger the pause (or vacation) and on the service policy, when the server returns from a pause and is available for service again.

In the context of urban traffic, modeling the queue length and the waiting time (delay) of vehicles is fundamental if one wants to study the performance of semaphorized intersections. Here we are concerned with semi-actuated intersections, which means that there are a main street and a secondary street and a sensor is placed in the secondary street, enabling the activation of the green signal and thus of the vehicles in this street to go through the intersection. The main difficulties involved in the analysis by means of the queuing theory come from the need of a good characterization of the circulating vehicles and drivers and from the fact that the desactivation of the server for random periods of time (red signal) has to be incorporated in the behavior of the queue. Due to this, essentially, the $M/M/1$, $M/G/1$ and $G/M/1$ models do not satisfactorily fit the waiting phenomena in these kinds of traffic intersections.

A detailed study of semaphorized intersections with a fixed period of green signal, which is not the case of semi-actuated signals, can be found in [Webster, 1958] where a formula of the delay of traffic which is much used in the traffic engineering practice is given. The traffic flow that reaches the intersection is assumed to follow a Poisson distribution and several parameters of the model are reduced to mean values which are obtained from the results of the $M/D/1$ and $M/D^X/1$ queues. Nevertheless, with such models, the regular but random desactivation of the served can not be well described. Indeed, as the signal alternates between red and green, modeling a semaphorized intersection is a problem lying in the class of queuing systems with server vacations [Doshi, 1986], with the particularity that the server remains inactive for random time durations. [Heidemann, 1994] proposes an analytic model that includes server vacations, starting from the assumption that the arrival process is Poissonian, that the intersection has a fixed cycle regulation, that the interval between departure of vehicles is constant and the traffic capacity is one way only. With these restrictions the probability generating functions for the measures of performance queue length and delay of a vehicle can be derived from the associated Markov chains. More recently [Alfa and Neuts, 1995] suggested the use of discrete time Markov arrival processes to describe the nature of platoons in the traffic flow.

In the present work a M/G/1 model for which the server occasionally takes a vacation and the service policy is non-gated time-limited is analyzed.

The term time-limited refers to the fact that the server is available to the queue for a maximum time duration at each visit (constant T_m). The term non-gated refers to the fact that clients that arrive while the server is active are candidates for service during this visit of the server in as much as the maximum service time T_m is not achieved. Clients are served in a FIFO regime and the server starts a vacation as soon as all clients in the queue are served or T_m expires, whatever occurs first. If the queue is empty when the server returns from a vacation it immediately starts a new vacation.

Our goal is to explore the theory of queues with server vacations, particularly the work by [Leung and Eisenberg, 1991], to find an approximate expression for the mean delay of a vehicle in the context of semi-actuated traffic using the comparison with the results obtained by numerical simulation of an intersection in [Simões *et al.*, 2002] to judge on the appropriateness of the proposed method.

2 An equation for the amount of work

For the class of models introduced above, the probability density function (pdf) of the amount of work at an arbitrary instant during a vacation period of the server is obtained by solving a functional equation that characterises the amount of work at the exact time the server starts a service period. Solving this equation, due to its complexity, is done by means of a numerical technique analogous to the one of [Weeks, 1966], based on the numerical inversion of the Laplace Transforms (LT).

The complementary of the distribution function of the duration of a service period (time between the beginning of service and the instant the queue becomes empty, assuming that T_m is never achieved) is approximated by a sum of Laguerre functions. Using the relation between the amount of work at the beginning of a service period and the duration of the server busy interval, the functional equation in transformed into a set of linear equations, from which the solution corresponds to the coefficients of the Laguerre functions in the expansion just mentioned.

Thus the amount of work in the queue at an arbitrary instant can be obtained from the equation that runs the amount of work at the instants the service starts serving the clients. From the decomposition of the amount of work and the PASTA property [Wolff, 1982], the mean waiting time can be deduced.

Notation:

$\bar{x}$, $\bar{x}^2$, $X^*(\cdot)$: mean, second moment and LT of the service time;

$\bar{\nu}$, $\bar{\nu}^2$, $V^*(\cdot)$: mean, second moment and LT of the duration of the vacation;

$\bar{u}_p$, $f_p(\cdot)$, $U_p^*(\cdot)$: mean, pdf and LT of the amount of work at the beginning of a service period;

$P_0(t)$: probability of the queue being empty at time t.

The following assumptions are made:

i) Clients arrive according to a Poisson process with parameter λ and the service time follows a general distribution for which the first two moments are finite;

ii) The system has an infinite waiting room;

iii) The system is in equilibrium and $\rho(=\lambda\bar{x}) < \dfrac{T_m}{T_m+\bar{\nu}}$;

iv) The duration of a vacation (random variable) is independent from the amount of work at the beginning of a service period.

The main theoretical result that we need when dealing with queues with server vacations is the stochastic decomposition property [Boxma and Groenendijk, 1987]: if the queue is in equilibrium, the LT of the amount of work at the beginning of a service period may be written as the product of the LT of the amount of work at the end of a service period, $U^*(s,T_m)$, by the LT of the amount of work that arrives during a vacation, $U_v^*(s)$. Making use of this property the major difficulty in the analysis of models that have a limited service time lies in the characterization of the amount of work at an arbitrary instant during a vacation period. In order to overcome this difficulty performing the following steps is required:

i) Set up the functional equation that characterizes the amount of work at the beginning of a service period. The stochastic decomposition property states that

$$U_p^*(s) = U_v^*(s)\cdot U^*(s,T_m)\,. \tag{1}$$

On the other hand one has $U_v^*(s) = V^*(\lambda - \lambda X^*(s))$ and

$$U^*(s,T_m) = e^{\hat{s}T_m}\left\{U_p^*(s) - \hat{s}\int_{y=0}^{T_m} e^{-\hat{s}y}P_0(y)dy\right\}\,, \tag{2}$$

where $\hat{s} = s - \lambda + \lambda X^*(s)$.

ii) Equation (2) is solved numerically, given that $1 - P_0(t)$ can be approximated by a weighted sum of Laguerre functions:

$$P_0(t) \overset{\text{sim}}{=} 1 - \sum_{n=0}^{N} a_n e^{-\frac{t}{2T}} L_n\left(\frac{t}{T}\right)\,.$$

Thus

$$P_0^*(s) \overset{\text{sim}}{=} 1 - \sum_{n=0}^{N} a_n \frac{s\left(s-\frac{1}{2T}\right)^n}{\left(s+\frac{1}{2T}\right)^{n+1}}\,. \tag{3}$$

The LT $U_p^*(s)$ is also approximated by means of Laguerre functions:

$$U_p^*(s) \overset{\text{sim}}{=} 1 - \sum_{n=0}^{N} a_n \frac{\hat{s}\left(\hat{s}-\frac{1}{2T}\right)^n}{\left(\hat{s}+\frac{1}{2T}\right)^{n+1}}\,. \tag{4}$$

The approximations given in (3) and (4) are used in equation (2), from which, using (1), one gets:

$$U_p^*(s) \overset{\text{sim}}{=} U_v^*(s).e^{\hat{s}T_m} \left\{ e^{-\hat{s}T_m} + U_p^*(s) - \sum_{n=0}^{N} a_n e^{-(\hat{s}+\frac{1}{2T})T_m} L_n\left(\frac{T_m}{T}\right) \right.$$

$$\left. - \int_{y=0}^{T_m} e^{-\hat{s}y} \sum_{n=0}^{N} a_n e^{-y/2T} \left[\frac{1}{2T} L_n\left(\frac{y}{T}\right) - L_n'\left(\frac{y}{T}\right)\right] dy \right\} \quad (5)$$

Notice that this equation is linear in the a_n for a given s with $Re(s) \geq 0$.

iii) The functional equation (5) is transformed into a linear system of equations, since, by taking $s = i\omega$ and using $N+1$ appropriated values for ω in equation (5), a set of $N+1$ linear equations is obtained (see [Weeks, 1966]). The coefficients a_n are known by solving this system.

iv) To end with, by using the decomposition of the amount of work and the PASTA[1] property [Wolff, 1982], the mean amount of work in the system as seen by a Poisson arrival is given by

$$\bar{u} = \frac{\lambda \overline{x^2}}{2(1-\rho)} + \sum_{n=0}^{N} (-1)^n (2T)(1-\rho) a_n - \rho\bar{\nu} + \rho\frac{\overline{\nu^2}}{2\bar{\nu}}\,. \qquad (6)$$

The mean waiting time of a client is obtained by applying Little's formula.

3 Application to the control of semi-actuated traffic

As mentioned in the introduction, traffic signals with semi-actuated regulation are frequently used in intersections which consist of a main street and a secondary street. The actuated phase serves the movement of vehicles in the secondary street. The control variables that lead the efficiency of a semi-actuated operation are the regulation plan of the semaphore and the placement of the sensor. The difficulty in applying the semi-actuated control is in the selection of an optimum combination of these operations. In the absence of a service call (non activation of the sensor) the green signal is always given to the non-actuated phase. As soon as the sensor is activated a change in the signals occurs. The time interval for this change to occur includes a yellow period followed by a period of “all red” (cleaning time). During the activation of the sensor the arrival of a vehicle in the actuated street extends the interval of green signal of this phase by an amount of time so that the minimum of green time is exceeded but not the maximum. It means that, in semi-actuated traffic, the green time is adapted to the demand, having a minimum and a maximum value. In this way, a larger number of vehicles is able to pass through the intersection per unit of time.

[1] *Poisson Arrival See Time Average*

In the present work the intersection illustrated in Fig. 1 is considered. The sensor is placed $5\,m$ away from the stopping line of the secondary street. The times given to the regulation of the two phases are shown in Table 1.

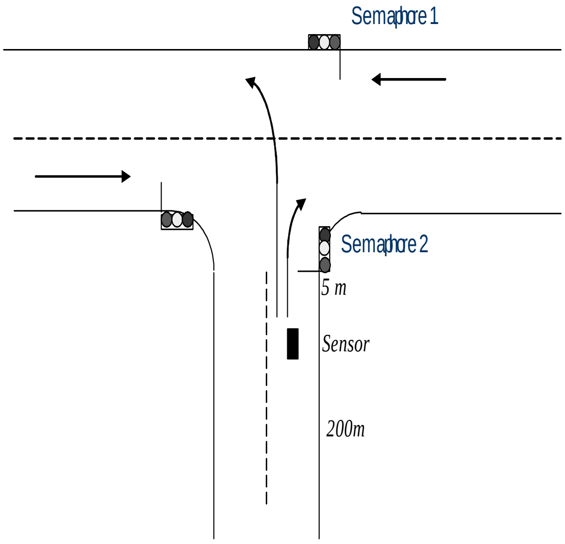


Fig. 1. Scheme of a semi-actuated intersection.

Table 1. Times given to the regulation of the two phases.

Time (sec.)	Semaphore 1	Semaphore 2
Green	20 to ∞	7 to 40
Yellow	3	3
Extension of green	–	4
All red	2	2

The degree of saturation, $x_{sat} = \rho \frac{T_m + \bar{\nu}}{T_m}$, represents the ratio between the mean number of vehicles that arrive during a cycle and the maximum number of vehicles that may pass through the intersection during that period of time. In the terminology of the queuing systems this parameter is known as the congestion index.

The mean waiting times estimated by the model presented in Section 2 (referred to as the analytical model) are shown in Fig. 2 as well as the average delays experienced by drivers according to the simulation (see [Simões *et al.*, 2002] for the detailed simulation study). A Dirac function with a mass at the point 27 and a Gaussian distribution with mean value equal to 2 and variance 0.04 are considered in the analytical model as the laws of the duration of a vacation (red period) and of the service time, respectively, since these are the best fit distributions in the case of semi-actuated urban traffic intersections.

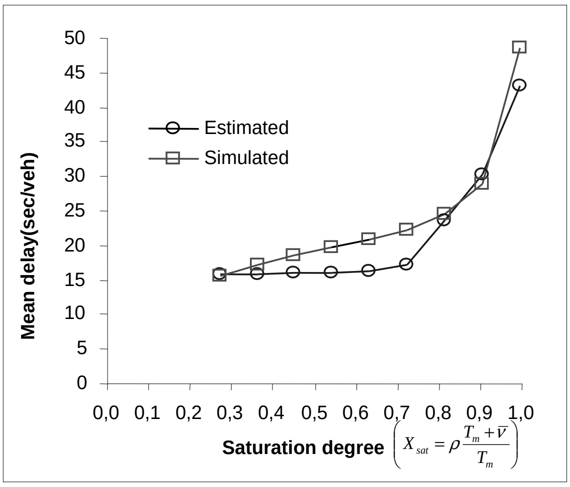


Fig. 2. Comparison between the mean delay estimated by the analytical model and the simulated mean delay ($\bar{\nu} = 27$ sec., $\bar{x} = 2$ sec. and $T_m = 43$ sec.).

The results suggest that, for approximately $x_{sat} > 0.7$, the analytical model gives good estimates of the mean delay of drivers. For $0.3 < x_{sat} < 0.7$, however, the estimates given by this model are smaller than those obtained by numerical simulation. This fact may be due to the diversity of reactions that is typical of drivers behavior and of interactions between vehicles but, most of all, the fact is that the duration of a vacation (red signal) is not really bounded, since it is extended until the activation of the sensor, which means it has no maximum value although it has a minimum.

It is important to remark that when dealing with the analytical model one should be aware of the importance of choosing adequate values for N and of the need of a high precision in the computations, as the numerical method explained here is very sensitive to precision errors. Difficulties in making these numerical procedures converging are also reported in the literature[Leung and Eisenberg, 1991] in the case of probability density functions with jumps or discontinuities (service times or durations of the vacations that are deterministic). In practice it is very much recommended to validate the outputs of the numerical procedures ensuring that the amplitudes of the a_n's are smaller than 10^{-8}.

4 Final comments

An analytical expression for the evaluation of an approximation of the mean delay of vehicles in semi-actuated traffic was found by applying systems of queues with server vacations theory, while previous expressions were known to be inappropriate for the semi-actuated case.

This procedure gives good approximations when the arrival flow is large, which was not possible with heuristic expressions commonly used in traffic engineering that had been developed for the fixed control case. The expressions that we give here provide realistic estimates of the mean delay particularly when the saturation index is below 70%, while for large traffic flows (congestion scenarios) the estimates they provide appear to be smaller than the real mean delays.

Having in mind improving the reliability of the results presented here and others that will be obtained in the future, the numerical properties of the relationship between N and T deserves a careful investigation, aiming to establish, for different distributions of the service durations, which values should be given to N and T in order to ensure good results when this method is applied.

References

[Alfa and Neuts, 1995]A.S. Alfa and M.F. Neuts. Modelling vehicular traffic using the discrete time Markovian Arrival Process. *Transportation Science*, 29(2), pages 109–117, 1995.

[Boxma and Groenendijk, 1987]O.J. Boxma and W.P. Groenendijk. Pseudo-conservation laws in cyclic-service systems. *J. Appl. Prob.*, 24, pages 949–964, 1987.

[Doshi, 1986]B.T. Doshi. Queueing systems with vacations - A survey. *Queueing Systems*, 1, pages 29–66, 1986.

[Heidemann, 1994]D. Heidemann. Queue length and delay distributions at traffic signals. *Transportation Research-B*, 28(5), pages 377–389, 1994.

[Leung and Eisenberg, 1991]K.K. Leung and M. Eisenberg. A single-server queue with vacations and non-gated time-limited service. *Performance Evaluation*, 12, pages 115–125, 1991.

[Simões *et al.*, 2002]M.L. Simões, P.M. Oliveira, and A.P. Costa. Análise probabilística do fluxo de tráfego num cruzamento semi-actuado. *Actas do IX Congresso Anual da SPE*, pages 111–124, 2002.

[Webster, 1958]F.V. Webster. *Traffic Signal Settings.* Road Research Laboratory, Road Research Laboratory 39, HMSO, London, 1958.

[Weeks, 1966]W.T. Weeks. Numerical inversion of Laplace transforms using Laguerre functions. *J. ACM*, 13, pages 419–426, 1966.

[Wolff, 1982]R.W. Wolff. Poisson arrivals see time averages. *Operations Research*, 30, pages 223–231, 1982.

Stability of the two queue system

Iain M. MacPhee and Lisa J. Müller

University of Durham
Department of Mathematical Science
Durham, DH1 3LE, UK
(e-mail: `i.m.macphee@durham.ac.uk, l.j.muller@durham.ac.uk`)

Abstract. We describe ergodicity and transience conditions for a general two queue system with multiple service regimes, a dedicated traffic stream for each queue, a further stream which can be routed to either queue and where completed jobs can be fed back into the queues. There is only one class of jobs but the service times and feedback probabilities depend upon the configuration of the servers. Several different levels of control of the service regimes are considered. We use the semi-martingale methods described in [Fayolle *et al.*, 1995] and our results generalise those of [Kurkova, 2001].

Keywords: controlled queue systems.

1 Introduction

In this paper we consider a system which has two queues with servers that can be configured in several ways. Our main aim is to identify conditions under which we can give a queue length dependent policy for choosing the service configurations that guarantees the stability of the system.

The queues have independent Poisson arrival streams with rates λ_i, $i = 1$, 2 and there is an independent Poisson arrival stream with rate λ of jobs that can be sent to either queue (we will call this the *routeable* stream). We assume all jobs are of the same class and are served in the order they join their queues but their service times depend upon their queue and the service scheme in force while they are being served. Under server configuration k, at most one job is in service at each non-empty queue and all jobs in queue i have independent, exponentially distributed service times with mean μ_{ki}^{-1}, $i = 1$, 2 (so the server configuration k and the destination of a routeable job determines its service distribution). We label the server configurations by $k = 1, \ldots, K$, the queue to which the routable stream is directed by $j = 1$, 2 so that the finite set $\mathcal{R}$ of overall *management regimes* has members $\eta = (k, j)$. In addition the system has Jackson-type feedback with probabilities that depend upon the current management regime. Any job that completes service at queue i under regime η independently enters queue i' with probability $p_{ii'}^{\eta}$, $i' = 1$, 2 or leaves the system with probability $p_{i0}^{\eta} \equiv 1 - p_{i1}^{\eta} - p_{i2}^{\eta} \geq 0$. We will assume throughout this paper that we can instantaneously switch between different management regimes at the instants just after changes to queue lengths.

Example The model described above includes as a special case the model with two servers, where server i can be used to process jobs at either queue which it does at rate μ_i. This gives four service regimes $s_1 s_2 = 1\,2,\ 2\,1,$ $1\,1$ and $2\,2$ (i.e. server 1 at queue 1, server 2 at queue 2; server 2 at queue 1, server 1 at queue 2; both servers at queue 1; both servers at queue 2). Given that the service rates are additive we get the pairs (μ_1, μ_2), (μ_2, μ_1), $(\mu_1 + \mu_2, 0)$ and $(0, \mu_1 + \mu_2)$ respectively.

The question we consider is whether for such a system with a given set of parameters, the management regime can be changed from time to time to ensure that the queue lengths remain stable or whether the queue lengths must grow indefinitely regardless of how the system is managed.

Similar systems but with fixed servers have been studied in the past using transform methods, often under strong symmetry assumptions on parameters, see [Feng *et al.*, 2002] and [Foley and McDonald, 2001] who give stability conditions for an n-dimensional JSQ model and carry out the large deviations analysis of system occupancy for the two dimensional system.

We define the model we consider in section 2 and state our results in section 3. We omit the proofs here to be able to describe the model in full length. The proof is done using the semi-martingale methods described in [Fayolle *et al.*, 1995] and can be found in [MacPhee and Müller,]. Our results generalise those of [Kurkova, 2001] as we consider multiple service regimes and do not require any symmetry.

2 Definitions

We now define the queueing system, its control, and the classes of control policies that we wish to investigate.

2.1 Events, blocks and control policies

As the Lyapunov function results we use are described in terms of discrete processes it is convenient to study a discrete time process which we now describe. To simplify comparison of the process dynamics under different management regimes we *uniformise* the continuous time jump process, following Serfozo [Serfozo, 1979], by choosing a constant $\rho \geq \max_k\{\lambda + \lambda_1 + \lambda_2 + \mu_{k1} + \mu_{k2}\}$ and introducing a fictitious *bell* event which has exponential inter-event times with rate $\rho - (\lambda + \lambda_1 + \lambda_2 + \mu_{k1} + \mu_{k2})$ at any given queue lengths when regime (k, j) is used (so the total event rate has the same value ρ in all states under all regimes). We now consider the uniformised discrete time process Ξ on state space $\mathbf{Z}_0^2 \equiv \{(x, y) \in \mathbf{Z}^2 : x \geq 0, y \geq 0\}$, obtained by considering the queue lengths at bell events, arrival times of new jobs and at service completions and consequent re-entry to queues. We will use $\alpha = (x, y) \in \mathbf{Z}_0^2$ to denote a typical state vector for Ξ.

It is also necessary to define the policies by which the management regimes at each state are selected. Our main interest will be in policies which choose the same regimes over large sets of states, specifically cone shaped blocks for which we need some notation. Let e_i denote the unit vector in the axis i direction and for non-zero $z \in \mathbf{R}^2$ let $|z|$ denote the length of z and $\arg_u(z)$ the argument relative to non-zero vector $u \in \mathbf{R}^2$ (the angle anticlockwise from u to z). For any non-zero u, $v \in \mathbf{R}^2$ let $\ell(u) = \{z \in \mathbf{R}^2 : z = tu, t > 0\}$ denote the half-line in the direction u and

$$\mathcal{C}(u,v) \equiv \{z \in \mathbf{R}^2 : |z| > 0,\ 0 < \arg_u(z) < \arg_u(v)\} \tag{1}$$

the cone swept anticlockwise from direction u to direction v. The closure of such a cone will be denoted $\bar{\mathcal{C}}(u,v)$. We give specific labels to the positive parts of the axes, $\mathcal{A}_i \equiv \ell(e_i)$ as we will consider them as blocks subsequently. It will also be convenient to define two special versions of the argument, one relative to each axis $\mathcal{A}_i$. Let $R : \mathbf{R}^2 \to \mathbf{R}^2$ be reflection in the line $z_1 = z_2$ i.e. $R(z_1, z_2) = (z_2, z_1)$ and define

$$\arg_1(z) = \arg_{e_1}(z)\ , \qquad \arg_2(z) = \arg_1\bigl(R(z)\bigr) \tag{2}$$

so $\arg_2(z)$ is the angle measured clockwise from e_2 to z.

A policy for controlling this discrete event system is a sequence $\Pi = \{\pi_n : n \geq 0\}$ of transition probabilities π_n from $\mathcal{H}_n$, the process history at time n, to $\mathcal{R}$, the set of regimes i.e. for any history α_0, η_0, $\ldots$, α_{n-1}, η_{n-1}, α_n the next action is selected according to the distribution $\pi_n(\alpha_0, \eta_0, \ldots, \alpha_n, \cdot)$. This definition includes non-stationary, non-Markov randomised policies though they offer no performance benefits when applied to stationary Markov processes, see e.g. Blackwell [Blackwell, 1965]. Let $\xi_i(n)$ denote the length of queue i at time n and $\xi(n) = (\xi_1(n), \xi_2(n))$. A policy Π along with an initial distribution for the queues determines a stochastic process $(\Xi, \Pi) = \{(\xi(n), \eta_n) : n \geq 0\}$ which will only be Markov when $\pi_n(\alpha_0, \eta_0, \ldots, \alpha_n, \cdot)$ is a distribution dependent only on α_n.

A policy Π which selects an action $a(\alpha)$ with probability 1 whenever the system state is α, where a is a map from $\mathbf{Z}_0^2$ to $\mathcal{R}$, is a deterministic stationary policy. Our main interest is in a class of these that we call *block pure policies*, denoted Π^b, where the state space $\mathbf{Z}_0^2$ is partitioned into a small number of disjoint blocks, always lines or cones, such that a is constant on each block $\mathcal{C}(u,v)$. We also investigate a generalisation of these, *block randomised* policies, denoted Π^r, where for each block the distribution $\pi_n^r(\alpha, \cdot)$ is the same at every state α in the block (so the Π^b are degenerate cases of the Π^r). With such policies the process (Ξ, Π^r) is Markov due to our assumptions about Poisson arrivals and exponential service times.

2.2 The queues and their mean drifts

The process (Ξ, Π) has bounded jumps, specifically $\pm e_i$ and $\pm(e_2 - e_1)$ and so all moments of its jump distributions exist under any policy but in this two dimensional case our results can be stated in terms of their first moments. For each regime η let

$$M^\eta = \mathbf{E}(\xi(n+1) - \xi(n) \mid \mathcal{H}_n, \pi_n = \eta) \tag{3}$$

denote the *mean drift* vector for any period when the policy selects regime η. We have, for $k = 1, \ldots, K$ at states $\alpha \in \mathbf{Z}^2_+ \equiv \{(x,y) \in \mathbf{Z}^2 : x > 0, y > 0\}$

$$\begin{aligned} M^\eta &= (M_1^\eta, M_2^\eta) \\ &= \begin{cases} \rho^{-1}\big(\lambda + \lambda_1 + \mu_{k2}p^\eta_{21} - \mu_{k1}p^\eta_{10},\ \lambda_2 + \mu_{k1}p^\eta_{12} - \mu_{k2}p^\eta_{20}\big),\, \eta = (k,1) \\ \rho^{-1}\big(\lambda_1 + \mu_{k2}p^\eta_{21} - \mu_{k1}p^\eta_{10},\ \lambda + \lambda_2 + \mu_{k1}p^\eta_{12} - \mu_{k2}p^\eta_{20}\big),\, \eta = (k,2) \end{cases} \end{aligned} \tag{4}$$

It is convenient to assume that when queue i is empty the policy selects a regime η chosen from among those with $\mu_{ki} = 0$ (this is equivalent to having non-idling servers). This ensures that equation (4) is also correct for histories ending in states $\alpha \in \mathcal{A}_1 \equiv \{(x,0) : x > 0\}$ and $\alpha \in \mathcal{A}_2 \equiv \{(0,y) : y > 0\}$ for such service regimes. We will sometimes use the notation M' and M'' to denote mean drifts for the system under appropriate regimes for $\mathcal{A}_1$ and $\mathcal{A}_2$ respectively.

Now consider any policy Π allowing randomisation. The mean drift of our process Ξ under Π when the current state is $\alpha \in \mathbf{Z}^2_+$ is a 2-dimensional vector M^Π lying in the convex set

$$\mathcal{M} = \left\{ \sum_\eta p_\eta M^\eta \,:\, p_\eta \in [0,1] \text{ and } \sum_\eta p_\eta = 1 \right\} \tag{5}$$

the convex hull of the regime mean drifts. The extreme points of $\mathcal{M}$ are a subset of the regime mean drifts M^η. When three or more of the M^η are distinct it may happen that the two-dimensional interior,

$$\mathrm{Int}_2(\mathcal{M}) \equiv \{z \in \mathcal{M} : B(z,\epsilon) \subset \mathcal{M} \text{ for some } \epsilon > 0\},$$

(where for $z \in \mathbf{R}^2_+$, $B(z,\epsilon) = \{z' \in \mathbf{R}^2 : |z - z'| < \epsilon\}$) is non-empty.

3 Classification of the system

The behaviour of the system depends on whether the convex set $\alpha + \mathcal{M}$ can be separated from the origin by a line through α. Any set of parameters for the process (Ξ, Π) falls into one the following four exclusive cases:

C1 $(0,0) = \underline{0} \notin \mathcal{M}$ and there exists a state $\alpha \in \mathbf{Z}^2_+$ and a line

$$L_v(\alpha) \equiv \{\beta \in \mathbf{R}^2 : v^T(\beta - \alpha) = 0\} \tag{6}$$

with normal vector v through α separating $\alpha + \mathcal{M}$ from the origin $\underline{0}$. If there exists one such $\alpha \in \mathbf{Z}^2_+$ then there is an infinite cone of such α.

C2 $\underline{0} \notin \mathcal{M}$ and there exists no $\alpha \in \mathbf{Z}_+^2$ and line $L_v(\alpha)$ which separates $\alpha + \mathcal{M}$ from $\underline{0}$.

C3 $\mathrm{Int}_2(\mathcal{M})$ is non-empty, $\underline{0} \in \mathcal{M}$ and there exists no $\alpha \in \mathbf{Z}_+^2$, $v \in \mathbf{R}^2$ such that the line $L_v(\alpha)$ separates $\alpha + \mathrm{Int}_2(\mathcal{M})$ from the origin.

C4 $\underline{0}$ is a boundary point of $\mathcal{M}$ and either $\mathrm{Int}_2(\mathcal{M}) = \emptyset$ or the tangent line to $\alpha + \mathcal{M}$ through α separates the origin from $\alpha + \mathrm{Int}_2(\mathcal{M})$ for each α in a cone within $\mathbf{Z}_+^2$.

See Figure 1 for examples of C1-C4.

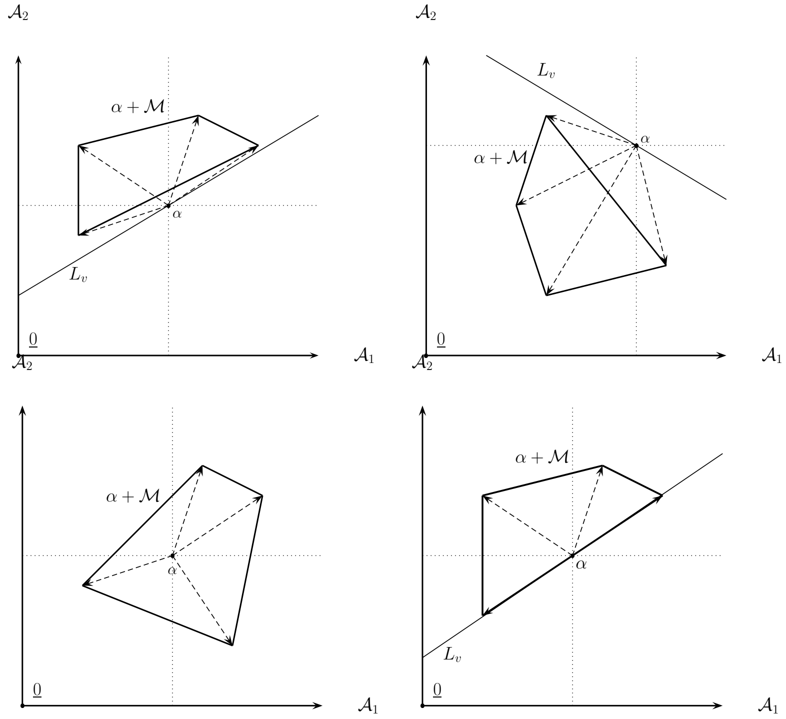


Fig. 1. From top left: C1, C2, and below C3, C4.

Note: the cases in C4 are critical but we will say very little about them in this paper.

We start stating our results by giving sufficient conditions for instability or stability respectively of the system under fully randomised controls in cases C1 and C2 respectively. Next we show that in case C3 there is always a block pure policy that makes $(\varXi, \varPi^b)$ ergodic and we also show that randomisation allows the use of fewer blocks. Finally in this section we consider some situations with even lower levels of control.

3.1 Fully randomised controls

The following two results apply when even the most general policy Π is used to control the queueing system. They imply that in cases C1 and C2 the control policy used does not affect the stability or otherwise of the process.

Theorem 1 *If $\underline{0} \notin \mathcal{M}$ and there exists an $\alpha \in \mathbf{Z}_+^2$ and $v \in \mathbf{R}^2$ such that the line $L_v(\alpha)$, see (6), separates $\alpha + \mathcal{M}$ from the origin $\underline{0}$ then the process (Ξ, Π) is unstable, in the sense that the total number of queued jobs almost surely goes to ∞ linearly in time for any policy Π.*

The conditions of the theorem can be pictured in an alternative way. Specifically there exists a state $\alpha \in \mathbf{Z}_+^2$ such that the line segment from $\underline{0}$ to α does not intersect $\alpha + \mathcal{M}$ (it follows that if there is any such pair α, v then there is an infinite cone of points α' such that $L_v(\alpha')$ separates $\underline{0}$ and $\alpha' + \mathcal{M}$).

Theorem 2 *If $\underline{0} \notin \mathcal{M}$ and there is no $\alpha \in \mathbf{Z}_+^2$, $v \in \mathbf{R}^2$ such that $L_v(\alpha)$ separates $\alpha + \mathcal{M}$ from $\underline{0}$ then (Ξ, Π) is stable, in the sense that the total number of queued jobs remains bounded in mean, under every policy Π.*

The alternative description of the conditions here is that for every $\alpha \in \mathbf{Z}_+^2$ the line segment joining $\underline{0}$ to α intersects $\alpha + \mathcal{M}$. From this it follows there is some $v \in \mathbf{R}_+^2$ such that $\underline{0}$ and $\alpha + \mathcal{M}$ are in the same halfspace created by $L_v(\alpha)$.

3.2 Block controls

In case C3 it does make a difference which policy is used for running the system. In fact we can show that block pure policies Π^b with at most a handful of blocks are adequate to ensure stability of the process. Under policies of this type the process (Ξ, Π^b) is Markov so we can now talk about ergodicity and transience.

Theorem 3 *If $\underline{0} \in Int_2(\mathcal{M})$ then there is a block pure policy Π^b with at most five blocks such that the Markov chain (Ξ, Π^b) is ergodic.*

Theorems 2 and 3 imply the following result.

Corollary 1 *If $\underline{0}$ is a boundary point of $\mathcal{M}$, $Int_2(\mathcal{M})$ is non-empty and there exists no $\alpha \in \mathbf{Z}_+^2$, $v \in \mathbf{R}^2$ such that $L_v(\alpha)$ separates $\alpha + Int_2(\mathcal{M})$ from $\underline{0}$ then there is a policy Π^b with at most three blocks such that (Ξ, Π^b) is ergodic.*

In Theorem 3 the number of blocks required to achieve ergodicity can be reduced if block randomised policies Π^r are used.

Corollary 2 *If $\underline{0} \in Int_2(\mathcal{M})$ and a block randomised policy Π^r is used then at most four blocks are necessary to ensure that (Ξ, Π^r) is ergodic.*

Example [Foley and McDonald, 2001] consider a model which has fixed servers (their service rate drops to 0 when their queues are empty), no feedback and is strictly JSQ. Their stability criterion for $N = 2$ queues is that $\rho_{\max} \leq 1$ where

$$\rho_{\max} = \max\{\lambda_1/\mu_1,\ \lambda_2/\mu_2,\ (\lambda + \lambda_1 + \lambda_2)/(\mu_1 + \mu_2)\ \}.$$

For the policy which sends the routable stream to the queue with minimum weighted work our model has two regimes depending upon where the routable traffic is sent and these have drift vectors

$$M^1 = \tfrac{1}{\rho}(\lambda + \lambda_1 - \mu_1, \lambda_2 - \mu_2) \text{ and } M^2 = \tfrac{1}{\rho}(\lambda_1 - \mu_1, \lambda + \lambda_2 - \mu_2).$$

As $\rho(M^2 - M^1) = (-\lambda, \lambda) \perp (1, 1)$ the line segment joining these two drift vectors has the form $z_1 + z_2 = (\lambda + \lambda_1 - \mu_1 + \lambda_2 - \mu_2)/\rho$ which can only intersect $\mathbf{R}^2_-$ when $\rho_{\max} < 1$. The case $\rho_{\max} = 1$ is critical and we see that our conditions are equivalent to those of Foley and McDonald in this case.

The simplicity of the classification based on the convex hull $\mathcal{M}$ confirms that this geometrical approach combined with the Lyapunov function method is a natural technique for studying stability of multi-queue systems though of course large deviations results like those [Foley and McDonald, 2001] are not achievable thisway.

3.3 Low levels of control

The results of [Fayolle *et al.*, 1995] can also be used to classify the process for any control policy that is block homogeneous for any small number of blocks. It soon becomes evident to anybody who attempts this that there are many ways for the process to remain stable and many more for it to be transient. To illustrate this we now spell out the possible behaviours of the queueing system with four blocks, specifically the axes $\mathcal{A}_1$, $\mathcal{A}_2$ and two cones, $\mathcal{C}_1 = \mathcal{C}(e_1, d) \cup \ell(d)$ and $\mathcal{C}_2 = \mathcal{C}(d, e_2)$ (see (1) for this notation), that partition $\mathbf{Z}^2_+$. The two cones are not assumed to be symmetric i.e. the vector $d \in \mathbf{R}^2_+$ need not be parallel to (1, 1).

We assume that in each of the $\mathcal{A}_i$ and $\mathcal{C}_i$, $i = 1, 2$ a single management regime is used (different blocks may have a common regime) with mean drift vectors M^1, M^2 in blocks $\mathcal{C}_1$, $\mathcal{C}_2$ respectively and M', M'' in blocks $\mathcal{A}_1$, $\mathcal{A}_2$ respectively. This assumption about and notation for the regimes on the $\mathcal{A}_i$ we will use in all further sections but the $\mathcal{C}_i$ are specific to this section.

We first label the M^i according to the angles φ_i they make relative to the axes $\mathcal{A}_i$, $i = 1, 2$. For each M^i angle $\varphi_i = 0$ is in the direction of $\mathcal{A}_i$ and φ_1 increases clockwise while φ_2 increases anticlockwise i.e. $\varphi_i = 2\pi - \arg_i(M^i)$. We label the directions of the M^i as **A** when $0 < \varphi_i < \pi$, **B** when $\pi \leq \varphi_i \leq \frac{3\pi}{2}$ and **D** when $\frac{3\pi}{2} < \varphi_i \leq 2\pi$. The various cases of this model are labelled with *label of* M^1/*label of* M^2 so a label **B**/**A** means M^1 has a positive y and a negative x component and M^2 has x component negative with y of either sign. Figure 2 illustrates this labelling scheme for the directions of the M^i from origins α_i.

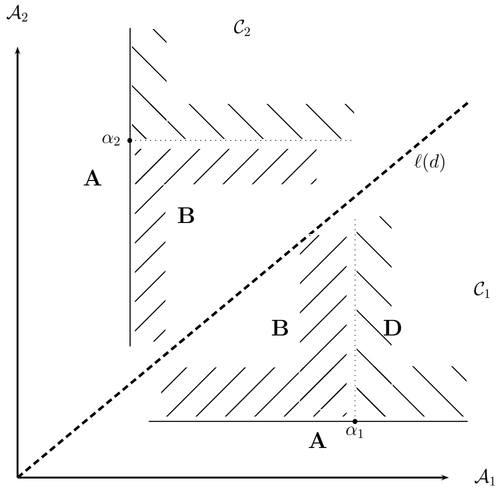


Fig. 2. Graphical explanation of the labels.

From the results in [Fayolle *et al.*, 1995] on the random walk in the positive quadrant, we have (in their terminology): (i) if a drift M^i has an $\mathbf{A}$ label then axis $\mathcal{A}_i$ is an *ergodic* face; (ii) that face will be *outgoing*, *ingoing* or *neutral* according to the sign of the *second vector field* (which is scalar in this case); (iii) if M^i has a $\mathbf{B}$ or $\mathbf{D}$ label then face $\mathcal{A}_i$ is transient and there is no second vector field. In this two dimensional case the sign of the second vector field depends only upon the angles of M' and M^1 for $\mathcal{A}_1$, M'' and M^2 for $\mathcal{A}_2$. As with the angles φ_i it is convenient to name the angles $\psi_1 = \arg_1(M')$, $\psi_2 = \arg_2(M'')$ that M', M'' make relative to axes $\mathcal{A}_1$, $\mathcal{A}_2$ respectively, so $\psi_i = 0$ is in the $\mathcal{A}_i$ direction and ψ_1 increases anticlockwise while ψ_2 increases clockwise. Now, following the sign of the second vector field, we modify the labels for M^i, $i = 1, 2$ to

$$\mathbf{A}^+ : \varphi_i + \psi_i < \pi, \quad \mathbf{A}^- : \varphi_i + \psi_i > \pi, \quad \mathbf{A}^0 : \varphi_i + \psi_i = \pi \tag{7}$$

Using this labelling system we can identify 25 different cases to deal with. It turns out that in many of the cases we get the same result for all choices of the two cones i.e. all slopes $d' \equiv d_2/d_1 \in (0, \infty)$ of the line $\ell(d)$ separating them. Theorem 4 classifies these invariant cases.

Theorem 4 *The system is*

(1) ergodic in cases $\mathbf{A}^-/\mathbf{A}^- \cup \mathbf{B}$, $\mathbf{B}/\mathbf{A}^-$, $\mathbf{B}/\mathbf{B}$ *with* $\left|\frac{M_x^1}{M_y^1}\right| > \left|\frac{M_x^2}{M_y^2}\right|$

(2) transient in cases $\mathbf{A}^+/\mathbf{A}\cup\mathbf{B}\cup\mathbf{D}$, $\mathbf{A}\cup\mathbf{B}\cup\mathbf{D}/\mathbf{A}^+$, $\mathbf{B}/\mathbf{B}$ *with* $\left|\frac{M_x^1}{M_y^1}\right| < \left|\frac{M_x^2}{M_y^2}\right|$, $\mathbf{D}/\mathbf{B}$, $\mathbf{B}/\mathbf{D}$, $\mathbf{D}/\mathbf{D}$;

(3) null recurrent in cases $\mathbf{A}^0/\mathbf{A}^0 \cup \mathbf{A}^+ \cup \mathbf{B}$, $\mathbf{A}^+ \cup \mathbf{B}/\mathbf{A}^0$, $\mathbf{B}/\mathbf{B}$ *with* $\left|\frac{M_x^1}{M_y^1}\right| = \left|\frac{M_x^2}{M_y^2}\right|$.

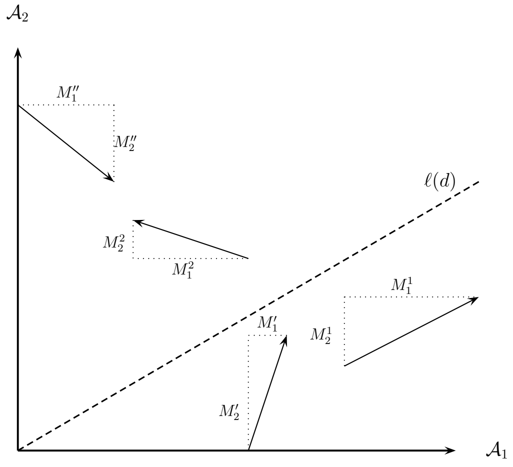


Fig. 3. Example of case $\mathbf{D}/\mathbf{A}^-$ where $\ell(d)$ is important.

For systems with no control over the service regimes there still may be some control over the routable traffic stream. The next theorem shows that there are sets of parameters such that a change to the slope of the switching line $\ell(d)$ can change Ξ from a transient to an ergodic process. We describe in detail only the case $\mathbf{D}/\mathbf{A}^0 \cup \mathbf{A}^-$, depicted in Fig. 3, as case $\mathbf{A}^0 \cup \mathbf{A}^-/\mathbf{D}$ is very similar. The relative slopes of M^1, $\ell(d)$ and M^2 are crucial so we label two key conditions:

E1: $M_2^1 < d' M_1^1$ (so $\ell(d)$ is steeper than M^1); E1$'$: $M_2^1 > d' M_1^1$;
E2: $-M_2^2 \leq d'(-M_1^2)$ (including cases with $M_2^2 \geq 0$ and implies $-M^2$ is not steeper than $\ell(d)$).

Theorem 5 *In case $\mathbf{D}/\mathbf{A}^0 \cup \mathbf{A}^-$ the ergodicity or non-ergodicity of the Markov chain Ξ also depends on the slope $d' > 0$ of the line $\ell(d)$ separating $\mathcal{C}_1$ and $\mathcal{C}_2$ as follows:*

(a) if E1 holds then Ξ is transient,
(b) if E1$'$ holds then Ξ's excursions into $\mathcal{C}_1$ have finite mean time and Ξ is
   (i) ergodic if E2 holds and M^2 is $\mathbf{A}^-$ or if E2 does not hold and $(-M_2^2)M_1^1 < M_2^1(-M_1^2)$ (so M^1 is steeper than $-M^2$);
   (ii) null recurrent if E2 holds and M^2 is $\mathbf{A}^0$ or if E2 does not hold and $(-M_2^2)M_1^1 = M_2^1(-M_1^2)$;
   (iii) transient if E2 does not hold and $(-M_2^2)M_1^1 > M_2^1(-M_1^2)$.

The case $\mathbf{A}^0 \cup \mathbf{A}^-/\mathbf{D}$ is simply the reflection of the above in the line $\ell(1,1)$.

Note: this theorem says nothing about the cases where M^1 is parallel to $\ell(d)$ but in practice this will not be a major problem if the slope of the line $\ell(d)$ is under user control.

References

[Blackwell, 1965]D. Blackwell. Discounted dynamic programming. *Annals of Mathematical Statistics*, pages 226–235, 1965.

[Fayolle *et al.*, 1995]G. Fayolle, V. A. Malishev, and M. V. Menshikov. *Topics in the Constructive Theory of countable Markov Chains.* Cambridge University Press, Cambridge, 1995.

[Feng *et al.*, 2002]W. Feng, K. Adachi, and A. Kowada. A two-queue and two-server model with a threshold-based control service policy. *European Journal of Operational Research*, pages 593–611, 2002.

[Foley and McDonald, 2001]R. D. Foley and D. R. McDonald. Join the shortest queue: stability and exact asymptotics. *Annals of applied Probability*, pages 569–607, 2001.

[Kurkova, 2001]I. A. Kurkova. A load-balanced network with two servers. *Queueing Systems*, pages 379–389, 2001.

[MacPhee and Müller,]I.M. MacPhee and L.J. Müller. Stability classification of a controlled multi-queue system with many service and routing regimes. *submitted to Queueing Systems.*

[Serfozo, 1979]R.F. Serfozo. An equivalence between continuous and discrete time markov decision processes. *Operations Research*, 27:616–620, 1979.

Part XIII

Reliability and Survival Analysis

Diagnostic tests for frailty

P. Economou and C. Caroni

Department of Mathematics
School of Applied Mathematics and Physical Sciences
National Technical University of Athens
9 Iroon Polytechniou, Zografou
155 80 Athens, Greece
(e-mail: polikon@math.ntua.gr, ccar@math.ntua.gr)

Abstract. A common way of allowing heterogeneity between individuals in models for lifetime data is to introduce an unobservable individual random effect Z. In a proportional hazards framework, the individual's hazard becomes $zh_b(t)$ where $h_b(t)$ is the baseline hazard. The random variable Z is often assumed to follow the Gamma or Inverse Gaussian distribution. We develop here diagnostic tests for these assumptions. One simple graphical diagnostic is based on the form of the unconditional survival function when $h_b(t)$ is assumed to be Weibull. Another plot uses a closure property of a family of frailty distributions, which implies that the frailty among survivors at time t has the same form as the original distribution of Z, with the same shape parameter but different scale parameter. In this method, we estimate the shape parameter at different times t and examine graphically whether it is constant. We give simulation results and examples to illustrate these methods.

Keywords: Lifetime data, frailty, proportional hazards, Burr distribution, Generalized Inverse Gaussian distribution.

1 Introduction

When modelling data obtained from time-to-event studies, it is often found that there is heterogeneity between individuals, over and above what can be accounted for by any available covariates. One common way of allowing for this heterogeneity is to introduce an unobservable individual random effect Z, the so-called frailty. This is usually assumed to operate in a proportional hazards framework, so that it acts multiplicatively on the baseline hazard function h_b which is common to all individuals. Thus the hazard function for an individual with frailty $Z = z$ is given by

$$h(t|z) = zh_b(t)$$

If there are also measured covariates, the model is usually extended to

$$h(t|z; X) = ze^{\beta^{'}\mathbf{x(t)}}h_b(t)$$

where $\mathbf{x}(.)$ is a q-dimensional vector of possibly time-dependent covariates.

Introducing heterogeneity in lifetime data by means of an unobserved quantity in this way was initiated by [Clayton, 1978], [Vaupel *et al.*, 1979] and [Hougaard, 1984]. The distribution of the random variable Z is often assumed to be Gamma [Vaupel *et al.*, 1979] or Inverse Gaussian [Hougaard, 1984]. As with any part of the process of statistical modelling, it is desirable to check that the assumed distribution is supported by the data. The purpose of the present paper is to develop diagnostic plots for the frailty distribution, with the emphasis on these two common choices, the Gamma and Inverse Gaussian. We will be assuming that the baseline hazard function h_b has been specified correctly and that the multiplicative proportional hazards frailty model is the proper one to describe the data.

2 Diagnostic plots for mixtures

From the proportional hazards assumption, it follows that the conditional survivor function for an individual with frailty z is

$$S(t|z) = [S_b(t)]^z$$

where S_b is the baseline survivor function. In particular, if the baseline model is taken to be Weibull (η, β), then the survivor function conditional on frailty z is

$$S(t|z) = exp(-zs)$$

where $s = (t/\eta)^\beta$. If Z has distribution function G on $(0, \infty)$, then the unconditional survivor function is given by

$$\int_0^\infty exp(-zs)dG(z)$$

If G is taken to be Gamma with shape and scale parameters both equal to ν (so that the mean is one), then

$$S(t) = \left(1 + \frac{s}{\nu}\right)^{-\nu} \tag{1}$$

(the Burr distribution), while taking G to be Inverse Gaussian with scale 1 and shape λ yields the survivor function

$$S(t) = exp\left(\lambda\left(1 - \sqrt{\frac{2s}{\lambda} + 1}\right)\right) \tag{2}$$

(In both cases, a constraint has been applied to the parameters of G to make the model identifiable. Other choices of constraint are possible but make no difference in principle.)

We now look for appropriate diagnostic plots to enable us to check that the assumption of one of these distributions is in fact correct. The idea is to plot some function of the non-parametric Kaplan-Meier estimate of survival, $\hat{S}(t)$, against some function of t, to obtain the characteristic shape associated with the distributions (1) and (2).

Taking logarithms of (1), and supposing that s/ν is large enough so that $log(1 + s/\nu) \approx log(s/\nu)$, we see that $-log\hat{S}(t)$ against $logt$ should give a straight line.

[Wolstenholme, 1999] suggests this plot for the Pareto distribution. This is the special case of the Burr distribution when the baseline hazard is exponential, which is a special case of the Weibull ($\beta = 1$). However, we observe that the above approximation is poor for the early failures. These give the plot a characteristic horizontal section, whose length depends on ν, disappearing as ν becomes small (high degree of heterogeneity) (Figure 1). When the frailty distribution is the Inverse Gaussian, taking logarithms of (2) suggests the plot of $log(-log\hat{S}(t))$ against $logt$ if λ is large. Under these circumstances, there is only a small degree of heterogeneity and the distribution will not be very different from the baseline Weibull distribution. This is the standard diagnostic plot used to check for the Weibull distribution. It gives a straight line with slope equal to the shape parameter β. On the other hand, suppose that λ is large. It is easy to show that in this case the Weibull(η, β) - Inverse Gaussian(1,λ) mixture tends to the Weibull with scale parameter $\eta/(2\lambda)^{1/\beta}$ and shape $\beta/2$. Thus the same plot gives a straight line with slope $\beta/2$. For intermediate values of λ, the plot should be curved with slope falling from β to $\beta/2$ as time increases. Examples are shown in Figure 2.

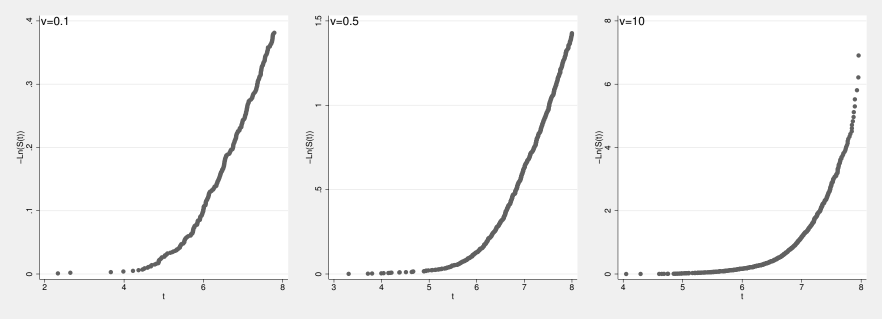


Fig. 1. Diagnostic plot for Burr distribution for various values of the parameter ν. 1000 simulations of samples of size 1000, baseline Weibull parameters 1000 (scale) and 2 (shape). Type I censoring at t=3000.

3 Closure property of the frailty distributions

To develop another kind of diagnostic plot, we start with a closure property of Gamma frailty in the proportional hazards model ([Vaupel *et al.*, 1979]).

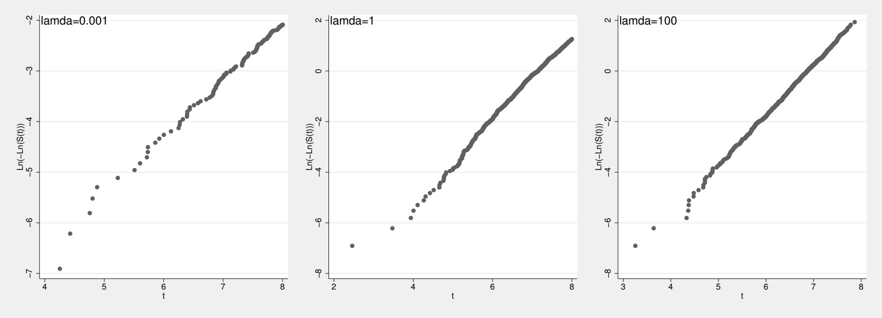


Fig. 2. Diagnostic plot for Weibull-Inverse Gaussian mixture for various values of the parameter λ. Details of simulations as in Figure 1.

Given that the frailty distribution among all individuals is Gamma with scale parameter κ and shape parameter λ, then the frailty distribution among the population of survivors at time t is again Gamma with the same shape parameter λ but different scale parameter given by $\kappa + H_b(t)$, where $H_b(t)$ is the cumulative baseline hazard function. This property can be generalized to a whole family of distributions.

3.1 Generalization of the Gamma frailty property

The Gamma frailty property given by [Vaupel *et al.*, 1979] can be generalized first to the case of available covariates. More specifically, given that the frailty distribution is Gamma(κ, λ), then the frailty distribution among survivors at time t, conditional on the value of the covariates, is Gamma $(\kappa + H_b^{\mathbf{x}}(t), \lambda)$, where $H_b^{\mathbf{x}}(t)$ is given by

$$H_b^{\mathbf{x}}(t) = \int_0^t e^{\ \beta' \mathbf{x}(u)} h_b(u) du.$$

The closure property is not a characterization of the Gamma distribution only. It is quite easy to show that a similar property holds also for the Inverse Gaussian and the Generalized Inverse Gaussian (GIG). Furthermore, a similar property holds for a whole class of distributions that belong to the exponential family [Hougaard, 1984]. Let frailty Z be a random variable with distribution $F(\alpha)$ on $(0, \infty)$, where α is the parameter vector, with p.d.f. of the form

$$f_z(z) = \frac{e^{-[z, g(z)][\eta_1(\alpha), \eta_2(\alpha)]'}}{\Phi(\alpha)} \ \xi(z)$$

which is an exponential family distribution with canonical statistics z and $g(z)$ [Shao, 1998]. For this frailty distribution the following theorem holds, which extends Hougaard's result by including covariates.

Theorem 1 *Given the frailty distribution $F(\alpha)$ with p.d.f. as above, then under the proportional hazards frailty model the frailty distribution among survivors at time t is again $F(.)$. The value of $\eta_1(\alpha)$, the element of the parameter vector corresponding to z, changes, but the components of $\eta_2(\alpha)$ do not. More specifically, the p.d.f. of frailty among survivors at time t is given by*

$$f_{Z_{|T>t}}(z) = \frac{e^{-[z,g(z)][\eta_1^*(\alpha),\eta_2(\alpha)]'}}{\Phi^*(\alpha)}\ \xi(z)$$

where $\eta_1^*(\alpha) = \eta_1(\alpha) + H_b^{\mathbf{x}}(t)$ and $\Phi^*(\alpha) = \Phi(\alpha)S_T(t)$.

The GIG distribution (and hence the Gamma and Inverse Gaussian distributions which it includes) belongs to the above class of the exponential family. Unfortunately, some other distributions, like the lognormal, which are also widely used as frailty distributions, do not belong to this class because they do not have z as a canonical statistic. This obstacle can be overcome by considering a generalized distribution, adding one more parameter [Hougaard, 1986] which will be zero initially. So, the Theorem can be applied to all distributions $F(\alpha)$ with p.d.f given by

$$f_z(z) = \frac{e^{-T(z)[\eta(\alpha)]'}}{\Phi(\alpha)}\ \xi(z)$$

where $T(z)$ does not contain z as a component, since the above distribution can be seen as a special case of $GF(\alpha,\beta)$ with p.d.f. given by

$$f_z(z) = \frac{e^{-[z,T(z)][\beta,\eta(\alpha)]'}}{\Phi_G(\alpha,\beta)}\ \xi(z)$$

for $\beta = 0$. $\Phi_G(\alpha,\beta)$ is the integral over the range of z, $R(z)$, of the numerator of the previous relationship. Applying the Theorem to the distribution $GF(\alpha,0)$ shows that the frailty distribution among the survivors at time t will be again GF but with parameter vector given by $(\alpha, H_b^{\mathbf{x}}(t))$.

3.2 Plots

Given that the frailty distribution has been chosen correctly, then our above Theorem shows that the vector $\eta_{\mathbf{2}}(\alpha)$ of the initial parameters does not change when we restrict our attention to the frailty distribution among those units that have survived until time t. Let $\hat{\eta}_{\mathbf{2i}}(\alpha)_{|T>t}$ denote component i of the maximum likelihood estimate of this vector among the survivors at time t. If our assumption of the frailty distribution is correct, then $\hat{\eta}_{\mathbf{2i}}(\alpha)_{|T>t}$ for any time t is an asymptotically unbiased estimator of the same quantity $\eta_{\mathbf{2i}}(\alpha)$. Therefore, a plot of $\hat{\eta}_{\mathbf{2i}}(\alpha)_{|T>t}$ against time should give a straight line parallel to the time axis.

For the Gamma(κ, λ) and Inverse Gaussian(κ, λ) distributions of frailty this plot reduces to $\hat{\lambda}_{|T>t}$ versus t since $\eta_{\mathbf{2}}(\kappa, \lambda) = \lambda$. For the GIG($\lambda, \delta, \gamma$), the proposed plots are of $\hat{\lambda}_{|T>t}$ and $\hat{\delta}_{|T>t}$ against t, since $\eta_{\mathbf{2}}(\lambda, \delta, \gamma) = (\lambda - 1, \delta^2)$ for this distribution.

In all cases, the maximum likelihood estimates of the model's parameters that are required for the plots are obtained by maximising the logarithm of the usual likelihood function for lifetime data ($t_i : i = 1, 2...n$)

$$L = \prod_{i=1}^{n} \left\{ h(t_i)^{\delta_i} S(t_i) \right\}$$

where δ_i is the censoring indicator which takes the value 1 if t_i is an observed lifetime and zero if it represents a right censored observation. The expressions for $S(t)$ are given above in (1) and (2) for the Gamma and Inverse Gaussian frailty distributions, respectively, and the hazard function $h(t)$ can be obtained as minus the derivative of $logS$. We first carry out this estimation using all the data. Then we select a sequence of convenient time points τ_j ($j = 1, 2...k$) and repeat the estimation k times, using in the ith estimation only those data points t_i satisfying $t_i \geq \tau_j$.

3.3 Simulations

To illustrate the method, we simulated a set of 1000 uncensored data points from the Burr distribution (Weibull-Gamma mixture) and produced the above plot based on repeated estimates of ν. Then we fitted the incorrect Weibull-Inverse Gaussian mixture to the same data and produced the corresponding plot (Figure 3). Next we repeated the exercise with the roles of the two distributions reversed. Thus we generated a set of data from the Weibull-Inverse Gaussian mixture and produced the plots for both the correct model and for the incorrect Burr distribution (Figure 4). In both cases, the plots

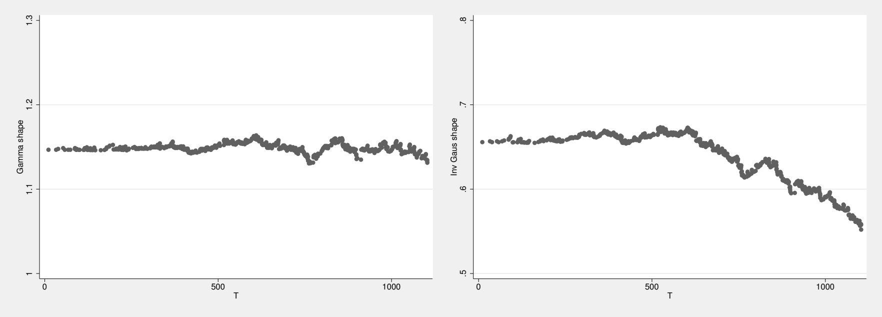


Fig. 3. Plot defined in Section 3.2, fitting (left) correct Burr distribution, (right) incorrect Weibull-Inverse Gaussian mixture to data generated from Burr ($\nu = 1$).

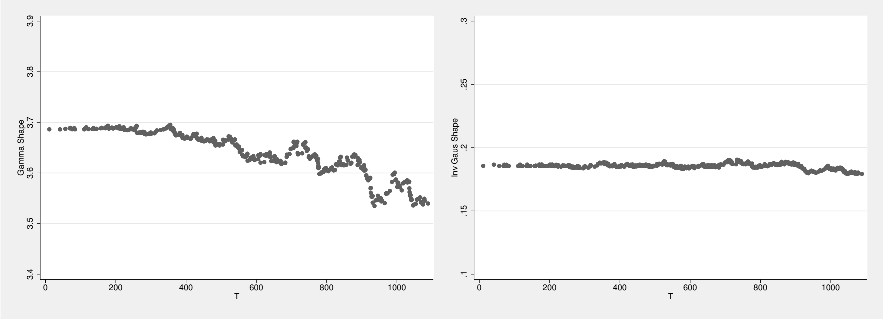


Fig. 4. Plot defined in Section 3.2, fitting (left) incorrect Burr distribution, (right) correct Weibull-Inverse Gaussian mixture to data generated from Weibull-Inverse Gaussian mixture ($\lambda = 0.5$).

discriminate extremely well between the two frailty distribution; the plot for the correct distribution is a horizontal straight line as predicted, but the plot for the incorrect distribution departs clearly from a horizontal line.

4 Example

For a real-data illustration of our methods, we used data on the duration of a treadmill test undertaken by 978 successive patients at a cardiac clinic in Athens. Figure 5 shows the simple diagnostic plots that were developed in Section 2. The plot for the Burr distribution, on the right, has the expected shape of a straight line preceded by a horizontal section. The plot for the Weibull-Inverse Gaussian mixture, on the left, is curved as expected, but the curvature is greater than it should be if this is the correct model. Figure 6 shows the diagnostic plots that were developed in Section 3. These indicate

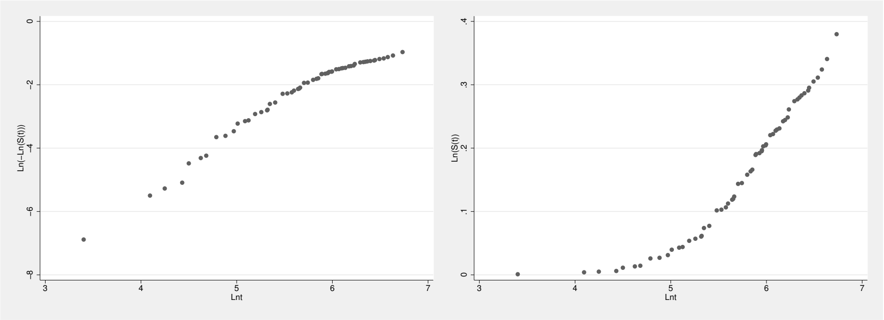


Fig. 5. Plots defined in Section 2 for data on 978 cardiac patients. Left: Weibull-Inverse Gaussian mixture; right: Burr.

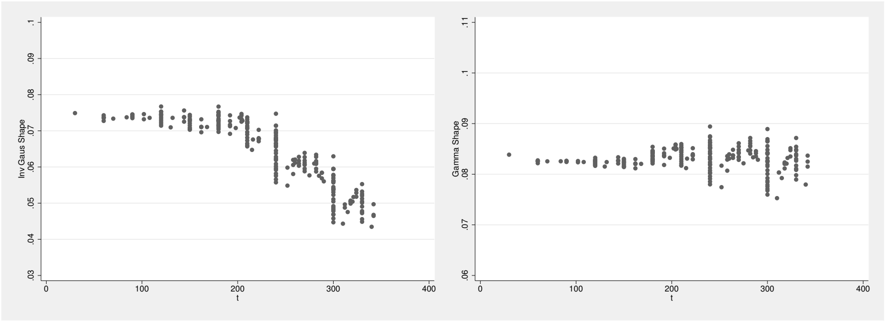


Fig. 6. Plots defined in Section 3 for data on 978 cardiac patients. Left: Weibull-Inverse Gaussian mixture; right: Burr.

clearly that the assumption of a Gamma distribution for frailty is acceptable, because the estimates of its shape parameter at different times are scattered about a horizontal line, but the Inverse Gaussian assumption is not.

5 Further research

Although graphical diagnostics have proved very useful in statistical modelling, it can also be valuable to have formal statistical tests for the presence of any frailty, thus showing whether or not it is necessary to use the models considered here. In these models, one parameter of the distribution controls both the presence and the degree of frailty. For example, when the frailty distribution is Gamma(ν, ν), the parameter ν controls the amount of frailty since $V(Z) = 1/\nu \to 0 \quad (\nu \to \infty)$. If the basic distribution is Weibull and the unconditional distribution is therefore Burr, the presence of any frailty can be examined by testing the null hypothesis $\nu = \infty$ (or $1/\nu = 0$) by likelihood-based methods applied to the Burr distribution. The theory for one of these methods, the score test, was given by [Crowder and Kimber, 1997] for multivariate lifetime data and the details for the univariate case which we are interested in by [Kimber, 1996]. (In fact, the test also holds for other Weibull mixtures, not just for Weibull-Gamma = Burr.) Other likelihood-based tests that can be applied include a Wald test and a likelihood ratio for this parameter. A difficulty that arises is that the null value of the parameter being tested falls on the boundary of the parameter space. In such cases, the distribution of minus twice the log likelihood ratio is not given by the usual chi-squared approximation. Instead, a mixture of chi-squared distributions usually applies. We intend to complete a study of the likelihood ratio test for this model and then carry out a simulation study to compare the properties of the different tests in order to recommend the best one for use in practice.

References

[Clayton, 1978]D.G. Clayton. A model for association in bivariate life tables and its application in epidemiological studies of familial tendency in chronic disease incidence. *Biometrika*, 65:141–151, 1978.

[Crowder and Kimber, 1997]M. Crowder and A. Kimber. A score test for the multivariate burr and other weibull mixture distributions. *Scandinavian Journal of Statistics*, 24:419–432, 1997.

[Hougaard, 1984]P. Hougaard. Life table methods for heterogeneous populations: Distributions describing the heterogeneity. *Biometrika*, 71:75–83, 1984.

[Hougaard, 1986]P. Hougaard. Survival models for heterogeneous populations derived from stable distributions. *Biometrika*, 73:387–396, 1986.

[Kimber, 1996]A. Kimber. A weibull-based score test for heterogeneity. *Lifetime Data Analysis*, 2:63–71, 1996.

[Shao, 1998]J. Shao. *Mathematical Statistics*. Springer-Verlag, New York, 1998.

[Vaupel *et al.*, 1979]J.A. Vaupel, K.G. Manton, and E. Stallard. The impact of heterogeneity in individual frailty on the dynamics of mortality. *Demography*, 16:439–454, 1979.

[Wolstenholme, 1999]L.C. Wolstenholme. *Reliability Modelling: A Statistical Approach*. Chapman and Hall/CRC, Boca Raton, 1999.

On Thresholds of Moving Averages With Prescribed On Target Significant Levels

A. R. Soltani, S. A. Al-Awadhi, and W. M. Al-Shemeri

Department of Statistics and OR
Faculty of Science, Kuwait University
P.o. Box 5969, Safat 13060 Kuwait
(e-Email: soltani@kuc01.kuniv.edu.kw, alawadi@kuc01.kuniv.edu.kw)

Abstract. Let $X_1, X_2 ...$ be a sequence of i.i.d random variables representing successive inputs to the moving average process,

$$Y_n = \frac{1}{K} \sum_{i=0}^{K-1} X_{n-i}.$$

The Y_n is off target by X_n if it exceeds a threshold. By introducing a two states Markov chain, we define "*on target significant level*" and establish a technique for evaluating the threshold corresponding to a prescribed on target significant level. It is proved that in such circumstances for exponential and normal inputs, the threshold is a linear function in the mean μ_{X_1}, where slops and intercepts are also specified. These relationships can be easily applied for estimating the thresholds.

Keywords: Moving Average, Threshold, On target significant level..

1 Introduction

Let $X_1, X_2, \cdots$ be a sequence of independent and identically distributed random variables, and let Y_n be the corresponding left sided moving average, as defined in the abstract. In practice, the input sequence $\{X_n\}$ may represent successive loads, excess loads, rain falls, water supply in successive periods, service time to the nth arrival, etc; and the moving averages are processes indicating accumulations of certain number of immediate prior inputs. Thus by taking into account $K-1$ immediate prior inputs to the nth input, the cumulative value corresponding to the n^{th} input is $\sum_{i=0}^{K-1} X_{n-i}$, $n = K, K+1, K+2, \cdots$, and $Y_n = \frac{1}{K} \sum_{i=0}^{K-1} X_{n-i}$, $n = K, K+1, K+2, \cdots$ is a sequence of moving averages. The process Y_n is off or on target at the commencement of the arrival of the $(n+1)^{th}$ input if $Y_n > L$ or $Y_n \leq L$ respectively. The threshold L is non-random and is considered as a parameter. Our aim in this article is to specify, or estimate, L so that the moving averages remains $(1-a)\%$, $0 < a < 1$, of times on target.

We prove that the status, off or on target, is indeed a two state Markov chain, and derive formulas for the transition probabilities in terms of the distribution of the inputs. This allows to define a prescribed "on target significant level" for the moving averages, and then proceed to introduce a method to achieve the aim. We have examined our method for exponential or normal inputs. Interestingly in these cases L turns out to be linear in the mean of the distribution of the inputs, μ_{X_1}. Point estimation and interval estimation can be easily established using the derived linear relationships.

The methodology and results presented in this article, we believe, can be applied in Reliability, Control Theory, System Assessments, and Hydrology. Moving averages are classical tools in time series, stochastic processes and scan statistics; and are basis for many linear and nonlinear models. Moving averages, in the content presented here, had not been treated in other works, to the best of the authors' knowledge. The threshold of moving averages, considered in this article, is different from the threshold moving average which is a nonlinear model, [G. and Gooijer, 1998]. Two-state Markov chains, in contents different from the one presented in this article, have been employed by different authors as underlying probability models of various hydrology events, [Vogel, 1987]. The works [Banifacio and Salas, 1999] and references therein are rich in providing applications of these types of probability techniques to hydrology data.

2 A Markov Chain

Let $X_1, X_2, \cdots$, and Y_n be as defined in the Introduction, Define

$$V_n = \begin{cases} 0, & Y_n > L \\ 1, & Y_n \leq L \end{cases}, n = K, K+1, \cdots$$

We recall that the situation $V_n = 0$ indicates that Y_n is off target by X_n, while $V_n = 1$ indicates that it is not. We prove below that $\{V_n\}$ is indeed a Markov chain and provide its transition probabilities.

Lemma 1. The process $V_n,\ n = K, K+1, \cdots$, is a Markov chain with transition probabilities.

$$P_{00} = \frac{\int_{-\infty}^{+\infty} [1 - F(KL - t)]^2 f_{T_{K-1}}(t) dt}{1 - F_{T_K}(KL)}, \quad K \geq 1, \tag{2.1}$$

$$P_{11} = \frac{\int_{-\infty}^{+\infty} [F(KL - t)]^2 f_{T_{K-1}}(t) dt}{F_{T_K}(KL)}, \quad K \geq 1, \tag{2.2}$$

where F is the distribution of X_1, and $T_K = X_1 + X_2 + ... + X_K,\ T_0 = 0$.

The Lemma 1 can be deduced through classical techniques in probability, so its proof is omitted here. By using the transition probabilities, the stationary distribution of the Markov Chain $\{V_n\}$ is easily given by

$$\pi_0 = \frac{P_{10}}{P_{10} + P_{01}}, \quad \pi_1 = \frac{P_{01}}{P_{10} + P_{01}}, \tag{2.3}$$

[Karlin and Taylor, 1998]. The return period of the state 0 and state 1 are respectively $m_{00} = \frac{1}{\pi_0}$, $m_{11} = \frac{1}{\pi_1}$, which specify the duration of successive visits to these states. Other duration are measured by $m_{01} = \frac{1}{1-P_{00}}$, $m_{10} = \frac{1}{1-P_{11}}$.

Now we are in a position to define "on target significant level".

Definition 1.1. We call the $(1-a)\%$ the "on target significant level" of the moving average process $\{Y_n\}$, where $a = \pi_0$ is the stationary probability of the state 0 of the Markov chain $\{V_n\}$.

3 Exponential And Normal Inputs

In this section we establish a relationship between the threshold L and the mean of the distribution of inputs, whenever the distribution is exponential or normal.

Let us assume loads $X_1, X_2, \cdots$ are i.i.d. exponentially distributed with parameter λ, $E(X_1) = 1/\lambda$. The following theorem specifies the appropriate threshold for the moving average to possess the on target $(1-a)\%$ significant level.

Theorem 3.1. If inputs $X_1, X_2, \cdots$ follow exponential distribution with parameter λ, then the least value L for the threshold to ensure $(1-a)\%$ on target significant level for the moving average Y_n is given by

$$L = \frac{\theta(a,K)}{K}(\frac{1}{\lambda}), \qquad (3.1).$$

where $\theta(a,K)$ is the positive solution to the equation

$$\pi_1(\theta,K) = 1-a, \qquad (3.2)$$

and $\pi_1(\theta,K)$ is given by (2.3) with

$$P_{00} = (K-1)\frac{N(\theta,K-2)}{(K-1)! - G(\theta,K-1)}\ , \quad \theta = \lambda K L, \qquad (3.3)$$

and

$$P_{11} = (K-1)\frac{G(\theta,K-2) + N(\theta,K-2) - \frac{2}{K-1}e^{-\theta}\theta^{K-1}}{G(\theta,K-1)}\ , \quad \theta = \lambda K L, \qquad (3.4)$$

where

$$G(\theta,K) = \int_0^\theta x^K e^{-x} dx\ , \quad N(\theta,K) = \int_0^\theta (\theta-x)^K e^{-(\theta+x)} dx.$$

Proof. The statement of the theorem indeed indicates the outline of the proof. By some algebraic simplification, the (2.1) and (2.2) will reduce to

(3.3) and (3.4) respectively. By examining later relations, we notice that K and $\theta = \lambda KL$ are parameters that are involved in transition probabilities. This gives $L = \frac{\theta}{K}(1/\lambda)$. But θ can be derived from (3.2) when the on target significant level is prescribed. Proof is complete.

Remark 3.1. For $K = 7$, we solved (3.2) for the $\theta(a, K)$ with different values of $1 - a$, using Mathematica 3.0, [Wolfram, 1991]. The solutions are given in Table 1. The transition and stationary probabilities are also plotted in terms of θ for $K = 7$, Figure 1. The threshold L in (3.1) is also plotted in terms of the mean $1/\lambda$, Figure 2. We notice from Fig. 2 that $\pi_1(\theta, 7)$ is strictly increasing, providing a unique solution for $\theta(a, 7)$.

1-a	0.9	0.8	0.7	0.6	0.5
$\theta(a,7)$	8.197	5.651	3.507	1.625	0

Table 1. Exponential Distribution; Significant Levels and Corresponding $\theta(a, 7)$ in (3.2).

Normal Distribution. Suppose the inputs X_1, X_2... are i.i.d normally distributed with mean μ and standard deviation σ. Interestingly, in this case also L is linear in μ. Details are given below.

Theorem 3.2. If inputs $X_1, X_2, \cdots$ follow normal distribution with mean μ and standard deviation σ, then the least value L for the threshold to ensure $(1 - a)\%$ on target significant level for the moving average Y_n is given by

$$L = \mu + \eta(a, K)\sigma, \tag{3.5}$$

where $\eta(a, K)$ is the solution to the equation

$$\pi_1(\eta, K) = 1 - a, \tag{3.6}$$

and $\pi_1(\eta, K)$ is given by (2.3) with

$$P_{00} = \frac{C(\eta, K)}{1 - \Phi(\sqrt{K}\eta)}, \qquad \eta = \frac{L - \mu}{\sigma}, \tag{3.7}$$

and

$$P_{11} = \frac{B(\eta, K)}{\Phi(\sqrt{K}\eta)}, \qquad \eta = \frac{L - \mu}{\sigma} \tag{3.8}$$

where

$$C(\eta, K) = \frac{1}{\sqrt{2\pi(K-1)}} \int_{-\infty}^{+\infty} [1 - \Phi(x)]^2 \, e^{-\frac{1}{2(K-1)}(x - K\eta)^2} dx,$$

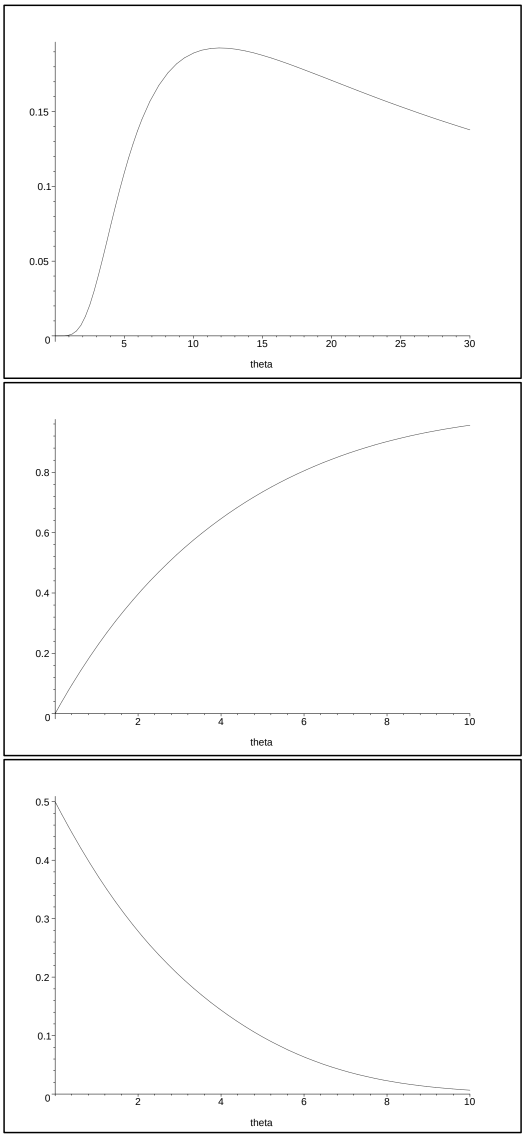


Fig. 1. Top Left: $P00(\theta, 7)$; Top Right: $P11(\theta, 7)$; Bottom: $\pi_0(\theta, 7)$.

and

$$B(\eta, K) = \frac{1}{\sqrt{2\pi(K-1)}} \int_{-\infty}^{+\infty} [\Phi(x)]^2 \ e^{-\frac{1}{2(K-1)}(x-K\eta)^2} dx,$$

Proof. In this case we note that the transition probabilities in (3.7) and (3.8) are expressed in terms of the parameter $\eta = \frac{L-\mu}{\sigma}$. So for given a, the

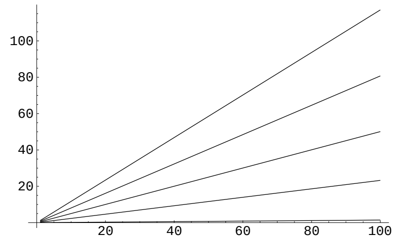


Fig. 2. Plots of L in terms of $1/\lambda$ for $a = 0.9,\ 0.8,\ 0.7,\ 0.6,\ 0.5$.

$\eta(a, K)$ in (3.5) is the solution to (3.6). The proof is complete.

Remark 3.2. For $K = 4$, the (3.6) is solved for $\eta(a, K)$ with different values for $1 - a$, using Mathematica. The version of Mathematica that we used did not solve the (3.6) directly, so we had to bypass this barrier by approximating the integrals involved in the equation by corresponding summations. The solutions are given in Table 2. The transition and stationary probabilities are also plotted in terms of η for $K = 4$, Figure 3. The threshold L in (3.1) is also plotted in terms of the mean μ for $\sigma = 1$, Figure 4.

Remark 3.3. The (3.1) and (3.5) can also be used estimation purposes when L is considered as an unknown parameter. It easily follows that for exponential and normal inputs, respectively

$$\hat{L} = \frac{\theta(a, K)}{K}\overline{x},$$

$$\hat{L} = \overline{x} + \eta(a, K)s.$$

Remark 3.4. Although the exponential and normal distributions were treated explicitly, the method, nevertheless, can be carried out for other distributions in order to identify or estimate the threshold parameter.

1-a	0.9	0.8	0.7	0.6	0.5
$\eta(a)$	0.65	0.47	0.28	0.14	0

Table 2. Normal Distribution; Significant levels and corresponding $\eta(a)$ in (3.6)

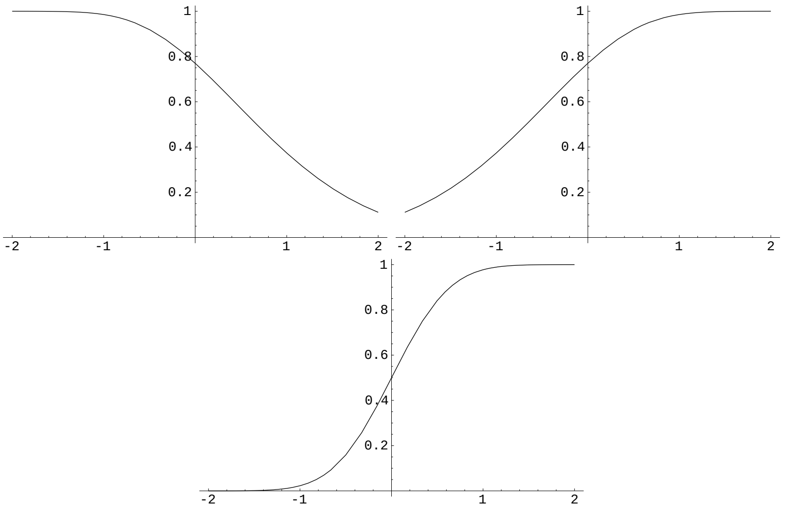


Fig. 3. Top Left: $P00(\eta, 4)$; Top Right: $P00(\eta, 4)$; Bottom Left: $\pi_1(\eta, 4)$

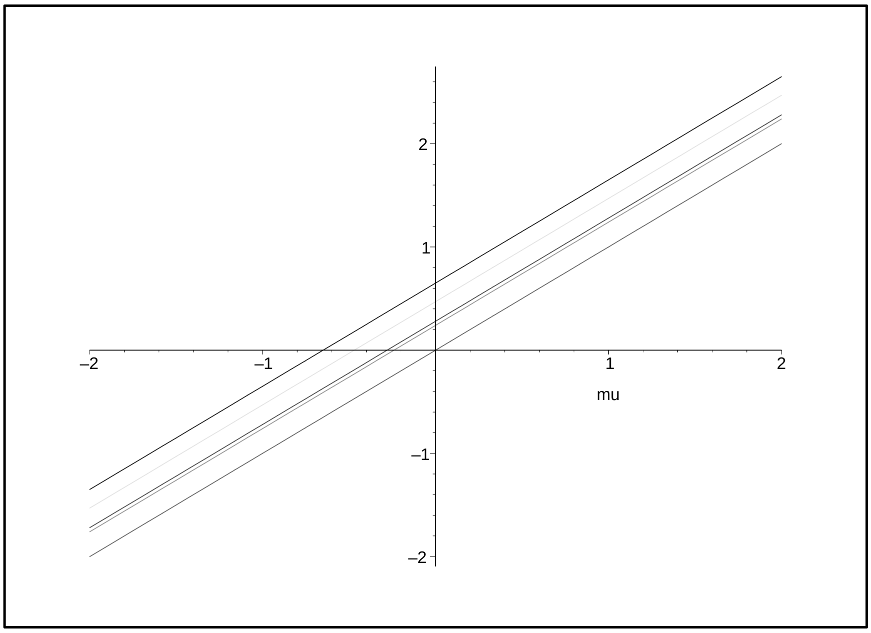


Fig. 4. Plots of $L(a, 4)$ in terms of μ for $\sigma = 1$ and $a = 0.9,\ 0.8,\ 0.7,\ 0.6,\ 0.5.$

References

[Banifacio and Salas, 1999]F. Banifacio and G. D. Salas. Return period and risk of hydrologic events. *Journal of Hydrologic Engineering*, pages 297–315, 1999.

[G. and Gooijer, 1998]Jan G. and De Gooijer. On threshold moving-average models. *Journal of Time Series Analysis*, pages 1–18, 1998.

[Karlin and Taylor, 1998]S. Karlin and H. M. Taylor. *An Introduction Stochastic Modelling*. Academic Press, New York, 1998.

[Vogel, 1987]R.M. Vogel. Reliability indices for water supply systems. *Journal of Water Resources Planning and Management*, pages 563–579, 1987.

[Wolfram, 1991]S Wolfram. *Mathematica 3.0*. Adison Wesley, California, Reaward City, 1991.

Weibull survivals with changepoints and heterogeneity

Ana-María Lara-Porras, Esteban Navarrete-Álvarez, Julia García-Leal, and
José-Manuel Quesada-Rubio

Department of Statistics and Operations Research, Faculty of Sciences,
University of Granada,
Campus de Fuentenueva, s/n, 18071, Granada, Spain
(e-mail: alara@ugr.es)

Abstract. This paper describes the proportional hazard models of right-censored survival data with Weibull distribution whose parameters vary in the time and the impact of individual heterogeneity being described by frailty. We obtain the equations of the maximum likelihood estimators for this model.

Keywords: proportional hazard models, Weibull distribution, frailty.

1 Introduction

The proportional hazard model is found among the most important models for the survival analysis data. The use of proportional hazard models for the study of survival times has received a great attention on the part of researchers, both in their theoretical aspect as well as in their interesting and numerous applications.

Likewise, an important characteristic of this type of investigations is that the data are frequently incomplete, meaning that the observation of survival time is not known for all individuals. These data are known as censored data.

[Aitkin and Clayton, 1980] and [Noura and Read, 1990] study the completely parametric models with a specified baseline hazard distribution. The first authors make use of exponential, Weibull, extreme value and generalized extreme value distributions. The second authors consider the piecewise model of the baseline hazard distribution and study the case of the Weibull distribution. In both works the presence of censored and uncensored observations is considered.

Piecewise models are based on the assumption that the parameters that characterize the base distribution vary with the passing of time. This is the reason why a partition of the time interval is introduced so as to maintain the base distribution, although with different parameters on each one of the intervals.

We have used this method considering that the survival time follows a Weibull distribution where the parameters chraracterizing the base-line distribution may vary with time for different intervals but remain constant at each interval. The points where the parameters change are called "changepoints".

Ordinary life table analyses implicitly assume that the population is homogeneous, an assumption which is usually unrealistic. It is more relevant to consider the population as a mixture of individuals with different risk, the heterogeneity being described by a quantity know as the frailty. Models for heterogeneity have been proposed, for example by [Vaupel *et al*, 1979] who introduced an unobserved quantity, that is the so-called frailty. This quantity described risk factors measurable or nonmeasurable, not included in the model.

The model to describe the population as a mixture assumes that to each individual corresponds a quantity, the frailty, describing the individual's relative risk.

In this work we consider the piecewise model for the Weibull distribution, the existence of censored observations and also we study the heterogeneity between individuals by a fraitly that we suppose follow a positive stable distribution.

2 Model construction

Let the survival time T be a nonnegative random variable that follows Weibull's distribution with a survivor function $S(t)$ and a hazard function $h(t)$. The heterogeneousness of the population is stated by a covariates vector $z = (z_1, z_2, \cdots, z_p)^T$ describing the characteristics of both the patient and the illness.

The hazard function depends in general on both time and on the set of covariates. The proportional hazard model separates these components as,

$$h(t; \mathbf{z}) = \gamma\rho(\rho t)^{\gamma-1} e^{\beta^T \mathbf{z}},$$

where the linear predictor $\beta^T z$ expresses the relative effect of the covariates z in terms of an unknown parameter vector $\beta = (\beta_1, \beta_2, \cdots, \beta_p)^T$.

The survivor function for these models is:

$$S(t; \mathbf{z}) = \exp\left[-(\rho t)^\gamma \exp\left(\beta^T \mathbf{z}\right)\right].$$

The hazard at time t conditional on x for a person with frailty x is assumed to be of form

$$h(t, x) = xh(t),$$

where the non-random function $h(t)$ common for all individuals is independent of x and describes the time effect.

Several authors have studied the model with gamma distributed frailties. We consider a stable positive distribution and whose scale factors have Laplace transform

$$L(s) = \mathrm{E}\left[\exp\left(-sx\right)\right] = \exp\left(-s^\alpha\right) \quad , \quad s \geq 0,$$

where $\alpha \in (0,1]$

Given the Fraitly x these expressions become

$$\begin{aligned} h(t;\mathbf{z},x) &= \gamma\rho(\rho t)^{\gamma-1} x e^{\beta^T \mathbf{z}} \quad , \\ S(t;\mathbf{z},x) &= \exp\left[-(\rho t)^{\gamma} x \exp\left(\beta^T \mathbf{z}\right)\right] . \end{aligned} \tag{1}$$

In this case, the corresponding survivor function is

$$S(t;\mathbf{z}) = \int \exp\left[-\Lambda(t) x e^{\beta^T \mathbf{z}}\right] f(x) dx, \tag{2}$$

where $\Lambda(t)$ is the cumulative hazard function.

If x follows a positive stable function, you get

$$S(t;\mathbf{z}) = \exp\left[-\left[\exp\left(\beta^T \mathbf{z}\right) \Lambda(t)\right]^{\alpha}\right] \tag{3}$$

where α is a parameter coming from the fraitly distribution.

We consider that the parameters that characterize the Weibull distribution can vary with time. Thus we divide the time axis in $k+1$ intervals, by using the changepoints $a_1, \cdots, a_k$. For convenience $a_0 = 0$ and $a_{k+1} = \infty$. In each interval (a_{j-1}, a_j) the distribution parameters take the values ρ_j and γ_j.

Denoting $g(t) = \ln \Lambda(t)$. In $a_{j-1} < t \leq a_j$, $g(t)$ becomes

$$g(t) = \ln\left[\rho_j t\right]^{\gamma_j} \quad , \quad j = 1, \ldots, k+1. \tag{4}$$

For $j = 1, \ldots, k$, due the continuity of $g(t)$ in the changepoints, has to be verified

$$\ln\left[\rho_j a_j\right]^{\gamma_j} = \ln\left[\rho_{j+1} a_j\right]^{\gamma_{j+1}} \quad , \quad j = 1, \ldots, k, \tag{5}$$

from where we derive that

$$\gamma_j = \gamma_1 \prod_{p=1}^{j-1} \frac{\ln\left[\rho_p a_p\right]}{\ln\left[\rho_{p+1} a_p\right]} \quad , \quad j = 2, \ldots, k+1. \tag{6}$$

Thus, for a survival time ending at $j-$th interval,

$$g(t) = \gamma_1 \ln\left(\rho_j t\right) \prod_{p=1}^{j-1} \frac{\ln\left[\rho_p a_p\right]}{\ln\left[\rho_{p+1} a_p\right]}. \tag{7}$$

For the $i-$th individual

$$g(t_i) = \sum_{j=1}^{k+1} c_{ij} \gamma_1 \ln\left(\rho_j t_i\right) \prod_{p=1}^{j-1} \frac{\ln\left[\rho_p a_p\right]}{\ln\left[\rho_{p+1} a_p\right]}, \tag{8}$$

where for $j = 1$, the product in p is omitted and c_{ij} is an indicator variable defined by:

$$c_{ij} = \begin{cases} 1 \text{ if } a_{j-1} < t_i \leq a_j \\ 0 \text{ otherwise} \end{cases}$$

with $i = 1, \ldots, N$ and $j = 1, \ldots, k+1$ where N represents the number of individuals.

Let $H_i = \exp \alpha g(t_i) + \beta^T \mathbf{z}$ and

$$h_i = H_i' = \alpha g'(t_i) H_i = \alpha H_i \prod_{j=1}^{k+1} \left[\frac{\gamma_1}{t_i} \prod_{p=1}^{j-1} \frac{\ln[\rho_p a_p]}{\ln[\rho_{p+1} a_p]} \right]^{c_{ij}}, \tag{9}$$

the survival and density functions can be expressed by

$$S(t_i; \mathbf{z}) = \exp[-H_i] \quad ; \quad f(t_i; \mathbf{z}) = h_i \exp[-H_i]. \tag{10}$$

3 Likelihood equations

Suppose that in a data set consisting of N observations, n are uncensored and m are censored. We define a censor indicator in the following manner:,

$$\omega_i = \begin{cases} 1 \text{ if the observation is uncensored} \quad (T_i = t_i) \\ 0 \text{ if it is censored} \quad (T_i > t_i) \end{cases}. \tag{11}$$

If a survival time observation is no censored contributes with $f(t)$ to the likelihood and if the observation is censored in time t, contributes with $S(t)$. Thus the likelihood function is,

$$l = \prod_{i=1}^{N} [f(t_i; \mathbf{z})]^{\omega_i} [S(t_i; \mathbf{z})]^{1-\omega_i}, \tag{12}$$

and the log-likelihood function,

$$L = \sum_{i=1}^{N} \{\omega_i \ln h(t_i; \mathbf{z}) + \ln S(t_i; \mathbf{z})\} =$$

$$\sum_{i=1}^{N} \left\{ \omega_i \left[\ln\alpha + \alpha \left(\sum_{s=1}^{p} \beta_s z_{is} + \sum_{j=1}^{k+1} c_{ij} \left(\gamma_1 \ln(\rho_j t_i) \prod_{p=1}^{j-1} \frac{\ln[\rho_p a_p]}{\ln[\rho_{p+1} a_p]} \right) \right) + \right. \right.$$

$$\left. \sum_{j=1}^{k+1} c_{ij} \ln \left(\frac{\gamma_1}{t_i} \prod_{p=1}^{j-1} \frac{\ln[\rho_p a_p]}{\ln[\rho_{p+1} a_p]} \right) \right] -$$

$$\left. \exp\alpha \left[\sum_{s=1}^{p} \beta_s z_{is} + \sum_{j=1}^{k+1} c_{ij} \left(\gamma_1 \ln(\rho_j t_i) \prod_{p=1}^{j-1} \frac{\ln[\rho_p a_p]}{\ln[\rho_{p+1} a_p]} \right) \right] \right\}. \tag{13}$$

where for $j=1$, the product in p is omitted.

The first derivatives of L with respect to the parameters β_s, γ, α and ρ are **:**

$$\frac{\partial L}{\partial \beta_s} = \alpha \sum_{i=1}^{N} z_{is}\left(\omega_i - H_i\right) \qquad \text{para} \qquad s = 1, \ldots, p \tag{14}$$

$$\frac{\partial L}{\partial \gamma_1} = \sum_{i=1}^{N} \alpha \left\{ \left(\omega_i - H_i\right) \left[\sum_{j=1}^{k+1} c_{ij} \ln\left(\rho_j t_i\right) \prod_{p=1}^{j-1} \frac{\ln\left[\rho_p a_p\right]}{\ln\left[\rho_{p+1} a_p\right]} \right] + \frac{\omega_i}{\gamma_1} \right\} \tag{15}$$

$$\frac{\partial L}{\partial \alpha} = \sum_{i=1}^{N} \left\{ \frac{\omega_i}{\alpha} + \left(\omega_i - H_i\right) \left[\sum_{s=1}^{p} \beta_s z_{is} + \sum_{j=1}^{k+1} c_{ij} \left(\gamma_1 \ln\left(\rho_j t_i\right) \prod_{p=1}^{j-1} \frac{\ln\left[\rho_p a_p\right]}{\ln\left[\rho_{p+1} a_p\right]} \right) \right] \right\} \tag{16}$$

$$\begin{aligned} \frac{\partial L}{\partial \rho_1} = \frac{1}{\rho_1} \sum_{i=1}^{N} & \left\{ \alpha\gamma_1 \left(\omega_i - H_i\right) \left(c_{i1} + \sum_{j=2}^{k+1} c_{ij} \frac{\ln\left(\rho_j t_i\right)}{\ln\left(\rho_2 a_1\right)} \prod_{p=2}^{j-1} \frac{\ln\left[\rho_p a_p\right]}{\ln\left[\rho_{p+1} a_p\right]} \right) + \right. \\ & \left. \frac{\omega_i}{\ln\left(\rho_1 a_1\right)} \sum_{j=2}^{k+1} c_{ij} \right\} \end{aligned} \tag{17}$$

$$\begin{aligned} \frac{\partial L}{\partial \rho_j} = \frac{1}{\rho_j} \sum_{i=1}^{N} & \left\{ \left(\omega_i - H_i\right) \frac{\alpha\gamma_1}{\left(\ln\left(\rho_j a_{j-1}\right)\right)^2} \frac{\prod_{p=1}^{j-1} \ln\left[\rho_p a_p\right]}{\prod_{p=2}^{j-1} \ln\left[\rho_p a_{p-1}\right]} \times \right. \\ & \left[c_{ij} \ln \frac{a_{j-1}}{t_i} + \frac{\ln \frac{a_{j-1}}{a_j}}{\ln\left[\rho_{j+1} a_j\right]} \sum_{p=j+1}^{k+1} c_{ip} \ln\left(\rho_p t_i\right) \prod_{r=j+1}^{p-1} \frac{\ln\left[\rho_r a_r\right]}{\ln\left[\rho_{r+1} a_r\right]} \right] + \\ & \left. \frac{\omega_i}{\ln\left(\rho_j a_{j-1}\right)} \left(-c_{ij} + \frac{\ln \frac{a_{j-1}}{a_j}}{\ln\left[\rho_j a_j\right]} \sum_{p=j+1}^{k+1} c_{ip} \right) \right\} \qquad \text{for} \qquad j = 2, \ldots, k+1 \; . \end{aligned} \tag{18}$$

4 Some questions about the equations resolution

The solving of these equations may be performed by general iterative methods, by directly employing statistical packages such as GLIM or we can study the behaviour of maximum-likelihood estimators through the simulation.

Upon the determination of the appropriate number of changepoints and their locations there are some graphic procedures. In practice it is enough, in most cases, to consider only one or two changepoints. Moreover, it must be indicated that the physical nature of the problem also sometimes permits on the location of the possible changepoints.

References

[Aalen, 1988]O. Aalen. Heterogeneity in Survival Analysis. *Statistics in Medicine*, 1121–1137, 1988.

[Aitkin and Clayton, 1980]M. Aitkin and D. Clayton. The Fitting of Exponential, Weibull and Extreme Value Distributions To Complex Censored Survival Data using GLIM. *Applied Statistics*, 156–163, 1980.

[Cox and Oakes, 1984]D.R. Cox and D. Oakes. *Analysis of Survival Data.* Chapman and Hall, 1984.

[Hougaard, 1984]P. Hougaard. Life Table Methods for Heterogeneous Populations: Distributions Describing the Heterogeneity. *Biometrika*, 75–83, 1984.

[Hougaard, 1986]P. Hougaard. Survival Models for Heterogeneous Populations Derived from Stable Distributions. *Biometrika*, 387–96, 1986.

[Hougaard, 1987]P. Hougaard. Modelling Multivariate Survival. *Scand. J. Statist.*, 291–304, 1987.

[Kalbfleisch and Prentice, 1980]J.D. Kalbfleisch and R.L. Prentice. *The Statistical Analysis of Failure Time Data.* John Wiley and Sons, New York, 1980.

[Lara-Porras, 1995]A.M. Lara-Porras. *Aportaciones a modelos de supervivencia: distribuciones base con puntos de cambio y covariables dependientes del tiempo*, Unpublished P.H.D. dissertation, University of Granada, Spain, 1995.

[Lara-Porras *et al*, 1998]A.M. Lara-Porras, J. García-Leal, and E. Navarrete-Álvarez. A Proportional Hazard Model with Time-Dependent Parameters and Covariates, *Journal of the Italian Statistical Society*, 233–242, 1998.

[Lara-Porras *et al*, 2000]A.M. Lara-Porras, J. García-Leal, and E. Navarrete-Álvarez. Simulation in a fully parametric proportional hazard model with changepoints, *Brazilian Journal of Probability and Statistics*, 113–122, 2000.

[Noura and Read, 1990]A.A. Noura and K.L.Q. Read. Proportional Hazard Changepoints Models in Survival Analysis, *Appl. Statist.*, 241–253, 1990.

[Vaupel *et al*, 1979]J.W. Vaupel, K.G. Manton, and E. Stallard. The Impact of Heterogeneity in Individual Frailty and the Dynamics of Mortality, *Demography*, 439–454, 1979.

Towards a filter-based EM-algorithm for parameter estimation of Littlewood's software reliability model

James Ledoux

INSA & IRMAR
20 avenue des Buttes de Cöesmes, CS 14315,
35043 Rennes Cedex, France
(e-mail: james.ledoux@insa-rennes.fr)

Abstract. In this paper, we deal with a continuous-time software reliability model designed by Littlewood. This model may be thought of as a partially observed Markov process. The EM-algorithm is a standard way to estimate the parameters of processes with missing data. The E-step requires the computation of basic statistics related to observed/hidden processes. In this paper, we provide finite-dimensional non-linear filters for these statistics using the innovations method. This allows to plan the use of the filter-based EM-algorithm developed by Elliott.
Keywords: Filtering, Hidden Markov process, Point process, Innovations method.

1 Introduction

A major issue in software reliability modeling is the calibration of the models from data. This is well documented in the so-called "black-box approach". We refer to [Ledoux, 2003] and references therein for details. To the best of our knowledge, no statistical procedure has been proposed in the architecture-based approach for assessing the reliability of a software. A standard model in this context was provided by Littlewood [Littlewood, 1975]. It has inspired most other works [Goseva-Popstojanova and Trivedi, 2001]. Littlewood proposed a Markov-type reliability model for modular softwares. For a software with a finite number of modules:

- the structure of the software is represented by a finite continuous time Markov chain $X = (X_t)_{t\geq 0}$ where X_t is the active module at time t. The generator of X is denoted by Q and its state space is assumed to be $\mathscr{U} := \{e_i, i = 1, \ldots, n\}$.
- When module e_i is active, the failure times are part of a homogeneous Poisson Process with intensity $\mu(i)$.
- When control switches from module e_i to module e_j, a failure may happen with probability $\mu(i, j)$.
- When a failure appears, the time to recover a safe state is neglected. A failure does not affect the execution dynamics of the software.
- All failure processes are assumed to be independent, given a sequence of activated modules.

Let us denote the number of observed failures over $[0,t]$ by N_t. It can be seen that $(N_t, X_t)_{t\geq 0}$ is a Markov process with state space $\mathbb{N} \times \mathscr{U}$. It has the following generator

$$A = \begin{pmatrix} D_0 & \mathbf{0} & \mathbf{0} & \cdots \\ D_1 & D_0 & \ddots & \ddots \\ \vdots & \ddots & \ddots & \ddots \end{pmatrix} \tag{1}$$

when the states are listed in lexicographic order and the matrices D_0 and D_1 are defined by

$$\begin{aligned} \text{if } j \neq i : D_0(j,i) &:= Q(j,i)(1-\mu(j,i)) \quad D_1(j,i) := Q(j,i)\mu(j,i), \\ D_0(i,i) &:= -\sum_{j\neq i} Q(j,i) - \mu(i) \quad D_1(i,i) = \mu(i). \end{aligned}$$

The nonnegative number $D_0(j,i)$ $(j \neq i)$ represents the rate at which X jumps from state e_i to e_j with no failure event. The entry $D_1(j,i)$ is the rate at which X jumps from state e_i to state e_j with the occurrence of one failure. Note that $Q = D_0 + D_1$. The distribution function of the counting variable N_t may be numerically evaluated using the uniformization technique. But this requires the knowledge of the non-negative parameter vector

$$\theta = \{D_k(j,i), \quad k = 0,1 \quad i,j = 1,\ldots,n\}.$$

In general, we can obtain a priori estimates for θ using procedures reported in [Goseva-Popstojanova and Trivedi, 2001]. They are based on data collected at earlier phases of the software life cycle (validation phases, integration tests,...). Sometimes, these estimates might appear to be rough estimates when the software is in operation. The only available data is the observation of failure events. In that perspective, the process (N, X) should be thought of as a partially observed Markov process or a hidden Markov process. The observed process is the failure point process $(N_t)_{t\geq 0}$ and the state or hidden process is the finite Markov process $(X_t)_{t\geq 0}$. The EM-algorithm is a standard way to estimate the parameters of hidden Markov processes. Elliott proposed a filter-based EM-algorithm in [Elliott *et al.*, 1995]. That is, the standard forward-backward form of the E-step of the algorithm is replaced by a single-pass procedure that involves finite-dimensional filters for various statistics related to the observed/hidden processes. The aim of this paper is to provide such finite-dimensional filters for Littlewood's model.

We point out that we deal with a failure point process that is a Markovian Arrival Process (MAP) as defined by Neuts [Neuts, 1989]. The Littlewood model has the (doubly stochastic) Poisson process (driven) modulated by a Markov process as a special instance (setting parameters $\mu(\cdot,\cdot)$ to 0). Statistical estimation for the MAP has been recently developed in the continuous-time context (see [Asmussen, 2000, Klemmm *et al.*, 2003, and references therein]. All these works use the forward-backward procedure. The numerical experiments reported in their studies show that EM-algorithm works well in general.

2 Finite-dimensional filters

Main notation and convention

- Vectors are column vectors. Row vectors are denoted by means of the transpose operator $(.)^{\mathsf{T}}$.
- $\mathbf{1}_k$ is a k-dimensional vector with each entry equals to one.
- We denote the left limit of function f at t by f_{t-}.
  For any function $t \mapsto f_t$, $\Delta f_t := f_t - f_{t-}$ for $t > 0$ is the jump of the function at time t. We set $\Delta f_0 := f_0$.
- The state space of X is $\mathscr{U} := \{e_i, i = 1, \dots, n\}$, where e_i is the ith vector of the canonical basis of $\mathbb{R}^n$. With this convention,

$$1_{\{X_t = e_i\}} = \langle X_t, e_i \rangle, \quad \mathbf{1}_n^{\mathsf{T}} X_t = 1$$

  where $\langle \cdot, \cdot, \rangle$ is the usual scalar product in $\mathbb{R}^n$.
- All processes are assumed to be defined on the same probability space $(\Omega, \mathcal{F}, \mathbb{P})$. The internal filtrations of processes N and (N, X) are denoted by $\mathbb{F}^N = (\mathbb{F}_t^N)_t$ and $\mathbb{F} = (\mathbb{F}_t)_t$ respectively. These filtrations are assumed to be complete.
- For any integrable adapted random process $(Z_t)_{t \geq 0}$, the conditional expectation $\mathbb{E}[Z_t \mid \mathbb{F}_t^N]$ is denoted by $\widehat{Z}_t$.

2.1 Basic material on the observed/hidden processes

We report here a semi-martingale representation of the basic statistics of the Littlewood's model for which filters will be derived. Due to the special structure of generator A of (N, X) (see (1)), N and the following counting processes are easily interpreted to be counters of specific transitions in (N, X)

$$\begin{aligned} N_t^{X,ji} &:= \sum_{0<s\leq t} \langle X_{s-}, e_i \rangle \langle X_s, e_j \rangle = \int_0^t \langle X_{s-}, e_i \rangle \langle e_j, dX_s \rangle \\ \mathcal{L}_t^{1,ji} &:= \sum_{0<s\leq t} \langle X_{s-}, e_i \rangle \langle X_s, e_j \rangle \Delta N_s = \int_0^t \langle X_{s-}, e_i \rangle \langle X_s, e_j \rangle dN_s \\ j \neq i \quad \mathcal{L}^{0,ji} &:= \sum_{0<s\leq t} \langle X_{s-}, e_i \rangle \langle X_s, e_j \rangle (1 - \Delta N_s) = N_t^{X,ji} - \mathcal{L}_t^{1,ji} \end{aligned}$$

and $N_t(x, y)$ is the number of transitions of (N, X) from state x to state y at time t. It is well known that [Bremaud, 1981]

$$M_t(y, x) := N_t(y, x) - \int_0^t A(y, x) \, 1_{\{(N_{s-}, X_{s-}) = x\}} ds$$

is a $\mathbb{F}$-*martingale*. In other words, the $\mathbb{F}$-semi-martingale (or Doob-Meyer here) decomposition of $N(y, x)$ is $N_t(y, x) = \int_0^t \mathbf{1}_{\{(N_{s-}, X_{s-}) = x\}} A(y, x) ds +$

$M_t(y,x)$. Then, it is easily seen that the $\mathbb{F}$-semi-martingale decomposition of the counting processes above are

$$\begin{aligned} N_t &= \int_0^t \lambda_s ds + \mathcal{M}_t \quad \text{with } \lambda_s := \mathbf{1}_n^\top D_1 X_{s-} && (2) \\ N_t^{X,ji} &:= \int_0^t Q(j,i)\langle X_{s-}, e_i\rangle ds + \mathcal{M}_t^{N^X(j,i)} \\ \mathcal{L}_t^{k,ji} &:= \int_0^t D_k(j,i)\langle X_{s-}, e_i\rangle d_s + \mathcal{M}_t^{\mathcal{L}_{k,ji}} \quad k = 0,1 && (3) \end{aligned}$$

where $\mathcal{M}, \mathcal{M}^{N^X(j,i)}, \mathcal{M}^{\mathcal{L}_{k,ji}}$ are $\mathbb{F}$-martingales.

The last statistics that we need, is the sojourn time of X in any state e_i in the interval $[0,t]$

$$\mathcal{O}_t^{(i)} := \int_0^t \langle X_{s-}, e_i\rangle ds$$

The basic semi-martingale decomposition of the Markov process X is [Bremaud, 1981]

$$X_t = \int_0^t QX_{s-}ds + \mathcal{M}_t^X. \tag{4}$$

We recognize in (4) and (2) a standard representation of a continuous-time hidden Markov process, with X as the state process and N the observed process. The observation and state "noises" are correlated here.

2.2 The EM-algorithm

We briefly explain the EM-algorithm for our continuous-time hidden Markov model. We refer to [Klemmm *et al.*, 2003] for full details. For a fixed parameter vector θ, we denote the underlying probability measure and associated expectation respectively by $\mathbb{P}_\theta$ and $\mathbb{E}_\theta$. X_0 or its probability distribution x_0 is assumed to be known. The observed data are supposed to be the inter-failure durations $\{t_1, \ldots, t_K\}$ where $t_K = t$. The likelihood function for the complete data (N, X) up to time t under $\mathbb{P}_\theta$ is

$$L_t(\theta; N, X) := \prod_{i,j=1,}^n D_1(j,i)^{\mathcal{L}_t^{1,ji}} \prod_{i,j=1,j\neq i}^n D_0(j,i)^{\mathcal{L}_t^{0,ji}} \prod_{i=1}^n e^{D_0(i,i)\mathcal{O}_t^{(i)}} \prod_{i=1}^n x_0(i)^{\langle X_0, e_i\rangle}.$$

The formulas for estimating θ from the observations $N_s, s \leq t$ are obtained using the following iterative procedure:

i) Initialization : Choose θ_0

ii) E-step. Set $\theta := \theta_l$. Compute the so-called pseudo-log-likelihood $Q(\cdot \mid \theta)$ defined by

$$Q(\theta^* \mid \theta) := \mathbb{E}_\theta\big[\log L_t(\theta^*; N, X) \mid \mathbb{F}_t^N\big] \tag{5}$$

where $\theta^* := \{D_k^*(j,i),\ i,j = 1, \ldots, n;\ k = 0,1\}$.

iii) M-step. Determine θ_{l+1} maximizing the function (5).
iv) Return in 2 until a stopping criterion is satisfied.

For the M-step, it is easily seen that

$$i,j=1,\dots,n \quad D_1^*(j,i)=\frac{\widehat{\mathcal{L}}_t^{1,ji}}{\widehat{\mathcal{O}^{(i)}}_t}, \quad D_0^*(j,i)=\frac{\widehat{\mathcal{L}}_t^{0,ji}}{\widehat{\mathcal{O}^{(i)}}_t}, \quad i\neq j. \tag{6}$$

An appealing property of the EM-algorithm is that the sequence of estimates $\{\theta_l, l\geq 0\}$ gives a nondecreasing values of the likelihood function with equality iff $\theta_{l+1}=\theta_l$ (under mild conditions). Note that the zero entries of D_ks are preserved by the procedure above.

As a result of the procedure above, we have to compute the estimates in (6). The standard way is to use the Baum-Welch implementation of the EM-algorithm (also referred to as the "forward-backward" technique). This is what is done in the previously mentioned works [Asmussen, 2000, Klemmm *et al.*, 2003]. Using the filter-based approach pioneering by Elliott [Elliott *et al.*, 1995], the estimates in (6) are computed from the filters given in Theorem 1. The basic difference with the standard Baum-Welch method is that only one pass through the data set is needed for the filter-based method.

2.3 The results

We use a trick proposed by Elliott. We compute the following filters

$$\widehat{N^{X,ji}X}_t,\ \widehat{\mathcal{O}^{(i)}X}_t \ \text{ and } \ \widehat{\mathcal{L}^{1,ji}X}_t$$

which turn to be finite-dimensional. Then, we have

$$\widehat{N^{X,ji}}_t=\mathbf{1}_n^\top\widehat{N^{X,ji}X}_t, \quad \widehat{\mathcal{O}^{(i)}}_t=\mathbf{1}_n^\top\widehat{\mathcal{O}^{(i)}X}_t, \quad \text{and} \quad \widehat{\mathcal{L}}_t^{1,ji}=\mathbf{1}_n^\top\widehat{\mathcal{L}^{1,ji}X}_t.$$

A filter equation for $\mathcal{L}_t^{0,ji}X_t$ ($j\neq i$) can be derived as that of Theorem 1 or using the fact that $\widehat{\mathcal{L}^{0,ji}X}_t=\widehat{N^{X,ji}X}_t-\widehat{\mathcal{L}^{1,ji}X}_t$.

Theorem 1 *Let $\widehat{\lambda}_t := \mathbf{1}_n^\top D_1\widehat{X}_{t-}$. The fundamental $\mathbb{F}^N$-martingale $(N_t - \int_0^t\widehat{\lambda}_s ds)_{t\geq 0}$ is denoted by $(\widehat{\mathcal{M}}_t^N)_{t\geq 0}$.*

i) *Estimator for the state. We have for any $t\geq 0$*

$$\widehat{X}_t=\widehat{X}_0+\int_0^t Q\widehat{X}_{s-}ds+\int_0^t\frac{D_1\widehat{X}_{s-}-\widehat{X}_{s-}\widehat{\lambda}_s}{\widehat{\lambda}_s}d\widehat{\mathcal{M}}_s^N. \tag{7}$$

ii) *Estimator for the number of jumps of X from e_i to e_j. We have for any $t\geq 0$*

$$\begin{aligned}\widehat{N^{X,ji}X}_t=&\int_0^t Q\widehat{N^{X,ji}X}_{s-}ds+\int_0^t Q(j,i)\langle\widehat{X}_{s-},e_i\rangle ds\,e_j\\&+\int_0^t\frac{D_1\widehat{N^{X,ji}X}_{s-}-\widehat{N^{X,ji}X}_{s-}\widehat{\lambda}_s}{\widehat{\lambda}_s}d\widehat{\mathcal{M}}_s^N.\end{aligned} \tag{8}$$

iii) *Estimator for the sojourn time to e_i. We have for any $t \geq 0$*

$$\widehat{\mathcal{O}^{(i)}X}_t = \int_0^t Q\widehat{\mathcal{O}^{(i)}X}_{s-}ds + \int_0^t \langle \widehat{X}_{s-}, e_i\rangle ds\, e_i \\ + \int_0^t \frac{D_1\widehat{\mathcal{O}^{(i)}X}_{s-} - \widehat{\mathcal{O}^{(i)}X}_{s-}\widehat{\lambda}_s}{\widehat{\lambda}_s} d\widehat{\mathcal{M}}^N_s. \tag{9}$$

iv) *Estimator for the number of joint transitions. We have for $t \geq 0$*

$$\widehat{\mathcal{L}^{1,ji}X}_t = \int_0^t Q\widehat{\mathcal{L}^{1,ji}X}_{s-}ds + \int_0^t D_1(j,i)\langle \widehat{X}_{s-}, e_i\rangle ds\, e_j \\ + \int_0^t \frac{D_1(j,i)\langle \widehat{X}_{s-}, e_i\rangle\, e_j + D_1\widehat{\mathcal{L}^{1,ji}X}_{s-} - \widehat{\mathcal{L}^{1,ji}X}_{s-}\widehat{\lambda}_s}{\widehat{\lambda}_s} d\widehat{\mathcal{M}}^N_s. \tag{10}$$

Remark 1 *The filters for the statistics of an MMPP may be obtained from the previous theorem. We have $D_1 = \mathrm{Diag}(\mu(i))$.*

Proof. A proof of (7) may be found in [Gravereaux and Ledoux, 2004]. In the sequel, $\mathcal{M}$ (resp. $\widehat{\mathcal{M}}$) will denote a generic $\mathbb{F}$ (resp. $\mathbb{F}^N$)-martingale. The proof of (9) is as follows. An integration by parts gives

$$\begin{aligned} \mathcal{O}^{(i)}_t X_t &= \int_0^t \mathcal{O}^{(i)}_{s-}dX_s + \int_0^t X_{s-}d\mathcal{O}^{(i)}_s + \underbrace{[\mathcal{O}^{(i)}, X]_t}_{0} \\ &= \int_0^t Q\mathcal{O}^{(i)}_s X_{s-}ds + \int_0^t \langle X_{s-}, e_i\rangle e_i ds + \mathcal{M} \quad \text{from (4).} \end{aligned} \tag{11}$$

The $\mathbb{F}^N$-optional projection of the equation above, is

$$\widehat{\mathcal{O}^{(i)}X}_t = \int_0^t Q\widehat{\mathcal{O}^{(i)}X}_{s-}ds + \int_0^t \langle \widehat{X}_{s-}, e_i\rangle e_i + \widehat{\mathcal{M}}. \tag{12}$$

The integral representation of $\mathbb{F}^N$-martingales says that $\widehat{\mathcal{M}}$ in the right hand side member above, has the form [Bremaud, 1981]

$$\int_0^t G^{(i)}_s d\widehat{\mathcal{M}}^N_s.$$

Thus, the proof will be complete if we prove that

$$G^{(i)}_s = \frac{D_1\widehat{\mathcal{O}^{(i)}X}_{s-} - \widehat{\mathcal{O}^{(i)}X}_{s-}\widehat{\lambda}_s}{\widehat{\lambda}_s}. \tag{13}$$

The product $N_t\widehat{\mathcal{O}^{(i)}X}_t$ has the form from an integration by parts

$$\begin{aligned} N_t\widehat{\mathcal{O}^{(i)}X}_t &= \int_0^t N_{s-}d\widehat{\mathcal{O}^{(i)}X}_s + \int_0^t \widehat{\mathcal{O}^{(i)}X}_{s-}dN_s + [N,\widehat{\mathcal{O}^{(i)}X}]_t \\ &= \int_0^t N_{s-}[Q\widehat{\mathcal{O}^{(i)}X}_s + \langle \widehat{X}_{s-}, e_i\rangle\, e_i]ds + \widehat{\mathcal{M}} \text{ from (12)} \\ &\quad + \int_0^t \widehat{\mathcal{O}^{(i)}X}_{s-}\widehat{\lambda}_s ds + \widehat{\mathcal{M}} \\ &\quad + \int_0^t G_s^{(i)}\widehat{\lambda}_s ds + \widehat{\mathcal{M}} \end{aligned} \tag{14}$$

Note that $\mathcal{O}^{(i)}$ has continuous paths so that $\Delta\mathcal{O}_s^{(i)} = 0$. Next, the product $N_t(\mathcal{O}_t^{(i)}X_t)$ is with an integration by parts

$$\begin{aligned} N_t\mathcal{O}_t^{(i)}X_t &= \int_0^t N_{s-}d(\mathcal{O}^{(i)}X)_s + \int_0^t \mathcal{O}_{s-}^{(i)}X_s dN_s \\ &= \int_0^t N_{s-}[Q\mathcal{O}_{s-}^{(i)}X_{s-} + \langle \widehat{X}_{s-}, e_i\rangle\, e_i]ds + \int_0^t \mathcal{O}_{s-}^{(i)}X_s dN_s \text{ from (11).} \end{aligned}$$

Let us compute the last term in the equality above:

$$\begin{aligned} \int_0^t \mathcal{O}_{s-}^{(i)}X_s dN_s &= \sum_{0<s\le t}\mathcal{O}_{s-}^{(i)}X_s\Delta N_s = \sum_j e_j \sum_k \int_0^t \mathcal{O}_{s-}^{(i)}d\mathcal{L}_s^{1,jk} \\ &= \int_0^t \mathcal{O}_{s-}^{(i)}\sum_j e_j \sum_k D_1(j,k)\langle X_{s-}, e_k\rangle ds + \mathcal{M} \text{ from (3)} \\ &= \int_0^t \mathcal{O}_{s-}^{(i)}D_1 X_{s-}ds + \mathcal{M}. \end{aligned}$$

Then, we deduce from the last equality that

$$N_t\mathcal{O}_t^{(i)}X_t = \int_0^t N_{s-}[Q\mathcal{O}_{s-}^{(i)}X_{s-} + \langle X_{s-}, e_i\rangle\, e_i]ds + \int_0^t \mathcal{O}_{s-}^{(i)}D_1 X_{s-}ds + \mathcal{M}.$$

The $\mathbb{F}^N$-optional projection of the previous formula leads to a second decomposition of the special semi-martingale $N_t\widehat{\mathcal{O}^{(i)}X}_t$

$$N_t\widehat{\mathcal{O}^{(i)}X}_t = \int_0^t N_{s-}[Q\widehat{\mathcal{O}^{(i)}X}_{s-} + \langle \widehat{X}_{s-}, e_i\rangle\, e_i]ds + \int_0^t D_1\widehat{\mathcal{O}^{(i)}X}_{s-}ds + \widehat{\mathcal{M}}. \tag{15}$$

We know that the bounded variations part of the decomposition of a special semi-martingale is unique. Then, we can identify the corresponding terms in the decompositions (14) and (15), that is the Lebesgue integrals. The expression (13) of the gain $G^{(i)}$ follows easily.

Formulas (10) and (8) are shown in the same way. Their proofs are not reported here for saving space.

3 Conclusion

The Littlewood's software reliability model may be thought of as a partially observed Markov process. The contribution of this paper is to provide finite-dimensional non-linear filters for various statistics associated with this model. These filters are the first step in view of implementing the filter-based form of the EM-algorithm proposed by Elliott [Elliott *et al.*, 1995]. We mention that basic extensions may be obtained following the guidelines of this paper. We can derive filters for the general class of MAP's. The case of occurrences of failures in clusters can also be included in the discussion. We just have to consider N as a multivariate point process of failures. From the numerical point of view, a second step in implementing the filter-based approach would be to find the so-called robust versions of the non-linear filtering equations obtained here. Then, robust numerical algorithms may be expected. We refer to [James *et al.*, 1996] for a detailed discussion of such a time discretization approach. In this perspective, we mention that it should be desirable to obtain Zakaï form for our filters. Such a form can be derived from our results. The details will be reported elsewhere.

References

[Asmussen, 2000]S. Asmussen. Matrix-analytic models and their analysis. *Scandinavian Journal of Statistics*, pages 193–226, 2000.

[Bremaud, 1981]P. Bremaud. *Point Processes and Queues.* Springer, 1981.

[Elliott *et al.*, 1995]R.J. Elliott, L. Aggoun, and J.B. Moore. *Hidden Markov Models.* Springer, 1995.

[Goseva-Popstojanova and Trivedi, 2001]K. Goseva-Popstojanova and K.S. Trivedi. Architecture-based approach to reliability assessment of software systems. *Performance Evaluation*, pages 179–204, 2001.

[Gravereaux and Ledoux, 2004]J.-B. Gravereaux and J. Ledoux. Poisson approximation for some point processes in reliability. *Advances in Applied Probability*, pages 455–470, 2004.

[James *et al.*, 1996]M.R. James, V. Krishnamurthy, and Le Gland F. Time discretization of continuous-time filters and smoothers for HMM parameter estimation. *IEEE Trans. Information Theory*, pages 593–605, 1996.

[Klemmm *et al.*, 2003]A. Klemmm, C. Lindemann, and M. Lohmann. Modeling IP traffic using the Batch Markovian Arrival Process. *Performance Evaluation*, pages 149–173, 2003.

[Ledoux, 2003]J. Ledoux. Chap. 12: Software reliability modeling. In H. Pham, editor, *Handbook of Reliability Engineering*, pages 213–234. Springer, London, 2003.

[Littlewood, 1975]B. Littlewood. A reliability model for systems with Markov structure. *Appl. Statist.*, pages 172–177, 1975.

[Neuts, 1989]M. F. Neuts. *Structured Stochastic Matrices of M/G/1 Type and Their Applications.* Marcel Dekker Inc., New-York and Basel, 1989.

On Two New Methods for Constructing Multivariate Probability Distributions with System Reliability Motivations

Jerzy Filus[1] and Lidia Filus[2]

[1] Department of Mathematics and Computer Science
Oakton Community College
Des Plaines, IL 60016, USA
(e-mail: : filusje@aol.com)

[2] Department of Mathematics
Northeastern Illinois University
Chicago, IL 60625, USA
(e-mail: l-filus@neiu.edu)

Abstract. Two distinct methods of construction of some interesting new classes of multivariate probability densities are described and applied. As common result of both procedures, two n-variate pdf classes are obtained. The classes are considered as multivariate generalizations of the classes of univariate Weibullian and gamma pdfs. Example of an application of the obtained n-variate pdfs to the problem of modeling the reliability of multicomponent systems with stochastically dependent life-times of their components is given. Possibility to construct an extension of the considered random vectors to stochastic processes is communicated. Application of the so obtained (ex-Weibullian) stochastic processes as highly non-Markovian but simple models for maintenance of systems, with a history of all past repairs recorded, is presented.

Keywords: multivariate probability density, system reliability and maintenance modeling, highly non-Markovian models, n-variate ex-exponential, ex-Weibullian, ex-gamma pdfs, pseudoaffine transformations on $\mathbf{R}^n$.

1 On pseudoaffine transformations

Suppose $T_1, T_2, \ldots, T_n$ are independent random variables and for each $i = 1, \ldots, n$, T_i has a pdf that belongs to one of the following four classes of probability distributions: Gaussian, exponential, Weibullian or gamma (i.e., all T_i 's are assumed to be in exactly one of the above classes). To any so defined random vector $(T_1, T_2, \ldots, T_n)$ apply a member from the following new class of $\mathbf{R}^n \rightarrow \mathbf{R}^n$ pseudoaffine transformations (see [Filus and Filus, 2001b], [Filus and Filus, 2003b]) defined by the following scheme (recall that the well known ordinary affine transformations in $\mathbf{R}^n$ are usually understood to be compositions of nonsingular linears and translations on $\mathbf{R}^n$):

$$
\begin{aligned}
X_1 &\overset{d}{=} \phi_0 T_1 + \psi_0, \\
X_2 &\overset{d}{=} \phi_1(X_1) T_2 + \psi_1(X_1), \\
&\ldots\ldots\ldots \\
X_n &\overset{d}{=} \phi_{n-1}(X_1, \ldots, X_{n-1}) T_n + \psi_{n-1}(X_1, \ldots, X_{n-1}),
\end{aligned}
\tag{1}
$$

where ϕ_0, ψ_0 are constants, with $\phi_0 \neq 0$, and the functions

$$\phi_1(x_1), \ldots, \phi_{n-1}(x_1, \ldots x_{n-1}), \psi_1(x_1), \ldots, \psi_{n-1}(x_1, \ldots, x_{n-1}),$$

called parameter functions, are assumed to be continuous at least with respect to each of their arguments $x_1, \ldots, x_{n-1}$ separately, whenever present. It is also assumed that, for $j = 1, \ldots, n-1$, $\phi_j(x_1, \ldots, x_j) \neq 0$. In general, especially in reliability applications of the models to appear in this text, both the following conditions: $\phi_j(0, \ldots, 0) = 1$, and $\psi_j(0, \ldots, 0) = 0$ should hold too. If $\psi_0 = \psi_1(x_1) = \psi_2(x_1, x_2) = \ldots = \psi_{n-1}(x_1, \ldots, x_{n-1}) \equiv 0$, then the scheme (1) reduces to the pattern that will be called '(diagonal) pseudolinear' as it is a generalization of linear mappings in $\mathbf{R}^n$. As it can easily be shown all the transformations (1) are easily reversible and the jacobians of their inverses have remarkably simple product form:

$$\partial(t_1, \ldots, t_n)/\partial(x_1, \ldots, x_n) = [\phi_0]^{-1} \cdot [\phi_1(x_1)]^{-1} \cdots [\phi^{n-1}(x_1, \ldots, x_{n-1})]^{-1}.$$

Our aim is to investigate the joint pdfs of the random vectors $(X_1, \ldots, X_n)$, which are the images of the random vectors $(T_1, \ldots, T_n)$ under the transformations (1). These can easily be obtained using standard methods, and accordingly to the class the distributions all T_i 's belong to, one obtains generalizations of those classes i.e., n-variate ex-normal , ex-exponential, ex-Weibullian or ex-gamma pdfs respectively.

The ex-normals (under the name "pseudonormals") were explored in [Filus and Filus, 2000], [Filus and Filus, 2001b], [Filus and Filus, 2001a] (see also [Kotz *et al.*, 2000], pages 217-218). The other classes will be investigated in this paper in association with system reliability and maintenance modeling.

2 The n-variate three parameter ex-Weibullian probability densities

Suppose the transformations (1) are applied to the random vectors $(T_1, \ldots, T_n)$ whose independent marginals are distributed according to, in general distinct, three parameter Weibull pdfs $f_1(t_1), \ldots, f_n(t_n)$ respectively. Thus, for $i = 1, \ldots, n$ we have:

$$f_i(t_i) = \begin{cases} (\gamma_i/\beta_i)(t_i-\alpha_i)^{\gamma(i)-1}\exp\left[-(t_i-\alpha_i)^{\gamma(i)}/\beta_i\right], & \text{for } t_i > \alpha_i, \\ 0, & \text{elsewhere,} \end{cases} \tag{2}$$

where the convention $\gamma(i) = \gamma_i$ is to be adopted. The densities (2) will also be denoted by $W(\alpha_i; \beta_i, \gamma(i))$. Using standard procedures one easily obtains the pattern, for ex-Weibullian pdfs of the random vectors $(X_1, \ldots, X_n)$ present in the formula (1), in the following factored form:

$$g(x_1, \ldots, x_n) = g_1(x_1) \cdot g_2(x_2|x_1) \cdots g_n(x_n|x_1, \ldots, x_{n-1}). \tag{3}$$

As it turns out, all the n factors are Weibullian pdfs. So the (initial) pdf of X_1 is

$$g_1(x_1) = W(\phi_0\alpha_1 + \theta_0; \beta_1(\phi_0)^{\gamma(1)}, \gamma_(1)), \tag{4}$$

while for each $j = 2, \ldots, n$, the conditional pdf $g_j(x_j|x_1, \ldots, x_{j-1})$ present in (3) is also Weibullian with respect to x_j alone i.e.,

$$g_j(x_j|x_1, \ldots, x_j) = W\left(\phi_{j-1}(x_1, \ldots, x_{j-1}) \cdot \alpha_j + \theta_{j-1}(x_1, \ldots, x_{j-1})\,;\right. \tag{5}$$
$$\left.\beta_j \cdot [\phi_{j-1}(x_1, \ldots, x_{j-1})]^{\gamma(j)}, \gamma(j)\right)$$

or, more concisely, as well as more generally (see further the "method of parameter replacement"), as:

$$g_j(x_j|x_1, \ldots, x_{j-1}) = W(A_j(x_1, \ldots, x_{j-1}); B_j(x_1, \ldots, x_{j-1}), \gamma(j)). \tag{6}$$

In practical situations the values $x_1, \ldots, x_{j-1}$ may often be considered 'fixed' (at the "time instant" j).

3 Pseudogamma probabitity densities

The pattern (1), when applied to the random vectors of independent random variables $(T_1, \ldots, T_n)$ distributed as three parameter gammas, produces other interesting class of joint probability distributions of the random vectors $(X_1, \ldots, X_n)$. As before, denote the pdfs of the n random variables T_i by $f_i(t_i)$ for $i = 1, \ldots, n$. This time we have:

$$f_i(t_i) = \begin{cases} \left[\Gamma(\gamma_i) \cdot (\beta_i)^{\delta_i}\right]^{-1}(t_i-\alpha_i)^{\delta(i)-1}\exp\left[-(t_i-\alpha_i)/\beta_i\right], & \text{for } t_i \geq \alpha_i, \\ 0, & \text{elsewhere,} \end{cases} \tag{7}$$

where the constants α_i are the shift parameters, and the positive reals β_i and $\delta(i)$ are the scale and the shape parameters respectively. Denote the pdfs $f_i(t_i)$ in (7) by $G(\alpha_i; \beta_i, \delta(i))$. The method of the construction of the joint pdf of any random vector $(X_1, \ldots, X_n)$ defined by (1) is exactly the same as that for the ex-Weibullians. The general formula for the joint pdf $g(x_1, \ldots, x_n)$ of $(X_1, \ldots, X_n)$ has also the factored form (3). Now $g_1(x_1)$ is the gamma pdf:

$$G(\theta_0 + |\phi_0|\alpha_1; |\phi_0|\beta_1; \delta(1)), \tag{8}$$

while for $j = 2, \ldots, n$, the conditional pdfs in (3) are:

$$g_j(x_j|x_1, \ldots, x_{j-1}) = G\left(\theta_{j-1}(x_1, \ldots, x_{j-1}) + |\phi_{j-1}(x_1, \ldots, x_{j-1})| \cdot \alpha_j;\right. \qquad \left. |\phi_{j-1}(x_1, \ldots, x_{j-1})| \cdot \beta_j; \delta(j)\right), \tag{9}$$

or in a more general form:

$$g_j(x_j|x_1, \ldots, x_{j-1}) = G(A_j(x_1, \ldots, x_{j-1}); B_j(x_1, \ldots, x_{j-1}); \delta(j)). \tag{10}$$

They are the ordinary three parameter gamma densities each considered as a function of the argument x_j only. For this reason the so obtained n-variate pdfs are proposed to be called ex-gamma.

4 Comments

A. Notice that in both the new pdf classes construction, described above, the vectors of shape parameters $(\gamma(1), \ldots, \gamma(n))$, $(\delta(1), \ldots, \delta(n))$ in ex-Weibullian and ex-gamma cases respectively are invariant with respect to the pseudoaffines (1). Therefore their values may stand as a criterion for classification of the ex-Weibullians or ex-gammas. In particular, the vector shape parameters $(1, \ldots, 1)$ uniquely determines the class of the two or one parameter ex-exponentials (the set theoretical intersection of ex-Weibullians and ex-gammas classes), while the vector $(2, \ldots, 2)$ determines subclass of ex-Rayleigh among the ex-Weibullians.

B. Occasionally, it is worth to mention an interesting theoretical fact that for any random vector $(T_1, \ldots, T_n)$ having ex-Weibullian or ex-gamma pdf its image $(X_1, \ldots, X_n)$ under (1) is also ex-Weibullian or ex-gamma respectively (see [Filus and Filus, 2003b] for more details).

C. Each of the considered above n-variate three parameters ex-Weibullian, as well as, each of the ex-gamma pdfs are uniquely determined by one of the two sets of the formulas i.e., by (1), (3), (4), (5) together with (2), or by (1), (3), (8), (9) with (7) respectively. The method described above will be called "transformation method". The use of the pseudoaffine transformations is mathematically an elegant way to define ex-Weibullians or ex-gammas. There is also another way to obtain the same pdfs, namely when the formula (1) in the above two lists is dropped. Moreover, significantly wider classes of ex-Weibullians and ex-gammas, that properly contain the corresponding classes defined by the transformation method may be obtained. This will happen when one replaces the defining formulas (5) and (9) by more general (6) and (10). Actually, in this case, both the classes of the pdfs are uniquely determined by a choice of the corresponding classes of functions $A_j(x_1, \ldots, x_{j-1})$, $B_j(x_1, \ldots, x_{j-1})$. The considered classes of the pdfs may even be more extended if, in (6) and (10) respectively, also the set of (constant) shape parameters $\gamma(j)$ is enlarged by properly chosen set of "shape parameter functions" $C_j(x_1, \ldots, x_{j-1})$. Therefore two distinct methods of the construction are available. The second method that relies on a proper conditioning, we propose to call "method of parameters replacement". The type of conditioning we apply somehow corresponds to the conditioning pattern used, for example, in [Arnold *et al.*, 1992], as well as in [Arnold and Strauss, 1988], [Arnold and Strauss, 1991] and in many other related papers (see references in the first cited position). On the other hand, those ideas essentially differ from the ones, described in our work. In the setting, outlined above, the following two rules make our conditioning method distinct from these presented in the above references: a) the predetermined order in conditioning (see formula (3)), with exactly $n-1$ conditional pdfs chosen to be specified, is imposed b) these $n-1$ conditional pdfs are always completed by exactly one (initial) marginal pdf ($g_1(x_1)$ in (3)). This is noteworthy that, using the method of parameter replacement, the resulting n-variate pdfs are uniquely characterized and constructed in a very simple way by (3), (4) and (6) in the Weibullian case, and by (3), (8) and (10) in the gamma case respectively. Briefly speaking, this second method of construction allows, in a largely "arbitrary" but unique way, to achieve the modeling goals simply by replacing some constant parameters in pdfs, say, $f_j(t_j)$ of the, already considered, independent random variables T_j, by properly chosen continuous functions of the arguments, say, $x_1, \ldots, x_{j-1}$, while 'formally' replacing t_j by x_j.

5 On reliability applications

Constructions of the new pdfs, carried out in this work, have their origin (see [Filus and Filus, 2003b]) in the set of problems associated with stochastic modeling of reliability of multicomponent parallel systems with stochastically dependent life times $X_1, \ldots, X_n$ of the components (for reliability references see for example [Barlow and Proschan, 1975]). As models for such systems the joint probability distributions of the component life times are frequently applied (see, for example [Freund, 1961], [Marshall and Olkin, 1967], [Lu, 1989], and others; see also [Filus, 1991]; for much more exhaustive references see [Kotz *et al.*, 2000]). Even as in the past more then four decades, numerous models in the form of multivariate probability distributions have been invented, various types of old and new physical or biological systems still require models of that type. The two classes of multivariate pdfs here presented are (to our best knowledge) new as both: the mathematical entities, and as a way of stochastic description of physical dependencies between the components. Roughly speaking, in the models of stochastic dependencies, presented, an assumed mechanism of system behavior relies on the following: if one (or more) of the system components, say, e_i fails then some survived component (or set of components) e_j ($j \neq i;\ i, j = 1, \ldots, n$) keeps a memory of the (random) time X_i of their mutual cooperation that affected conditional pdf $g_j(x_j|x_i)$ (one of those given by (5), (6) or (9), (10)) of its life time X_j, given $X_i = x_i < X_j$. It is assumed that the component e_i by its activity changes the environment or work conditions of the component e_j. The continuous influence of e_i on e_j causes either an improvement or a deterioration in the components e_j functioning, so that these changes, during the time $X_i = x_i$, cause the life time X_j of e_j to become statistically longer or shorter than its "original" life times, say, T_j, under "laboratory conditions" (i.e., in an absence of any other component influences). The underlying Weibullian and gamma conditional pdfs were already discussed in this text. Notice also that the laboratory condition life times $T_1, \ldots, T_n$ may be considered to be independent Weibullian or gamma as those described in Section 1. Here, physical act of installation of the set of separate components into a real system may be thought off as, in a way, corresponding to the mathematical relationship (1) between the random vectors $(T_1, \ldots, T_n)$ and $(X_1, \ldots, X_n)$. For a more exhaustive description of such systems together with a stochastic reasoning, on how to model them, see [Filus and Filus, 2003b].

6 On ex-Weibullian stochastic processes

The dimension n of the space $\mathbf{R}^n$ associated with the pattern of the pseudoaffine transformations (1) may be extended, in a natural way, to infinity (i.e., by letting $n \to \infty$). In such a case the infinite version of (1) may be specified as follows:

$$\begin{aligned} X_1 &\overset{d}{=} \phi_0 T_1 + \psi_0, \\ &\ldots\ldots\ldots \\ X_j &\overset{d}{=} \phi_{j-1}(X_1, \ldots, X_{j-1}) T_j + \psi_{j-1}(X_1, \ldots, X_{j-1}), \qquad (11) \\ &\ldots\ldots\ldots \end{aligned}$$

where $j = 2, 3, \ldots$

Using (11) one obtains new classes of stochastic processes $\{X_1, X_2, \ldots\}$ corresponding to some well known processes $\{T_1, T_2, \ldots\}$ chosen. When assuming that all the random variables $T_1, T_2, \ldots$ are independent Weibullian a class of ex-Weibullian random processes $\{X_j\}$, with discrete time $j = 1, 2, \ldots$ is obtained. Notice that, also in a more general case, if all the parameter functions $\phi_{j-1}(\cdot)$, $\psi_{j-1}(\cdot)$ depend on X_{j-1} only, while all the input random variables $T_1, T_2, \ldots$ are independent, the obtained stochastic process (including the ex-Weibullian) will be Markovian (see Proposition 1 in [Filus and

Filus, 2003a]). For such a Markovian case new extensions of the (discrete and continuous time) normal, in particular extensions of the Wiener stochastic processes, are presented in [Filus and Filus, 2003a]. On the other hand, a variety of non-Markovian cases are available too. In the next, an application of the above ex-Weibullian model to some maintenance problems associated with repairable systems will be presented. For this purpose the stochastic processes, chosen as models, will be deliberately assumed to be highly non-Markovian in the sense that all the parameter functions in the defining formula (11) will essentially depend on all the 'previous' random variables $X_1, \ldots, X_{j-1}$.

7 The maintenance models

Suppose, that after each failure, a system is repaired with a possibility of a choice among a finite number of kinds of repair available. These repairs differ each other by a quality of the repair on one side and by costs on the other. For simplicity, the state of the system at any time is assumed to be known. Also, the time- length of repairs are not included in this simplified setting. Let the stochastically dependent times of system functioning between $(j-1)$-th and j-th failure be modeled by X_j, $j = 1, 2, \ldots$ One of the basic features of the emerging new methodology is the following. Suppose that, for some j, a $(j-1)$-th failure occurred. Also suppose, all the "maintenance history" of the system performance i.e., the times $X_1, X_2, \ldots, X_{j-1}$ of work between the previous failures, and the corresponding sequence of kinds of repair $r_1, r_2, \ldots, r_{j-2}$ applied, is recorded. One of the main questions, that may arise at this point, can be stated as follows: what would be the pdf (or just an expectation) of the time X_j 'from now' to the next failure, if an r_{j-1}-th kind of the repair would be chosen? To get an answer, in the considered framework, one of the Weibull conditional pdfs $g(x_j|x_1, \ldots, x_{j-1})$ of X_j, given by (5) or (6) may be applied as a proposed model. In particular, one may consider the following conditional pdf:

$$\begin{aligned} & g_j(x_j|x_1, \ldots, x_{j-1}) = \qquad (12) \\ & \left[\lambda\left(1 + a_{1,k(1)}x_1^{\beta_1} + a_{2,k(2)}x_2^{\beta} + \cdots + a_{j-1,k(j-1)}x_{j-1}^{\beta}\right)\right] x_j^{\gamma-1} \\ & \exp\left\{-\left[\lambda\left(1 + a_{1,k(1)}x_1^{\beta_1} + a_{2,k(2)}x_2^{\beta} + \cdots + a_{j-1,k(j-1)}x_{j-1}^{\beta}\right)\right] x_j^{\gamma}\right\}, \end{aligned}$$

where all the coefficients present in (12) are positive, and each coefficient $a_{i,k(i)}$ depends on the choice of $r_{k(i)}$-th kind of repair that took place directly after an i-th failure, $i = 1, \ldots, j-1$. If one seeks the best policy for choices of the repairs after the failures a set of optimization problems emerges. In particular, a possible aim, that may be considered, would be to balance system efficiency (in sense of maximizing length of the times $X_1, X_2, \ldots$) against total cost of the repairs, in order to attain a maximal expected profit from the systems exploitation. Other model, an alternative to (12), can also be considered using the following class of the conditional (Weibullian in x_j) pdfs:

$$g_j(x_j|x_1,\ldots,x_{j-1}) = \tag{13}$$
$$\left[\lambda \exp\left(b_{1,k(1)}x_1^{\beta_1} + b_{2,k(2)}x_2^{\beta} + \cdots + b_{j-1,k(j-1)}x_{j-1}^{\beta}\right)\right] x_j^{\gamma-1}$$
$$\exp\left[-\lambda \exp\left[\left(b_{1,k(1)}x_1^{\beta_1} + b_{2,k(2)}x_2^{\beta} + \cdots + b_{j-1,k(j-1)}x_{j-1}^{\beta}\right)\right] x_j^{\gamma}\right],$$

where the coefficients $b_{i,k(i)}$, $(i = 1,\ldots,j-1)$ are arbitrary (possibly also negative). Somewhat simplified versions of the models (12), and (13) one obtains if only the conditional expectations of the life times are of interest. Then, for $j = 2, 3, \ldots$ we have the regressions

$$E[X_j|x_1,\ldots,x_{j-1}] = \tag{14}$$
$$\left[\lambda\left(1 + a_{1,k(1)}x_1^{\beta_1} + a_{2,k(2)}x_2^{\beta} + \cdots + a_{j-1,k(j-1)}x_{j-1}^{\beta}\right)\right]^{-1/\gamma} \Gamma(1+1/\gamma),$$

and

$$E[X_j|x_1,\ldots,x_{j-1}] = \tag{15}$$
$$\left\{\lambda \exp\left[b_{1,k(1)}x_1^{\beta_1} + b_{2,k(2)}x_2^{\beta} + \cdots + b_{j-1,k(j-1)}x_{j-1}^{\beta}\right]\right\}^{-1/\gamma} \Gamma(1+1/\gamma).$$

Obviously the expectations (14), (15) correspond to the pdfs (12), (13) respectively. Both cases simplify to the exponential cases when $\gamma = 1$.

8 Analytic examples

Because of the space limitation we only mention that numerous nice examples of the new bivariate pdfs with easy analytical calculations can be given. For more on that we refer readers to [Filus and Filus, 2003b].

References

[Arnold and Strauss, 1988]B.C. Arnold and D.J. Strauss. Bivariate distributions with exponential conditionals. *Journal of the American Statistical Association*, 83:522–527, 1988.

[Arnold and Strauss, 1991]B.C. Arnold and D.J. Strauss. Bivariate distributions with conditionals in prescribed exponential families. *Journal of the Royal Statistical Society, Series B*, 53:365–375, 1991.

[Arnold *et al.*, 1992]B.C. Arnold, E. Castillo, and J.M. Sarabia. *Conditionally Specified Distributions.* Lecture Notes in Statistics 73, Springer, New York, 1992.

[Barlow and Proschan, 1975]R.E. Barlow and F. Proschan. *Statistical theory of reliability and life testing.* Holt, Rinehart and Winston, Inc., New York, 1975.

[Filus and Filus, 2000]J. Filus and L. Filus. A class of generalized multivariate normal densities. *Pakistan J. Statist.*, 16:11–32, 2000.

[Filus and Filus, 2001a]J. Filus and L. Filus. On some bivariate pseudonormal densities. *Pakistan J. Statist.*, 17:1–19, 2001.

[Filus and Filus, 2001b]J. Filus and L. Filus. On the n-variate pseudonormal distributions. *Preprint of the ICCS 7-th Conference,* Lahore, Pakistan, Jan. 2-5, 2001.

[Filus and Filus, 2003a]J. Filus and L. Filus. Construction of some new stochastic processes. *Technical Report,* No. 03-12-19, Department of Mathematics, Northeastern Illinois University, Chicago, USA, 2003.

[Filus and Filus, 2003b]J. Filus and L. Filus. On two new methods for constructing multivariate probability distributions with system reliability motivations. *Technical Report,* No. 03-01-28, Department of Mathematics, Northeastern Illinois University, Chicago, USA, 2003.

[Filus, 1991]J.K. Filus. On a type of dependencies between weibull life times of system components. *Reliability Engineering and System Safety*, 31:267–280, 1991.

[Freund, 1961]J.E. Freund. A bivariate extension of the exponential distribution. *J. Amer. Statist. Assoc.*, 56:971–977, 1961.

[Kotz *et al.*, 2000]S. Kotz, N. Balakrishnan, and N.L. Johnson. *Continuous multivariate distributions. Vol. 1.* Wiley-Interscience, New York, second edition, 2000.

[Lu, 1989]J. Lu. Weibull extensions of the freund and Marshall-Olkin bivariate exponential models. *IEEE Transactions on Reliability*, 38:615–619, 1989.

[Marshall and Olkin, 1967]A.W. Marshall and I. Olkin. A multivariate exponential distribution. *J. Amer. Statist. Assoc.*, 62:30–44, 1967.

A reliability system governed by a LDQBD process

Pérez-Ocón, Rafael[1], Montoro-Cazorla, Delia[2], and Ruiz-Castro, Juan Eloy[1]

[1] Departamento de Estadística e I.O. Universidad de Granada. Spain
(e-mail: rperezo@ugr.es, jeloy@ugr.es ; http://www.ugr.es/ sim jeloy)
[2] Departamento de Estadística e I.O. Universidad de Jaén. Spain
(e-mail: dmontoro@ujaen.es)

Abstract. We study an n-unit system. The system functions as long as there is one unit online and the others in warm standby. When a unit fails it goes to repair. There is a repairman. The units are repaired following the arrival order. The operational and repair times follow phase-type distributions. The warm-standby units have a lifetime exponentially distributed. We construct the Markov model that govers the system and calculate performance measures. The mathematical expressions are algorithmically and computationally implemented, using the Matlab programme.

Keywords: Reliability, Availiability, Markov process, Rate of occurrence of failures (Rocof), Level-Dependent-Quasi-Birth-and-Death process.

1 Introduction

The literature on reliability systems concerning with Markov processes is related to systems with units having exponentially distributed lifetimes or extensions of it, such as Erlang, generalized Erlang or hyperexponential. It is known that the phase-type distributions (PH-distributions) constitute a large class that contains all the previous ones. This class has been studied in detail by [Neuts, 1981] and it has been recently applied in reliability by [Pérez-Ocón and Montoro-Cazorla, 2004a], [Neuts *et al.*, 2000], [Pérez-Ocón and Montoro-Cazorla 2004b].

When PH-distributions are involved in the modelization of systems, the generator of the Markov model that governs the system in certain finite cases has a tri-diagonal block structure, which characterizes the classes of quasi-birth-and-death processes (QBD processes) and level-dependent quasi-birth-and-death processes (LDQBD processes).These processes have been studied in [Latouche and Ramaswami, 1999] and oftenly considered in queueing theory ([Bright and Taylor, 1997], [Naoumov, 1997] and references therein). However, we have no information concerning the application of these processes in reliability theory. Recently, a multiple cold standby system involving PH distributions and governed by a QBD process has been studied by [Pérez-Ocón and Montoro-Cazorla, 2004b] . In the present paper we extend that work considering the system in warm standby, being the lifetime of the units in standby exponentially distributed. The stochastic process that governs the system results then a LDQBD process.

For this system, the stationary probability vector, the availability, and the rate of occurrence of failures are calculated. In addition, the distributions of the up and down periods are determined. The steady-state probability vector is calculated following the general methodology provided by [Naoumov, 1997] for solving linear systems with tri-diagonal block matrices. A numerical example is presented

We summarize the following definitions used in the paper.

Definition 1 *The distribution $H(\cdot)$ on $[0,\infty[$ is a phase-type distribution (PH-distribution) with representation (α, T), if it is the distribution of the time until absorption in a Markov process on the states {1,. . . ,m,m+1} with generator by blocks*

$$\begin{pmatrix} T & T^0 \\ 0 & 0 \end{pmatrix}$$

and initial probability vector (α, α_{m+1}),where α is a row m-vector. We assume that the states {1,. . . ,m} are all transient and m+1 absorbent. The distribution $H(\cdot)$ is given by

$$H(x) = 1 - \alpha \exp(Tx)e, \qquad x \geq 0.$$

It will be denoted that $H(\cdot)$ follows a $PH(\alpha, T)$ distribution.

Definition 2 *A level-dependent quasi-birth-and-death process (LDQBD process) on the state space $E = \{(i,j), 0 \leq i \leq n, 1 \leq j \leq m\}$, is a Markov process the infinitesimal generator of which is given by*

$$Q = \begin{pmatrix} B_{0,0} & B_{0,1} & & & & \\ B_{1,0} & A_1^1 & A_0^1 & & & \\ & A_2^2 & A_1^2 & A_0^2 & & \\ & & \ddots & \ddots & \ddots & \\ & & & A_2^{n-1} & A_1^{n-1} & B_{n-1,n} \\ & & & & B_{n,n-1} & B_{n,n} \end{pmatrix} \tag{1}$$

The general definition of these type of processes can be modified depending on the boundary behavior. If we put $A_0^k = A_0, A_1^k = A_1, A_2^k = A_2$, $k = 1, 2, \ldots, n-1$, we get a QBD process.

Definition 3 *If A and B are rectangular matrices of dimensions $m_1 \times m_2$ and $n_1 \times n_2$ respectively, their Kronecker product $A \otimes B$ is the matrix of dimensions $m_1 n_1 \times m_2 n_2$, written in compact form as $(a_{ij}B)$.*

2 The model

Let us consider a repairable n-system, with one unit online and the rest in one of the following three situations: in warm standby, being repaired or waiting for repair. There is one repairman, which serves following the arrival order of the units. The unit online has a lifetime distributed as a $PH(\alpha, T)$ with m operational phases. The units in warm standby have lifetime distributed following $\exp(\lambda_s)$. The repair time follows a distribution $PH(\beta, S)$ with k repairing phases. The repair is as good as new. These times are independent. If there is a unit online and a repair is completed, it goes to standby. When all the units are non-operational and a repair is completed, the repaired unit becomes the unit online.

For introducing a Markov model it is necessary to identify exponentially distributed states in the evolution of the system. These will be the operational and repair phases. Thus, the states will be triplets indicating theses phases and the number of non-operational units. The state spaces is given by $S = S_1 \cup S_2 \cup S_3$,with

$$\begin{aligned}
S_1 &= \{(0, j), 1 \leq j \leq m\},\\
S_2 &= \{(i, j, l), 1 \leq i \leq n-1, 1 \leq j \leq m, 1 \leq l \leq k\},\\
S_3 &= \{(n, l), 1 \leq l \leq k\},
\end{aligned}$$

where i denotes the number of non-operational units, j the operational phase of the online unit, and l the repair phase of the unit under repair. The system macro-states are given in the set $S = \{i, \, i = 0, 1, \ldots, n\}$.

The infinitesimal generator, Q, is calculated from the transition rates among the macro-states. This generator is composed of blocks and the matrix is like the one given in (1), with the blocks in (2).

In the expressions below, the matrix I denotes the identity matrix of appropriate order.

$$\begin{aligned}
B_{0,0} &= T - (n-1)\lambda_s I,\\
B_{0,1} &= \left[T^0\alpha + (n-1)\lambda_s I\right] \otimes \beta,\\
B_{1,0} &= I \otimes S^0,\\
A_2 &= I \otimes S^0\beta,\\
A_1^{(i)} &= [T \oplus S] - (n-i-1)\lambda_s I, \quad i = 1, 2, \ldots, n-1\\
A_0^{(i)} &= \left[T^0\alpha \otimes I\right] + (n-i-1)\lambda_s I, \quad i = 1, 2, \ldots, n-2\\
B_{n-1,n} &= T^0 \otimes I,\\
B_{n,n-1} &= S^0\alpha \otimes \beta,\\
B_{n,n} &= S.
\end{aligned} \tag{2}$$

3 Stationary probability vector

We use $\pi = (\pi_0, \pi_1, \ldots, \pi_{n-1}, \pi_n)$ to denote the stationary-probability vector, which satisfies the matricial equation $\pi Q = 0$, subject to the normalization condition $\pi e = 1$.

To solve resulting system we use previous results ([Naoumov, 1997], Proposition 18). It is obtained that the stationary vector can be recursively obtained in terms of π_0 and rate matrices as $\pi_j = \pi_0 \prod_{k=0}^{j-1} R_k,\ j = 1, \ldots, n$, being

$$\begin{aligned} R_{n-1} &= -B_{n-1,n} B_{n,n}^{-1}, \\ R_{n-2} &= -A_0^{(n-2)} (A_1^{(n-1)} + R_{n-1} B_{n,n-1})^{-1}, \\ R_{j-1} &= -A_0^{(j-1)} \left(A_1^{(j)} + R_j A_2 \right)^{-1}, \quad j = n-2, \ldots, 2 \\ R_0 &= -B_{0,1} (A_1^{(1)} + R_1 A_2)^{-1} \end{aligned}$$

The vector π_0 is determined by the equation $\pi_0 (B_{0,0} + R_0 B_{1,0}) = \mathbf{0}$ subjected to the normalization condition $\pi_0 \left(\sum_{j=0}^{n} \prod_{i=0}^{j-1} R_i \right) e = 1$.

4 Performance measures

The performance measures will be given by means of the stationary probability vector and, consequently, from the matrices R. Below two of these measures appear, the availability and several rates of occurrence of failures: for the unit online and for the system.

The availability of the system is the probability that the system will be operational, thus:

$$A = \sum_{i=0}^{n-1} \pi_i e = \pi_0 \left(\sum_{i=0}^{n-1} \prod_{k=0}^{i-1} R_k \right) e = 1 - \pi_n e.$$

We now calculate the rate of occurrence of failures for the unit online, whose expression results:

$$v_1 = \pi_0 T^0 + \pi_0 \left(\sum_{i=1}^{n-1} \prod_{k=0}^{i-1} R_k \right) (T^0 \otimes e).$$

The mean number of times that the system is down per unit time.is given by

$$\nu_2 = \pi_{n-1} (T^0 \otimes e) = \pi_0 \left(\prod_{k=1}^{n-2} R_k \right) (T^0 \otimes e).$$

5 Distributions of the up and down periods

It is useful to know the distribution of the times during the system is operational or is being repaired in the long run. These are of special importance in systems that require a high reliability. We will show that these random times follow PH-distributions. In the references we have found different ways to define an up period. One is the timespan between the point at which all the units are initially operational (macro-state 0) and the point at which all the units are not operational by first time (macro-state n). Another definition is the timespan between the instant in which an unit completes its repair while the others are non-operational (the system enters the macro-state $n-1$ from n) and the instant in which for the first time the system is non-operational (enters the macro-state n). For calculating the distribution function of this time, we consider a modified Markov process from the original, with the same operational macro-states and identifying the non-operational macro-states in a new absorbent macro-state that will be denoted by n^*. The up period is the time up to the absorption by the macro-state n^*, and thus the distribution will be a PH-distribution. The generator Q^* of this new Markov process is derived from the expression (1) where the block $B_{n,n-1}$ is a null row vector, $B_{n,n} = 0$ and $B_{n-1,n}$ is replaced by the column vector $B_{n-1,n}e$.

The representation of the up period is (γ_u,L_u), matrix L_u being the one calculated from Q^* eliminating the row and the column corresponding to the macro-state n^*. The initial conditions need to be chosen so as to reflect the physical conditions of the system at time $t = 0$. If all units are operational at this point, the initial vector can be chosen as $(\alpha, 0, \ldots, 0)$. Choosing this definition we focus on the initial warranty period of the system, that is, the time to system failure given that initially all the units are operational. However, if we consider the second definition of the up period given above, the initial vector can be chosen as $(0, \ldots, 0, \alpha \otimes \beta)$. It is possible to express the initial condition in terms of the stationary probability vector, then, the initial vector considering the first definition above can be chosen as

$$\gamma_u = \left[\frac{\pi_0}{\pi_0 e}, \mathbf{0}\right].$$

The operational mean time is

$$MTTF = -\gamma_u L_u^{-1} e.$$

The down period begins when the only operational unit fails (the rest are in repair or waiting for repairing), and finishes at the point when the first repair is completed. This period follows a $PH(\gamma_d, S)$, where γ_d is determined as follows. Let $\gamma_d(l)$ $1 \leq l \leq k$, be the stationary probability that the unit under repairing occupies the phase l. The system initiates its down-period in the infinitesimal interval $(t, t + dt)$ with probability $\pi_n(T^0 \otimes e)dt$, then,

	$n = 10$		
π_0	0.0004	*	*
π_n	0.0463	0.0142	0.0115

	$n = 50$		
π_0	*	*	*
π_n	0.0470	0.0144	0.0117

	$n = 20$		
π_0	*	*	*
π_n	0.0468	0.0144	0.0117

	$n = 100$		
π_0	*	*	*
π_n	0.0471	0.0144	0.0117

Table 1. Stationary probabilities π_0, π_n for different values of n

$$\gamma_d(l) = \frac{\sum_j \pi_{n-1}(j,l)T_j^0}{\pi_{n-1}(T^0 \otimes e)}, \quad 1 \leq l \leq k,$$

being $\pi_{n-1}(j,l)$ the probability that at any time $n-1$ components of the system are down with the unit online in phase j and the unit under repair in phase l. The initial vector yields then

$$\gamma_d = (\gamma_d(l))_{1 \leq l \leq k}.$$

The mean time that the system remains down is given by

$$MTTD = -\gamma_d S^{-1} e.$$

6 Numerical application

In this section we apply the calculations performed above to a practical case, preserving the notation of the previous ones. We consider the following representations for the PH-distributions of the operational and repair times.

$$\alpha = (1,0,0) \qquad\qquad \beta = (1,0,0)$$

$$\mathbf{T} = \begin{pmatrix} -0.0027 & 0.0027 & 0 \\ 0 & -0.008 & 0.008 \\ 0 & 0 & -0.02878 \end{pmatrix}, \mathbf{S} = \begin{pmatrix} -0.02 & 0.02 & 0 \\ 0.01 & -0.08 & 0.07 \\ 0.005 & 0 & -0.1 \end{pmatrix}.$$

Let us study the behavior of the system defined in Section 2 with these numerical values for different number of units n. The stability of the measures in terms of the number of units is calculated. The failure rate for the units in standby will be $\lambda_s = 0.03$.

In Table 1 we present the values of π_0 and π_n for different values of n, showing that for n close to 20 the probabilities remain stable when n increases. The values of π_0 are very close to $\mathbf{0}$ for $n \geq 20$, and the ones corresponding to π_n tends to (0.0470,0.0144,0.0117) when n increases.

These values indicate that there are frequently non-operational units, and the system is down with a probability close to 7.31% when there a few units.

n	A	v_1	v_2	$MTTF$	ρ	L
10	0.928	0.002	0.001	828.193	0.999	8.659
20	0.927	0.002	0.001	842.439	1	18.534
30	0.927	0.002	0.001	851.241	1	28.402
50	0.927	0.002	0.001	862.998	1	48.128

Table 2. Performance measures for different number of units

In Table 2 we present the performance measures that have been introduced in previous sections. We use ρ to denote the utilization factor, that is, the proportion of time that the repairman is busy, and L denotes the mean number of non-operational units. MTTF is the mean time of the up period.

The availability decreases slightly when the number of units increases, and stabilizes at around the 92.7%. The different rate of occurrence of failures change slowly with n. The mean number of units failing per unit time increases softly; for example, for a system with 30 units, the mean time between two consecutive unit failures is about 75.757 t.u. The utilization factor of the repairman is very near to 1, so that the repair system is almost saturated, and the mean number of units in the repair channel consequently increases. The mean number of total failure per unit time of the system is 0.001.

Final note. Taking $\lambda_s = 0$ in our model we have an 1-out-of-n-system, where one unit is online and the others in cold standby. Thus, the stochastic process that governs the system is a quasi-birth-and-death process (QBD process), that has been studied in [Pérez-Ocón and Montoro-Cazorla, 2004a].

Acknowledgement. This research was supported by the Proyecto MTM 2004-03672 from the Ministerio de Ciencia y Tecnologia, Spain.

References

[Bright and Taylor, 1997]L. Bright and P.G. Taylor. Equilibrium distributions for level-dependent quasi-birth-and-death processes. *In S.R. Chakravarty and A.S. Alfa,Eds., Matrix-analytic methods in stochastic models*, pages 359–375, 1997.

[Latouche and Ramaswami, 1999]G. Latouche and V. Ramaswami. *Introduction to Matrix Analytic Methods in Stochastic Modeling.* ASA-SIAM, 1999.

[Naoumov, 1997]V. Naoumov. Matrix-multiplicative approach to quasi-birth-and-death processes analysis. *In S.R. Chakravarty and A.S. Alfa, Eds., Matrix-analytic methods in stochastic models*, pages 87–106, 1997.

[Neuts *et al.*, 2000]M.F. Neuts, R. Pérez-Ocón, and I. Torres-Castro. Repairable models with operating and repair times governed by phase type distributions. *Advances in Applied Probability*, pages 468–479, 2000.

[Neuts, 1981]M.F. Neuts. *Matrix Geometric Solutions in Stochastic Models. An Algorithmic Approach.* John Hopkins, Univ. Press., 1981.

[Pérez-Ocón and Montoro-Cazorla, 2004a]R. Pérez-Ocón and D. Montoro-Cazorla. A multiple system governed by a quasi-birth-and-death process. *Reliability Engineering & System Safety*, pages 187–196, 2004.

[Pérez-Ocón and Montoro-Cazorla, 2004b]R. Pérez-Ocón and D. Montoro-Cazorla. Transient analysis of a repairable system using phase type distributions and geometrics processes. *IEEE Transactions in Reliability*, pages 185–192, 2004.

Estimation Methods for Accelerated Failure Time Model

Zhezhen Jin

Department of Biostatistics
Mailman School of Public Health
Columbia University
722 West 168th Street
New York, NY 10032, USA
(e-mail: zj7@columbia.edu)

Abstract. In the literature, a lot of effort has been devoted to develop effective estimation and inference methods for the accelerated failure time (AFT) model for right censored data. In the talk, we will give a review on the recent development on the estimation and inference methods for the AFT model based on the work in [Jin *et al.*, 2003] and [Jin *et al.*, 2004].

Keywords: Accelerated failure time model, Right censoring, Rank estimation, Least squares method.

Right censored data are common in many scientific fields. The right censored data consist of (X_i, Y_i, δ_i), $i = 1, \cdots, n$, where X is a p-dimensional covariate, $Y = \min\{T, C\}$, with T being the response variable and C being the censoring variable, and $\delta = 1\{T \leq C\}$ being the indicator of censoring.

The accelerated failure time (AFT) model is of the same form as usual linear regression model:

$$\log T_i = X_i^T \beta_0 + \epsilon_i \tag{1}$$

where β_0 is the unknown true $p \times 1$ parameter of interest and ϵ_i $(i = 1, \cdots, n)$ are unobservable independent random errors with a common but completely unspecified distribution function. (Thus, the mean of ϵ is not necessarily 0). The AFT model is an attractive alternative to the popular Cox proportional regression model, [Cox, 1972].

Several approaches have been proposed for the estimation and inference on the AFT model in the literature. Rank-based methods were studied [Tsiatis, 1990], [Wei *et al.*, 1990], [Lai and Ying, 1991], [Lai and Ying, 1992], [Lin and Geyer, 1992], [Ying, 1993], [Fygenson and Ritov, 1994], among many others. Least squares based and M-estimation methods were investigated by [Miller, 1976], [Buckley and James, 1979], [Koul *et al.*, 1981], [Ritov, 1990] and [Lai and Ying, 1991], among many others. Despite theoretical advances, all these approaches are numerically complicated and difficult to implement, especially when the number of covariates is large. These are due to the non-differentiability and non-monotonicity of the estimating functions. Furthermore, the covariance matrices of the estimators are rather difficult to obtain because they involve nonparametric estimation of the underlying unknown density function for ϵ.

Recently, we have developed new rank-based and least squares estimation and inference method for the AFT model [Jin *et al.*, 2003], [Jin *et al.*, 2004]. In [Jin *et al.*, 2003], a class of rank-based estimating functions are developed. The functions are monotone and can be easily solved by linear programming technique. The covariance matrix of the parameter estimators are obtained by a resampling method. In [Jin *et al.*, 2004], a numerically easy to implement least squares method is developed and a resampling method sharing the similar spirit in the rank-estimation is also proposed.

In the talk, we will give a review on the recent development on the estimation and inference methods for the AFT model.

References

[Buckley and James, 1979]I.V. Buckley and I. James. Linear regression with censored data. *Biometrika*, pages 429–436, 1979.

[Cox, 1972]D.R. Cox. Regression models and life-tables (with discussion). *J. R. Statist. Soc. Ser. B*, pages 187–220, 1972.

[Fygenson and Ritov, 1994]M. Fygenson and Y. Ritov. Monotone estimating equations for censored data. *Annals of Statist.*, pages 732–746, 1994.

[Jin *et al.*, 2003]Z. Jin, D.Y. Lin, L.J. Wei, and Z. Ying. Rank-based inference for the accelerated failure time model. *Biometrika*, pages 341–353, 2003.

[Jin *et al.*, 2004]Z. Jin, D.Y. Lin, and Z. Ying. On least-squares regression with censored data. *Manuscript*, 2004.

[Koul *et al.*, 1981]H. Koul, V. Susarla, and J. Van Ryzin. Regression analysis with randomly right-censored data. *Ann. Statist.*, pages 1276–1288, 1981.

[Lai and Ying, 1991]T.L. Lai and Z. Ying. Large sample theory of a modified buckley-james estimator for regression analysis with censored data. *Ann. Statist.*, pages 1370–1402, 1991.

[Lai and Ying, 1992]T.L. Lai and Z. Ying. Linear rank statistics in regression analysis with censored or truncated data. *Journal of Multivariate Statistics*, pages 13–45, 1992.

[Lin and Geyer, 1992]D.Y. Lin and C.J. Geyer. Computational methods for semiparamtric linear regression with censored data. *J. Computational and Graphical Stat.*, pages 77–90, 1992.

[Miller, 1976]R. G. Miller. Least squares regression with censored data. *Biometrika*, pages 449–464, 1976.

[Ritov, 1990]Y. Ritov. Estimation in linear regression model with censored data. *Ann. Statist.*, pages 354–372, 1990.

[Tsiatis, 1990]A.A. Tsiatis. Estimating regression parameters using linear rank tests for censored data. *Ann. Statist.*, pages 354–372, 1990.

[Wei *et al.*, 1990]L.J. Wei, Z. Ying, and D.Y. Lin. Linear regression analysis of censored survival data based on rank tests. *Biometrika*, pages 845–851, 1990.

[Ying, 1993]Z. Ying. A large sample study of rank estimation for censored regression data. *Ann. Statist.*, pages 76–99, 1993.

Adaptive Design for Clinical Trials

Mark Chang

Millennium Pharmaceuticals, Inc., Cambridge, MA 02139,USA
(e-mail: Mark.Chang@Statisticians.org)

Abstract. Adaptive design is a trial design that allows modifications to some aspects of the trial after its initiation without undermining the validity and integrity of the trial. Adaptive design makes it possible to discover and rectify inappropriate assumptions in trial designs, lower development costs and reduce the time to market. It has become very attractive to the pharmaceutical industries. In this paper, adaptive designs for clinical trials with multiple endpoints including binary, ordinal, normal, and survival responses are studied using computer simulations.
Keywords: Adaptive design, Sequential design, Adaptive randomization.

1 Overview of Adaptive Design

Drug development is a sequence of complicated decision-making processes. Options are provided at each stage and decisions are dependent on the prior information and the probabilistic consequence of each action (decision) taken. This requires the trial design to be flexible such that it can be modified during the trial process. Adaptive design emerges for this reason and has become very attractive to pharmaceutical industries. An adaptive design is a design that allows modifications to some aspects of the trial after its initiation without undermining the validity and integrity of the trial. The following are the examples of modifications to a trial.

- Sample size re-estimation
- Early stopping due to efficacy or futility
- Adaptive randomization
- Dropping inferior treatment groups

There are several methods available for adaptive designs such as the Fisher's combination of independent p-values [Bauer and Kohne, 1994], Brownian motion [Lan and Demets, 1988] [Lin et al., 1999], conditional power approach [Proschan and Hunsberger 1995], [Babb and Rogatko, 2004], and approach using down-weighting later-stage data [Cui et al., 1999] have been used in group sequential and adaptive designs. However, in this paper, we will discuss the use of computer trial simulation (CTS) for adaptive design. CTS provides a unique and powerful tool for achieving the optimal design. An overall process of an adaptive design is depicted in figure 1.

[0] Invited paper for International Symposium on Applied Stochastic Models and Data Analysis, France, May 2005.

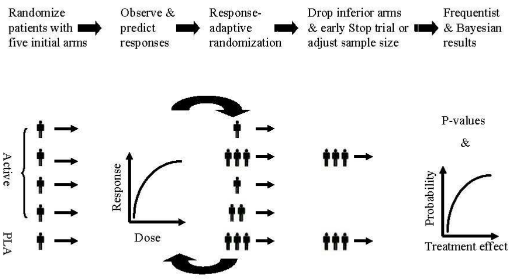

Randomize patients with five initial arms
Observe & predict responses
Response-adaptive randomization
Drop inferior arms & early Stop trial or adjust sample size
Frequentist & Bayesian results
Active
PLA
Response
Dose
P-values
&
Probability
Treatment effect

Fig. 1. Overview of Adaptive Design.

2 Utility-Based Trial Objective

A clinical trial typically involves multiple endpoints such as efficacy, safety and cost. Therefore a single measure, i.e., utility index, which summaries the effects of major endpoints is desirable. The trial objective then becomes to find the dose or treatment with the maximum response probability (rate). The response probability is defined as $Pr(u >= c)$ where u is utility index and c is a threshold. The utility index is the weighted average of trial endpoints such as safety and efficacy. The weights and the threshold are often determined by experts in the relevant field.

3 Dose-Response Model

The response of an ongoing trial can be modeled using a function. We find that the following so-called hyper-logistic function can be used model many different response shapes. The hyper-logistic function is defined by the probability of response

$$\Pr(x = 1) = (a_1 \exp(a_2 x) + a_3 \exp(-a_4 x))^{-a_5}$$

The modeling can be on a continual basis, i.e., the model is updated when new response data become available. This approach refers to the continual re-assessment method (CRM), which can be either Bayesian or frequentist based method. If the observed the responses are used as the basis for an adaptation instead of modeled or predicted response, we call it null-model approach.

4 Adaptation Rules

4.1 Randomization Rules

It is desirable to randomize more patients to superior treatment groups. This can be accomplished by increasing the probability of assigning a patient to the treatment group when the evidence of responsive rate increases in a group. The response-adaptive randomization rule can be Randomized-Play-the-Winner (RPW) [Rosenberger and Lachin, 2002], or Utility offset model.

Response-adaptive randomization requires unblinding the data, which may not feasible at real time. There is often a delayed response, i.e., randomizing the next patient before knowing responses of previous patients. Therefore, it is practical to unblind the data several times during the trial, i.e., group sequential response-adaptive randomization, instead of fully sequential adaptive randomization.

4.2 Early Stopping Rules

It is desirable to stop trial when the efficacy or futility of the test drug becomes obvious during the trial. To stop a trial prematurely, we provide a threshold for the number of subjects randomized and at least one of the following:

(1) Utility rules: The difference in response rate between the most responsive group and the control group exceeds a threshold and the corresponding two-sided 95% naïve confidence interval lower bound exceeds a threshold.

(2) Futility rules: The difference in response rate between the most responsive group and the control is lower than a threshold and the corresponding two-sided 90% naïve confidence interval upper bound is lower a threshold.

4.3 Rules for Dropping Losers

In addition to the response-adaptive randomization, you can also improve the efficiency of a trial design by dropping some inferior groups (losers) during the trial. To drop a loser, we provide two thresholds for (1) maximum difference in response rate between any two dose levels, and (2) the corresponding two-sided 90% naïve confidence lower bound. We may choose to retain all the treatment groups without dropping a loser, and/or to retain the control group with a certain randomization rate for the purpose of statistical comparisons between the active groups and the control.

4.4 Sample Size Adjustment

Sample size determination requires anticipation of the expected treatment effect size defined as the expected treatment difference divided by its standard deviation. It is not uncommon that the initial estimation of the effect size turns out to be too large or small, which consequently leads to an underpowered or overpowered trial. Therefore, it is desirable to adjust the sample size according to the effect size for an ongoing trial.

The sample size adjustment is determined by a power function of treatment effect size, i.e.,

$$N = N_0 \left(\frac{E_{0\,\max}}{E_{\max}} \right)^a \tag{1}$$

where N is the newly estimated sample size, N_0 the initial sample size, and a a constant. The effect size $E_{\max}$ is defined as

$$E_{\max} = \frac{p_{\max} - p_1}{\sigma^2}; \; \sigma^2 = \bar{p}(1 - \bar{p}); \; \bar{p} = \frac{p_{\max} + p_1}{2};$$

$p_{\max}$ and p_1 are the maximum response rates, respectively, and the control response rate, and $E_{0\,\max}$ is the initial estimation of $E_{\max}$.

5 Response-Adaptive Randomizations

The conventional randomization refers to any randomization procedure with a constant treatment allocation probability such as simple randomization. Unlike the conventional randomization, response-adaptive randomization is a randomization in which the probability of allocating a patient to a treatment group is based on the response of the previous patients. The purpose is to improve the overall response rate in the trial. There are many different algorithms such as random-play-the-winner (RPW), the utility-offset model and the maximum utility model.

5.1 Random-Play-the-Winner (RPW)

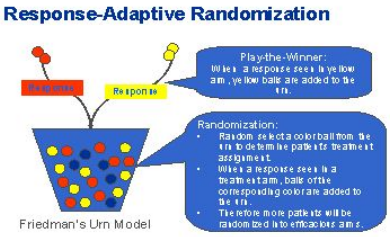

Response-Adaptive Randomization
Response
Response
Play-the-Winner:
When a response seen in yellow arm, yellow balls are added to the urn.
Randomization:
• Random select a color ball from the urn to determine patients treatment assignment.
• When a response seen in a treatment arm, balls of the corresponding color are added to the urn.
• Therefore more patients will be randomized into efficacious arms.
Friedman's Urn Model

Fig. 2. Random-Play-the-Winner

The generalized RPW denoted by $RPW(n_1,\ n_2, ...,\ n_k;\ m_1,\ m_2, ...,\ m_k)$ can be described as follows.

(i) Place n$_i$ balls of the ith color (corresponding to the ith treatment) into a urn ($i = 1, 2, ..., k$), where k is number of treatment groups. There are initially $N = \sum n_i$ balls in the urn.

(ii) Randomly choose a ball from the urn. If it is the ith color, assign the next patient to the ith treatment group.

(iii) Add m$_k$ balls of the ith color to the urn for each response observed in the ith treatment. This creates more chances for choosing the ith treatment.

(iv) Repeat Steps (ii) and (iii).

When $n_i = n$ and $m_i = m$ for all i, we simply write $RPW(n, m)$ for $RPW(n_1,\ n_2, ...,\ n_k;\ m_1,\ m_2, ...,\ m_k)$.

5.2 Utility-Offset Model (UOM)

To have a high probability of achieving target patient distribution among the treatment groups, the probability of assigning a patient to a group should be proportional to the corresponding predicted or observed response rate minus the proportion of patients that have been assigned to the group.

5.3 Maximum Utility Model (MUM)

Maximum utility model for the adaptive-randomization always assigns the next patient to the group that has the highest response rate based on current estimation of either the observed or model-based predicted response rate.

6 Null Model versus Model Approach

It is interesting to compare model and null-model approaches. When sample size is larger than 20 per group, there is no obvious advantage by using the model-based method with respect to the precision and accuracy (Table 1). Therefore, null-model approach will be used in the subsequent simulations.

Dose Level	1	2	3	4	5
Target rate	0.02	0.07	0.37	0.73	0.52
Simulated rate	0.02	0.07	0.36	0.73	0.52
Predicted rate	0.02	0.07	0.40	0.65	0.41
Standard deviation	0.00	0.02	0.11	0.09	0.04
Number of subjects	1.02	2.48	12.6	20.5	13.4

Table 1. Comparisons of Simulations Results

7 Test Statistic

It is very interesting to know that the choice of test statistics for hypothesis tests is very flexible if the analysis is carried out through computer simulations. The only requirement is that the test statistic should be a monotonic function of both the treatment effect δ and sample size n, which can be, for example, $\sqrt{n}\delta$ the treatment difference or the effect size $\frac{\sqrt{n}\delta}{\sigma}$, where σ is the standard deviation of δ. Using computer simulation, it is easy to generate the distributions of the test statistic under the null hypothesis and alternative hypothesis or any other specified conditions for the monitoring purpose.

8 Bias in Rate Estimation and Alpha Adjustment

The commonly used estimators that are based on the assumption of independent samples are often biased in the case of adaptive design. The bias could be as much as 20

The α-adjustment is required when (i) there are multiple comparisons with more than two groups are involved, (ii) There are interim looks, i.e., early stopping for futility or efficacy, or (iii) There is a response-dependent sampling procedure such as response-adaptive randomization and unblinded sample size re-estimation. When samples or observations from the trial are not independent, the response data is no longer normally distributed. Therefore, the p-value from a normal distribution assumption should be adjusted or equivalently the alpha should be adjusted if the p-value is not adjusted. For the same reason, the other statistic estimates from normal assumption should also be adjusted.

9 Simulation Examples

To investigate the effect of the adaptations, we will compare the classic, group sequential and adaptive designs with regards to their operating characteristics using computer simulations. In what follows, each example represents a different trial design. All simulations are performed using ExpDesign Studio (www.CTriSoft.net) [CTriSoft, Intl. 2005]

Examples 1 to 3 will use the following scenario: Assume a phase II oncology trial with two treatment groups¿ The primary endpoint is tumor response (PR and CR) and the estimated response rates for the two groups are 0.2 and 0.3 respectively. We use simulation to calculate the sample size required, given that one-sided alpha = 0.05 and power = 80%.

Example 1: Conventional Design with Two Parallel Treatment Groups

A classic fixed sample size design with 600 subjects will have a power of 81.4% at one-sided $\alpha = 0.025$. The total number of responses per trial is 150 based on 10,000 simulations.

Example 2: Flexible design with Sample Size Re-estimation

Power of a trial is heavily dependent on the estimated effect size; therefore it is desirable to have a design that allows modification of sample size at some point during the trial. Let us re-design the trial in example 1 such that it allows a sample size-re-estimation and then study the robustness of the design.

In order to control the family-wise error rate (FWE) at 0.025, the alpha must be adjusted to 0.023 which can be obtained by computer simulation under the null hypothesis. The average sample size is 960 under the null hypothesis. Using the algorithm for sample size re-estimation (1), where $E_{0\,\max} = 0.1633$ and $a = 2$, the design has 92% power with an average sample size of 821.5.

Now assume the initial effect sizes are not 0.2 versus 0.3 for the two treatment groups. Instead, they are 0.2 and 0.28 respectively. We want to know what the power of the flexible design pertains. Keep everything the same (Also keep Eo_max 0.1633), but change the response rates to 0.2 and 0.28 for the two dose levels and run the simulation again. It turns out that the design has 79.4% power with an average sample size of 855.

Given the two response rates 0.2 and 0.28, the design with a fixed sample size of 880 has a power of 79.4%. We can see that there is a saving of 25 patients by using the flexible design. If the response rates are 0.2 and 0.3, for 92.1% power, the required sample size is 828 with the fixed sample size design, which means that the flexible design saves 6-7 subjects. A flexible design increases power when observed effect size is less than expected, while a traditional design with a fixed sample size either increases or decreases the power regardless of the observed effect size when the sample increases.

Example 3: Adaptive Design Permitting Early Stopping and Sample Size Re-estimation

It is some time desirable to have a design permitting both early stopping and sample size modification.

With an initial sample size of 700 subjects, a grouping size of 350, and a maximum sample size of 1000. The one-sided adjusted alpha is found to be 0.05. The simulation results are presented in the following.

The maximum sample size is 700. The trial will stop if 350 or more are randomized and one of the following criteria is met. (1) The efficacy (utility) stopping criterion: The maximum difference in response rate between any dose and the control is larger than 0.1 with the lower bound of the two-sided 95% naive confidence interval larger than or equal to 0.0. (2) The futility stopping criterion: The maximum difference in response rate between any dose and the control is smaller than 0.05 with the upper bound of the one-sided 95% naive confidence interval smaller than 0.1. The sample size will be re-estimated at the time when there are 350 subjects randomized.

When the null hypothesis is true ($p_1 = p_2 = 0.2$), the average total number of subjects for each trial is 398.8. The probability of early stopping for efficacy is 0.0096. The probability of early stopping for futility is 0.9638.

When the alternative hypothesis is true ($p_1 = 0.2,\ p_2 = 0.3$), the average total number of subjects for each trial is 543.5. The total number of responses per trial is 136. The probability of correctly predicting the most responsive dose level is 0.985 based on observed rates. The probability of early stopping for efficacy is 0.6225. The probability of early stopping for futility is 0.1546. The power for testing the treatment difference is 0.842.

Examples 4 to 6 are for the same scenario of the six arm study with response rates 0.5, 0.4, 0.5, 0.6, 0.7, and 0.55 for the 6 dose levels from dose 1 to 6, respectively.

Example 4: Conventional Design with Multiple Treatment Groups

With 800 subjects, 0.5 response rate under Ho, and grouping size of 100, we found the one-sided adjusted α to be 0.0055. The total number of responses per trial is 433. The probability of correctly predicting the most responsive dose level is 0.951 based on observed rates. The power for testing the maximum effect comparing any dose level to the control is 80%. The powers for comparing each of the 5 dose levels to the control are 0, 0.008, 0.2, 0.796, and 0.048, respectively.

Example 5: Response-Adaptive Design with Multiple Treatment Groups

To further investigate the effect of Random-Play-the-Winner randomization RPW(1,1), a design with 800 subjects, grouping size of 100, and a response rate of 0.2 under null hypothesis is simulated. The one-sided adjusted α is found to be 0.016. Using this adjusted alpha and response rates 0.5, 0.4, 0.5, 0.6, 0.7, and 0.55 for the dose levels 1 to 6, respectively, the simulation indicates that design trial has 86% power and 447 responders per trial on average. In comparison to 80% power and 433 responders for the design with simple randomization RPW(1,0), the adaptive randomization is superior in both power and number of responders. The simulation results also indicate there are biases in the estimated mean response rates in all dose levels except dose level 1, where a fixed randomization rate is used.

Dose level	1	2	3	4	5	6
Response rate	0.5	0.5	0.5	0.5	0.5	0.5
Observed rate	0.50	0.49	0.49	0.49	0.49	0.49

Table 2. Design with RPW(1,1) under H$_o$

Dose level	1	2	3	4	5	6
No. of subjects	200	74	100	133	176	116
Response rate	0.5	0.4	0.5	0.6	0.7	0.55
Observed rate	0.50	0.39	0.49	0.59	0.7	0.54

Table 3. Design with RPW(1,1) under H_a

The average total number of subjects for each trial is 800. The total number of responses per trial is 446.8. The probability of correctly predicting the most responsive dose level is 0.957 based on observed rates. The power for testing the maximum effect comparing any dose level to the control (dose level 1) is 0.861 at a one-sided significant level (alpha) of 0.016. The powers for comparing each of the 5 dose levels to the control are 0, 0.008, 0.201, 0.853, and 0.051, respectively.

Example 6: Adaptive Design with Dropping Losers

Implementing the mechanism of dropping loser can also improve the efficiency of a design. With 800 subjects, grouping size of 100, a response rate of 0.2 under the null hypothesis, and fixed randomization rate in dose level 1 at 0.25, an inferior group (loser) will be dropped if the maximum difference in response between the most effective group and the least effective group (loser) is larger than 0 with the lower bound of the one-sided 95% naive confidence interval larger than or equal to 0. Using the simulation, the adjusted alpha is found to be 0.079. From the simulation results below, more biases can be observed with this design. The design has 90% power with 467 responders. The probability of correctly predicting the most responsive dose level is 0.965 based on observed rates. The powers for comparing each of the 5 dose levels to the control (Dose level 1) are 0.001, 0.007, 0.205, 0.889, and 0.045, respectively. The design is superior to both RPW(1,0) and RPW(1,1).

Dose level	1	2	3	4	5	6
Response rate	0.5	0.5	0.5	0.5	0.5	0.5
Observed rate	0.50	0.46	0.46	0.46	0.46	0.46

Table 4. Bias in Rate with dropping losers under H_o

10 Summary

From classic design to group sequential design to adaptive design, each step forward has an increased complexity and at the same time improves the efficiency of clinical trials. Adaptive design can increase the number of responses

Dose level	1	2	3	4	5	6
No. of subjects	200	26	68	172	240	95
Response rate	0.5	0.4	0.5	0.6	0.7	0.55
Observed rate	0.50	0.37	0.46	0.57	0.69	0.51

Table 5. Bias in Rate with dropping losers under H_a

in a trial and provide more benefits to the patient in comparison to the classic design. With sample size re-estimation, an adaptive design can preserve the power even when the initial estimations of treatment effect and its variability are inaccurate. In the case of a multiple-arm trial, dropping inferior arm or response-adaptive randomization can improve the efficiency of a design dramatically. Finding analytic solutions for adaptive designs is theoretically challenging. However, computer simulation makes it easier to achieve an optimal adaptive design. It allows a wide range of test statistics as long as they are monotonic functions of treatment effects. Adjusted alphas and p-values due to response-adaptive randomization and other adaptations with multiple comparisons can be easily determined using computer simulations. Unbias in point estimation with adaptive design has not completely revolved yet using computer simulations. However, the bias can be ignored in practice by using a proper grouping size (cluster) such that there are only a limited number of adaptations (< 8).

References

[Bauer and Kohne, 1994]Bauer, P. and Kohne, K. (1994). Evaluation of experiments with adaptive interim analyses. Biometrics 50, 1029-1041. Correction in Biometrics 52, 380.

[Babb and Rogatko, 2004]Babb, J.S. and Rogatko, A. (2004). Bayesian methods for cancer phase I clinical trials, Advances in Clinical Trial Biostatistics, Nancy L. Geller (ed.), Marcel Dekker, Inc, 2004.

[CTriSoft, Intl. 2005]CTriSoft, Intl. (2005). ExpDesign Studio Manual, Lexington, MA, USA.

[Cui et al., 1999]Cui, L., Hung, H.M.J. and Wang, S.J. (1999). Modification of sample size in group sequential clinical trials. Biometrics 55, 853-857.

[Lan and Demets, 1988]Lan, K.K.G. and Demets D.L. (1988), Discrete sequential boundaries for clinical trials. Biometriku (1988), 70, 3, pp, 659-663.

[Lin et al., 1999]Lin, D.Y., Tao, Q. and Ying, Z. (1999), A general theory on stochastic curtailment for censored survival data. JASA, (1999) Vol. 94, No. 446.

[O'Quigley et al., 1990]O'Quigley, J., Pepe, M., and Fisher, L. (1990). Continual reassessment method: A practical design for phase I clinical trial in cancer, Biometrics 46:33-48.

[Proschan and Hunsberger 1995]Proschan, M.A. and Hunsberger, S.A. (1995). Design extension of studies based on conditional power. Biometrics 51, 1315-1324.

[Rosenberger and Lachin, 2002]Rosenberger, W.F. and Lachin J.M. (2002). Randomization in Clinical Trials, John Wiley & Sons, Inc., New York.

Bayesian Analysis for Markers and Degradation

Mei-Ling Ting Lee[12], Maria Shubina[1], and Alan Zaslavsky[13]

[1] Biostatistics Department, Harvard School of Public Health, Boston, USA
[2] Channing Laboratory, Brigham and Women's Hospital, Boston, USA
(e-mail: meiling@channing.harvard.edu)
[3] Department of Health Care Policy, Harvard Medical School, Boston, USA

Abstract. Incorporating marker information into analysis of lifetime data is a topic treated in the current literature in many different ways. In this paper we apply a Bayesian approach to the model introduced by Whitmore, Crowder and Lawless in 1998. In their model they assumed that observable marker process and a latent "true" degradation process together follow a bivariate Wiener process with marker value available at the failure time with censoring. Using data augmentation method for the latent degradation for surviving subjects we construct a full Bayesian model with a closed form posterior distribution. As a sampling procedure we use Metropolis-Hastings within Gibbs algorithm. The model and estimating procedure are applied to a simulated data set from the original article by Whitmore, Crowder and Lawless in order to evaluate the performance of our algorithm. Our method appears to work well, while allowing also to incorporate prior information on the parameters of the model, which can be available from previous studies in similar populations.

Keywords: Marker, Degradation, Latent Models, Bayesian Inference.

1 Introduction

1.1 Markers, degradation and thresholds

Many articles in the literature have focused on incorporating auxiliary information, such as markers, in modelling lifetime data. For good reviews, see Fleming, Prentice, Pepe and Glidden [Fleming *et al.*, 1994], Lefkopoulou and Zelen [Lefkopoulou and Zelen, 1995], Jewell and Kalbfleisch [Jewell and J.D. Kalbfleisch, 1996], Shi, Taylor and Munoz [Shi *et al.*, 1996], among others. On the basis of both proportional and additive hazards models, Lin, Fleming and DeGruttola [Lin *et al.*, 1997] incorporated a time-varying covariate marker as a marker process and considered a variety of models for the marker process.

Another school of thought, represented by the work of Whitmore [Whitmore, 1979], [Whitmore, 1995], Doksum and Hoyland [Doksum and Hoyland, 1992], Doksum and Normand [Doksum and Normand, 1995], Lu [Lu, 1995], and Whitmore and Schenkelberg [Whitmore and Schenkelberg, 1997], considers several models that relate the occurrences of failure events directly to an observable degradation process. An assumption in many of these models is that an event occurs when observable degradation reaches a threshold. Hence, these models were also referred to as "first-passage time", or "first-hitting time" models (Lee and Whitmore [Lee and Whitmore, 2003]).

Instead of a single observable degradation process to model event occurrences, Whitmore, Crowder and Lawless [Whitmore *et al.*, 1998] (abbreviated herein as WCL) introduced the joint distribution of an observable marker process and an unobservable degradation process. Specifically, they assume that the observable marker process and a latent but unobservable "true" degradation process together follow a bivariate Wiener process. This bivariate model deals with more realistic situations where failure is not deterministically related to an observable marker. The bivariate model also allows us to evaluate the reliability of the observed marker values in the assessment of the latent degradation of a subject. Although most of the earlier papers assume that degradation follows a Wiener process, other forms of degradation processes have recently been considered. Lawless and Crowder [Lawless and Crowder, 2004] used a gamma increment process; and Aalen and Gjessing [Aalen and Gjessing, 2004] modelled survival data using an Ornstein-Uhlenbeck process.

1.2 Motivation of the proposed Bayesian methods

Most of the papers listed above formulated their models and estimation procedures using a conventional frequentist paradigm. On the basis of the univariate degradation model discussed by Lu [Lu, 1995], Pettit and Young [Pettit and Young, 1999] adopt a conventional Bayesian approach. Using uniform priors for the threshold level and proper priors for parameters of degradation, Pettit and Young derived inferences for parameters of both the degradation process and the threshold level to make predictions regarding future events. They used a Gibbs sampler to sample from the posterior distributions of model parameters and estimated predictive distributions of failure times. For a newly enrolled subject, they estimated future degradation levels at different times using estimated parameters averaged over all samples. They applied their methods to a simulated dataset obtained from Lu [Lu, 1995] and compared their results to those obtained by using ML estimators. The paper by Pettit and Young, however, considered only the univariate degradation model.

In this article, we consider the use of Bayesian inference procedures for joint modelling of marker and degradation processes using the bivariate methodology introduced by Whitmore, Crowder, and Lawless (WCL) [Whitmore *et al.*, 1998]. Unlike the conventional Bayesian methods adopted by Pettit and Young, we needed to incorporate the data augmentation technique into our Bayesian models in order to derive the likelihood function in closed form.

We use a full Bayesian approach to make inferences on parameters of both the marker and degradation processes. Also, for surviving subjects, we can model the distribution of residual survival times. For newly enrolled subjects, we can predict their failure times. For subjects expected to survive until a given time with a given marker value, we can predict degradation levels. We applied our model to a simulated dataset from WCL.

2 Short Review of the Bivariate Marker and Degradation Model

The bivariate threshold model introduced by WCL assumes that every subject is represented by a path of a bivariate Wiener process $\mathrm{W}(\tau) = \{X(\tau), Y(\tau)\}$, $\tau > 0$, with initial values $\mathrm{W}(0) = \{X(0), Y(0)\} = \{0, 0\}$, drift $\mu = (\mu_X, \mu_Y)$ with nonnegative μ_X, and covariance matrix $\Sigma = \begin{vmatrix} \sigma_{XX} & \sigma_{XY} \\ \sigma_{XY} & \sigma_{YY} \end{vmatrix}$. The component $X(\tau)$ represents the latent process of an unobservable *degradation* (disease) state of a subject and component $Y(\tau)$ denotes a *marker* process that is correlated with the degradation process $X(\tau)$. The strength of the association between the two components of the bivariate Wiener process is described by the correlation coefficient ρ. The subject fails when the degradation process $X(\tau)$ reaches a failure threshold $a > 0$ for the first time. We denote this first-hitting time by the random variable S. It is well known that, when $X(\tau)$ follows a Wiener process, its first-hitting time S has an inverse Gaussian (IG) distribution with corresponding parameters (see, e.g., Chhikara and Folks[Chhikara and Folks, 1989]).

Each subject is observed during a fixed period of time $[0, t]$ with one of two possible outcomes:
(1) **failing subject** – subject fails at some time $s \in [0, t]$;
(2) **surviving subject** – subject is alive and censored at the time t.

For surviving subjects, the marker component $Y(\cdot)$ is measured at the end of the observation period t. For failing subjects, the marker component $Y(\cdot)$ is measured at the failure time.

As a result, the observed data consists of the following forms.
(1) For **failing** subjects:

i) failure time $S = s < t$,the first-passage time for the degradation X,
ii) value $y = Y(s)$, of the marker component Y at the failure time s ,
iii) $X(s) = a$, since failure is the first–passage to the threshold a;

(2) For **surviving** subjects:

i) time $t < S$, which implies that $X(\tau) < a$ for $\tau \in [0, t]$,
ii) value $y(t)$ of the marker component Y at the time t.

The model also assumes that the latent degradation component has a nonnegative drift $\mu_X \geq 0$ so as to ensure that all subjects will fail within a finite time period with probability 1.

The corresponding probability distributions for failing and surviving subjects were derived in WCL.

3 The Proposed Bivariate Model with Data Augmentation for the Degradation Process

Two approaches are possible to relate survival information to the latent degradation.

i) Consider only the observed data, and obtain a marginal p.d.f. of the marker component for a surviving subject (see [Whitmore *et al.*, 1998], 2.10). This strategy will result in a rather complicated combination of the p.d.f. and c.d.f. of normal distributions with different means and covariances;

ii) Alternatively, one can augment unobserved degradation values for surviving subjects and treat these values as additional parameters.

In this article, we will take the second approach and construct a full Bayesian model based on complete likelihood function.

3.1 Data augmentation

Assume that there are n independent subjects, and each subject could be observed during a fixed time period $[0, T_i]$, $i = 1, \ldots, n$. Let S_i denote the random failure time variable for the ith subject, $i = 1, \ldots, n$. If the ith subject failed at time $S_i = s_i \leq T_i$, the marker value $Y(s_i) = y_i$ is measured. If the ith subject did not fail during the observation period, then the survival time S_i is unobserved, and the marker value is measured at T_i with $Y(T_i) = y_i$. Let $\delta_i = I(S_i < T_i)$, $i = 1, \ldots, n$, be a *censoring* indicator.

We define a stopping time for the ith subject as $t_i = \begin{cases} s_i, \text{ if } \delta_\mathrm{i} = 1, \\ T_i, \text{ if } \delta_\mathrm{i} = 0. \end{cases}$

Thus, for n subjects, there are three vectors of length n of completely observed data, as follows.

- a vector of stopping times $t = (t_1, t_2, \ldots, t_n)$,
- a vector of censoring indicators $\delta = (\delta_1, \delta_2, \ldots, \delta_n)$,
- a vector $y = (y_1, y_2, \ldots, y_n)$ of values of the Y–component of the Wiener process $\mathrm{W}(\tau)$ at stopping times.

The latent degradation component X is observed only for the failed ($\delta_i = 1$) subjects and is equal to the failure threshold a. Thus the observed data are

$$D_{obs} = (t,\, y,\, \delta,\, (X(t_i) : \delta_i = 1)).$$

To get the "complete" data, we augment the observed data with latent degradation levels for surviving subjects as described below.

The augmented vector of the X–components for all n subjects is defined as $x = (x_1, x_2, \ldots, x_n)$, where $x_i = \begin{cases} a, & \text{if } \delta_\mathrm{i} = 1, \\ \text{augmented value } x_i, & \text{if } \delta_\mathrm{i} = 0. \end{cases}$

Thus, through data augmentation, we get an additional parameter vector $x_s = \{x_i : \delta_i = 0\}$ of length $n - k$, where $k = \sum_{i=1}^{n} \delta_i$.

3.2 Likelihood function for augmented data

Using the "complete" data consisting of $D = (t,\ x,\ y,\ \delta)$, we can easily derive from WCL (2.5), (2.7) and (2.12), and the condition $\mu_X \geq 0$, the likelihood function for the augmented set of parameters μ, Σ, a, x_s:

$$L\ (\mu, \Sigma, a,\ x_s \mid D_{obs}) = \\ \times \prod_{\delta_i=1} p_f(y_i, t_i \mid \mu, \Sigma, a) \prod_{\delta_i=0} p_s(x_i, y_i, t_i \mid \mu, \Sigma, a) I(\mu_X \geq 0). \tag{1}$$

Examination of densities $p_f(\cdot \mid \cdot)$ and $p_s(\cdot \mid \cdot)$ shows that they are overparameterized, and, without loss of generality, we can fix the failure threshold $a = 1$. To simplify notation, we denote the bivariate vector $\mathrm{w}_i = (x_i, y_i),\ i = 1, \ldots, n$. Thus

$$L\ \ (\mu, \Sigma, a,\ x_s \mid D_{obs}) = L(\mu, \Sigma,\ x_s \mid D_{obs}) = \\ = \frac{a^k}{(2\pi)^n |\Sigma|^{n/2}} \prod_{i=1}^{n} t_i^{-1-\delta_i} \cdot \exp\left(-\sum_{i=1}^{n} \frac{(\mathrm{w}_i - t_i\mu)\Sigma^{-1}(\mathrm{w}_i - t_i\mu)'}{2t_i} \right) I(\mu_X \geq 0) \\ \times \prod_{\delta_i=0} \left(\left(1 - \exp\left(-\frac{2a(a - x_i)}{t_i \sigma_{XX}} \right) \right) I(x_i \leq a) \right). \tag{2}$$

The introduction of augmented latent variables resulted in a likelihood function with a closed form and three groups of parameters, namely a 2–dimensional vector of drift parameters μ, a 2×2 symmetric positive definite covariance matrix Σ, and a $(n - k)$–dimensional vector of augmented degradation values x_s.

4 The Prior and Posterior Distributions

To facilitate a Bayesian inference procedure, we need to specify prior distributions. Note, that conditionally on μ and Σ the prior distribution of augmented vector x_s is fully accounted for by the Wiener process model. We propose to use independent prior distributions for each group of parameters μ and Σ because they are related to different features of the trajectories of the Wiener process: drift describes the average path, whereas variance is responsible for the variability of each individual path. Therefore, the joint prior distribution has the following form:

$$\pi(\mu,\ \Sigma,\ x_s) = \pi(\mu)\pi(\Sigma)\pi(x_s \mid \mu,\ \Sigma). \tag{3}$$

Taking into account that the distribution of x_s is defined by the model, the joint posterior distribution has the form

$$p_{post}(\mu,\, \Sigma,\, x_s \mid D_{obs}) \propto L(\mu, \Sigma,\, x_s \mid D_{obs}) \cdot \pi(\mu)\pi(\Sigma). \tag{4}$$

It could be shown that, under some weak restrictions on observed data, and proper $\pi(\mu)$ and $\pi(\Sigma)$, the joint posterior distribution will be proper.

The proper prior distributions could be made by the choice of hyperparameters to be noninformative or informative, depending on availability of a priori information on marker behavior or/and patient population. Since the main part of the likelihood has a Gaussian form (though truncated for μ_X in the current formulation of the model), we suggest using traditional prior distributions for the class of Gaussian models:

$$\pi(\mu) \propto \exp\left(-\frac{1}{2}(\mu-\mu_0)\Sigma_0^{-1}(\mu-\mu_0)'\right) I(\mu_X \geq 0), \tag{5}$$

$$\pi(\Sigma)\text{simInverse Wish}_2(l, R). \tag{6}$$

5 Predictive distributions of survival times

For surviving subject i, the predictive distribution for residual time s_i^{res} is :

$$p(s_i^{res} \mid D_{obs}) = \iiint p(s_i^{res} \mid \mu, \Sigma, x_i) p_{post}(\mu, \Sigma, x_s \mid D_{obs}) d\mu\, d\Sigma\, dx_s. \tag{7}$$

where $p(s_i^{res} \mid \mu,\, \Sigma,\, x_i) = \dfrac{a-x_i}{\sqrt{2\pi\sigma_{XX}(s_i^{res})^3}} \exp\left(-\dfrac{(a-x_i-\mu_X s_i^{res})^2}{2\sigma_{XX}s_i^{res}}\right).$

The integral over the measure generated by the joint posterior distribution of the parameters can be estimated as a mean of the density $p(s_i^{res} \mid \mu,\, \Sigma,\, x_i)$ over a sample from the joint posterior distribution.

For a newly enrolled subject, the predictive distribution for survival time $p(s \mid D_{obs})$ have the same analytical forms as for residual survival times with initial degradation level $x(0) = 0$. The same estimation procedure applies.

6 Computational Implementation

We designed our computational model to be analytically convenient for the implementation of a Gibbs sampler to draw samples from the joint posterior distribution. We suggest using a form of Gibbs sampling, that allows sampling from conditional distributions for blocks of variables. There are three natural groups of parameters: μ, Σ, and x_s. In section 4 we specified the joint prior distribution for the parameters. From the analytical form of the joint posterior distribution we can see that the conditional posterior distributions for parameters μ and Σ are the products of respective conditional posterior distributions with flat priors and the respective prior distributions. The conditional posterior distribution for the vector of augmented values x_s is a product of conditional posterior distributions of its components.

6.1 Conditional posterior distributions for $\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$, and $\boldsymbol{x_s}$

It can be easily shown that the conditional posterior distribution for μ with a flat improper prior is a truncated ($\mu_X \geq 0$) bivariate normal distribution with location parameter $\mu_{flat} = \bar{\mathrm{w}}/\bar{t}$, covariance parameter $\Sigma_\mu = \frac{1}{n\bar{t}}\Sigma$, where $\bar{\mathrm{w}} = \frac{1}{n}\sum_{i=1}^{n} \mathrm{w}_i$, $\bar{t} = \frac{1}{n}\sum_{i=1}^{n} t_i$.

For priors (5) and (6) the conditional posterior distribution for μ is also a truncated ($\mu_X \geq 0$) bivariate normal with location parameter $\mu'_{post} = \Sigma_{\mu_{post}}(\Sigma_\mu^{-1}\mu'_{flat} + \Sigma_0^{-1}\mu'_0)$, covariance parameter $\Sigma_{\mu_{post}} = (\Sigma_\mu^{-1} + \Sigma_0^{-1})^{-1}$.

The convenient way to specify the prior distributions for covariances is to specify them for the inverse matrices. Let denote $\Sigma^{-1} = I\Sigma = \begin{vmatrix} i\sigma_{XX} & i\sigma_{XY} \\ i\sigma_{XY} & i\sigma_{YY} \end{vmatrix}$, so that $\rho = -i\sigma^2_{XY}/\sqrt{i\sigma_{XX}i\sigma_{YY}}$. The kernel for the posterior conditional distribution of $I\Sigma$ with a flat prior can be written as

$$K(I\Sigma) = |I\Sigma|^{\frac{n}{2}} \prod_{\delta_i=0}\Big(1 - \exp\Big(-\frac{2a(a-x_i)i\sigma_{XX}(1-\rho^2)}{t_i}\Big)\Big) \tag{8}$$

$$\times \exp\Big(-\frac{1}{2}\mathrm{tr}\big(I\Sigma \cdot SE\big)\Big), \quad \text{where} \quad SE = \sum_{i=1}^{n} \frac{1}{t_i}(w_i - t_i\mu)'(w_i - t_i\mu).$$

For a Wishart prior for $I\Sigma$ corresponding to (6) with hyperparameters l,and R the kernel for the posterior conditional distribution of $I\Sigma$ is

$$K_{\mathrm{post}}\,(\mathrm{I}\Sigma) = |I\Sigma|^{\frac{n+l-3}{2}} \prod_{\delta_i=0}\left(1 - \exp\Big(-\frac{2a(a-x_i)i\sigma_{XX}(1-\rho^2)}{t_i}\Big)\right) \tag{9}$$

$$\times\ \exp\left(-\frac{1}{2}\mathrm{tr}\left(I\Sigma \cdot R_{\mathrm{post}}^{-1}\right)\right), \quad \text{where} \quad R_{\mathrm{post}} = \left(SE + R^{-1}\right)^{-1}.$$

The kernel for a conditional posterior distribution of the component x_i of the augmented vector x_s for a surviving subject i ($\delta_i = 0$) is

$$K(x_i) = \exp\left(-\frac{(x_i - \mu_{X.Y}(t_i))^2}{2t_i\sigma_{XX.Y}}\right)\left(1 - \exp\left(-\frac{2a(a-x_i)}{t_i\sigma_{XX}}\right)\right) \times I(x_i \leq a),$$

$$\text{where} \ \ \mu_{X.Y} = \mu_X t_i + \sigma_{XY}/\sigma_{YY}(y_i - \mu_Y t_i), \ \sigma_{XX.Y} = \sigma_{XX}(1-\rho^2). \tag{10}$$

6.2 Sampling Schemes for the Conditional Posterior Distributions

As a sampling scheme we propose to use Metropolis-Hastings within Gibbs (MHwG) algorithm, described in Section 6.2 of Chib and Greenberg [Chib and Greenberg, 1995], for three "natural" groups of parameters: μ, Σ, and augmented values x_s of the process component $X(t)$ at the censoring time t. It can be applied to a joint distribution that has one of its conditional distributions in an analytical form that makes it difficult to develop a direct sampling procedure. The MHwG algorithm is a Gibbs sampler that allows sampling the intractable conditional distribution using a Metropolis-Hastings algorithm, whereas all other conditional distributions are sampled directly. For the joint posterior distribution (4) with suggested priors the covariance matrix of the Wiener process $\varSigma$ has an intractable conditional posterior distribution, drift parameter μ could be sampled directly. For independent components of vector x_s, we construct a convenient rejection sampling scheme.

7 Analysis of Simulated Data Set

In order to test the performance of our Bayesian scheme, we applied it to the simulated dataset obtained from WCL (see Table 5.1) and compared their ML estimates of parameters μ and $\varSigma$ to the estimates based on the introduced Bayesian procedure. This dataset was generated by simulating a Wiener process $W = \{X, Y\}$ with parameters $(\mu_X, \mu_Y, \sigma_{XX}, \sigma_{YY}, \rho) = (.1, 1., .4^2, .1^2, .75)$. Fifty sample paths were generated by running steps with time increments $dt = .01$ until the cumulative sums exceeded the threshold level 1 or the number of steps reached 1000, which was equivalent to truncating the paths at time $T = 10$. The generated dataset contains 12 truncated observations.

Since we were interested in comparing our inference procedure to the maximum likelihood estimation of parameters from WCL, we needed to specify a noninformative set of priors. We chose the parameters for the priors (5) and (6) to make them noninformative comparing to the data. It can be shown that the location parameter $\mu_0 = (0, 0)$ and the covariance matrix $\varSigma_0 = \begin{pmatrix} 100 & 0 \\ 0 & 100 \end{pmatrix}$ for (5), and $l = 3$ with the "covariance" matrix $R = 100.0 \begin{pmatrix} 1 & -\rho \\ -\rho & 1 \end{pmatrix}$ with $\rho = 0$ for (6) will be sufficiently noninformative for the WCL dataset. The correlation 0 here corresponds to an a priori hypothesis of no association between the marker and the degradation.

The MHwG algorithm was implemented in S-Plus.

We ran one simulation chain of 12,000 iterations, starting with overdispersed initial values. We considered the first 2000 iterations a "warm-up run" and used the next $10,000$ iterations for inference . The plots of traces of simulated values for parameters μ, augmented latent degradation levels for survivors at the time of censoring, as well as variances and correlation looked rather homogeneous, and allowed us to conclude that the simulation chain has converged.

In Table 1 we present the true parameter values, the ML estimates from WCL, and the values of parameters μ_X, μ_Y, $\sigma_X = \sqrt{\sigma_{XX}}$, $\sigma_Y = \sqrt{\sigma_{YY}}$, and ρ estimated from the simulated Markov chain, We remind that the likelihood function in WCL is the likelihood for the observed data. Bayesian estimates of parameters are sample means based on all simulated samples, even for parameters with large autocorrelation, because it was shown by S. N. MacEachern and L. M. Berliner [MacEachern and Berliner, 1994] that subsampling leads to less efficient estimates. Numbers in parenthesis represent the standard errors of the estimates based on an estimate of the inverse observed information matrix for ML estimates from WCL, and the estimates of the standard deviations of the posterior distributions of parameters. Based on the full sampled chain we calculated standard deviations as square roots from variance estimates. Median estimations are also based on the full chain. High density intervals are estimated by 250–th and 9750–th respective order statistics.

Table 1. True Values and Estimates for the Parameters of the Process

	True	**Estimates**				
	Values	**from WCL**	**Bayesian**			
Parameter		**ML(SD)**	**Mean(SD)**	**Median**	**2.5%**	**97.5%**
μ_X	0.1	0.120(0.023)	0.121(0.022)	0.122	0.077	0.164
μ_Y	1.0	1.012(0.005)	1.012(0.005)	1.012	1.002	1.022
σ_X	.4	0.364(0.039)	0.354(0.036)	0.352	0.289	0.424
σ_Y	.1	0.089(0.008)	0.088(0.008)	0.088	0.075	0.107
ρ	.75	0.737(0.063)	0.721(0.063)	0.729	0.600	0.828

Analysis of autocorrelation functions of parameter samples for μ, σ_{XX}, σ_{YY}, ρ and x_s indicates that samples of location parameters μ and augmented degradation levels x_s are relatively uncorrelated, whereas samples for σ_{XX}, σ_{YY}, and ρ have significant autocorrelations.

To check the stability of the behavior of the samples of σ_{XX}, σ_{YY} and ρ we analyzed subsamples of σ_X, σ_Y and ρ with lags 20 and 50, which are practically uncorrelated. The results are presented in Table 2. It can be seen that mean values and standard deviations are practically unchanged. Histograms for subsamples, which are not presented here, are also similar to those based on full simulated sample.

Table 2. Comparison of Estimates σ_X, σ_Y and ρ by subchains

	True	**Estimates**			
	Values	**from WCL**	**Full Chain**	**Every 20th**	**Every 50th**
Parameter		**ML(SD)**	**Mean(SD)**	**Mean(SD)**	**Mean(SD)**
σ_X	.4	0.364(0.039)	0.3539(0.0358)	0.3529(0.0355)	0.3515(0.0344)
σ_Y	.1	0.089(0.008)	0.0882(0.0084)	0.0878(0.0082)	0.0881(0.0088)
ρ	.75	0.737(0.063)	0.7208(0.0633)	0.7206(0.0616)	0.7227(0.0596)

References

[Aalen and Gjessing, 2004]O.O. Aalen and H.K. Gjessing. Survival models based on the Ornstein-Uhlenbeck process. *LIDA*, pages 407–423, 2004.

[Chhikara and Folks, 1989]R. S. Chhikara and J. L. Folks. *The Inverse Gaussian Distribution: Theory, Methodology, and Applications.* Marcel Dekker, 1989.

[Chib and Greenberg, 1995]S. Chib and E. Greenberg. Understanding Metropolis-Hastings algorithm. *The American Statistician*, pages 238–242, 1995.

[Doksum and Hoyland, 1992]K. Doksum and A. Hoyland. Models for variable-stress accelerated testing experiments based on Wiener processes and inverse Gaussian distribution. *Technometrics*, pages 74–82, 1992.

[Doksum and Normand, 1995]K.A. Doksum and S.-L. Normand. Gaussian models for degradation processes–Part I: Methods for the analysis of biomarker data. *Lifetime Data Analysis*, pages 131–144, 1995.

[Fleming *et al.*, 1994]T.R. Fleming, R.L. Prentice, M.S. Pepe, and D. Glidden. Surrogate and auxiliary endpoints in clinical trials, with potential applications in cancer and AIDS research. *Statistics in Medicine*, pages 167–178, 1994.

[Jewell and J.D. Kalbfleisch, 1996]N.P. Jewell and J.D. J.D. Kalbfleisch. Marker process in survival analysis. *Lifetime Data Analysis*, pages 15–29, 1996.

[Lawless and Crowder, 2004]J.F. Lawless and M.J. Crowder. Covariates and random effects in a gamma process model with application to degradation and failure. *Lifetime Data Analysis*, pages 213–227, 2004.

[Lee and Whitmore, 2003]M.-L. T. Lee and G. A. Whitmore. First hitting time models for lifetime data. In N. Balakrishnan and C.R. Rao, editors, *Handbook of Statistics*, pages 537–543, 2003.

[Lefkopoulou and Zelen, 1995]M. Lefkopoulou and M. Zelen. Intermediate clinical events, surrogate markers and survival. *LIDA*, pages 73–85, 1995.

[Lin *et al.*, 1997]D.Y. Lin, T.R. Fleming, and V. DeGruttola. Estimating the proportion of treatment effect explained by surrogate marker. *Statistics in Medicine*, pages 1515–1527, 1997.

[Lu, 1995]J. Lu. *A Reliability Model Based on Degradation and Lifetime Data.* McGill University, Montreal, Canada, 1995.

[MacEachern and Berliner, 1994]S. N. MacEachern and L. M. Berliner. Subsampling the Gibbs sampler. *The American Statistician*, pages 188–190, 1994.

[Pettit and Young, 1999]L.I. Pettit and K.D.S. Young. Bayesian analysis for inverse Gaussian lifetime data with measures of degradation. *J. Statist. Comput. Simul.*, pages 217–234, 1999.

[Shi *et al.*, 1996]M. Shi, J.M.G. Taylor, and A. Munoz. Models for residual time to AIDS. *Lifetime Data Analysis*, pages 31–49, 1996.

[Whitmore and Schenkelberg, 1997]G.A. Whitmore and F. Schenkelberg. Modelling accelerated degradation data using Wiener diffusion with a time scale transformation. *Lifetime Data Analysis*, pages 27–45, 1997.

[Whitmore *et al.*, 1998]G.A. Whitmore, M.J. Crowder, and J.F. J.F. Lawless. Failure inference from a marker process based on a bivariate Wiener model. *Lifetime Data Analysis*, pages 229–251, 1998.

[Whitmore, 1979]G.A. Whitmore. An inverse Gaussian model for labour turnover. *Journal of the Royal Statistical Society, Series A*, pages 468–478, 1979.

[Whitmore, 1995]G.A. Whitmore. Estimating degradation by a Wiener diffusion process subject to measurement error. *LIDA*, pages 307–319, 1995.

Part XIV

Spatial Processes

A Statistical Analysis of the 2D Discrete Wavelet Transform

Alexandru Isar[1], Sorin Moga[2], and Xavier Lurton[3]

[1] University POLITEHNICA
2 Bd. V. Parvan,
1900 Timisoara, Romania
(e-mail:isar@etc.utt.ro)

[2] GET - ENST Bretagne
Technopôle de Brest Iroise, CS 83818,
29238 BREST Cedex, France
(e-mail: sorin.moga@enst-bretagne.fr)

[3] IFREMER - Centre de Brest
Service Acoustique et Sismique (TMSI/AS)
BP 70 Plouzané 29820 France
(e-mail: lurton@ifremer.fr)

Abstract. The aim of this paper is a complete statistical analysis of the two dimensional discrete wavelet transform, 2D DWT. This analysis represents a generalization of the statistical analysis of the 1D DWT, already reported in literature. The probability density function, the correlation and the first two moments of the coefficients of the 2D-DWT are computed. The asymptotic behaviour of this transform is also studied. The results obtained were used to design a new denoising system dedicated to the processing of SONAR images.

Keywords: Discrete Wavelet Transform, Asymptotic analysis, convergence speed.

1 Introduction

The 2D DWT is a very modern mathematical tool. It is used in compression (JPEG 2000), denoising and watermarking applications. To exploit all its advantages, it must be carefully analyzed. The aim of this paper is the study of this transform from the statistical point of view. Such a complete study was not already reported.

2 The 2D DWT

In this paper the most commonly used 2D DWT is considered. It is built with separable orthogonal mother wavelets, having a given regularity. At every iteration of the DWT, the lines of the input image (obtained at the end of the previous iteration) are low-pass filtered with a filter having the impulse response m_0 and high-pass filtered with the filter m_1. Then the lines of the two images obtained at the output of the two filters are decimated with a factor of 2. Next, the columns of the two images obtained are low-pass filtered with m_0 and high-pass filtered with m_1. The columns of those four images are also decimated with a factor of 2. Four new sub-images (representing the result of the current iteration) are generated. The first one, obtained after two low-pass filterings, is named approximation sub-image (or LL image). The others three are named detail sub-images: LH, HL and HH. The LL image represents the input for the next iteration. In the following, the coefficients of the DWT will be noted with ${}_xD_m^k$, where x represents the image who's DWT is computed, m represents the iteration index (the resolution level) and $k = 1$, for the HH image, $k = 2$, for the HL image, $k = 3$, for the LH image and $k = 4$, for the LL image. These coefficients are computed using the following relation:

$$ {}_xD_m^k\left[n,p\right] = \left\langle x\left(\tau_1,\tau_2\right), \psi_{m,n,p}^k\left(\tau_1,\tau_2\right)\right\rangle \tag{1}$$

where the wavelets can be factorized:

$$\psi_{m,n,p}^k\left(\tau_1,\tau_2\right) = \alpha_{m,n,p}^k\left(\tau_1\right)\cdot\beta_{m,n,p}^k\left(\tau_2\right) \tag{2}$$

and the two factors can be computed using the scale function $\varphi\left(\tau\right)$ and the mother wavelets $\psi\left(\tau\right)$ with the aid of the following relations:

$$\alpha_{m,n,p}^k\left(\tau\right) = \begin{cases}\varphi_{m,n}\left(\tau\right), \, k = 1,4\\ \psi_{m,n}\left(\tau\right), \, k = 2,3\end{cases} \tag{3}$$

$$\beta_{m,n,p}^k\left(\tau\right) = \begin{cases}\varphi_{m,n}\left(\tau\right), \, k = 2,4\\ \psi_{m,n}\left(\tau\right), \, k = 1,3\end{cases} \tag{4}$$

where:

$$\varphi_{m,n}\left(\tau\right) = 2^{-\frac{m}{2}}\varphi(2^{-m}\tau - n) \tag{5}$$

$$\psi_{m,n}\left(\tau\right) = 2^{-\frac{m}{2}}\psi\left(2^{-m}\tau - n\right) \tag{6}$$

3 The pdfs of the wavelet coefficients

These pdfs can be computed following the description of the 2D DWT given in the previous paragraph. In fact each sub-image has its own pdf. The pdfs computation is based on the relation between the pdfs of the random variables from the input and the output of a digital filter. This is a sequence of convolutions which number is equal with the number of the filter coefficients. The pdfs of the wavelet coefficients, ${}_xD_m^k$, can be expressed with the aid of the pdf of the input image, x, using the relation, [1]:

$$f_{{}_xD_m^k}\left(a\right) = \star_{q_1=1}^{M(k)}\ldots\star_{r_m=1}^{M_0} f_d(k, q_1, r_1, \ldots, q_m, r_m, a) \tag{7}$$

where:

$$f_d\left(k, q_1, ..., r_m, a\right) = G\left(k, q_1, ..., r_m\right) f_x\left(G\left(k, q_1, ..., r_m\right) a\right) \tag{8}$$

and:

$$G\left(k, q_1, ..., r_m\right) = \frac{1}{F\left(k, q_1, r_1\right) \prod\limits_{l=2}^{m} m_0\left[q_l\right] m_0\left[r_l\right]} \tag{9}$$

where:

$$F\left(k, q_1, r_1\right) = \begin{cases} m_0\left[q_1\right] m_0\left[r_1\right], & for \quad k = 4 \\ m_0\left[q_1\right] m_1\left[r_1\right], & for \quad k = 3 \\ m_1\left[q_1\right] m_0\left[r_1\right], & for \quad k = 2 \\ m_1\left[q_1\right] m_1\left[r_1\right], & for \quad k = 1 \end{cases} \tag{10}$$

M_0 represents the length of the impulse response $m_0, M_1$the length of m_1 and the numbers of the first two groups of convolutions in relation (7) are given by the relation:

$$M\left(k\right) = \begin{cases} M_0, & for \quad k = 4 \\ M_0, & for \quad k = 3 \\ M_1, & for \quad k = 2 \\ M_1, & for \quad k = 1 \end{cases} and \; N\left(k\right) = \begin{cases} M_0, & for \quad k = 4 \\ M_1, & for \quad k = 3 \\ M_0, & for \quad k = 2 \\ M_1, & for \quad k = 1 \end{cases} \tag{11}$$

In conformity with (7), each pdf of the wavelet coefficients is a sequence of convolutions. Hence, the random variable representing the wavelet coefficients can be written like a sum of independent random variables. So, the central limit theorem can be applied. This is the reason why the pdf of the wavelet coefficients tends asymptotically to a Gaussian, when the number of convolutions in (7) (the DWT iterations number) tends to infinity. This number depends on the mother wavelets used and on the number of iterations of the DWT. For mother wavelets with a long support, this number becomes large very fast (for a small number of iterations). The mother wavelet with the shortest support is the Haar mother wavelets. We have computed, using the relation (7), the pdfs of the coefficients of the 2D DWT of an image, containing a noise distributed following a $log - gamma$ distribution, using the Haar mother wavelets. The support of the mother wavelets used in practice is longer than the support of the Haar mother wavelets, considered in this theoretical case. The difference between the pdfs of the wavelet coefficients obtained after the second iteration and Gaussians is small in this case. So, after two iterations, the pdfs of the wavelet coefficients can be considered Gaussians. For the first two iterations, heavy-tailed models must be considered. Finer analysis, measuring the distance between the real pdfs and Gaussians, are performed in [Foucher *and al.*, 2001], [Achim *and al.*, 2003] and [Xie *and al.*, 2002].

4 The correlation of the wavelet coefficients

The input image, x, represents, in general, the sum of the useful image, s, and of the noise image, n. Because these two random signals are not correlated, the correlation of the wavelet coefficients of the image x, is the sum of the correlations of the wavelet coefficients of the useful image and of the noise image. The correlation function of the wavelet coefficients can be computed using the following relation:

$$\mathit{\Gamma}_{{}_xD_m^k}\left[n_1,n_2,p_1,p_2\right]=E\left\{{}_xD_m^k\left[n_1,p_1\right]\left({}_xD_m^k\left[n_2,p_2\right]\right)^*\right\}$$

$$=\int_{R^4}E\left\{x\left(\tau_1,\tau_2\right)\right\}\cdot\psi_{m,n_1,p_1}^{k*}\left(\tau_1,\tau_2\right)\cdot\psi_{m,n_2,p_2}^{k}\left(\tau_3,\tau_4\right)d\tau_1d\tau_2d\tau_3d\tau_4 \qquad (12)$$

or:

$$\mathit{\Gamma}_{{}_xD_m^k}\left[n_1,n_2,p_1,p_2\right]=\frac{1}{4\pi^2}\int_{R^2}\gamma_x\left(2^{-m}\nu_1,2^{-m}\nu_2\right)\cdot$$

$$\cdot\left|\alpha_2\left\{\psi^k\left(\nu_1,\nu_2\right)\right\}\right|^2\cdot e^{-j\left[\nu_1\left(n_2-n_1\right)+\nu_2\left(p_2-p_1\right)\right]}d\nu_1d\nu_2 \qquad (13)$$

where the first factor under the integral from the right hand side represents the power spectral density of the input image and the second factor represents the power spectral density of the one dimensional mother wavelets used. In the following, the influence of each of these two factors will be analyzed. For the beginning, the influence of the first factor is considered. If the input image is a white noise, with a known variance, z, it can be written:

$$\gamma_n\left(2^{-m}\nu_1,2^{-m}\nu_2\right)=z \qquad (14)$$

and the expression of the wavelet coefficients of the input noise image correlation function becomes:

$$\mathit{\Gamma}_{{}_nD_m^k}\left[n_1,p_1\right]=z\cdot\delta\left[n_1\right]\cdot\delta\left[p_1\right] \qquad (15)$$

This relation was obtained applying some very well known results from harmonic analysis: the Wiener-Hincin identity and the symmetry theorem. A magic property of the orthogonal wavelet bases (the samples of the correlation functions of the corresponding mother wavelets and scaling functions, taken at integer moments, are discrete-time unit impulses) was also used. Hence, the correlation of the wavelet coefficients of a white noise image do not depends on the regularity of the one dimensional mother wavelets used. The same result can be obtained taking in (13) the limit for m (number of iterations) tending to infinity. Indeed, under the integral from the right hand side of (13), only the power spectral density of the input image depends on m. After the limit computation, this function becomes a constant, like in the case when the input image is a white noise. Asymptotically, the 2D DWT transforms every colored noise into a white one. Hence this transform can be regarded as a whitening system, for any regularity of the one dimensional mother wavelets used. So, the wavelet coefficients sequences of the noise component of the input image are white noise sub-images, having the same variance. In the following, some considerations about the influence of the second factor of the product under the integral from the right hand side of the relation (13), will be made. This second factor takes into account the specific of the one dimensional mother wavelets used. It explains how the regularity of the wavelet decomposition affects the coefficients correlation. It can be proved that the convergence speed to a white noise (when m tends to infinity) increases when the regularity (the length of the filters m_0 and m_1) increases. So, the convergence speed to a Gaussian white noise can be increased using one dimensional mother wavelets with higher regularity. The first and second order moments of the wavelet coefficients can be computed using the following relations.

$$E\left\{{}_xD_m^k\left[n_1,p_1\right]\right\}=E\left\{\int_{R^2}x\left(\tau_1,\tau_2\right)\cdot\psi_{m,n_1,p_1}^{k*}\left(\tau_1,\tau_2\right)d\tau_1 d\tau_2\right\}= \tag{16}$$

$$=\begin{cases}0, k=1,2,3\\ 2^m\cdot\mu_x, k=4\end{cases}$$

Only the means of the images formed with the approximation wavelet coefficients are not nulls. The mean of the DWT of the noise component of the input image is given by the relation:

$$E\left\{{}_nD_m^k\left[n_1,p_1\right]\right\}=\begin{cases}0, k=1,2,3\\ -2^m\cdot\mu_n, k=4\end{cases} \tag{17}$$

In practice the number of iterations of the DWT is important. The dimensions of the image built with the approximation wavelet coefficients obtained after the last iteration are smalls. This is the reason why this image is not filtered in the denoising applications based on the use of the DWT. The variance of the wavelet coefficients of the noise component can be computed using the relation:

$$\sigma^2_{{}_xD_m^k}=E\left\{\left|{}_xD_m^k\left[n_1,p_1\right]\right|^2\right\}=\mathit{\Gamma}_{{}_xD_m^k}\left(0,0\right)=$$
$$=\frac{1}{4\pi^2}\int_{R^2}\gamma_x\left(2^m\nu_1,2^m\nu_2\right)\cdot\left|\alpha_2\left\{\psi^k\left(\nu_1,\nu_2\right)\right\}\right|^2 d\nu_1 d\nu_2$$

The DWT of the input noise component, n, has a variance given by:

$$\sigma^2_{{}_nD_m^k}=\begin{cases}z, & k=1,2,3\\ z-2^{2m}\mu_n^2, & k=4\end{cases} \tag{18}$$

This variance is constant for all the images formed using detail wavelet coefficients. Hence, it can be estimated using the first HH image. This estimation can be used for the filtering of any other detail image, formed with the detail wavelet coefficients obtained at any iteration. The correlation of the DWT of s is given by:

$$\Gamma_{{}_sD^k_m}\left[n_1,p_1\right] = 2^{2m} \cdot \Gamma_s\left[2^m n_1, 2^m p_1\right] \tag{19}$$

its mean by:

$$E\left\{{}_sD^k_m\left[n_1,p_1\right]\right\} = \begin{cases} 0, & k = 1,2,3 \\ 2^m \cdot \mu_s, & k = 4 \end{cases} \tag{20}$$

and its variance, by:

$$\sigma^2_{{}_sD^k_m} = 2^{2m} \cdot \sigma^2_s \tag{21}$$

So, the variance of the detail wavelet coefficients sequences obtained starting from the useful component of the input image increases when the iteration index increases. All the relations established in this paragraph were used in [Isar and Moga, 2004], for the design of a denoising system for SONAR images.

5 Conclusion

A complete analysis of the 2D DWT was reported. It is proved that the 2D DWT asymptotically converges to the 2D Karhunen-Loève transform. So, the DWT of a colored noise image, with a given probability density function, converges asymptotically to a white Gaussian noise. This is a generalization of the results reported in [Isar *and al.*, 2002], where the case of the 1D DWT was considered. Another reference for the statistical analysis of the 1D DWT is [Pastor and Gay, 1995]. The asymptotic analyses of 1D DWT and 2D DWT have similar results. The pdfs of both wavelet transforms converge asymptotically to Gaussians. Both wavelet transforms converge asymptotically to the corresponding Karhunen-Loève transforms, for any regularity of the one dimensional mother wavelets used. The convergence speed to a Gaussian white noise can be improved increasing the regularity of the one dimensional mother wavelets used. Both wavelet transforms convert a white noise into a white noise with the same variance. All the other results of the statistical analyses of the 1D DWT and 2D DWT (pdfs, correlations, moments) are also similar. Based on the statistical analysis reported in this paper, a new denoising system was built in [Isar and Moga, 2004]. Its performances for the treatment of the SONAR images are also reported. This statistical analysis can be used for compression or watermarking purposes also. Statistical analyses of other wavelet transform will be reported soon.

References

[Foucher *and al.*, 2001]Samuel Foucher, Gozé Bertin Bénié, Jean-Marc Boucher, "Multiscale MAP Filtering of SAR images", *IEEE Transactions on Image Processing*, vol. 10, no.1, January 2001, 49-60.

[Achim *and al.*, 2003]Alin Achim, Panagiotis Tsakalides and Anastasios Bezerianos, "SAR Image Denoising via Bayesian Wavelet Shrinkage Based on Heavy-Tailed Modeling", *IEEE Transactions on Geoscience and Remote Sensing*, Vol. 41, No. 8, August 2003, 1773-1784.

[Xie *and al.*, 2002]Hua Xie, Leland E. Pierce and Fawwaz T. Ulaby, "Statistical Properties of Logarithmically Transformed Speckle", *IEEE Transactions on Geoscience and Remote Sensing*, vol. 40, no. 3, March 2002, 721-727.

[Isar and Moga, 2004]A. Isar, S. Moga, Le débruitage des images SONAR en utilisant la transformée en ondelettes discrète à diversité enrichie, Rapport de recherche, LUSSI-TR-2004-4, Département Logiques des Usages, Sciences Sociales et Sciences de l'Information, Laboratoire Traitement Algorithmique et Matériel de la Communication, de l'Information et de la Connaissance, CNRS FRE 2658, ENST-Bretagne, 2004.

[Isar *and al.*, 2002]A. Isar, A. Cubitchi, M. Nafornita, Algorithmes et techniques de compression, Ed. Orizonturi Universitare Timisoara, 2002.

[Pastor and Gay, 1995]D. Pastor, R. Gay, "Décomposition d'un processus stationnaire du second ordre. Propriétés statistiques d'ordre 2 des coefficients d'ondelettes et localisation fréquentielle des paquets d'ondelettes", Traitement du signal, vol. 12, no. 5, pp. 393-420, 1995.

Stochastic Restoration Of Local-Scale Meteorological Records Under Urban Heat Island Signal

Paulo Lucio[1] and Ricardo Deus[2]

[1] CGE - Centre of Geophysics of Évora
Rua Romão Ramalho, 59, 7000-554, Évora – Portugal
(e-mail: pslucio@uevora.pt, ricardo.deus@meteo.pt)
[2] IM - Instituto Meteorologia
Rua C do Aeroporto 1749-077
Lisboa, PORTUGAL

Abstract. One of the biggest constraints to study meteorological fields is due to the fact that the ground-based meteorological network does not operate over a common time period of adequate length. In general, the biggest drawback is that recorded data available must be gap-filled and quality controlled (coherent and consistent) to provide a reliable continuous reference daily/monthly/yearly time series. Hence, this paper is addressed to procedures for reconstruction and evaluation of extremes air temperature time series obeying a sequential strategy divided in two moments: (1) the interpolation considering the cross-correlation and the autocorrelation time-memory; and (2) the spatial interpolation procedure based upon the "optimum distance" between stations. The latter is accomplished subdividing areas of a 2D region into triangles (simplex) to assess the interpolation structure that make use of the altitude of stations as a weight correction factor. Hence, an integrated model for the restoration of time series was developed, which conjugates small-scale space-time interaction between meteorological stations. To validate this work-algorithm, the diagnostic of extreme air temperatures was accomplished based on the analysis of daily time series (1941-2001) from eighteen meteorological stations placed in the Lisboa (Portugal) metropolitan region. As expected, this innovator and robust reconstruction method has good performance, since more information is introduced in the decision-making system.

Keywords: autocorrelation, bias, barycentric coordinates, statistical quality control, time series reconstruction.

1 Introduction

In general, the biggest drawback in climate time series research and diagnostic is that recorded data available must be gap-filled and quality controlled (coherent and consistent) to provide a reliable continuous reference daily ⇒ monthly ⇒ yearly time series (control or reference series). Hence, this manuscript is addressed to stochastic procedures for reconstruction and evaluation (quality control) of extremes air temperature time series. For this reason, we have created a Time Series Reconstruction via Integrated (Interactive) Modelling algorithm: MIRS, an integrated model for the restitution of time series, which conjugates small-scale "space-time interaction" between meteorological stations, Fig.1. It is basically subdivided in two major steps: (1) The temporal linear interpolation, considering the time-memory; and (2) The spatial linear interpolation based upon the "optimum distance" between stations. The time series have been recovered and the empirical mean squared error (MSE) has been evaluated, taking into account the comparison between records in local neighbourhood (time and space small-scale) presenting the same climatic (seasonal) characteristics.

Throughout this work some predictors for spatiotemporal processes will be derived. It is realised on bases of linear techniques and assuming some conditions for the processes under study. First, it is important to formalise the notion of spatiotemporal process. Hence, consider a random function $\{Z(s,t) : s \in D \subseteq \Re^n; t = 0, \mp 1. \mp 2, ...\}$, realizations of a spatiotemporal stochastic process. Thus, for fixed t,$\{Z(s,t) : s \in D\} \equiv \{Z_t(s) : s \in D\}$ is a purely spatial processes and for a fixed location $s \in D$, $\{Z(s,t) : t = 0, \mp 1. \mp 2, ...\} \equiv \{Z_s(t) : t = 0, \mp 1. \mp 2, ...\}$is a time series. Hence, a spatiotemporal stochastic process is simply an infinite, possibly correlated, sequence of spatial processes in time or vice-versa. For our purpose we will make the distinction between space-partial trajectories $\{Z(t) : t = 0, \mp 1. \mp 2, ...\}$ (reconstruction on temporal domain) and time-partial trajectories $\{Z(s) : s \in D\}$ (reconstruction on spatial domain) of the spatiotemporal stochastic process [Kyriakidis and Journel, 1999].

Let $\hat{Z}(x_0)$ the predictor of a random function on partial trajectories based on the realisations $\{Z(x_1), Z(x_2), ..., Z(x_2)\}$; the prediction error associated is defined as $\varepsilon(x_0) = \hat{Z}(x_0) - Z(x_0)$ and the mean squared error, which is interconnected with the prediction's quality, is $MSE\left[\hat{Z}(x_0)\right] = E\left[\hat{Z}(x_0) - Z(x_0)\right]^2$. The best prediction function in terms of the minimum MSE [Graybill, 1976] is given by:

$$\psi_0\left[Z(x_1), Z(x_2), ..., Z(x_n)\right] = E\left[Z(x_0)|Z(x_1), Z(x_2), ..., Z(x_n)\right],$$

and the best linear prediction function is:

$$\psi_0^*\left[Z(x_1), Z(x_2), ..., Z(x_n)\right] = \lambda_0 + \sum_{i=1}^{n} \lambda_i Z(x_i),\ \lambda_i \in \Re,\ i = 1, 2, ..., n,$$

where $\hat{Z}(x_0) = \lambda_0 + \sum\limits_{i=1}^{n} \lambda_i Z(x_i)$. Moreover, it is well-known that the MSE can be written as follows: $MSE\left[\hat{Z}(x_0)\right] = Var\left[\hat{Z}(x_0) - Z(x_0)\right] + B^2\left[\hat{Z}(x_0) - Z(x_0)\right]$. An optimum predictor must be unbiased and ψ_0^* is unbiased $B\left[\hat{Z}(x_0) - Z(x_0)\right] = 0$, since

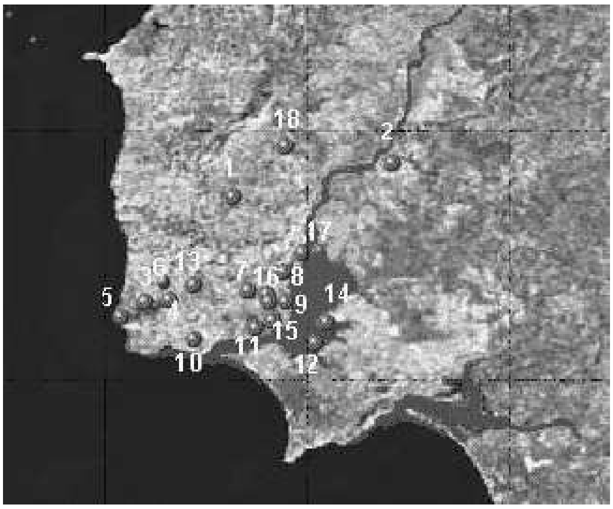


Fig. 1. Map of Lisboa metropolitan area with the location of the automatic meteorological stations of the urban network. Identification of the eighteen meteorological stations and their respective altitudes: 1. Torres-Vedras/Dois-Portos (ID: #139) – 150m; 2. Salvaterra de Magos (ID: #141) – 5m; 3. Colares Sarrazola (ID: #148) – 55m; 4. Sintra (ID: #149) – 200m; 5. Cabo da Roca (ID: #150) – 142m; 6. Paiã/Escola-Agrícola (ID: #153) – 70m; 7. Sacavém (ID: #155) – 9m; 8. Cabo Ruivo (ID: #157) – 16m; 9. Sassoeiros (ID: #160) – 50m; 10 – Lisboa/Tapada-da-Ajuda (ID: #162) – 37m; 11. Lavradio (ID: #166) – 6m; 12. Sintra/Granja (ID: #532) – 134m; 13. Montijo/Base-Aérea (ID: #534) – 14m; 14. Lisboa/Geofísico (ID: #535) – 77m; 15. Lisboa/Portela (ID: #536) – 103m; 16. Alverca/Base-Aérea (ID: #537) – 2m; 17. Ota/Base-Aérea (ID: #539) – 40m; 18. Lisboa/Gago-Coutinho (ID: #579) – 104m.

$$E\left\{\psi_0\left[Z(x_1), Z(x_2), ..., Z(x_n)\right]\right\} = E\left\{E\left[Z(x_0)|Z(x_1), Z(x_2), ..., Z(x_n)\right]\right\} = E\left[Z(x_0)\right]$$

Then ψ_0^* is the best linear unbiased predictor (BLUP) of $Z(x_0)$. Therefore, minimise the MSE is reduce the variance of prediction: $Var\left[\varepsilon(x_0)\right] = Var\left[\hat{Z}(x_0) - Z(x_0)\right]$ and $\{\varepsilon(x_0)\}$, $\forall x_0$ unsampled points, determine series of uncorrelated random variables, supposed to be a zero mean and constant variance - white noise. Moreover, $\hat{Z}(x_0) = \lambda_0 + \sum_{i=1}^{n} \lambda_i Z(x_i) + \varepsilon(x_0)$, where $\sum_{i=1}^{n} \lambda_i Z(x_i)$ is considered the large scale trend surface (1st order component in time or space domain; defined by the wide meaning neighbourhood influence zone) and $\varepsilon(x_0)$ the local component (2nd order factor) or residuals.

2 Daily Reconstruction

2.1 Temporal Domain

The reconstruction model is based on: (1) the use of the own series for filling records without information, considering the strong daily relationship of the internal variation between minimum and maximum air temperature, this association can be verified performing the cross-correlation function analysis between both air temperature attributes for each station in time t = t_0, the antecedent t - 1 and the subsequent $t + 1$ values (two-day time influence); (2) and (3) the use of the serial correlation, considering the strong connection between a record in time t, the antecedent t-2, t-1 and the subsequent $t + 1$, $t + 2$ values (four-day time influence). Hence, the autocorrelation and the partial autocorrelation functions are applied for each series of data, whenever an isolated missing value is found, taking into consideration a second order autoregressive (AR(2)) model [Box *et al.*, 1994].

The coefficient of correlation is used as a measure of the strength of linear association between both variables, a measure of the interdependence of two random variables that ranges in value from -1 to +1, indicating perfect negative correlation at -1, absence of correlation at zero, and perfect positive correlation at +1. The cross-correlation function is a standard method of estimating the degree to which two series are correlated.

Particularly, in this first reconstruction phase, three lags (days) are considered, taking into account the presumed strong linear association between a record in timet, the previous t - 1 and the subsequent $t + 1$ observed values:

$$\beta_1(t) = \lambda_1 \cdot TMAX(t-1) + \lambda_2 \cdot TMAX(t) + \lambda_3 \cdot TMAX(t+1)$$

$$\alpha_1(t) = \lambda_1 \cdot TMIN(t-1) + \lambda_2 \cdot TMAX(t) + \lambda_3 \cdot TMIN(t+1),$$

$$\hat{X}(t) = TMIN(t) = [\alpha_1(t) + \beta_1(t)] + \varepsilon(t) \tag{1}$$

$$\beta_2(t) = \lambda_1 \cdot TMIN(t-1) + \lambda_2 \cdot TMIN(t) + \lambda_3 \cdot TMIN(t+1),$$

$$\alpha_2(t) = \lambda_1 \cdot TMAX(t-1) + \lambda_2 \cdot TMIN(t) + \lambda_3 \cdot TMAX(t+1),$$

$$\hat{Y}(t) = TMAX(t) = [\alpha_2(t) + \beta_2(t)] + \varepsilon(t) \tag{2}$$

where $\lambda_1 = \frac{\hat{\rho}(-1)}{(\hat{\rho}(-1)+\hat{\rho}(0)+\hat{\rho}(1))}$, $\lambda_2 = \frac{\hat{\rho}(0)}{(\hat{\rho}(-1)+\hat{\rho}(0)+\hat{\rho}(1))}$, $\lambda_3 = \frac{\hat{\rho}(1)}{(\hat{\rho}(-1)+\hat{\rho}(0)+\hat{\rho}(1))}$ and $\varepsilon(t)$ denote the empirical series of uncorrelated random variables (residues), whose the ensemble is supposed to be a white noise. It is well known that this prediction method is not optimum at all, since it considers that the attributes are correlated when a linear change in one variable is associated with a change in another one - an unrealistic assumption for daily temperature extremes. The serial correlation is the correlation of a variable with itself over successive time intervals. In climatology we use serial correlation to determine how well the past climate could predicts the future climate and impacts. When the correlation is calculated between a series and a lagged version of itself it is called autocorrelation. The autocorrelation is a correlation coefficient. However, instead of correlation between two different variables, the correlation is between two values of the same variable at times. A high correlation is likely to indicate a periodicity in the signal of the corresponding time duration. The partial autocorrelations, like autocorrelations, are correlations between sets of ordered data pairs of a time series; partial autocorrelations measure the strength of relationship with other terms being accounted for. In practice, for daily data, only two lags are necessary to be considered, taking into account the presumed strong association between a record in time t, the previous (t - 2, t - 1) and the subsequent (t + 1, t + 2) observed values (to get supplementary available information – backward and forward second order autoregressive AR(2) predictor):

$$\alpha(t) = \frac{\hat{\varphi}(-2).X(t-2) + \hat{\varphi}(-1).X(t-1) + \hat{\varphi}(1).X(t+1) + \hat{\varphi}(2).X(t+2)}{\hat{\varphi}(-2) + \hat{\varphi}(-1) + \hat{\varphi}(1) + \hat{\varphi}(2)}$$

$$\Downarrow$$

$$\alpha(t) = \lambda_1.(X(t-2) + X(t+2)) + \lambda_2.(X(t-1) + X(t+1)),$$

$$\hat{X}(t) = \alpha(t) + \varepsilon(t), \tag{3}$$

where$\lambda_1 = \frac{\hat{\varphi}(2)}{2\hat{\varphi}(1)+2\hat{\varphi}(2)}$, $\lambda_2 = \frac{\hat{\varphi}(1)}{2\hat{\varphi}(1)+2\hat{\varphi}(2)}$ and the ensemble $\varepsilon(t)$ is supposed to be a white noise process. The partial autocorrelation at a lag κ is the correlation between residuals at time t from an autoregressive model and observations at lag κ with terms for all intervening lags present in the autoregressive model. The PACF associated to a stochastic process is defined as a sequence of $\hat{\varphi}(\kappa)$'s obtained by the resolution of the Yule-Walker equations for κ = 1,2,3,...:

$$\alpha(t) = \frac{\hat{\varphi}(-2).X(t-2) + \hat{\varphi}(-1).X(t-1) + \hat{\varphi}(1).X(t+1) + \hat{\varphi}(2).X(t+2)}{\hat{\varphi}(-2) + \hat{\varphi}(-1) + \hat{\varphi}(1) + \hat{\varphi}(2)}$$

$$\Downarrow$$

$$\alpha(t) = \lambda_1.(X(t-2) + X(t+2)) + \lambda_2.(X(t-1) + X(t+1)),$$

$$\hat{X}(t) = \alpha(t) + \varepsilon(t), \tag{4}$$

where$\lambda_1 = \frac{\hat{\varphi}(2)}{2\hat{\varphi}(1)+2\hat{\varphi}(2)}$, $\lambda_2 = \frac{\hat{\varphi}(1)}{2\hat{\varphi}(1)+2\hat{\varphi}(2)}$ and the ensemble $\varepsilon(t)$ is supposed to be a white noise process. It is furthermore well known that these of prediction linear methods (2) and (3) are not favourable at all, while they consider that the extremes attributes are correlated when a linear change in one day is associated with a change in the adjacent two days - an improbable postulation, mainly considering severe events. However, we also believe that is worthwhile to make an effort in this direction for regular time series reconstruction.

We considered a "bivariate linear interpolation", for reconstructing daily extreme air temperatures, since for each t - 1 ,t and t + 1 three values (2 of TMIN and 1 of TMAX or vice versa) were available; once verified the strong correlation (in phase - same day) between TMIN and TMAX and a less strong or even weak one (out of phase). If we had used an in phase model we would have a colinearity problem due to the great dependence between the meteorological variables, implying a certain redundancy. So the 1-day lag would be strongly satisfactory for our "bivariate linear interpolation". It was done in the practice, and the equations 3 and 4 are reliable interpolations scheme to reconstruct these kind of meteorological variables taking into account the coupled phenomena.

When all these linear interpolation approaches are applicable and appropriate (the weighted correlations are statistically significant) the decision-making criterion is based on the minimum empirical MSE among the ensembles. These temporal stochastic reconstructions were achieved for overall meteorological stations data series. In Fig.2 we cover the residuals graphical summary that includes: histogram with an overlaid normal curve, box-plot, 95% confidence intervals for the means and 95%confidence intervals for the median. The graphical summary also displays a table of descriptive statistics. No more than 0.3% of the missing values (Tab.1) records were recovered considering the temporal domain! Hence, let us move now to the spatial approach.

2.2 Spatial Domain

The reconstruction model is based on two influence factors: (1) the Euclidean and angular distances, defining a triangle-based network and the areas of each elementary cell (2D simplex); and (2) the Euclidian distance between stations and the respective difference between altitudes (heights). The objective of the first scheme is to construct elementary cells (simplex structure) by triangulation of the convex sub-region S $\subseteq \Re^2$ and the interpolating tool is adopted using the areas of a region subdivided in triangles. The Voronoi region of an object is the region of space closer to the given object than to any other object of the sample. The set of Voronoi triangulations for a set of spatial objects, called a Voronoi diagram (also known as a Dirichlet tessellation or Thiessen polygons), provides a partition of a point-pattern according to its spatial structure. Features of this kind can also be used for analysis of the underlying point process. In practice, the triangulated network of a sub-region of the two-dimensional convex envelope must be determined. Then, let $m \geq 3$ events of a random sample over a sub-region S $\subseteq \Re^2$ and assume that the two-dimensional convex hull of this sub-region has area |A| and that the partition produces (M = 2m - ν - 2) triangles, where ν is the number of extreme points of the two-dimensional partition of the unity$A$, with areas$|A_1|, |A_2|, ..., |A_{2m-\nu-2}|$, respectively. Hence, for this schematic prediction we employ the following topological concepts. The tessellation and oriented areas - A closed ensemble K of the $n-$dimensional space $\Re^n$ is convex if for any $x \in K$, $y \in K$ and $0 \leq \lambda \leq 1$, the linear combination$\{\lambda x + (1-\lambda)\, y \in K\}$. A point $\varpi \in K$ is an extreme point of K if it may not be written as a convex combination of κ different elements of ϖ. A two-dimensional convex envelope of a finite set $C = \{p_1, p_2, ..., p_m\}$ of m

events of $\Re^n$ is defined as the set of all the convex combinations of elements:

$$conv(C) = \{\sum_{i=1}^{m} \lambda_i p_i | \lambda_i \geq 0 \text{ and } \sum_{i=1}^{n} \lambda_i = 1\}. \tag{5}$$

The polygon $C = \{p_1, p_2, ..., p_m\}$ is convex if, and only if, each internal angle is convex, *i.e.*, if each triangle p_{i-1}, p_i, p_{i+1} has the same polygon orientation. A triangle (the 2D simplex structure) defines a coordinate system in the plane (Farin, 1993). Let us consider p_{i-1}, p_i, p_{i+1} non-collinear events onto a triangle $\mathit{\Delta} \subset \Re^2$. Each point $p \in \mathit{\Delta} \subset \Re^2$ can be written as a unique linear combination satisfying

$$\{\sum_{i=1}^{3} \lambda_i p_i : \lambda_i \geq 0 \text{ and} \sum_{i=1}^{3} \lambda_i = 1\}. \tag{6}$$

The parameters $(\lambda_1, \lambda_2, \lambda_3)$ are the barycentric coordinates of p (the relative centroids) in relation to(p_{i-1}, p_i, p_{i+1}). For $(p, p_{i-1}, p_i, p_{i+1})$ with $p = (x, y)$ and$p_j = (x_{ji}, y_{ji})\,, j = i-1, i, i+1$, the parameters $(\lambda_{i-1}, \lambda_i, \lambda_{i+1})$ satisfying some initial conditions are solutions of the system represented below:

$$\begin{cases} \lambda_{i-1} x_{i-1} + \lambda_i x_i \ + \lambda_{i+1} x_{i+1} = x \\ \lambda_{ii-1} y_{i-1} + \lambda_i y_i + \lambda_{i+1} y_{i+1} = y \\ \lambda_{i-1} + \lambda_i + \lambda_{\mathrm{i+1}} \ = 1 \end{cases} . \tag{7}$$

The determinant $(\mathit{\Delta})$ of the solution matrix of the system above is the scalar 2S,

$$\mathit{\Delta} = \begin{vmatrix} x_{i-1} & x_i & x_{i+1} \\ y_{i-1} & y_i & y_{i+1} \\ 1 & 1 & 1 \end{vmatrix} = 2S, \tag{8}$$

where S is the area of the triangle (p_{i-1}, p_i, p_{i+1}). The values of each one of the elements $(\lambda_{i-1}, \lambda_i, \lambda_{i+1})$ can be determined by Cramer's rule:

$$\lambda_{i-1} = \frac{S_{p\ p_i\ p_{i+1}}}{S_{p_{i-1}\ p_i\ p_{i+1}}} = \frac{S_{i-1}}{S}, \lambda_i = \frac{S_{p_{i-1}\ p\ p_{i+1}}}{S_{p_{i-1}\ p_i\ p_{i+1}}} = \frac{S_i}{S}, \lambda_{i+1} = \frac{S_{p\prime_{-1}\ p_i\ p}}{S_{p_{i-1}\ p_i p_{i+1}}} = \frac{S_{i+1}}{S}. \tag{9}$$

The system (10) determines oriented areas. The weights $\lambda_j, j = i-1, i, i+1$ are positive if and only if $S_j, j = i-1, i, i+1$, and S has the same orientation (signal). The barycentric linear interpolation can be used to determine, for continuous phenomena, the unknown values at unsampled points in the spatial point pattern. The barycentric interpolation - Neighbour relationships can also be weighted. Weights based on barycentric coordinates are the subject of this section. Given an ensemble $C = \{x_1, x_2, ..., x_m\}$ of events and real values$f(x_i), i = 1, ..., m$, a piecewise linear function F(x) , defined inside an adequate domain D, such as$F(x) = \sum f(x_i),\ i = 1, ..., m,$ can be obtained. The natural choice for this domain D is a $conv(C)$. However, given a point$x \in conv(C)$, the calculation of F(x) is not obvious. The basic idea is to write$x \in conv(C)$ as a disjoint union of an ensemble of triangles (the simplex). Based on the construction of this triangle network, given $x \in conv(C)$, it can be determined whether x belongs to a particular triangle p_{i-1}, p_i, p_{i+1}, and then F(x) can be computed using equations (7), (8) and (9). The fundamental step in this approach consists of solving the related problem of the triangulation of an ensemble $C = \{x_1, x_2, ..., x_m\}$ based on the construction of $conv(C)$. Notice that, on a two-dimensional space, the triangulation does not exhibit the property of unicity (*i.e.* there are several ways to triangulate a convex network). However, it is possible to determine the optimal number of triangles on the $conv(C)$. Hence, a robust first-order scheme based on barycentric coordinates is used to interpolate the observations at elementary cell vertices on a denser grid. For each unsampled location, the values are evaluated and updated by linear interpolation using the values at the vertices of the triangle. Notice that the precision of the linear interpolation can be estimated with the same properties as the kriging methodology; and without loss of generality the variogram model can be considered linear.

When this approach is legitimate (the weights can be determined, *i.e.*, we can define a triangular network do interpolate the unknown inner point) the stochastic reconstruction is achieved for a particular meteorological station (Fig.3), presenting missing values, represented in the barycentric coordinates system [Lucio and Brito, 2004]. In effect, the spatial linear interpolation is the representation of the data as a parametric model plus a random process function of space$\hat{Z}(s) = \mu(s) + \varepsilon(s)$. The parametric model $\mu(s) = \sum\limits_{i=1}^{n} \lambda_i Z(s_i)$, representing the smooth variation and $\varepsilon(s)$ the deviations from$\mu(s)$.

To clarify the interpolation scheme, consider an ensemble and suppose that we would like to know an attribute value over the point $P = (303.8825\ ,\ -141.2425)$, this point is undoubtedly inside the triangle with vertices $P_{i-1} = (-2436.0075\ ,$ $3814.9875)$, $P_i = (1959.3325,\ 348.9375)$, $P_{i+1} = (172.7925,\ -4022.4825)$, with total area equal to $12,703,497$. The barycentric coordinates satisfy the systems (10), and the output of the program (algorithm in S-Plus) gives the solution: $\lambda_1 = \frac{3,180,636}{12,703,497} = 0.2503748$, $\lambda_2 = \frac{3,946,190}{12,703,497} = 0.3106381$, $\lambda_3 = \frac{5,576,670}{12,703,497} = 0.438987$, with $\sum\limits_{j=i-1}^{i+1} \lambda_j = 1$. Taking an associated continuous function into account, *e.g.* temperature, we obtain an estimate for each point using barycentric interpolation. This estimate value is given by the expression $\hat{Z}(s) = \sum\limits_{j=i-1}^{i+1} \lambda_j Z(s_j)$, where $Z(s_j)$ is the attribute observed on P_j. Hence, in our illustration, let $Z\left(P_{i-1}\right) = 18^o$C, $Z\left(P_i\right) = 20^o$C and $\mathrm{Z(P_{i+1})}$ = 25°C:

$$Z(\hat{P}) = 0.2503748 \times 18^o\mathrm{C} + 0.3106381 \times 25^o\mathrm{C} + 0.438987 \times 20^o\mathrm{C} = 21.05244^o\mathrm{C}.$$

The second phase considers the Euclidian distance between stations: $\mathit{\Delta}\delta^2$ and the respective difference between altitudes (heights): $\mathit{\Delta}z^2$, calculated for each pair of stations ($\mathit{\Delta}z_{j,k} = z_k - z_j$,$\mathit{\Delta}\delta_{j,k} = \sqrt{\left(x_k - x_j\right)^2 + \left(y_k - y_j\right)^2}$,$j, k = 1, ..., 18$) to construct the matrix $\mathit{\Omega} = \frac{1}{\mathit{\Delta}h^2}$ (the inverse of the hypotenuse: $\mathit{\Delta}h^2 = \mathit{\Delta}\delta^2 + \mathit{\Delta}z^2$) that is used as an issue to recover the missing values and factor corrector for the barycentric interpolation. The value to be predicted for a station meteorological ($\kappa \in K$, where K is a closed ensemble), which has no record in a day t is based on the weighted sum ($\frac{\omega_{[i,j]}}{\sum\limits_j \omega_{[j,\kappa]}}$) of the values of the records of all other available stations ($i \neq j = 1, ..., m$) at the same instant t using the idea of "nearest neighbourhood" and the symmetric weight matrix is given by:

$$\Omega = \begin{bmatrix} 0 & \omega_{[2,1]} & \cdots & \omega_{[m,1]} \\ \omega_{[1,2]} & 0 & \cdots & \omega_{[m,2]} \\ \vdots & \vdots & \ddots & \vdots \\ \omega_{[1,m]} & \omega_{[2,m]} & \cdots & 0 \end{bmatrix}. \tag{10}$$

In fact, we consider that height is a dependent variable of longitude ($x \in K$) and latitude ($y \in K$) in terrain surface (this idea is widespread used in mapping) and apply the information $\lambda_{\kappa,j} = \frac{\omega_{[i,j]}}{\sum\limits_j \omega_{[j,\kappa]}}$as an improvement factor after 2D linear interpolation based on barycentric coordinates [Farin, 1993]. Consequently,

$$\hat{Z}(\kappa) = \hat{Z}_0(\kappa) + \sum_j \lambda_{\kappa,j} Z(j) + \varepsilon(\kappa). \tag{11}$$

where $\hat{Z}_0(\kappa)$ is the result of the barycentric interpolation (when applicable, otherwise it is zero) and $Z(j)$ are the contributors (with valuable data) meteorological stations.

Moreover, our method allows us to determine an interpolation criterion, similar to kriging methodology [Kyriakidis and Journel, 1999] since the variogram has to be linear, based on the barycentric coordinates in influence zones. The MSE can be considered independent (uncorrelated) and approximately zero-centred. In addition, they give us an idea about the spatial interpolation misfit based on the variance of prediction. This stochastic reconstruction was achieved for overall meteorological stations data series. In practice, the available records of a station are used to predict the extreme air temperature attributes of the missing value records, considering the neighbourhood and the own station historical records.

As a result of these applications all the series were recovered, except for the first six months of 1997 (181 days without available observations), observed data do not exist in any station or at least it is presumed to not exist. So, we now consider the monthly model identification and characterisation of extreme time series making use of an appropriated forecast model, which might be extrapolated to high levels of the climatological process: the autoregressive integrated moving average (ARIMA) modelling (cf. [Box and Jenkins, 1976]).

THE SCHEME:

<u>DAILY DATA</u> ⇒ <u>MONTHLY DATA</u>

TMIN ⇒ Min {TMIN}, Mean {TMIN} and Max {TMIN},

TMAX ⇒ Min {TMAX}, Mean {TMAX} and Max {TMAX}.

In this work, no more than the extremes was analysed. This stochastic reconstruction was achieved for overall meteorological stations data series and the reference period of validation was 1992-1996.

3 Conclusions

These time series recoverable approach is a very simple way to offer high efficiency results for a low computational cost. Furthermore, this alternative method allows barycentric interpolation of the unsampled points into a two-dimensional simplex (triangular) framework. Moreover, our method allows us to determine an interpolation criterion, similar to kriging methodology since the variogram has to be linear, based on the barycentric coordinates in influence zones. Nevertheless, we can identify two main sources of uncertainties:

i) The induced error when applying the autocorrelation function in the reconstruction of the daily series varies between -3^{o}C and 3^{o}C;

ii) The error generated when estimating values of the air temperature for missing values record considering the spatial reconstruction depends on certain expected conditions. When few stations contribute for filling the gaps, the associate error presents fail values, e.g. for the last 5 years only two stations present records (Lisboa/Geofísico - urban and Lisboa/Gago-Coutinho - suburban); the result under the “heat island effect” may overestimate (contaminate) the calculated values for the other stations;

iii) The integration of new parametrization on the spatial interpolation procedure, like land declination, ocean/river distance can reduce the error associated with this step.

ACKNOWLEDGEMENTS The authors wish to thank the contributions of the Portuguese Meteorological Institute (IM – Portugal). P. S. Lucio was sponsored by a grant from FCT (Portugal). Grateful thanks to the referees.

References

[Box and Jenkins, 1976]G.E.P. Box and G.M. Jenkins. *Time series analysis: forecasting and control.* Holden-Day, San Francisco, 1976.

[Box *et al.*, 1994]G.E.P. Box, G.M. Jenkins, and G.C. Reinsel. *Time series analysis: forecasting and control.* Prentice Hall, 1994.

[Farin, 1993]G. E. Farin. *Curves and surfaces for computed aided geometrical design.* Academic Press, London, 1993.

[Graybill, 1976]F. A. Graybill. *Theory and application of the linear model.* Duxbury Press, Massachusetts, 1976.

[Kyriakidis and Journel, 1999]P. C. Kyriakidis and A. G. Journel. Geostatistical space-time models: a review. *Mathematical Geology*, pages 651–684, 1999.

[Lucio and Brito, 2004]P. S. Lucio and N. L. Brito. Detecting spatial randomness: A stat-geometrical alternative. *Mathematical Geology*, pages 79–99, 2004.

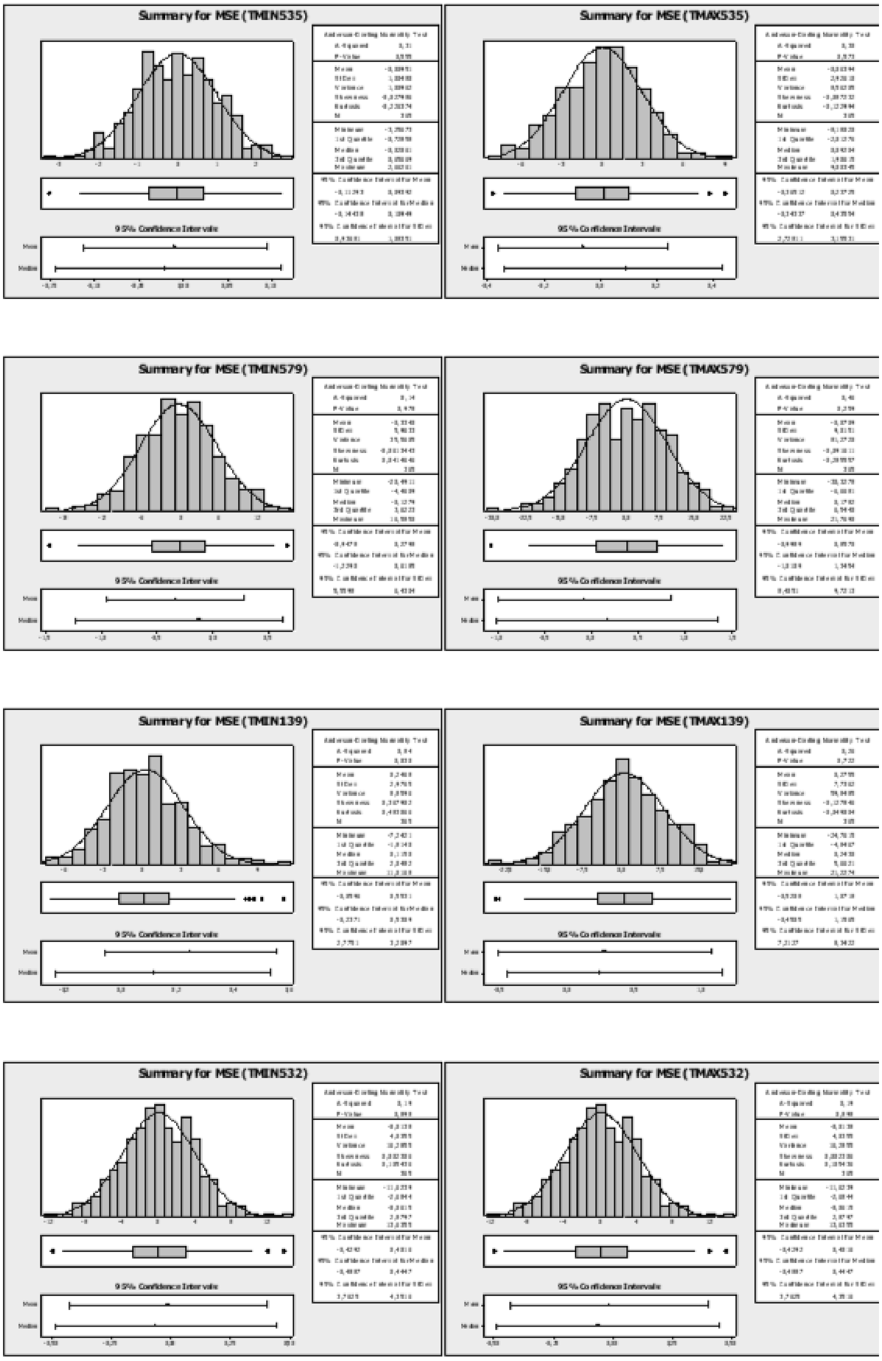
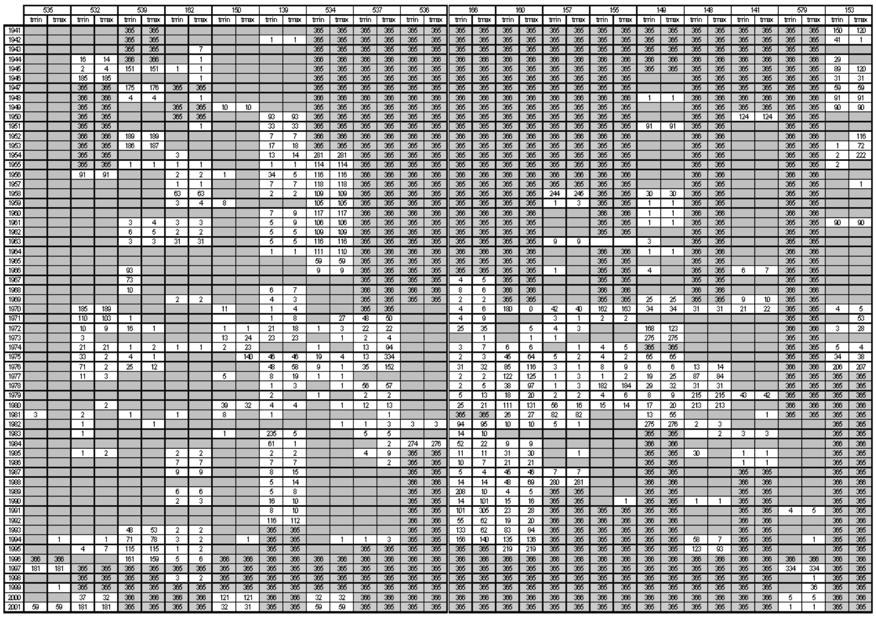


Fig. 2. The residuals graphical summary for TMIN (left) and TMAX (right): Lisboa/Geofísico (535), Lisboa/Gago-Coutinho (579), Torres-Vedras/Dois-Portos (139) and Sintra/Granja (532).



	535		532		539		162		150		139		534		537		536		166		160		157		155		149		148		141		579		153	
	tmin	tmax	tmin	tmax	tmin	tmax	tmin	tmax	tmin	tmax	tmin	tmax	tmin	tmax	tmin	tmax	tmin	tmax	tmin	tmax	tmin	tmax	tmin	tmax	tmin	tmax	tmin	tmax	tmin	tmax	tmin	tmax	tmin	tmax	tmin	tmax
1941					365	365							365	365	365	365	365	365	365	365	365	365	365	365	365	365	365	365	365	365	365	365	365	365	150	120
1942					365	365					1	1	365	365	365	365	365	365	365	365	365	365	365	365	365	365	365	365	365	365	365	365	365	365	41	1
1943					365	365		7					365	365	365	365	365	365	365	365	365	365	365	365	365	365	365	365	365	365	365	365	365	365		
1944			16	14	366	366		1					366	366	366	366	366	366	366	366	366	366	366	366	366	366	366	366	366	366	366	366	366	366	29	
1945			2	4	151	151	1	1					365	365	365	365	365	365	365	365	365	365	365	365	365	365	365	365	365	365	365	365	365	365	89	120
1946			185	185				1					365	365	365	365	365	365	365	365	365	365	365	365	365	365			365	365	365	365	365	365	31	31
1947			365	365	175	176	365	365					365	365	365	365	365	365	365	365	365	365	365	365	365	365			365	365	365	365	365	365	59	59
1948			366	366	4	4		1					366	366	366	366	366	366	366	366	366	366	366	366	366	366	1	1	366	366	366	366	366	366	91	91
1949			365	365			365	365	10	10			365	365	365	365	365	365	365	365	365	365	365	365	365	365			365	365	365	365	365	365	90	90
1950			365	365			365	365			93	93	365	365	365	365	365	365	365	365	365	365	365	365	365	365			365	365	124	124	365	365		
1951			365	365				1			33	33	365	365	365	365	365	365	365	365	365	365	365	365	365	365	91	91	365	365			365	365		
1952			366	366	189	189					7	7	366	366	366	366	366	366	366	366	366	366	366	366	366	366			366	366			366	366		116
1953			365	365	186	187					17	18	365	365	365	365	365	365	365	365	365	365	365	365	365	365			365	365			365	365	1	72
1954			365	365			3				13	14	281	281	365	365	365	365	365	365	365	365	365	365	365	365			365	365			365	365	2	222
1955			365	365	1	1	1	1			1	1	114	114	365	365	365	365	365	365	365	365	365	365	365	365			365	365			365	365	2	
1956			91	91			2	2	1		34	5	116	116	366	366	366	366	366	366	366	366	366	366	366	366			366	366			366	366		
1957							1	1			7	7	118	118	365	365	365	365	365	365	365	365	365	365	365	365			365	365			365	365		1
1958							63	63			2	2	109	109	365	365	365	365	365	365	365	365	244	245	365	365	30	30	365	365			365	365		
1959							3	4	8				105	105	365	365	365	365	365	365	365	365	1	3	365	365	1	1	365	365			365	365		
1960											7	9	117	117	366	366	366	366	366	366	366	366			366	366	1	1	366	366			366	366		
1961					3	4	3	3			5	9	106	106	365	365	365	365	365	365	365	365			365	365	1	1	365	365			365	365	90	90
1962					6	5	2	2			5	5	109	109	365	365	365	365	365	365	365	365			365	365			365	365			365	365		
1963					3	3	31	31			5	5	116	116	365	365	365	365	365	365	365	365	9	9			3		365	365			365	365		
1964											1	1	111	110	366	366	366	366	366	366	366	366			366	366	1	1	366	366			366	366		
1965													59	59	365	365	365	365	365	365	365	365			365	365			365	365			365	365		
1966					93								9	9	365	365	365	365	365	365	365	365	1		365	365	4		365	365	6	7	365	365		
1967					73										365	365	365	365	4	5	365	365			365	365			365	365			365	365		
1968					10						6	7			366	366	366	366	8	6	366	366			366	366			366	366			366	366		
1969							2	2			4	3			365	365	365	365	2	2	365	365			365	365	25	25	365	365	9	10	365	365		
1970			185	189					11		1	4			365	365			4	6	180	0	42	40	162	163	34	34	31	31	21	22	365	365	4	5
1971			110	103	1						1	8		27	48	50			4	9			3	1	2	2							365	365		53
1972			10	9	16	1			1	1	21	18	1	3	22	22			25	35		5	4	3			168	123					366	366	3	28
1973			3						13	24	23	23		1	2	4				1		1	1				275	275					365	365		
1974			21	21	1	2	1	1	2	23			1		13	94			3	7	6	6		1	4	5	365	365					365	365	5	4
1975			33	2	4	1				140	46	46	19	4	13	334			2	3	46	64	5	2	4	2	65	65					365	365	34	38
1976			71	2	25	12					48	58	9	1	35	152			31	32	85	116	3	1	8	9	6	6	13	14			366	366	206	207
1977			11	3					5		8	19	1	1					2	2	122	125	1	3	1	2	19	25	87	84			365	365	365	365
1978											1	3		1	56	57			2	5	38	97	1	3	182	184	29	32	31	31			365	365	365	365
1979											2		1		2	2			5	13	18	20	2	2	4	6	8	9	215	215	43	42	365	365	365	365
1980				2					39	32	4	4		1	12	13			25	21	111	131	56	16	15	14	17	20	213	213			366	366	366	366
1981	3		2		1		1		8		1					1			365	365	26	27	82	82			13	55				1	365	365	365	365
1982			1			1								1	1	3	3	3	94	95	10	10	5	1			275	276	2	3					365	365
1983			1						1		235	5			5	5			14	10							365	365		2	3	3			365	365
1984											61	1				2	274	276	52	22	9	9					366	366							366	366
1985			1	2			2	2			2	2			4	9	365	365	11	11	31	30		1			365	365	30		1	1			365	365
1986							7	7			7	7				2	365	365	10	7	21	21					365	365			1	1			365	365
1987							9	9			8	15					365	365	5	4	46	46	7	7			365	365			365	365			365	365
1988											5	14					366	366	14	14	48	69	280	281			366	366			366	366			366	366
1989							6	6			5	8					365	365	208	10	4	5	365	365			365	365			365	365			365	365
1990							2	3			16	10					365	365	14	101	15	16	365	365		1	365	365	1	1	365	365			365	365
1991											8	10					365	365	101	305	23	28	365	365	365	365	365	365			365	365	4	5	365	365
1992											116	112					366	366	55	62	19	20	366	366	366	366	366	366			366	366			366	366
1993					48	53	2	2			365	365					365	365	133	62	83	84	365	365	365	365	365	365			365	365			365	365
1994		1		1	71	78	3	2		1	365	365		1	1	3	365	365	156	140	135	136	365	365	365	365	365	365	58	7	365	365		1	365	365
1995			4	7	115	115	1	2			365	365			365	365	365	365	365	365	219	219	365	365	365	365	365	365	123	93	365	365			365	365
1996	366	366			161	159	5	6	366	366	366	366	366	366	366	366	366	366	366	366	366	366	366	366	366	366	366	366	366	366	366	366	366	366	366	366
1997	181	181	365	365	365	365	365	365	365	365	365	365	365	365	365	365	365	365	365	365	365	365	365	365	365	365	365	365	365	365	365	365	334	334	365	365
1998			365	365	365	365	3	2	365	365	365	365	365	365	365	365	365	365	365	365	365	365	365	365	365	365	365	365	365	365	365	365		1	365	365
1999		1	365	365	365	365	365	365	365	365	365	365	365	365	365	365	365	365	365	365	365	365	365	365	365	365	365	365	365	365	365	365		36	365	365
2000			37	32	366	366	366	366	121	121	366	366	32	32	366	366	366	366	366	366	366	366	366	366	366	366	366	366	366	366	366	366	5	5	366	366
2001	59	59	181	181	365	365	365	365	32	31	365	365	59	59	365	365	365	365	365	365	365	365	365	365	365	365	365	365	365	365	365	365	1	1	365	365

Fig. 3. Original meteorological data time series size. In darkish empty cells – complete year of records with good-quality information for TMIN and/or TMAX; in white and darkish filled cells – number of records without information in a year and "complete" year without records, respectively. Note that the data series from Lisbon/Gago-Coutinho (579) has no data for the period that precedes 1982.

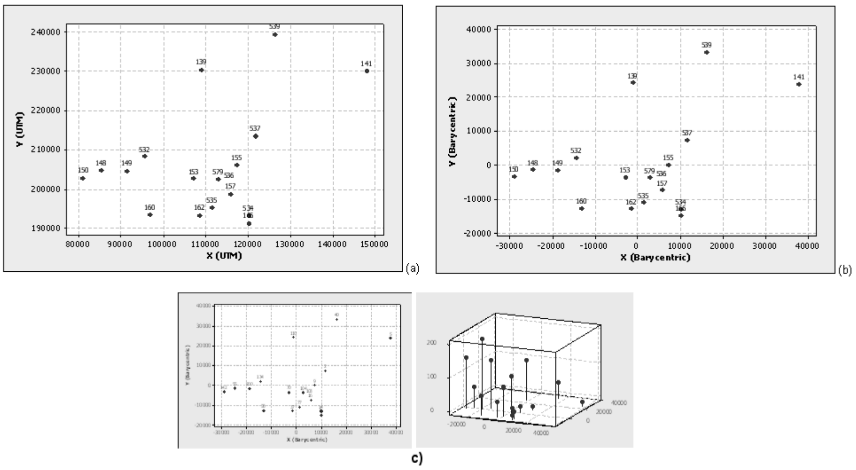


Fig. 4. The geographical position of the eighteen meteorological stations (the data labels) in (a) the UTM transform system and (b) the barycentric coordinates system. The altitude (data label on the left side) representation of the network superimposed over the barycentric coordinates system (c).

Prediction and Conditional Simulation of a 2D Lognormal Diffusion Random Field ⋆

Ramón Gutiérrez[1], Concepción Roldán[2], Ramón Gutiérrez-Sánchez[1], and José M. Angulo[1]

[1] Department of Statistics and Operations Research
University of Granada. Campus de Fuentenueva, s/n,
E-18071 Granada, Spain
(e-mail: rgjaimez@ugr.es, ramongs@ugr.es, jmangulo@ugr.es)

[2] Department of Statistics and Operations Research
University of Jaén. Las Lagunillas, s/n,
E-23071 Jaén, Spain
(e-mail: iroldan@ujaen.es)

Abstract. This paper describes techniques for estimation, prediction and simulation of two-parameter lognormal diffusion random fields which are diffusions on each coordinate and satisfy a particular Markov property.

Keywords: Diffusion Random Field, Kriging, Lognormal Diffusion Process.

1 Introduction

Lognormal random fields represent the technically more complex stage of lognormal modelling. Problems as parameter estimation, lognormal simple kriging, estimation based on lognormal maximum entropy, among others, are generally undertaken by simply considering the lognormal random field as the exponential transformation of a Gaussian random field, without reference to any specific diffusion structure. This latter approach, however, constitutes an important alternative in relation to modelling, parameter estimation and inference, analysis of first passage through barriers, associated Îto equations and derivation of discrete simulation schemes, etc.

Among the contribution to theoretical foundations for diffusion random fields, see [Nualart, 1983]. In this context, [Gutiérrez *et al.*, 2004] considered lognormal random field models which are diffusions on each coordinate. Involving exogenous factors affecting the drift term, the drift and diffusion coefficients, which characterize a two-parameter lognormal diffusion under certain conditions, were estimated by maximum likelihood. For data on a regular grid, an alternative method was proposed to estimate the diffusion coefficient.

In this work, the estimates of the drift and the diffusion coefficients given in [Gutiérrez *et al.*, 2004] are used for obtaining predictions and conditional

⋆ *This work has been partially supported by projects BFM2002-03633 and BFM2002-01836 of the DGI, Spain.*

simulations. The contents are organized as follows. First, the 2D lognormal random field model is introduced. Second, estimation of the drift and diffusion coefficients based on a discrete finite set of data is given. Finally, aspects related to kriging and conditional simulation are addressed and illustrated.

2 Lognormal Diffusion Random Fields

Lognormal diffusion processes are commonly used in the analysis of economic variables. When the parameter space is a subset of $\mathbf{R}_{+}^{2}$, [Nualart, 1983] introduced a class of two-parameter random fields which are diffusions on each coordinate and satisfy a particular Markov property related to partial ordering in $\mathbf{R}_{+}^{2}$. Using this theory, we can introduce a 2D lognormal diffusion random field as follows.

Let $\left\{X\left(\mathbf{z}\right):\mathbf{z}=(s,t)\in I=[0,S]\times[0,T]\subset\mathbf{R}_{+}^{2}\right\}$ be a positive-valued Markov random field, defined on a probability space $(\Omega,\mathcal{A},P)$, where $X\left(0,0\right)$ is assumed to be constant or a lognormal random variable with $E\left[\ln X\left(0,0\right)\right]=\phi_0$ and $var\left(\ln X\left(0,0\right)\right)=\sigma_0^2$. The distribution of the random field is determined by the following transition probabilities:

$$\begin{aligned}&P\left(B,(s+h,t+k)\mid\left(x_1,x,x_2\right),\mathbf{z}\right)=\\&\quad P\left[X\left(s+h,t+k\right)\in B\mid X\left(s,t+k\right)=x_1,X\left(\mathbf{z}\right)=x,X\left(s+h,k\right)=x_2\right],\end{aligned}$$

where $\mathbf{z}=(s,t)\in I$, $h,k>0$, $(x_1,x,x_2)\in\mathbf{R}_{+}^{3}$ and B is a Borel subset. We suppose that the transition densities exist and are given by

$$\begin{aligned}&g(y,(s+h,t+k)\mid\left(x_1,x,x_2\right),\mathbf{z})\\&=\frac{1}{y\sqrt{2\pi\sigma_{\mathbf{z};h,k}^{2}}}\exp\left\{-\frac{1}{2}\left(\frac{\ln\left(\frac{yx}{x_1x_2}\right)-m_{\mathbf{z};h,k}}{\sigma_{\mathbf{z};h,k}}\right)^{2}\right\},\end{aligned}$$

for $y\in\mathbf{R}_{+}$, with

$$m_{\mathbf{z};h,k}=\int_s^{s+h}\int_t^{t+k}\tilde{a}\left(\sigma,\tau\right)d\sigma d\tau,\quad\sigma_{\mathbf{z};h,k}^{2}=\int_s^{s+h}\int_t^{t+k}\tilde{B}\left(\sigma,\tau\right)d\sigma d\tau,$$

and $\tilde{a}$, $\tilde{B}$ being continuous functions on I. Under these conditions we can assert that $\{X\left(\mathbf{z}\right):\mathbf{z}\in I\}$ is a lognormal diffusion random field. The one-parameter drift and diffusion coefficients associated are given by

$$\begin{aligned}a_1\left(\mathbf{z}\right)x&:=\left(\tilde{a}_1\left(\mathbf{z}\right)+\frac{1}{2}\tilde{B}_1\left(\mathbf{z}\right)\right)x,\quad B_1\left(\mathbf{z}\right)x^2:=\tilde{B}_1\left(\mathbf{z}\right)x^2,\\a_2\left(\mathbf{z}\right)x&:=\left(\tilde{a}_2\left(\mathbf{z}\right)+\frac{1}{2}\tilde{B}_2\left(\mathbf{z}\right)\right)x,\quad B_2\left(\mathbf{z}\right)x^2:=\tilde{B}_2\left(\mathbf{z}\right)x^2,\end{aligned}$$

where

$$\tilde{a}_1(s,t) = \int_0^t \tilde{a}\,(s,\tau)\,d\tau, \quad \tilde{B}_1(s,t) = \int_0^t \tilde{B}\,(s,\tau)\,d\tau,$$
$$\tilde{a}_2(s,t) = \int_0^s \tilde{a}\,(\sigma,t)\,d\sigma, \quad \tilde{B}_2(s,t) = \int_0^s \tilde{B}\,(\sigma,t)\,d\sigma,$$

for all $\mathbf{z} = (s,t) \in I$, $x \in \mathbf{R}_+$.

The random field $\{Y(\mathbf{z}) : \mathbf{z} \in I\}$ defined as $Y(\mathbf{z}) = \ln X(\mathbf{z})$ is then a Gaussian diffusion random field, with $\tilde{a}$ and $\tilde{B}$ being, respectively, the drift and diffusion coefficients, and $\tilde{a}_1$, $\tilde{a}_2$, $\tilde{B}_1$ and $\tilde{B}_2$ being the corresponding one-parameter drift and diffusion coefficients. Furthermore, if $\mathbf{z}, \mathbf{z}' \in I$, $\mathbf{z} = (s,t)$, $\mathbf{z}' = (s',t')$, then

$$m_Y(\mathbf{z}) := E[Y(\mathbf{z})] = \phi_0 + \int_0^s \int_0^t \tilde{a}\,(\sigma,\tau)\,d\sigma d\tau,$$
$$\sigma_Y^2(\mathbf{z}) := var\,(Y(\mathbf{z})) = \sigma_0^2 + \int_0^s \int_0^t \tilde{B}\,(\sigma,\tau)\,d\sigma d\tau,$$
$$c_Y(\mathbf{z},\mathbf{z}') := cov\,(Y(\mathbf{z}), Y(\mathbf{z}')) = \sigma_Y^2(\mathbf{z} \wedge \mathbf{z}'),$$

where we write $\mathbf{z} \wedge \mathbf{z}'$ for $(s \wedge s', t \wedge t')$, with '$\wedge$' denoting the minimum.

Under suitable regularity conditions, it is possible to obtain a SPDE formulation for a two-parameter diffusion RF. In fact, we need hypotheses **I-V** to be satisfied, in order to apply *Theorem 2.8* of [Nualart, 1983]. These hypotheses and the uniqueness of solution have been proved by the authors to hold for the lognormal diffusion RF considered. Thus, there exists a two-parameter Wiener RF $\{W(\mathbf{z}) : \mathbf{z} \in I\}$ (adjoining, if it is necessary, a new probability space) such that $\{X(\mathbf{z}) : \mathbf{z} \in I\}$ is the only diffusion RF satisfying the following partial SPDE:

$$\frac{\partial^2 X(s,t)}{\partial s \partial t} - X^{-1}(s,t)\frac{\partial X(s,t)}{\partial s}\frac{\partial X(s,t)}{\partial t} - \frac{\partial a_2(s,t)}{\partial s}X(s,t) =$$
$$\left(\frac{\partial B_2(s,t)}{\partial s} + B_1(s,t)\,B_2(s,t)\right)^{1/2} X(s,t)\frac{\partial^2 W(s,t)}{\partial s \partial t}.$$

This aspect is not essential for the approach considered in this work, although it provided an alternative interesting interpretation of the RF formulation considered.

Henceforth we will assume that the conditions usually considered for estimation of the drift and diffusion coefficients in the one-parameter case hold; that is, $P[\ln X(0,0) = \phi_0] = 1$ (i.e. $\sigma_0^2 = 0$) and $\sigma_Y^2(\mathbf{z}) = \tilde{B}st$, $\mathbf{z} = (s,t) \in I$.

3 Estimation of the Drift and Diffusion Coefficients

Let $\{X(\mathbf{z}) : \mathbf{z} \in I\}$ be a lognormal diffusion random field. Data $\mathbf{X} = (X(\mathbf{z}_1), ..., X(\mathbf{z}_n))^t$ are assumed to be observed at known spatial locations $\mathbf{z}_1 = (s_1,t_1)$, $\mathbf{z}_2 = (s_2,t_2), ..., \mathbf{z}_n = (s_n,t_n) \in I$. Let $\mathbf{x} = (x_1, x_2, ..., x_n)^t$ be a sample. Let us consider the log-transformed $n-$dimensional random vector, $\mathbf{Y} = (Y(\mathbf{z}_1), Y(\mathbf{z}_2), ..., Y(\mathbf{z}_n))^t = (\ln X(\mathbf{z}_1), \ln X(\mathbf{z}_2), ..., \ln X(\mathbf{z}_n))^t = \ln \mathbf{X}$, and the log-transformed sample, $\mathbf{y} = (y_1, y_2, ..., y_n)^t = \ln \mathbf{x}$. We denote

$$\mathbf{m}_Y = (m_Y(\mathbf{z}_1), ..., m_Y(\mathbf{z}_n))^t, \quad \Sigma_Y = \left(\sigma_Y^2(\mathbf{z}_i \wedge \mathbf{z}_j)\right)_{i,j=1,\dots,n}.$$

3.1 MLE for the Drift and Diffusion Coefficients Using Exogenous Factors

Suppose that the drift coefficient $\tilde{a}$ of Y is a linear combination of several known functions, set $\{h_1(\mathbf{z}), ..., h_p(\mathbf{z}) : \mathbf{z} \in I\}$, with real coefficients ϕ_1,..., ϕ_p :

$$\tilde{a}(\mathbf{z}) = \sum_{\alpha=1}^{p} \phi_\alpha h_\alpha(\mathbf{z}), \quad \mathbf{z} \in I.$$

Defining, for $\mathbf{z} = (s, t) \in I$,

$$f_0(\mathbf{z}) = 1, \quad f_\alpha(\mathbf{z}) = \int_0^s \int_0^t h_\alpha(\sigma, \tau)\, d\sigma d\tau, \quad \alpha = 1, ..., p,$$

the mean of Y is given by

$$m_Y(s, t) = \phi_0 + \sum_{\alpha=1}^{p} \phi_\alpha \int_0^s \int_0^t h_\alpha(\sigma, \tau)\, d\sigma d\tau = \sum_{\alpha=0}^{p} \phi_\alpha f_\alpha(\mathbf{z}).$$

Thus, denoting $\mathbf{F} = (\mathbf{f}_0, \mathbf{f}_1, ..., \mathbf{f}_p)$, with $\mathbf{f}_\alpha = (f_\alpha(\mathbf{z}_1), f_\alpha(\mathbf{z}_2), ..., f_\alpha(\mathbf{z}_n))^t$, for $\alpha = 0, 1, ..., p$, and $\boldsymbol{\phi} = (\phi_0, \phi_1, ..., \phi_p)^t$, we have

$$\mathbf{m}_Y = (\phi_0 \mathbf{f}_0 + \phi_1 \mathbf{f}_1 + ... + \phi_p \mathbf{f}_p) = \mathbf{F}\boldsymbol{\phi}.$$

Let us write

$$\Sigma_Y = \tilde{B}\mathbf{M} := \tilde{B} \begin{pmatrix} s_1 t_1 & (s_1 \wedge s_2)(t_1 \wedge t_2) & \cdots & (s_1 \wedge s_n)(t_1 \wedge t_n) \\ (s_1 \wedge s_2)(t_1 \wedge t_2) & s_2 t_2 & \cdots & (s_2 \wedge s_n)(t_2 \wedge t_n) \\ \vdots & \vdots & \ddots & \vdots \\ (s_1 \wedge s_n)(t_1 \wedge t_n) & (s_2 \wedge s_n)(t_2 \wedge t_n) & \cdots & s_n t_n \end{pmatrix}.$$

With this notation, the MLE for the drift and diffusion coefficients are, respectively,

$$\boldsymbol{\phi}^* = \left(\phi_0^*, \phi_1^*, ..., \phi_p^*\right)^t = \left(\mathbf{F}^t \mathbf{M}^{-1} \mathbf{F}\right)^{-1} \mathbf{F}^t \mathbf{M}^{-1} \ln \mathbf{x} \tag{1}$$

and

$$\tilde{B}^* = \frac{1}{n} (\ln \mathbf{x} - \mathbf{m}_Y^*)^t \mathbf{M}^{-1} (\ln \mathbf{x} - \mathbf{m}_Y^*), \tag{2}$$

where $\mathbf{m}_Y^* = \mathbf{F}\boldsymbol{\phi}^*$.

3.2 Estimation of the Drift and Diffusion Coefficients from Data on a Regular Grid

Suppose now that the data are obtained on a regular grid in $\mathbf{R}_+^2$. Let $\mathbf{z} = (s,t)$ be a point in a set S of locations included in the regular grid and let us denote the 2D four-point increment of Y by

$$Y\left(\varDelta_{hk}\left(\mathbf{z}\right)\right) = Y\left(s+h,t+k\right) - Y\left(s,t+k\right) - Y\left(s+h,t\right) + Y\left(s,t\right),$$

for $h,k>0$. Taking into account that the variance of this increment,

$$var\left(Y\left(\varDelta_{hk}\left(\mathbf{z}\right)\right)\right) = \sigma^2_{\mathbf{z};h,k} = \int_s^{s+h}\int_t^{t+k}\tilde{B}\left(\sigma,\tau\right)d\sigma d\tau = \tilde{B}hk,$$

does not depend on the location $\mathbf{z}$, but only on the area hk, the diffusion coefficient $\tilde{B}$ can be estimated using a similar approach to Matheron's estimator for the variogram (see, for example, [Cressie, 1993]), considering here 2D four-point increments, as follows.

Under the implicit condition that $\mathbf{z}_i = (s_i,t_i) < \mathbf{z}_j = (s_j,t_j)$, we denote

$$[\mathbf{z}_i,\mathbf{z}_j] = \left\{\left(s_i,t_i\right),\left(s_i,t_j\right),\left(s_j,s_i\right),\left(s_j,t_j\right)\right\}.$$

The estimator, for $\mathbf{z} = (s,t)$, is

$$\begin{aligned}&\widehat{var}\left(Y\left(\varDelta_{hk}\left(\mathbf{z}\right)\right)\right)\\ &= \frac{1}{\left|N\left(hk\right)\right|}\sum_{N(hk)}\left(Y\left(s+h,t+k\right) - Y\left(s,t+k\right) - Y\left(s+h,t\right) + Y\left(s,t\right)\right.\\ &\quad \left.-m_Y\left(s+h,t+k\right) + m_Y\left(s,t+k\right) + m_Y\left(s+h,t\right) - m_Y\left(s,t\right)\right)^2,\end{aligned}$$

where

$$N\left(hk\right) \equiv \left\{\left(\mathbf{z}_i,\mathbf{z}_j\right) : [\mathbf{z}_i,\mathbf{z}_j] \in S, \left(s_j-s_i\right)\left(t_j-t_i\right) = hk, \quad i,j=1,...,n\right\}$$

and $\left|N\left(hk\right)\right|$ is the number of different elements of $N\left(hk\right)$. If the mean is unknown, it can be estimated using (1) by $m_Y^*\left(\mathbf{z}\right) = \sum_{\alpha=0}^{p}\phi_\alpha^* f_\alpha\left(\mathbf{z}\right)$.

4 Numerical Examples

In this section we describe some numerical examples illustrating estimation for a lognormal diffusion random field under the approaches considered and an example of prediction and conditional simulation. First, using simulated data on a regular grid, the two estimation methods for the diffusion coefficient respectively described in Sections 3.1 and 3.2 are compared, considering the case of known non constant mean (for the associated Gaussian random field). Second, we obtain a conditional simulation for a lognormal diffusion random field.

The parameter space considered is $I = [0, 1.65] \times [0, 1.05]$. Realizations are generated on a regular 19×19 grid, S, with SW corner at the origin $(0, 0)$ and NE corner at point (1.65,1.05). Parameter estimates, kriging predictions and simulations are obtained on this grid based on the data $\mathbf{X}$, consisting of the values corresponding to the 7×7 regular grid, subset determined by the same corner points. We will obtain unconditionally simulated realizations by the method of unconstrained simulation described in [Christakos, 1992].

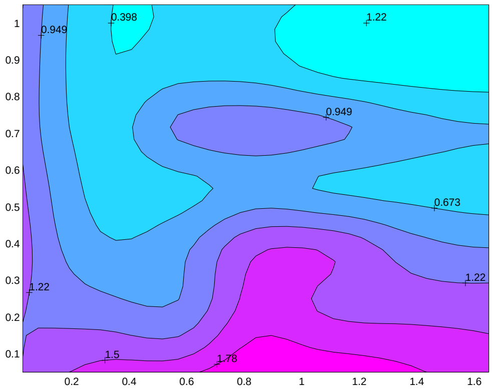


Fig. 1. Contour-level plot of 49 values generated (simulation 1) for the lognormal diffusion random field (known non constant mean case)

Sim. no.	$\tilde{B}^*$	$\tilde{B}^{**}$	Sim. no.	$\tilde{B}^*$	$\tilde{B}^{**}$
1	1.1115	0.8199	9	0.9911	0.8236
2	1.0605	0.5584	10	0.9909	0.4597
3	1.2060	1.0004	11	1.0107	0.4792
4	1.2153	0.5457	12	0.9016	0.3990
5	1.1324	0.8595	13	0.8914	1.5703
6	0.8309	0.6103	14	1.1850	0.9163
7	0.6138	0.3944	15	0.9870	1.2419
8	1.3243	0.4456	16	1.0684	1.1750

Table 1. Estimates of $\tilde{B}$ by the two methods considered, for 16 simulations of the lognormal diffusion random field (known non constant mean case)

We consider a lognormal diffusion random field with non constant mean, with $\phi_0 = 0.25$, $\tilde{a}(\mathbf{z}) = -2$, for all $\mathbf{z} \in I$, and $\tilde{B} = 1$. Table 1 gives the estimates of $\tilde{B}^*$ and $\tilde{B}^{**}$ obtained for 16 independent unconstrained simulations for this random field, assuming that the mean of the associated Gaussian random field is known.

From the results obtained in both cases studied, we can observe that the maximum likelihood estimation method overall provides more accurate estimates for the diffusion coefficient than the alternative method based on evaluation of 2D four-point increments. A similar behavior has been observed in several other cases studied by the authors. Lack of stability in the estimate $\tilde{B}^*$ can be possibly overcome by robust estimation of the slope of $\widehat{var}\left(Y\left(\Delta_{hk}\left(\mathbf{z}\right)\right)\right)$ vs. hk instead of using the least-squares approach.

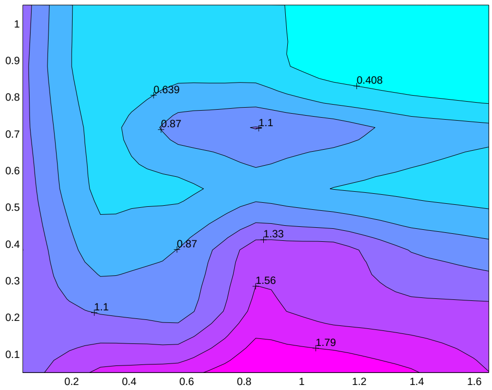


Fig. 2. Contour-level plot of the 361 predictions obtained by simple lognormal kriging using the 49 values plotted in Figure 1

As for simulation, we have considered a practical method for generating conditional simulations that combines unconditional simulation and kriging, described in [Yuh-Ming and Hugh Ellis, 1997]. This technique yields an unbiased conditional simulation (with respect to sample data) and reproduces conditional variances. We can summarize the procedure as follows:

Step 1 Predict $\left\{\widehat{y}\left(\mathbf{z}_i\right) : \mathbf{z}_i \in S\right\}$ based on the data $\mathbf{Y}$ and on the predictor of simple lognormal kriging.

Step 2 Calculate unconditionally simulated realizations $\left\{y^u\left(\mathbf{z}_i\right) : \mathbf{z}_i \in S\right\}$ based on the method of unconstrained simulation and using the estimates given in (1) and (2).

Step 3 Calculate the set of predictions $\left\{\widehat{y^u}\left(\mathbf{z}_i\right) : \mathbf{z}_i \in S\right\}$ based on the data $\left\{y^u\left(\mathbf{z}_i\right) : \mathbf{z}_i \in G\right\}$ and on the predictor of simple lognormal kriging.

Step 4 Calculate conditional simulation realizations of Y by

$$y^c\left(\mathbf{z}_i\right) = y^u\left(\mathbf{z}_i\right) + \left[\widehat{y}\left(\mathbf{z}_i\right) - \widehat{y^u}\left(\mathbf{z}_i\right)\right], \quad \forall \mathbf{z}_i \in S.$$

Step 5 Calculate conditional simulation realizations of X by

$$x^c\left(\mathbf{z}_i\right) = \frac{\exp\left\{y^u\left(\mathbf{z}_i\right)\right\}\exp\left\{\widehat{y}\left(\mathbf{z}_i\right)\right\}}{\exp\left\{\widehat{y^u}\left(\mathbf{z}_i\right)\right\}} \equiv \frac{x^u\left(\mathbf{z}_i\right)\widehat{x}\left(\mathbf{z}_i\right)}{\widehat{x^u}\left(\mathbf{z}_i\right)}, \quad \forall \mathbf{z}_i \in S.$$

For the example of prediction and conditional simulation we consider the previous diffusion. That is, a lognormal diffusion random field with non constant mean, $\phi_0 = 0.25$, $\tilde{a}(\mathbf{z}) = -2$, for all $\mathbf{z} \in I$, and $\tilde{B} = 1$. Using the 49 values obtained from simulation 1 (see Figure 1) we have obtained $\tilde{B}^* = 1.1115$ and using this estimate we have calculated 19×19 predictions by simple lognormal kriging. The results are plotted in Figure 2.

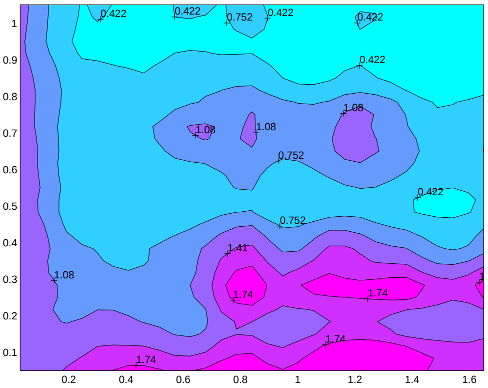


Fig. 3. Contour-level plot of the 361 simulations obtained by conditional simulation using the 49 values plotted in Figure 1

Figure 3 displays a contour-level plot for the 19×19 conditional simulation realization based on the data of simulation 1, and Figure 4 displays the original contour-level plot.

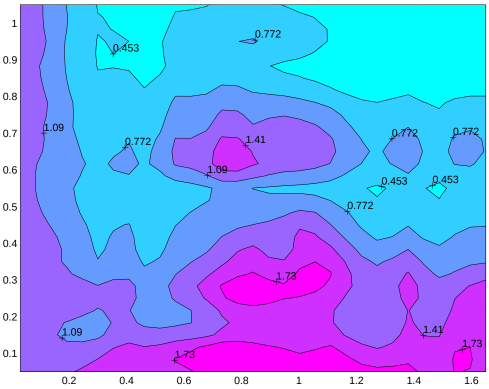


Fig. 4. Contour-level plot of the 361 values (including the 49 values used for estimating $\tilde{B}$) generated (simulation 1) for the 2D lognormal diffusion considered

5 Conclusions

In this paper we study prediction and conditional simulation for a 2D lognormal diffusion random field, including exogenous factors in its formulation. This is an important case of random fields which are not intrinsically stationary, then well-known related techniques cannot be applied. Such models are useful to represent diffusion-type positive valued characteristics, like pollutant indicators in environmental studies. The approach considered allows us to use well-known techniques for estimation and prediction, such as simple kriging, and for conditional simulation.

Possible extensions under investigation by the authors include consideration of non-constant diffusion-type values at the boundary axes as well as higher-dimension spatial and spatio-temporal formulations. Also, development of validation techniques in this context would be most important for real applications.

References

[Christakos, 1992]G. Christakos. *Random Field Models in Earth Sciences.* Academic Press, San Diego, 1992.

[Cressie, 1993]N. Cressie. *Statistics for Spatial Data.* Wiley & Sons, New York, 1993.

[Gutiérrez *et al.*, 2004]R. Gutiérrez, C. Roldán, R. Gutiérrez-Sánchez, and J.M. Angulo. Estimation and prediction of a 2D lognormal diffusion random field. *Stochastic Environmental Research and Risk Assessment*, in press, 2004.

[Nualart, 1983]D. Nualart. Two-parameter diffusion processes and martingales. *Stochastic Processes and their Applications*, 15:31–57, 1983.

[Yuh-Ming and Hugh Ellis, 1997]L. Yuh-Ming and J. Hugh Ellis. Estimation and simulation of lognormal random fields. *Computers and Geosciences*, 23(1):19–31, 1997.

Approximating processes of stochastic diffusion models under spatial uncertainty

L. Jódar, J. C. Cortés, P. Sevilla, and L. Villafuerte

Instituto de Matemática Multidisciplinar
Universidad Politécnica de Valencia
46071 Valencia, Spain
(e-mail: ljodar,jccortes,pabsepe@mat.upv.es, lauvilal@doctor.upv.es)

Abstract. This paper deals with the construction of approximate numerical processes of mixed diffusion models under spatial uncertainty in the diffusion coefficient and the source term. After discretization, the stochastic discrete problem is solved using a stochastic separation of the variables method.
Keywords: diffusion, discrete approximation, stochastic process.

1 Introduction

Mathematical models are useful to describe reality up to certain point. Individual behaviour may be erractic, but aggregate behaviour is often quite predictable. Spatial uncertainty is frequent in Geostatistics descriptions of natural variables. Examples of such variables are, pressure, temperature and wind velocity in the atmosphere, concentrations of pollutants in a contamined site, see [Chilés and Delfiner, 1999]. Wave propagation problems in random media have been studied in [Keller, 1963]. A different approach to numerical stochastic methods for diffusion models where the spatial uncertainty is a Brownian motion is developed in [Kloeden and Platen, 1992] using Ito stochastic calculus. In this paper we study stochastic diffusion problems of the form

$$u_t = [p(x)u_x]_x + F(x,t) \quad , \quad 0 < x < 1 \quad , \quad t > 0 \tag{1}$$

$$a_1 u(0,t) + a_2 u_x(0,t) = 0 \quad , \quad t > 0, \quad |a_1| + |a_2| > 0, \tag{2}$$

$$b_1 u(1,t) + b_2 u_x(1,t) = 0 \quad , \quad t > 0, \quad |b_1| + |b_2| > 0, \tag{3}$$

$$u(x,0) = f(x) \quad , \quad 0 \leq x \leq 1, \tag{4}$$

where the diffusion coefficient $p(x)$ is assumed to be a stochastic process and for each t fixed, $F(x,t)$ is also a stochastic process. Here $f(x)$ is a deterministic function and h_1 and h_2 are constants. Chance of randomness can affect in any of the following ways:

- uncertainty as to the diffusion properties of the medium which the diffusion takes place,
- random variations of the internal influences of the system undergoing diffusion,

- random external sources to the medium in which the diffusion takes place.

For the particular case where $F(x,t) = 0$ and $p(x)$ is a constant random variable, problem has been recently treated in [Cortés *et al.*, 2005a].

This paper is organized as follows. Section 2 studies random discrete Sturm-Liouville problem and random discrete Fourier series. In section 3 the way for obtaining an exact series solution process is summarized and problem (1)-(4) is discretized and an explicit solution process of the stochastic discretized model is given by means of a random eigenfunctions method. Section 4 includes an illustrative example.

2 Random discrete Sturm-Liouville problems

For the sake of clarity in the presentation we recall some concepts, notations and results related to the mean square stochastic calculus, that may be found in [Soong, 1973]. Let $(\Omega, \mathcal{F}, P)$ be a probability space. A real random variable (r.v.) $Y : \Omega \to R$ is said to be continuous if its distribution function F_Y is continuous and almost everywhere differentiable. In this case, its density function is defined by

$$g_Y(y) = \frac{dF_Y(y)}{dy}.$$

If Y satisfies the additional property

$$E\left[Y^2\right] = \int_{-\infty}^{+\infty} y^2 g_Y(y)dy < +\infty, \tag{5}$$

then Y is said to be a second order random variable (2-r.v.) and the integral in (5) is the expectation of Y^2. If $\{p(x)\}_{x\in I}$ is a real stochastic process on the probability space $(\Omega, \mathcal{F}, P)$, we say that it is a second order process (2-s.p.), if $E\left[p^2(x)\right] < +\infty$, for all $x \in I$.

Throughout this paper a random variable will mean a 2-r.v. and a stochastic process will denote a 2-s.p. If $\{p(x)\}_{x\in I}$ is a 2-s.p., its covariance function is the deterministic function $\Gamma_{pp}(r,s) = E\left[p(r)p(s)\right] - E\left[p(r)\right]E\left[p(s)\right]$, for $(r,s) \in I \times I$. If Y is a 2-r.v., then $\|Y\| = \sqrt{E\left[Y^2\right]}$ is a norm and the set of all 2-r.v.'s endowed with this norm is a Banach space denoted by L_2, [Soong, 1973, chap.4]. From the Cauchy-Schwarz property in L_2, we recall that if X and Y are two 2-r.v.'s in L_2, then

$$\|XY\| \leq \|X\| \, \|Y\|. \tag{6}$$

A sequence of 2-r.v.'s $\{Y_n\}$ converges in mean square (m.s.) to a 2-r.v. Y as $n \to \infty$ if

$$\lim_{n\to\infty} \|Y_n - Y\|^2 = \lim_{n\to\infty} E\left[|Y_n - Y|^2\right] = 0. \tag{7}$$

This type of convergence is called mean square convergence. A 2-s.p. $\{p(x)\}_{x\in I}$ is m.s. continuous if, for $x, x+\tau \in I$, one satisfies

$$\lim_{\tau\to 0} \|p(x+\tau)-p(x)\| = 0. \tag{8}$$

Let N be the set of all natural numbers including zero. If $a < b$ are two natural numbers in N, we denote $N(a,b) = \{a, a+1, \ldots, b\}$.

Let $p(i)$, $r(i)$ be 2-r.v., $p(i) : \Omega \to R$, $r(i) : \Omega \to R$ such that

$$\left.\begin{array}{l} p(i)(\omega) > 0 \ , \ \omega \in \Omega \ , \ i \in N(0,K) \\ r(i)(\omega) > 0 \ , \ \omega \in \Omega \ , \ i \in N(1,K) \end{array}\right\} \tag{9}$$

and let α, β be real numbers. If Δ denotes the forward difference operator defined by $\Delta u(i) = u(i+1) - u(i)$, then a boundary value problem of the form

$$\left.\begin{array}{c} \Delta\left(p(i-1)\Delta u(i-1)\right) + \lambda r(i)u(i) = 0 \ , \ i \in N(1,K) \\ u(0) = \alpha u(1) \quad , \quad u(K+1) = \beta u(K) \ , \end{array}\right\} \tag{10}$$

is called a random discrete Sturm-Liouville problem. Note that for each event $\omega \in \Omega$, the problem

$$\left.\begin{array}{c} \Delta\left(p(i-1)(\omega)\Delta u(i-1)\right) + \lambda r(i)(\omega)u(i) = 0 \ , \ i \in N(1,K) \\ u(0) = \alpha u(1) \quad , \quad u(K+1) = \beta u(K) \ , \end{array}\right\} \tag{11}$$

is a deterministic discrete Sturm-Liouville problem, see [Agarwal, 1991, p.663]. A problem (11) has exactly K real eigenvalues $\lambda_m(\omega)$, $1 \le m \le K$, which are distinct, and corresponding to each eigenvalue $\lambda_m(\omega)$ there exist an eigenfunction $\phi_m(i)(\omega)$, $i \in N(1,K)$. These eigenfunctions $\phi_m(i)(\omega)$, $1 \le m \le K$ are mutually orthogonal with respect to weight function $r(i)(\omega)$, i.e.,

$$\sum_{l=1}^{K} r(l)(\omega)\phi_\mu(l)(\omega)\phi_\nu(l)(\omega) = 0 \quad , \quad \text{if} \ \ \mu \neq \nu. \tag{12}$$

In particular, these eigenfunctions $\phi_m(i)(\omega)$ are linearly independent on the set $N(1,K)$. Eigenpairs $(\lambda_m(\omega), \phi_m(i)(\omega))$ of the Sturm-Liouville problem (11) for each $\omega \in \Omega$, are easily computed as eigenpairs of the matrix eigenvalue problem

$$R^{-1}(\omega)A(\omega)u = \lambda u \quad , \tag{13}$$

where

$$R(\omega) = \operatorname{diag}\left(r(1)(\omega), r(2)(\omega), \ldots, r(K)(\omega)\right) \quad , \tag{14}$$

and if we denote

$$\left.\begin{array}{l} s(i)(\omega) = p(i)(\omega) + p(i-1)(\omega) \ , \ i \in N(1,K) \\ \overline{s}(1)(\omega) = \quad s(1)(\omega) - \alpha p(0)(\omega) \quad , \overline{s}(K)(\omega) = s(K)(\omega) - \beta p(K)(\omega) \end{array}\right\} , \tag{15}$$

$A(\omega)$ is the symmetric tridiagonal matrix

$$A(\omega)=\begin{bmatrix} \overline{s}(1)(\omega) & -p(1)(\omega) & 0 & \cdots & 0 \\ -p(1)(\omega) & s(2)(\omega) & -p(2)(\omega) & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & -p(K-2)(\omega) & s(K-1)(\omega) & -p(K-1)(\omega) \\ 0 & \cdots & 0 & -p(K-1)(\omega) & \overline{s}(K)(\omega) \end{bmatrix} \tag{16}$$

Thus, the eigenvalues and eigenfunctions of the random discrete Sturm-Liouville problem (10) are random variables whose statistical properties are determined by those of the random coefficients. See [Boyce, 1960] for the treatment of analogous of continuous Sturm-Liouville stochastic problems.

Under previous hypotheses and notation, if $\{u(i); 1 \leq i \leq K\}$ is a finite sequence of r.v.‘s, or a discrete stochastic processes defined on a common sample space Ω, then for each $\omega \in \Omega$ the function $\{u(i)(\omega); i \in N(1,K)\}$ admits a series representation

$$u(i)(\omega) = \sum_{m=1}^{K} c_m(\omega)\phi_m(\omega)(i) \quad , \qquad i \in N(1,K) \tag{17}$$

where

$$c_m(\omega) = \frac{\sum_{i=1}^{K} r(i)(\omega)\phi_m(i)(\omega)u(i)(\omega)}{\sum_{i=1}^{K} r(i)(\omega)\left(\phi_m(i)(\omega)\right)^2} \quad , \tag{18}$$

is called the m-th discrete Fourier coefficient of $u(i)(\omega)$ with respect to $\{\phi_m(i)(\omega);\ 1 \leq m \leq K\}$ and (17) is the discrete Fourier series of the deterministic function $\{u(i)(\omega);\ i \in N(1,K)\}$, see [Agarwal, 1991, p.675].

Summarizing, under hypotheses (9) the random discrete Sturm-Liouville problem (10) admits exactly K real random eigenvalue variables $\lambda_m,\ 1 \leq m \leq K$ and K real random eigenfunction variables $\phi_m(i),\ 1 \leq m \leq K$, so that each realization of problem (10), for $\omega \in \Omega$ fixed, described by (11), represents a deterministic discrete Sturm-Liouville problem. In a analogous way, given a discrete stochastic process $\{u(i);\ i \in N(1,K)\}$ defined on Ω, the m-th Fourier coefficient c_m defined by (18) is a random variable and

$$u(i) = \sum_{m=1}^{K} c_m\phi_m(i) \quad , \tag{19}$$

is the random Fourier series representation of the process $\{u(i);\ i \in N(1,K)\}$ with respect to the random eigenpairs $(\lambda_m, \phi_m(i))$ of the random Sturm-Liouville problem (10).

3 Approximating stochastic diffusion processes

An exact theoretical series solution process $u(x,t)$ of problem (1)-(4) of form

$$u(x,t) = \sum_{n\geq 1} \varphi_n(x) b_n(t), \tag{20}$$

can be obtained using a random continuous eigenfunction method, where $(\lambda_n, \varphi_n(x))$ is the random normalized eigenpair sequence associated to the Sturm-Liouiville problem

$$[p(x)X']' + \lambda X = 0, \quad 0 < x < 1 \tag{21}$$

$$a_1 X(0) + a_2 X'(0) = 0, \tag{22}$$

$$b_1 X(1) + b_2 X'(1) = 0, \tag{23}$$

and $b_n(t)$ is the random variable defined by

$$b_n(t) = \alpha_n e^{-\lambda_n t} + \int_0^t e^{-\lambda_n (t-s)} \gamma_n(s) ds, \tag{24}$$

$$\gamma_n(t) = \int_0^1 F(x,t) \varphi_n(x) dx, \tag{25}$$

$$\alpha_n = \int_0^1 f(x) \varphi_n(x) dx. \tag{26}$$

Under appropriate hypotheses it can be proved that $u(x,t)$ given by (20)-(26) is a well defined mean square convergent series, termwise mean square partially differentiable satisfying (1)-(4). In order to prove this fact it is necessary to find bounds of the eigenpairs $(\lambda_n(\omega), \varphi_n(x,\omega))$ of each deterministic realization

$$\left.\begin{array}{l} [p(x)(\omega)X']' + \lambda X = 0, \quad 0 < x < 1, \\ a_1 X(0) + a_2 X'(0) = 0, \\ b_1 X(1) + b_2 X'(1) = 0, \end{array}\right\}$$

using results of section 10.12 and 10.13 of [Birkhoff and Rota, 1965], the ideas developed in [Weinberger, 1965, p.135-137] and [Cortés *et al.*, 2005b] for deterministic eigenfunction method. For the sake of limitation in the extension of this paper we omit technical details of the proof, under the hypotheses

$$F(x,t) \quad \text{m.c. continuous} \tag{27}$$

$$F(x,t) \quad \text{twice m.c. differentiable with respect to} \quad x \tag{28}$$

$$\int_0^1 \frac{\partial^2 F}{\partial x^2}(x,t) dx \quad \text{uniformly bounded in}\, L_2 \,\text{with respect to}\; t \in [0,\infty[\tag{29}$$

$$a_1 F(0,t) + a_2 F_x(0,t) = 0, \tag{30}$$

$$b_1F(1,t)+b_2F_x(1,t)=0. \tag{31}$$

We considering a stochastic discretization of the problem (1)-(4). Let us subdivide the domain $[0,1]\times[0,+\infty[$ into equal rectangles of sides $\Delta x=h$, $\Delta t=k$, and introduce coordinates of a typical mesh point $P(ih,jk)$; let us also put $u(ih,jk)=U(i,j)$, $F(ih,jk)=F(i,j)$ and $f(ih)=f(i)$. Let us approximate the partial derivatives

$$u_t(ih,jk)\approx\frac{U(i,j+1)-U(i,j)}{k}\quad;\quad u_x(ih,jk)\approx\frac{U(i+1,j)-U(i,j)}{h}$$

$$\begin{aligned}\left[p(x)u_x\right]_x(ih,jk)\approx{}&\frac{1}{h^2}\\ &\{p(i)U(i+1,j)-(p(i)+p(i-1))U(i,j)\\ &+p(i-1)U(i-1,j)\},\end{aligned}$$

and consider the discrete stochastic partial difference mixed problem

$$\begin{aligned}&-a\left\{p(i)U(i+1,j)-\left[p(i)+p(i-1)\right]U(i,j)+p(i-1)U(i-1,j)\right\}\\ &\qquad+\left[U(i,j+1)-U(i,j)\right]=kF(i,j)\quad,i\in N(1,K)\quad,\quad j\geq 0\end{aligned} \tag{32}$$

$$\alpha U(1,j)=U(0,j)\quad,\quad j\geq 0\quad, \tag{33}$$

$$U(K+1,j)=\beta U(K,j)\quad,\quad j\geq 0\quad, \tag{34}$$

$$U(i,0)=f(i)\quad,\quad 1\leq i\leq K\quad, \tag{35}$$

$$\left.\begin{array}{l} a=\frac{k}{h^2}=\frac{\Delta t}{(\Delta x)^2}\ ,\ h=\frac{1}{k}\ ,\ 1\leq i\leq K\quad j\geq 0\\ \alpha=\frac{a_2}{a_2-a_1h}\qquad,\qquad \beta=\frac{b_2-b_1h}{b_2}.\end{array}\right\} \tag{36}$$

Firstly, we seek solutions of (32) with $F=0$, of the form

$$U(i,j)=H(i)G(j)\quad,\quad 1\leq i\leq K\quad j\geq 0\quad, \tag{37}$$

satisfying (33)-(34),

$$\begin{aligned}&-a\{p(i)H(i+1)-\left[p(i)+p(i-1)\right]H(i)\\ &\qquad+p(i-1)H(i-1)\}G(j)=-H(i)\left[G(j+1)-G(j)\right],\end{aligned} \tag{38}$$

$$H(0)=\alpha H(1)\quad,\quad H(K+1)=\beta H(K)\quad. \tag{39}$$

Adding the term $a\lambda G(j)H(i)$ to both members of (38), where λ is a real parameter, the resulting equation can be written in the form

$$\begin{aligned}&-a\left\{p(i)H(i+1)-\left[p(i)+p(i-1)-\lambda\right]H(i)+p(i-1)H(i-1)\right\}G(j)\\ &\qquad+H(i)\left\{G(j+1)-(1-a\lambda)G(j)\right\}=0.\end{aligned} \tag{40}$$

Note that equation (40) holds true if

$$p(i)H(i+1)-\left[p(i)+p(i-1)-\lambda\right]H(i)+p(i-1)H(i-1)=0,\ 1\leq i\leq K\quad, \tag{41}$$

and

$$G(j+1)-(1-a\lambda)G(j)=0 \quad , \quad j\geq 0 \quad . \tag{42}$$

Note that equation (41) together with (39) defines a random discrete Sturm-Liouville problem. Under hypothesis (9), such problem admits exactly K real distinct random eigenvalues functions λ_m, $1\leq m\leq K$, and corresponding to each eigenvalues r.v. λ_m there exists an eigenfunction r.v. $\phi_m(i)$, $1\leq i\leq K$. Now let us seek a solution process of the unhomogeneous problem (32)-(35) of the form

$$U(i,j)=\sum_{n=1}^{K}\phi_n(i)b_n(j) \quad , \tag{43}$$

where $b_n(j)$ are r.v. to be determined for $1\leq n\leq K$, $j\geq 0$ and $\{\phi_n(\cdot)\}_{n=1}^{K}$ are chosen so that they are orthonormal with respect to the weight function $r(i)=1$.

Let us take j fixed, and using the results of the section 2, let us consider the random discrete Fourier series expansion of the process $F(\cdot,j)$, see (17)-(19), given by

$$F(i,j)=\sum_{n=1}^{K}\gamma_n(j)\phi_n(i) \quad ; \quad \gamma_m(j)=\sum_{n=1}^{K}F(n,j)\phi_m(n) \quad , \tag{44}$$

with $1\leq i\leq K$, $j\geq 0$. Substituting (43) and (44) into (32) and taking account that $(\lambda_m,\phi_m(\cdot))$ is a random eigenpairs of problem (10), it follows that

$$\sum_{n=1}^{K}\left[b_n(j+1)-(1-a\lambda_n)b_n(j)-k\gamma_n(j)\right]\phi_n(i)=0 \quad . \tag{45}$$

Note that (45) holds if $b_n(j)$ satisfies the random difference equation

$$b_n(j+1)-(1-a\lambda_n)b_n(j)=k\gamma_n(j) \quad , \quad 1\leq n\leq K \quad , \quad j\geq 0 \quad . \tag{46}$$

The solution of the analogous deterministic problem, see [Agarwal, 1991, p.68], suggests the solution

$$b_n(j)=(1-a\lambda_n)^j b_n(0)+\sum_{l=0}^{j-1}k(1-a\lambda_n)^{j-1-l}\gamma_n(l) \quad , \quad j\geq 1 \quad . \tag{47}$$

¿From the initial condition $U(i,0)=f(i)$, one gets

$$b_n(0)=\sum_{i=1}^{K}f(i)\phi_n(i)=\alpha_n \quad , \quad 1\leq n\leq K \quad , \tag{48}$$

and by (43), (47), (48) one gets the approximating stochastic process

$$U(i,j)=\sum_{n=1}^{K}\alpha_n(1-a\lambda_n)^j\phi_n(i)+k\sum_{n=1}^{K}\sum_{l=0}^{j-1}(1-a\lambda_n)^{j-1-l}\gamma_n(l)\phi_n(i) \quad . \tag{49}$$

Once we have the approximating stochastic diffusion process (49) we may compute the expectation $E\left[U(i,j)\right]$ and the variance $V\left[U(i,j)\right]$ assuming the knowledge of the N-density function of both process $p(x)$ and $F(x,t)$, for a fixed value of t. In the following section an illustrative example is included. From the computational point of view the stability condition requires an appropriate size of the parameter $a = \frac{k}{h^2} = \frac{\Delta t}{(\Delta x)}$ so that

$$|1 - a\lambda_n| < 1, \quad 1 \leq n \leq K. \tag{50}$$

This condition guaranties that the values of any realization of the discrete process $U(i,j)$ remain bounded.

4 Numerical example

Let us consider the stochastic problem

$$\begin{gathered} u_t = \left[p(x)u_x\right]_x + 4tv^3 \sin\left(\frac{3\pi}{2}x\right) \quad , \quad 0 < x < 1 \quad , \quad t > 0 \\ u(0,t) = 0 \quad , \quad t > 0, \\ u_x(1,t) = 0 \quad , \quad t > 0, \\ u(x,0) = 1 \quad , \quad 0 \leq x \leq 1, \end{gathered}$$

where $p(x) = v + \cos(vx)$, v is an uniform random variable defined on the interval $[0,1]$ and with the notation of section 3 we have $\alpha = 0$, $\beta = 1$. In the following tables we compare the expectation value $\widehat{E}\left[U(x,t)\right]$ and the variance $\widehat{V}\left[U(x,t)\right]$ of the discrete approximate process $U(i,j)$ given by (49) at the points $\left\{\left(\frac{i}{10}, 1\right); 1 \leq i \leq 9\right\}$ taking an appropriate time discretization $\Delta t = 1/400$ and several different space discretization $\Delta x = h$ so that the stability condition (50) is satisfied. This allows the comparison of the computed values in order to show the changes with respect to the uncertain variable x.

Note that the discrete approximate process $U(i,j)$ in our example is a function of the random variable v. Hence

$$E\left[U\left(i,j\right)\right] = \int_0^1 U\left(i,j\right)(v)dv \tag{51}$$

$$\begin{aligned} V\left[U\left(i,j\right)\right] &= E\left[U\left(i,j\right)^2\right] - \left(E\left[U\left(i,j\right)\right]\right)^2 \\ &= \int_0^1 U\left(i,j\right)^2(v)dv - \left(\int_0^1 U\left(i,j\right)(v)dv\right)^2 \end{aligned} \tag{52}$$

In the tables the numerical integration of previous expressions (51) and (52) are performed using composite Simpson's rule with 10 points.

(x,t)	$\widehat{E}\,[U(x,t)]$	$\widehat{u}(i,j)$	$\widehat{V}\,[U(x,t)]$
$(1/10,1)$	0.0068	0.0055	1.9×10^{-5}
$(2/10,1)$	0.0135	0.0108	7.5×10^{-5}
$(3/10,1)$	0.0198	0.0159	1.6×10^{-4}
$(4/10,1)$	0.0257	0.0206	2.7×10^{-4}
$(5/10,1)$	0.0310	0.0249	3.9×10^{-4}
$(6/10,1)$	0.0356	0.0286	5.1×10^{-4}
$(7/10,1)$	0.0393	0.0316	6.2×10^{-4}
$(8/10,1)$	0.0421	0.0339	7.1×10^{-4}
$(9/10,1)$	0.0439	0.0353	7.7×10^{-4}

Table 1. Numerical results for $K = 40$, $a = 1/4$.

(x,t)	$\widehat{E}\,[U(x,t)]$	$\widehat{u}(i,j)$	$\widehat{V}\,[U(x,t)]$
$(1/10,1)$	0.0066	0.0053	1.8×10^{-5}
$(2/10,1)$	0.0130	0.0104	7.2×10^{-5}
$(3/10,1)$	0.0192	0.0153	1.5×10^{-4}
$(4/10,1)$	0.0249	0.0198	2.6×10^{-4}
$(5/10,1)$	0.0300	0.0239	3.7×10^{-4}
$(6/10,1)$	0.0344	0.0275	4.9×10^{-4}
$(7/10,1)$	0.0380	0.0304	5.9×10^{-4}
$(8/10,1)$	0.0406	0.0325	6.8×10^{-4}
$(9/10,1)$	0.0423	0.0339	7.3×10^{-4}

Table 2. Numerical results for $K = 80$, $a = 1/4$.

Acknowledgements.This work has been supported by the Spanish M.C.Y.T. grant DPI 2004-08383-C03-03.

References

[Agarwal, 1991]R. P. Agarwal. *Difference Equations and Inequalities.* Marcel Dekker, New York, 1991.

[Birkhoff and Rota, 1965]G. Birkhoff and G.C. Rota. *Ordinary Differential Equations.* John–Wiley, New York, 1965.

[Boyce, 1960]W. E. Boyce. Random eigenvalues problems. In A. T. Bharucha-Reid, editor, *Probabilistics Methods in Apllied Mathemactics*, pages 1–73, 1960.

[Chilés and Delfiner, 1999]J. Chilés and P. Delfiner. *Geostatistics. Modelling Spatial Uncertainty.* John–Wiley, London, 1999.

[Cortés *et al.*, 2005a]J.C. Cortés, P. Sevilla-Peris, and L. Jódar. Analytic-numerical approximating processes of diffusion equation with data uncertainty. *Computers Math. Appl.*, 2005.

[Cortés *et al.*, 2005b]J.C. Cortés, P. Sevilla-Peris, and L. Jódar. Constructing approximate diffusion processes with uncertain data. *Appl. Num. Math.*, 2005.

[Keller, 1963]J. B. Keller. Stochastic equations and wave propagation in random media. In *Simp. Appl. Math.*, pages 145–170, 1963.

[Kloeden and Platen, 1992]P. E. Kloeden and E. Platen. *Numerical Solution of Stochastic Differential Equations*. Springer, Berlin, 1992.

[Soong, 1973]T. T. Soong. *Random Differential Equations in Science and Engineering*. Academic–Press, New York, 1973.

[Weinberger, 1965]H. F. Weinberger. *Partial Differential Equations*. Blaisdell Pub. Co., New York, 1965.

Part XV

Time Series

Nonparametric Regression in Time Series with Errors-in-Variables

Dimitris Ioannides[1] and Philippos Alevizos[2]

[1] Department of Economics
University of Macedonia
54006 Thessaloniki, Greece
(e-mail: dimioan@uom.gr)

[2] Department of Mathematics
University of Patras
26500 Patras, Greece
(e-mail: philipos@math.upatras.gr)

Abstract. In this paper, we study the nonparametric estimation of the regression function for dependent data with measurement errors in responses and covariates. The usual assumption in the errors-in-variables problem of indepedent errors can be replaced by dependent errors when the data are time series. Both cases are examined, and it is considered for first time the effect of measurement errors in responses when we are estimating nonparametrically the regression function.

Keywords: Deconvolution, nonparametric estimation, $\alpha-$mixing, noisy observations, regression function, uniform convergence.

1 Intoduction

Let $\{(X_i, Y_i)\}$, $i \geq 1$, be a strictly stationary process, where (X_i, Y_i) takes values in $\mathbb{R}^d \times \mathbb{R}$, $d \geq 1$, and has probability density function (pdf) $f(x, y)$. Consider the deconvolution model

$$Z_i = Y_i + \eta_i, \quad \text{and} \quad S_i = X_i + \epsilon_i, \tag{1}$$

where the noise processes $\{\eta_i\}$, and $\{\epsilon_i\}$, $i \geq 1$, are independent of the processes $\{Y_i\}$ and $\{X_i\}$, $i \geq 1$, respectively. In addition, we assume that the marginal distributions of the noise processes $\{\eta_i\}$, $i \geq 1$, and $\{\epsilon_i\}$, $i \geq 1$, are known, and also the components $\epsilon_{i1}, ..., \epsilon_{id}$ of the random vector ϵ_i are indentically distributed according to a r.v. ϵ. Models of this type and the deconvolution problems to which they lead arise in a variety of contexts in economic statistics, biostatistics, and various other fields. For example, if $d = 1$ in (1), X_i may represent the true income of a household at time i measured with error ϵ_i, Y_i its expenditures for some good which is subject to the measurement error η_i, and S_i, Z_i its measured income and expenditures, respectively. The interested reader may find additional applications of this problem in [Carroll *et al.*, 1995].

On the basis of the observations $(Z_1, S_1), ..., (Z_n, S_n)$, the problem is that of providing nonparametric estimate of the kth conditional moment function

$m(k;x) = E(Y^k/X = x)$, where Y and X are distributed as the r.v.'s Y_i and X_i, respectively. For the special case, $k = 1$, this problem was extensively studied in the literature and also when the covariates X_i are measured with some noise (i.e. $\eta_i \equiv 0, \epsilon_i \neq 0$). See, for example, [Carroll *et al.*, 1995], [Fan and Masry, 1992] and [Ioannides and Alevizos, 1997].

Here, we investigate the more complicate deconvolution model defined as in (1).

If we were using a Nadaraya-Watson type estimator, this problem could not be solved since for $k = 1$ the noise could not be extracted from the responses. Instead to use a Nadaraya-Watson type estimator, we construct first an estimator $\hat{f}_n(y/x)$ for the conditional density of Y given X, $f_{Y/X}(y/x)$, on the basis of our osbervations $(Z_1, S_1), ..., (Z_n, S_n)$. Then one natural estimator of $m(k;x)$ is obtained if we integrating apropriate the quantity $y^k\hat{f}_n(y/x)$ with respect to y. Because the type of this estimator was first introduced for uncontaminated data by [Roussas, 1969], we call it Roussas's estimator. In order to construct an estimator for $f_{Y/X}(y/x)$, the introduction of some notation and related concepts is necessary. Let $\widetilde{\Phi}_{K_1}(t)$ and $\widetilde{\Phi}_{K_2}(\tau)$ be the Fourier transforms of the univariate kernel density functions $\widetilde{K}_1(x)$ and $\widetilde{K}_2(y)$, and let $\widetilde{\Phi}_\epsilon(t)$ and $\widetilde{\Phi}_\eta(\tau)$ be the characteristic functions of the noise variables ϵ and η, respectively. Then, as in [Fan, 1991], we define the corresponding deconvoluting kernel functions by the following relations,

$$\widetilde{W_{1n}}(u) = \frac{1}{2\pi}\int_{\mathrm{I\!R}} e^{-iut}\frac{\widetilde{\Phi}_{K_1}(t)}{\widetilde{\Phi}_\epsilon(\frac{t}{h_n})}dt, \quad \widetilde{W_{2n}}(v) = \frac{1}{2\pi}\int_{\mathrm{I\!R}} e^{-iv\tau}\frac{\widetilde{\Phi}_{K_2}(\tau)}{\widetilde{\Phi}_\eta(\frac{\tau}{h_n})}d\tau, \qquad (2)$$

where $0 < h_n \downarrow 0$. Thus the deconvoluting nonparametric estimator for the conditional density function is given by:

$$\hat{f}_n(y/x) = \frac{\hat{f}_{2n}(x,y)}{\hat{f}_{1n}(x)}, \qquad (3)$$

where $\hat{f}_{1n}(x) = \frac{1}{nh_n^d}\sum_{i=1}^n W_{1n}(\frac{x-S_i}{h_n})$ and
$\hat{f}_{2n}(x,y) = \frac{1}{nh_n^{d+1}}\sum_{i=1}^n W_{1n}(\frac{x-S_i}{h_n})\widetilde{W}_{2n}(\frac{y-Z_i}{h_n})$ with $W_{1n}(x) = \prod_{j=1}^d \widetilde{W}_{1n}(x_j)$.

Consequently the Roussas's estimator for the kth conditional moment is defined as follows:

$$\begin{aligned} m_n(k;x) &= \int_{-B_n}^{B_n} y^k\hat{f}_n(y/x)dy \\ &= \frac{1}{h_n}\sum_{i=1}^n \frac{W_{1n}(\frac{x-S_i}{h_n})\int_{-B_n}^{B_n} y^k\widetilde{W}_{2n}(\frac{y-Z_i}{h_n})dy}{\sum_{i=1}^n W_{1n}(\frac{x-S_i}{h_n})}, \end{aligned} \qquad (4)$$

where B_n goes to infinity as $n \to \infty$. We integrate the quantity $y^k \widetilde{W}_{2n}(\frac{y-Z_i}{h_n})$ from $-B_n$ to B_n, since this is not in general integrable. The proposed estimator can be used in certain prediction problems.

[Ioannides, 1999] proved that the modal regression estimator can be used for extracting the noises from both variables Y and X. [Hannan, 1963] and [Robinson, 1986] treating this problem in the case by which $m(1;x) = E(Y/X = x)$ is the simple linear regression model. This paper attempts to study this problem in a more general setting using the Roussas's estimator (4).

In most publications on nonparametric deconvolution problems, a distinction is made between the case, where the noise characteristic functions $\widetilde{\Phi}_\epsilon(t)$ and $\widetilde{\Phi}_\eta(\tau)$ decay for large $|t|$ and $|\tau|$ either algebraically (*ordinary smooth case*) or exponentially (*supersmooth case*).

In the case by which the noise variables follow an ordinary smooth distribution, one of our main results is that the rates of the uniform strong convergence for the Roussas's estimator in (4) is of order $\max\{(\frac{logn}{nh_n^{[(d+2\beta)+1+2\beta']}})^{\frac{1}{2}}, h_n\}$ with β and β' positive numbers greater than 1 denoted the degree of smoothness of the noise variables ϵ and η, respectively. See, also, Assumption (A5) in the Appendix. This rates is better than the rate $\max\{(\frac{logn}{nh_n^{2[(d+2\beta)+1+2\beta']}})^{\frac{1}{4}}, h_n\}$ found for the modal estimator in [Ioannides, 1999]. In the noiseless case our rate becomes of order $\max\{(\frac{logn}{nh_n^{d+1}})^{\frac{1}{2}}, h_n\}$, which is essentially the optimal rate of $\hat{f}_{2n}(x,y)$ for estimating the pdf $f(x,y)$, and it is slight weaker than the optimal rate $\max\{(\frac{logn}{nh_n^d})^{\frac{1}{2}}, h_n\}$, obtained by the Nadaraya-Watson estimator.

The case where the noise variable has a super smooth distribution can be treated similarly, but the a.s. convergence rate is expected to be of logarithmic order.

Another interesting aspect of this paper is that we are not dealing only with independent measurement errors, but in general we allow them to be dependent. Usually in nonparametric deconvolution problems it is assumed that the noise process cosists from i.i.d. r.v.'s, avoiding correlated noise as is considered by [Hannan, 1963] and [Robinson, 1986]. Assuming that the joint stochastic process $\{(X_i, \epsilon_i, Y_i, \eta_i)\}$, $i \geq 1$, is a strong mixing process and the noise process $\{(\epsilon_i, \eta_i)\}$, $i \geq 1$ either consists from i.i.d. r.v.'s or dependent indentical r.v.'s we are proving the strong consistency of Roussas's estimator with the above rates under some mixing conditions which are weaker for the i.i.d measurement errors case.

This paper is organized as follows. The main result, Theorem 3.1, is given in Section 3, while some preparatory lemmas are given in Section 2. All the asssumptions made in this paper are given at the end of the paper in Appendix.

2 Some preliminary results

Set

$$R_n^{B_n}(k;x) = \frac{1}{h_n^{d+1}} \sum_{i=1}^{n} W_{1n}(\frac{x-S_i}{h_n}) \int_{-B_n}^{B_n} y^k \widetilde{W}_{2n}(\frac{y-Z_i}{h_n})dy, \tag{5}$$

then the Roussas's estimator $m_n(k;x)$ can be written as

$$m_n(k;x) = \frac{R_n^B(k;x)}{\hat{f}_n(x)}.$$

Now, for $k > 0$, denote by C_k the Cube in $\mathbb{R}^d$ which is the Cartesian product of d copies of $[-k,k]$. Then working similar as in [Roussas, 1990], dividing $[-k,k]$ into b_n subintervals each of length δ_n, and taking J_{nl}, $l = 1,...,N$ the sets into which C_k is divided. Let x_{nl} arbitrary points in J_{nl}. Pick k sufficiently large, so that $J \subset C_k$, J compact subinterval of $\mathbb{R}^d$. Then, clearly,

$$\begin{aligned} |m_n(k;x) - m(k;x)| \leq |\hat{f}_n^{-1}|\{&|ER_n(k;x) - ER_n^{B_n}(k;x)| \\ &+ |R_n^{B_n}(k;x) - R_n^{B_n}(k;x_{nl})| \\ &+ |ER_n^{B_n}(k;x) - ER_n^{B_n}(k;x_{nl})| \\ &+ |ER_n^{B_n}(k;x) - R(k;x)| + |R(k;x)||\hat{f}_n(x) - f(x)| \\ &+ |R_n^{B_n}(k;x_{nl}) - ER_n^{B_n}(k;x_{nl})|\}, \end{aligned} \tag{6}$$

with $R(k;x) = \int_{\mathbb{R}} y^k f(x,y)dy$, and
$R_n(k;x) = \frac{1}{h_n^{d+1}} \sum_{i=1}^{n} K_1(\frac{x-X_i}{h_n}) \int_{\mathbb{R}} y^k \widetilde{K}_2(\frac{y-Y_i}{h_n})dy$.

Lemma 2.1. Under Assumptions (A1)(ii)-(iv), and (A6)(ii), one has

$$|E\hat{R}_n(k;x) - E\hat{R}_n^{B_n}| \leq cB_n^{k-s},$$

for all x in $\mathbb{R}^d$, and $s > k$.

Lemma 2.2. Under Assumptions (A2), and (A5), it holds: $|R_n^{B_n}(k;x) - R_n^{B_n}(k;x')|$ and $|E\hat{R}_n^{B_n}(k;x) - E\hat{R}_n^{B_n}(k;x')|$ are bounded by $c_1 B_n^k h_n^{-(d(\beta+1)+\beta'+2)} \sum_{i=1}^{d} |x_i - x_i'|$, for any x, $x' \in \mathbb{R}^d$, and $c_1 > 0$.

Lemma 2.3. Under Assumptions (A1)(ii)-(iii), it holds

$$|E\hat{R}_n(k;x) - R(k;x)| \leq c_2 h_n,$$

for all $x \in \mathbb{R}^d$, and some $c_2 > 0$.

Lemma 2.4. (i) Under Assumptions (A1)(i)-(iv), (A4), (A5), and if the noise processes $\{\epsilon_i\}$ and $\{\eta_i\}$, $\{i \geq 1\}$, consists from i.i.d. r.v.'s, then

$$\lim_{n\to\infty} \sup_{\mathbb{R}^d} nh_n^{d(1+2\beta)+1+2\beta'} Var(R_n^{B_n}(k;x)) < c',$$

for some $c' > 0$.

(ii) Under the additional Assumption (A1)(v), one has

$$\lim_{n\to\infty} \sup_{\mathbb{R}^d} nh_n^{d(1+2\beta)+1+2\beta'} Var(R_n^{B_n}(k;x)) < c',$$

for some $c' > 0$.

Lemma 2.5. Under Assumptions (A1), (A4), (A5) and (A6) one has

$$|R_n^{B_n}(x_{nl}) - ER_n^{B_n}(x_{nl})| = O\left[(\frac{logn}{nh_n^{d(1+2\beta)+1+2\beta'}})^{\frac{1}{2}} \right], \quad \text{a.s.}$$

Lemma 2.6. Under Assumptions (A1), (A4), (A5) and (A6) one has

$$|\hat{f}_n(x_{nl}) - E\hat{f}_n(x_{nl})| = O\left[(\frac{logn}{nh_n^{d(1+2\beta)+1+2\beta']}})^{\frac{1}{2}} \right], \quad \text{a.s.}$$

3 Main Result

The main result of this paper is the following theorem whose proof is a consequence of the preliminary results established. More precisely, one has:

3.1 Theorem

Under Assumptions (A1)-(A6), then

$$sup_{x\in J}|m_n(k;x) - m(k;x)| \leq O(h_n) + O\left[(\frac{logn}{nh_n^{d(1+2\beta)+1+2\beta'}})^{\frac{1}{2}} \right], \quad \text{a.s.}$$

Proof: The proof follows from Lemmas 2.1-2.6, in conjunction with the relation (6) using the same technique as in [Roussas, 1990] and [Ioannides, 1999]. ■

4 Appendix

The basic assumptions under which the a.s. uniform convergence of $m_n(k;x)$ is established.

Assumption (A1)

(i) The process $\{(X_i, Y_i, \epsilon_i, \eta_i)\}$, $i \geq 1$, is strictly stationary.
(ii) The processes $\{(X_i, Y_i)\}$, $i \geq 1$, and $\{(\epsilon_i, \eta_i)\}$, $i \geq 1$, are independent.
(iii) The processes $\{\epsilon_i\}$, $i \geq 1$, and $\{\eta_i\}$, $i \geq 1$, are independent.
(iv) The process $\{(X_i, Y_i, \epsilon_i, \eta_i)\}$, $i \geq 1$, is $\alpha-$mixing with mixing coefficient $\alpha(i) = O(i^{-k})$, $k > 1 + \frac{2}{\delta}, \delta > 0$.
(v) The process $\{(X_i, Y_i, \epsilon_i, \eta_i)\}$, $i \geq 1$, is $\alpha-$mixing with mixing coefficient $\alpha(i)$ satisfying the requirement $\frac{1}{h_n^{(2d\beta+2\beta')}} \sum_{j=h_n^{-d-1}}^{\infty} \alpha(j)^{\frac{2}{2+\delta}} < \infty$, for $h_n \to 0$, and $c_n \to \infty$, as $n \to \infty$.

Assumption (A2)

(i) The probability density $f(x)$ of X satisfies the Lipschitz condition of order 1 on $\mathbb{R}^d$.
(ii) $\inf_{x \in J} |f(x)| > 0$, where J is a compact subset of $\mathbb{R}^d$.
(iii) The quantity $R(k;x)$ satisfies the Lipschitz condition of order 1 on $\mathbb{R}^d$.

Assumption (A3)

The kernel functions $\widetilde{K}_i(.)$, $i = 1, 2$ are bounded probability density functions on $\mathbb{R}$ with $\int_{\mathbb{R}} |u| \widetilde{K}_1(u) du < \infty$ and $\int_{\mathbb{R}} |v| \widetilde{K}_2(v) dv < \infty$.

Assumption (A4)

(i) The characteristic functions $\widetilde{\Phi}_\epsilon(t)$ and $\widetilde{\Phi}_\eta(\tau)$ satisfy $\widetilde{\Phi}_\epsilon(t) \neq 0, \quad \widetilde{\Phi}_\eta(\tau) \neq 0$ for all t and τ.
(ii) $|t|^{\beta} |\widetilde{\Phi}_\epsilon(t)| > d, \quad d > 0, \quad |t|^{\beta'} |\widetilde{\Phi}_\eta(\tau)| > d', \quad d' > 0$, for large t and τ, and for some positive β , β'.
(iii) The joint characteristic function of ϵ_1 and ϵ_j is ordinary smooth of order 2β.
(iv) The joint characteristic function of η_1 and η_j is ordinary smooth of order $2\beta'$.

Assumption (A5)

The characteristic functions $\widetilde{\Phi}_{K_1}(t)$ and $\widetilde{\Phi}_{K_2}(\tau)$ satisfy the requirements: $\int |t|^{1+\beta} \widetilde{\Phi}_{K_1}(t) dt < \infty, \quad \int |\tau|^{1+\beta'} \widetilde{\Phi}_{K_2}(\tau) d\tau < \infty$.

Assumption (A6)

(i) $E|Y|^s \leq \infty$, for $s > 1$, and $x \in E \subset \mathbb{R}^d$.
(ii) $sup_{x \in \mathbb{R}^d} [\int |y|^s f(x,y) dy] < \infty$, for some $s > 1$.

References

[Carroll *et al.*, 1995]R.J. Carroll, D. Ruppert, and L.A. Stefanski. *Measurement Error in Nonlinear Models*. Chapman Hall, Britain, 1995.

[Fan and Masry, 1992]J. Fan and E. Masry. Multivariate regression estimation with error-in-variables. *J. Multivariate Anal.*, pages 237–271, 1992.

[Fan, 1991]J. Fan. On the optimal rates of convergence for nonparametric deconvolution problems. *Ann. Statist.*, pages 1257–1272, 1991.

[Hannan, 1963]E.J. Hannan. Regression for time series with errors of measurement. *Biometrica*, pages 293–302, 1963.

[Ioannides and Alevizos, 1997]D. Ioannides and P. Alevizos. Nonparametric regression with errors in variables and applications. *Statist. Probab. Letters*, pages 35–43, 1997.

[Ioannides, 1999]D. Ioannides. Estimating the conditional mode of a stationary process from noise observations. *Metrika*, pages 19–35, 1999.

[Robinson, 1986]P.M. Robinson. On the errors-in-variables problem for time series. *J. Multivariate Anal.*, pages 240–250, 1986.

[Roussas, 1969]G.G. Roussas. Nonparametric estimation of the transition distribution of a markov process. *Ann. of Math. Stat.*, pages 1386–1400, 1969.

[Roussas, 1990]G.G. Roussas. Nonparametric regression estimation under mixing conditions. *Stochastic Processes and their Applic.*, pages 107–116, 1990.

Classification of GARCH Time Series: A Simulation Study

Thomas Kalantzis and Demetrios Papanastassiou

University of Macedonia,
Dept. of Applied Informatics,
156 Egnatia Street,
54006 Thessaloniki, Greece
(e-mail: tkalant@uom.gr, papanast@uom.gr)

Abstract. We examine a discrimination rule for time series data generated by a GARCH(1,1) process that classifies a sample into a group in terms of its unconditional variance. A simulation study indicates that our rule is more efficient than a benchmark rule in all cases, except from a narrow range of alternatives lying on the right side of the null.

Keywords: Local heteroscedasticity, Discrimination rule, Likelihood ratio.

1 Introduction

In analyzing high frequency financial time series data, the common practice is to examine the first differences of the logged observations, known as returns. Contrary to the raw prices, returns are considered to be more amenable to statistical manipulations. Under some fundamental economic hypotheses, they form a sample of uncorrelated second order stationary series.

However, if we look at a typical returns plot of reasonable length, we shall observe clusters of different variation, which, at a first sight, may cast some doubt on the issue of the conventional equal variance perception. The main characteristics of this idiosyncratic regular local heteroscedasticity are captured by the widely used GARCH models introduced by [Bollerslev, 1986]. For a returns series, the variance is of practical interest, since it is widely considered as a measure of the risk involved on investing on the particular stock, [Tsay, 2002].

If we want to classify such a series in terms of its variance into one of two groups, in principle we can treat it as an independent sample from identically distributed observations and apply the usual discriminant function, see [Johnson and Wichern, 1992]. However, because of the presence of the local heteroscedasticity, and the fact that independence and normality are challenged both on empirical and theoretical basis, we were motivated to seek discrimination rules which take these facts into account.

In this paper we introduce a likelihood ratio type discrimination rule to classify a GARCH(1,1) process into two categories. Since it is the unconditional long term variance which is mainly of interest, the test concentrates on this aspect. Methods and theory for discriminating processes on an overall basis, mainly of the linear type, are reviewed by [Taniguchi and Kakizawa, 2000] and [Shumway and Stoffer, 2000].

In the sequel, in Section 2 we discuss the discrimination rule for an appropriately parameterized GARCH(1,1) model. In Section 3 we present the results of a simulation study comparing our approach against the benchmark rule of independent and identically distributed, iid, observations. The final section summarizes our conclusions and suggestions.

2 The GARCH(1,1) Discrimination Rule

Let $Y_t, t = 1, 2, ..., n$, be a set of normally distributed iid observations. Suppose one samples from either of two groups, $G_1 : N(0, \sigma_1^2)$ or $G_2 : N(0, \sigma_2^2)$. The conventional likelihood ratio based rule, see [Johnson and Wichern, 1992], states that

classify the sample as belonging to G_1 when $ln\frac{L_1}{L_2} \geq 0$,

while

classify the sample as belonging to G_2 when $ln\frac{L_1}{L_2} < 0$, (1)

where L_j is the sample likelihood value, supposing it comes from G_j, $j = 1, 2$. More precisely, the discriminant function is

$$\ln\frac{L_1}{L_2} = -\frac{T}{2}\ln\frac{\sigma_1^2}{\sigma_2^2} - \frac{1}{2}\sum_{t=1}^{t=T} y_t^2\left(\frac{1}{\sigma_1^2} - \frac{1}{\sigma_2^2}\right). \tag{2}$$

On the other hand, suppose our data are generated by a stationary GARCH(1,1) process, [Bollerslev, 1986],

$$\begin{aligned} Y_t &= u_t, \\ u_t &= \varepsilon_t h_t^{1/2}, \varepsilon_t \overset{iid}{\text{sim}} N(0,1), \\ h_t &= a_0 + a_1 u_{t-1}^2 + b_1 h_{t-1}, \end{aligned} \tag{3}$$

ε_t independent of h_t, and $a_0 > 0$, $a_1, b_1 \geq 0$, are constant parameters. It is easy to see that the unconditional variance of Y_t is

$$\sigma^2 = E(Y_t^2) = \frac{a_0}{1 - a_1 - b_1},$$

and that, although Y_t are uncorrelated, they are not independent and normally distributed, see [Hamilton, 1994], amongst many others. Since σ^2 is the parameter of our prime interest, we reparameterize the model in terms of σ^2, writing

$$h_t = \sigma^2(1 - a_1 - b_1) + a_1 u_{t-1}^2 + b_1 h_{t-1}. \tag{4}$$

Group G_j, $j = 1, 2$, is described as the set of all possible GARCH(1,1) models that have the same variance σ_j^2, $j = 1, 2$. We are interested to allocate a sample $Y_t, t = 1, 2, ..., T$, to one of the G_1 and G_2 groups in terms only of its variance σ^2. The a_1 and b_1 parameters, parameterizing the dynamic behavior of the conditional variance, are a sort of nuisance parameters.

The likelihood based rule will remain as in (1), but the likelihood ratio in (2) is modified to take into account the special form of our heteroscedastic data. Noting that the conditional distribution of Y_t given h_t is a normal $N(0, h_t)$, the decomposition of the likelihood function of a time series process yields the discriminant function

$$\ln \frac{L_1}{L_2} = -\frac{1}{2}\sum_{t=1}^{T}\left(\ln \frac{h_{1t}}{h_{2t}}\right) - \frac{1}{2}\sum_{t=1}^{T} y_t^2 \left(\frac{1}{h_{1t}} - \frac{1}{h_{2t}}\right). \tag{5}$$

3 Simulation Study

We carried out a simulation study to assess the GARCH discriminant function in (5) against (2), which we consider as a sort of benchmark rule. The experimental data come from the GARCH(1,1) model in (3) with its conditional variance reparameterized as in (4). Examining real daily or weekly series of stock or exchange rate returns, we calculated their free variance to be of the order of $5 \cdot 10^{-5}$. We considered that as a typical variance value of real life data, and in our experiments we set the variance of group G_1 equal to this value, that is $\sigma_1^2 = 5 \cdot 10^{-5}$. The remaining parameters a_1 and b_1 take a range of values within what is considered as typical in the relative literature. We mention that the condition for (3) to be stationary is $a_1 + b_1 < 1$. Usually, in real series applications, the sum of a_1 and b_1 lies close to 1, and a_1 is always smaller than b_1. When $a_1 + b_1 = 1$ the model is still stationary, but with infinite variance and therefore makes no sense for our study.

Models	0	1	2	3	4	5
Parameters						
a_1	0.00	0.10	0.10	0.40	0.40	0.40
b_1	0.00	0.50	0.80	0.50	0.55	0.58
Sum	0.00	0.60	0.90	0.90	0.95	0.98

Table 1. Models tested in the simulation

The criterion to assess our findings was the error rate $P(2|1)$, that is the probability to allocate a sample to G_2 when it truly comes from G_1. These probabilities are reported in the corresponding tables and are calculated by repeating the same experiment 300 times. Factors we felt that might be of influence in the efficiency of the discrimination rules were the sample size

T, the magnitude of the alternative variance in G_2, and the combination of the a_1 and b_1 values. The simulation study was designed to take into consideration all these factors. In Table 1 we present only a few selected combinations of a_1 and b_1 values from those examined, declared as models 0 to 5. Practically, Model 0 is an iid series.

σ_2^2	1	2	3	4	4.5	4.9	5.1	5.5	6	7	8	9
series length												
GARCH rule												
300	.000	.053	.313	.500	.550	.567	.417	.393	.350	.300	.277	.250
1000	.000	.007	.127	.400	.513	.560	.417	.383	.337	.280	.227	.193
benchmark rule												
300	.000	.000	.003	.107	.317	.487	.420	.240	.107	.020	.003	.000
1000	.000	.053	.313	.013	.143	.460	.397	.160	.010	.000	.000	.000

Table 2. Error rates for Model 5. Alternative variance values should be multiplied by 10^{-5}. The true variance of the series is $5 \cdot 10^{-5}$

Before presenting our results, we clarify the computational flow of our procedure. Once we had in hand a series from G_1, we maximized L_1 with respect to a_1 and b_1 considering σ^2 known and equal to $\sigma_1^2 = 5 \cdot 10^{-5}$. Next, we maximized L_2 for a_1 and b_1 setting now $\sigma^2 = \sigma_2^2$, one of the alternatives. A conjugate gradient routine was written to maximize the loglikelihoods, after transforming a_1 and b_1 so that the restrictions, $a_1, b_1 \geq 0$ and $a_1 + b_1 < 1$ were fulfilled. If maximization of both L_1 and L_2 was terminated successfully, then (5) was calculated and the series was classified into G_1 or G_2 accordingly.

The most definite of our conclusions is that both rules perform better as alternative variance σ_2^2 takes values further away from σ_1^2. Also, the $P(2/1)$ error rate improves with the sample size, and this can be seen for the case of Model 3 in Table 2. Since the general pattern of $P(2/1)$ is the same for either $T = 300$ or $T = 1000$, for reasons of space economy, we report more detailed results in Table 3 only for $T = 1000$.

Concerning the effect of the sum $\alpha_1 + \beta_1$, the error rate for both rules increases as $\alpha_1 + \beta_1$ approaches unity. For models with the same sum, the rate is worse for larger α_1, see for instance Model 2 versus Model 3 in Table 3. This can be explained by the fact that larger α_1 implies wider local variance bursts.

Regarding the relative performance of the GARCH rule against the benchmark rule, which is of the main interest in our study, there is not a clear pattern for the whole range of alternatives. The GARCH rule is always better than the benchmark for σ_2^2 smaller than the true $\sigma_1^2 = 5 \cdot 10^{-5}$. For a range of alternatives from $5.1 \cdot 10^{-5}$ to approximately $10 \cdot 10^{-5}$, the benchmark rule outperforms the GARCH rule. This can be seen graphically in Fig.1 for the

σ_2^2:	1	2	3	4	4.5	4.9	5.1	5.5	6	7	8	9
GARCH	rule											
model 0	.000	.000	.003	.020	.133	.447	.413	.170	.133	.000	.000	.000
model 1	.000	.000	.000	.023	.190	.473	.410	.210	.070	.003	.000	.000
model 2	.010	.000	.000	.117	.316	.483	.447	.310	.197	.073	.027	.000
model 3	.000	.007	.127	.400	.513	.560	.417	.383	.337	.280	.227	.193
model 4	.000	.090	.363	.500	.570	.563	.420	.387	.350	.280	.260	.240
model 5	.090	.487	.663	.740	.783	.787	.193	.177	.167	.143	.110	.080
benchmark	rule											
model 0	.000	.000	.000	.013	.143	.460	.397	.160	.010	.000	.000	.000
model 1	.000	.000	.000	.030	.197	.480	.410	.207	.057	.003	.000	.000
model 2	.003	.000	.000	.137	.350	.500	.423	.297	.163	.037	.033	.000
model 3	.000	.033	.290	.051	.610	.677	.303	.257	.227	.197	.167	.133
model 4	.010	.307	.557	.710	.737	.747	.233	.213	.200	.180	.163	.133
model 5	.370	.690	.790	.827	.840	.853	.140	.133	.117	.103	.087	.083

Table 3. Error rates for models 0 to 5. Alternative variance values should be multiplied by 10^{-5}. The true variance of the series is $5 \cdot 10^{-5}$.

case of Model 3. The superiority of the benchmark rule grows larger as the sum $\alpha_1 + \beta_1$ approaches unity.

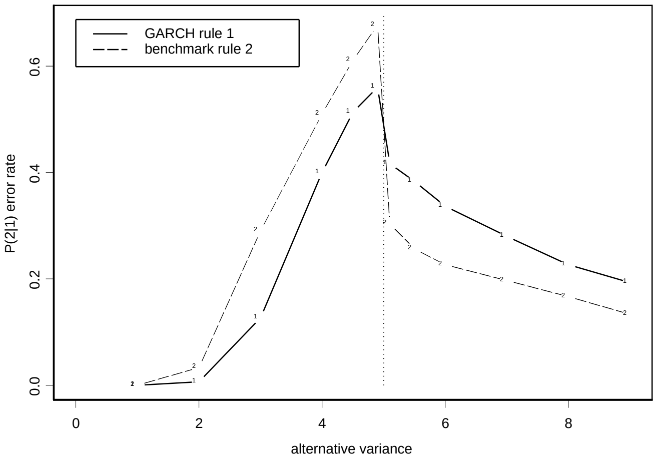


Fig. 1. Error rates for Model 3. Variances should be multiplied by 10^{-5}.

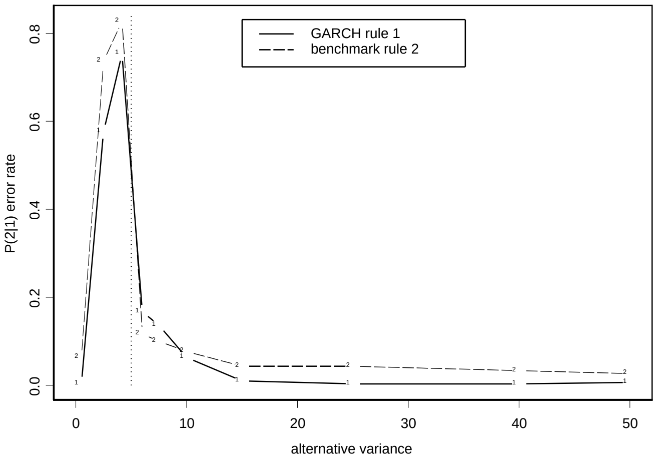


Fig. 2. Error rates for Model 5. Variances should be multiplied by 10^{-5}.

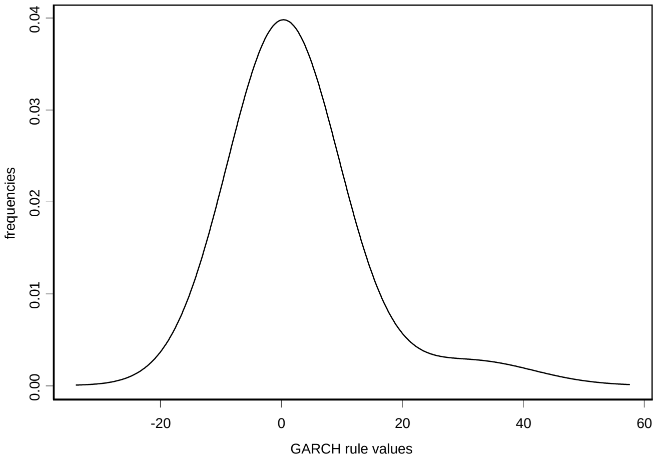


Fig. 3. Smoothed frequency curve from 2000 replications from the GARCH rule, Model 5, $\sigma_2^2 = 7 \cdot 10^{-5}$.

Moving farther to the right of σ_1^2, the pattern is reversing. Fig.2 illustrates the case for Model 5. Note that error rates in this interval are not reported in Table 3. We can not explain this behavior. Finally, Fig.3 gives a smoothed plot of the distribution of (5) for Model 5 with $\sigma_2^2 = 7 \cdot 10^{-5}$. The simulated distribution was plotted from 2000 replications.

4 Conclusions

We conducted an empirical study classifying GARCH(1,1) time series data on the basis of their unconditional variance. The procedure may prove useful to classify financial returns data into different risk groups.

The GARCH rule is better than the benchmark rule, except from a small range of alternatives starting from the null σ_1^2 and going approximately up to $\sigma_2^2 = 2\sigma_1^2$. This is a point that deserves further investigation, and a proper derivation of the distribution of (5) may shed some light.

Rule (5) generalizes easily for higher order GARCH models, although for most applications a simple GARCH(1,1) suffices. Experience with real data could allow us to cross examine rule (5) with other risk classifying criteria, such as the β coefficient value provided by financial econometric theory.

References

[Bollerslev, 1986]T. Bollerslev. Generalized autoregressive conditional heteroscedasticity. *Journal of Econometrics*, pages 307–327, 1986.

[Hamilton, 1994]J.D. Hamilton. *Time Series Analysis.* Princeton University Press, Chichester, 1994.

[Johnson and Wichern, 1992]R.A. Johnson and D.W. Wichern. *Applied Multivariate Statistical Analysis.* Prentice-Hall International, New York, 1992.

[Shumway and Stoffer, 2000]R.H. Shumway and D.S. Stoffer. *Time Series Analysis and Its Applications.* Springer-Verlag, New York, 2000.

[Taniguchi and Kakizawa, 2000]M. Taniguchi and Y. Kakizawa. *Asymptotic Theory of Statistical Inference for Time Series.* Springer-Verlag, New York, 2000.

[Tsay, 2002]R.S. Tsay. *Analysis of Financial Time Series.* Wiley-Interscience, 2002.

Time Series Analysis, Control Charts: An Industrial Application

Amor Messaoud, Claus Weihs, and Franz Hering

Fachbereich Statistik,
Universität Dortmund, Germany
(e-mail: messaoud@statistik.uni-dortmund.de)

Abstract. In this work, time series analysis and control charts are used to devise a real-time monitoring strategy in a BTA deep-hole-drilling process. BTA deep-hole-drilling is used to produce holes with high length to diameter ratio, good surface finish and straightness. The process is subject to dynamic disturbances usually classified as either chatter vibration or spiralling. In this work, we will focus on chatter which is dominated by single frequencies. The results showed that the proposed monitoring strategy can detect chatter and that some alarm signals are related to changing physical conditions of the process.

Keywords: Drilling process, Time series, Control charts.

1 Introduction

Deep hole drilling methods are used for producing holes with a high length-to-diameter ratio, good surface finish and straightness. For drilling holes with a diameter of 20 mm and above, the BTA (Boring and Trepanning Association) deep hole machining principle is usually employed. The working principle is shown in Figure 1. The process is subject to dynamic disturbances usually classified as either chatter vibration or spiralling. Chatter leads to excessive wear of the cutting edges of the tool and may

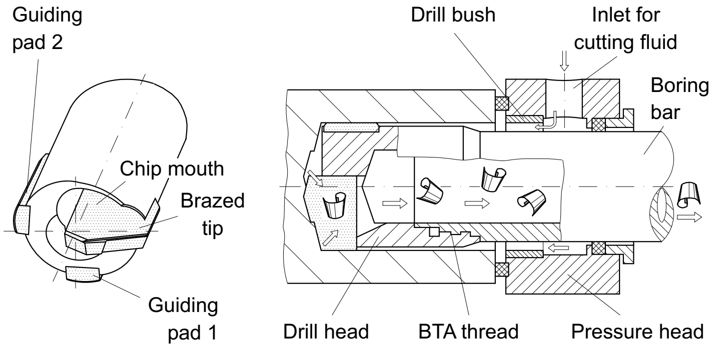


Fig. 1. BTA deep hole drilling, working principle

also damage the boring walls. Spiralling damages the workpiece severely. The defect of form and surface quality constitutes a significant impairment of the workpiece. As the deep hole drilling process is often used during the last production phases of expensive workpieces, process reliability is of primary importance and hence disturbances should be avoided. Therefore, it is necessary that a process monitoring system be devised to detect dynamic disturbances.

In this work, we will focus on chatter which is dominated by single frequencies, mostly related to the rotational eigenfrequencies of the boring bar. Therefore, we propose to monitor the amplitude of the relevant frequencies in order to detect chatter vibration as early as possible. Firstly, models that describe the process are reviewed in section 2. In section 3, the proposed monitoring strategy is discussed. Time series analysis is used in section 4 in order to identify the transition to chatter and to check basic assumptions of the application of control charts. Finally, the control charts are applied to real data in section 5.

2 Process models

[Weinert *et al.*, 2002] used the van der Pol equation to describe the transition from stable operation to chatter in one frequency

$$\frac{d^2M(t)}{dt^2} + h(t)(b^2 - M(t)^2)\frac{dM(t)}{dt} + w^2M(t) = W(t), \qquad (1)$$

where $t \in [0, \infty)$, $M(t)$ is the drilling torque, $b \in \mathbb{R}$, the frequency $w \in [200{,}2500]$, $h(t) : \mathbb{R} \to \mathbb{R}$ is an integrable function and $W(t)$ is a white noise process. [Theis, 2004] described the main features of the variation of the amplitudes of the relevant frequencies, using a logistic function. He showed that his approximation is directly connected to the proposed model. In fact, he considered $M(t)$ as a harmonic process

$$M(t) = R(t)cos(w + \phi),$$

where ϕ is the corresponding phase. He showed that

$$2\frac{dR(t)}{dt} + h(t)R(t)\left(b^2 - \frac{R(t)^2}{2}\right) = \frac{W(t)}{w}. \qquad (2)$$

is the amplitude-equation for the differential equation in (1) if there is only one frequency present in the process. From equation (2), the observed variation in amplitude of the relevant frequencies may be described by

$$R_t = (1 + a_t)R_{t-1} - a_t b_t R_{t-1}^3 + \varepsilon_t, \qquad (3)$$

where a_t and b_t are time varying parameters and ε_t is normally distributed with mean 0 and variance σ_ε^2.

3 Monitoring the residuals

For the monitoring procedure, the model given by equation (3) is approximated by its linear autoregressive part

$$R_t = (1 + a_t)R_{t-1} + \varepsilon_t,$$

and this AR(1) model is used to calculate the residuals. In fact, it is known that the nonlinear term $-a_t b_t R_{t-1}^3$ becomes important when there is chatter. The empirical evidence of this approximation is studied in section 4 using real data. The idea behind residual control charts is if the AR(1) model fits the data well, the residual will be approximately independent. Then, traditional control charts designed to monitor independent data can be applied to the residuals. Generally, residual control charts are designed for processes where stationarity in the steady state is assumed, which means that a unique model parameter for the whole process is used. For this reason a window of the T recent observations is used to estimate parameters a, β and σ_ϵ of the linear regression model

$$R_t = \beta + (1 + a)R_{t-1} + \epsilon_t, \tag{4}$$

where β is included because there is a general shift in the amplitudes after depth 35 mm due to a change in the physical conditions of the process, see section 4.2. The residuals are calculated using

$$e_t = R_t - (1 + \hat{a}_{t-1})R_{t-1} - \hat{\beta}_{t-1}, \tag{5}$$

where $\hat{a}_{t-1}$ and $\hat{\beta}_{t-1}$ are estimates of the regression parameters a and β at time $t-1$. The choice of $\hat{\beta}_{t-1}$ and $\hat{a}_{t-1}$ is motivated by the fact that using the estimated parameters at time t to calculate the residuals and to set the control limits may rather serve to mask changes than to detect them, see [Messaoud *et al.*, 2004b]. In this work, two control charts are considered: the residual Shewhart and a nonparametric EWMA based on standardized sequential ranks.

3.1 The residual Shewhart

The residual Shewhart control chart operates by plotting residuals e_t given by equation (5). It signals that the process is out of control at time t when e_t is outside UCL and LCL, given by

$$LCL = -k\hat{\sigma}_{\varepsilon,t-1} \text{ and } UCL = k\hat{\sigma}_{\varepsilon,t-1},$$

where $\hat{\sigma}_{\varepsilon,t-1}$ is the estimated standard deviation of the regression 4 at time $t-1$ and k is a constant. The choice of k is discussed later. Also, we used $\hat{\sigma}_{\varepsilon,t-1}$ to avoid the masking problem.

For the residual Shewhart charts, it is assumed that the residuals are normally distributed. Thus, the statistical properties of these charts are exact only if this assumption is satisfied. In practice, it is well known that this assumption rarely holds. Therefore, a distribution-free control chart, the EWMA based on sequential ranks, is used to monitor the process.

3.2 The EWMA chart based on sequential ranks

[Hackl and Ledolter, 1992] consider a nonparametric control chart procedure for individual observations that use the "standardized rank" of the observations among the recent group of T observations. For this chart, the sequential rank R_t^* is the rank of e_t among the most recent T ($T > 1$) observations e_t, e_{t-1}, …, e_{t-T+1}. That is,

$$R_t^* = 1 + \sum_{i=t-T+1}^{t} I(e_t > e_i),$$

where $I(.)$ is the indicator function. The standardized sequential rank $R_t^{(T)}$ is defined as

$$R_t^{(T)} = \frac{2}{T}\left(R_t^* - \frac{T+1}{2}\right).$$

The control statistic Q_t is the exponentially weighted moving averages (EWMA) of standardized ranks, computed as follow

$$Q_t = (1-\lambda)Q_{t-1} + \lambda R_t^{(T)},$$

where $Q_{t,1}$ is a starting value usually set equal to zero, and $0 < \lambda < 1$ is a smoothing parameter. The two sided EWMA chart signals that the process is out-of-control when Q_t is outside $-h$ and h defined to be equal $\pm H\sigma_Q$, where σ_Q and H are the standard deviation of Q_t and a constant, respectively. The choice of h is discussed later. For more details about this chart, see [Hackl and Ledolter, 1992] and [Messaoud *et al.*, 2004b].

4 Time series analysis of the residuals

[Messaoud *et al.*, 2004b] used the two control charts to monitor the variation in amplitudes of frequency 703 Hz, which is among the eigenfrequencies of the boring bar. The data, 1662 observations, are obtained in an experiment with feed $f = 0.185$ mm, cutting speed $v_c = 90$ m/min and amount of oil $\dot{V}_{oil} = 300$L/min. For more details, see [Weinert *et al.*, 2002]. In this experiment chatter is dominated by the frequency 703 Hz. Figure (2) shows the amplitude of frequency 703 Hz. The transition from stable operation to chatter occurs before depth 300 mm. Indeed, by eye inspection, the effect of chatter in this experiment is apparent on the bore hole wall after depth

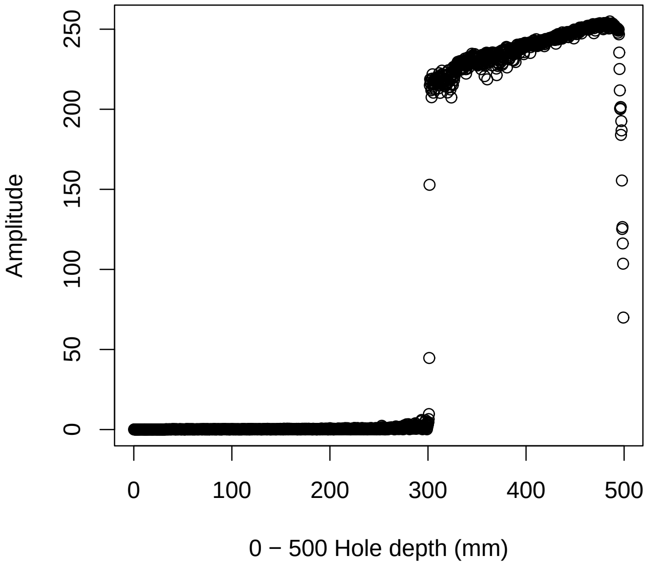


Fig. 2. Amplitude of frequency 703 Hz

300 mm. Therefore, only the first 1000 observations (depth $\leq$ 300 mm) are considered. Figure (3) shows the residuals calculated using equation (5). Note that the first 100 residuals are calculated using

$$e_t = R_t - (1 + \hat{a}_{100})R_{t-1} - \hat{\beta}_{100},$$

where $\hat{a}_{100}$ and $\hat{\beta}_{100}$ are estimates of the regression parameters a and β at time 100.

4.1 Transition from stable state to chatter

In order to investigate the ability of the different control charts to detect chatter, it is important to identify the transition from stable operation to chatter. For this reason, [Messaoud *et al.*, 2004b] studied the mean and variance of frequency 703 Hz. Moreover, the authors applied the Teräsvirta-Lin-Granger statistical test for nonlinear dependence in the residuals, see [Teräsvirta *et al.*, 1993]. As mentioned the nonlinear term $-a_t b_t R^3_{t-1}$ of model given by equation (3) becomes important when the process is unstable. The nonlinearity test is used for residual nonlinear structure, after linear structure has been removed by fitting the AR(1) model. The idea behind this

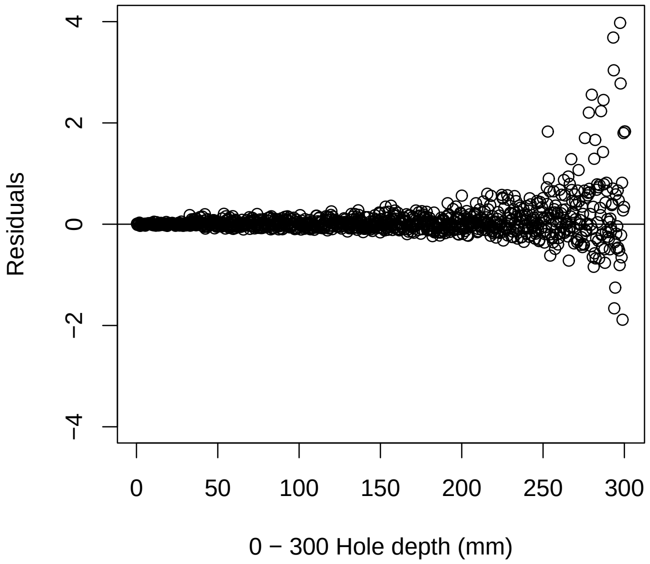


Fig. 3. Plot of the residuals

test is by fitting the linear AR(1) model to the data, the inherent nonlinearity structure has been swept into the residuals. The authors used different time windows of length 100 observations to test for neglected nonlinearity for the regression (3). The results confirms that the nonlinear term $-a_t b_t R_{t-1}^3$ is not important when the process is stable and showed that a change occurs in the process at depth 252.91 mm. This change may indicate the presence of chatter or that chatter will start in a few seconds.

4.2 Independence and normality assumptions of the residuals

[Messaoud *et al.*, 2004b] used the Ljung-Box test in order to check the independence assumption of the residuals. In fact, if the AR (1) model fits the data well, the residuals will be "approximately" independent. This is a basic assumption for the application of the two control charts. In fact, it is known that the performance of control charts is affected by the autocorrelation in the observations. In our process, the presence of autocorrelation in the residuals is destructive to the success of the proposed quality control process. Furthermore, the authors checked the normality assumption using the Shapiro-Wilks test. This assumption is very important only for the Shewhart chart, see section 3. The results shows that the residuals are independent. However, the hypothesis of normality is rejected.

5 Choice of the control charts parameters and results

Knowing that the transition to chatter occurs at depth 252.91 mm, only the first 900 observations (depth $\leq$ 270 mm) are considered for the application of the different control charts. For the reference sample, usually sample of 100-200 observations is used in SPC applications. In this work, the $T = 100$ recent observations $R_{t-T+1}, \ldots, R_t$ are used to estimate the parameters of the AR(1) model and to calculate the residuals. A larger sample cannot be used because the monitoring procedures should start before depth 35 mm (observation 120). In fact, chatter may be observed after that depth because the guiding pads of the BTA tool leave the starting bush, which will be discussed next.

5.1 Choice of the control charts parameters

Usually, the performance of control charts are evaluated by the average run length (ARL). The run length is defined as the number of observations that are needed to exceed the control limit for the first time. The ARL should be large when the process is statistically in-control (in-control ARL) and small when a shift has occurred (out-of-control ARL).

The parameters of the different control charts are selected so that all control charts have the same in-control ARL equal to 370. This choice should avoid many false alarm signals because all control charts are applied to 900 observations. A value $k = 2.95$ is used for the residual Shewhart control charts. For the EWMA chart, we used $\lambda = 0.1$, 0.3 and 0.5. The corresponding values for h are respectively 0.349, 0.629 and 0.786.

5.2 Results

Table 1 shows the out of control signals for depth $\leq$ 270 mm . Table 1 shows that all control charts (except the EWMA charts with λ=0.1 and 0.3) signal at 32 $\leq$ depth $\leq$ 35 mm. As mentioned before the guiding pads leave the starting bush approximately at depth 32 mm, which induce an increase in the process mean and variance for the amplitude of the frequency 703 Hz. This increase explains that all control charts have picked up these changes very quickly. All control charts (except the EWMA charts with λ=0.1 and 0.3) signal at 110 $\leq$ depth $\leq$ 125 mm . It is known that depth 110 mm is approximately the position where the tool enters the bore hole completely. Theis (2004) noted that this might lead to changes in the dynamic process because the boring bar is slightly thinner than the tool and therefore the pressures in the hole may change. The important out of control signals are produced at 250 $\leq$ depth $\leq$ 255 mm. As discussed, it is showed that the transition from stable operation to chatter have occurred at depth 252.91 mm. Therefore, in this experiment chatter may be avoided if corrective actions are taken after this signal.

Table 1. Out of control signals of the different control charts applied to the amplitude of frequency 703 Hz using window length T=100 (depth $\leq$270 mm)

Hole depth (mm)	Observation number	Residual Shewhart	EWMA		
			$\lambda = 0.1$	$\lambda = 0.3$	$\lambda = 0.5$
≤32	≤107	0	0	0	0
32-35	108-117	2	0	0	1
35-45	118-150	7	14	4	2
45-70	151-249	1	0	1	1
70-110	250-366	1	0	0	0
110-125	370-416	1	0	0	1
125-200	417-665	4	8	3	2
200-250	666-832	5	1	0	0
250-255	833-849	2	1	3	2
255-260	850-865	0	0	0	0
260-270	866-898	1	0	0	0
Total		24	24	10	9

Note: The shaded lines refer to the the three physical conditions of the process (i.e., guiding pads leave the starting bush, the tool is completely in the hole and transition from stable operation to chatter)

In this experiment, the EWMA control chart with λ=0.5 is the best, and should be chosen among the three EWMA charts considered in this work. Indeed, only 9 out of control signals are produced and all changes of the physical conditions of the process are detected. In practice, a procedure to choose the smoothing parameter λ is required. As noted in section 5.3, the Residual Shewhart control chart produces more signals than the EWMA control chart with λ=0.5. This may be due to its sensitivity to the normality assumption.

5.3 Multivariate monitoring

In this work, the results showed that chatter can be detected only by monitoring the variation in amplitudes of frequency 703 Hz. This conclusion is expected because this frequency is the relevant frequency in this experiment. However, in practice there are more relevant frequencies and chatter may be observed at the beginning of the drilling process immediately after the guiding pads have left the starting bush, with high and low frequencies, see [Weinert *et al.*, 2002]. Thus, an SPC procedure that monitors all the relevant frequencies is necessary. [Messaoud *et al.*, 2004a] used a multivariate distribution-free EWMA control chart to monitor the drilling process. This chart is based on sequential rank of data depth measures. The results showed that it can detect chatter vibrations.

6 Conclusion

This work showed that using time series analysis and control charts, a reliable on-line monitoring system in the BTA process is proposed. The results showed that the proposed monitoring strategy detect chatter and that some out-of-control signals are related to physical conditions of the process (i. e. guiding pads leave the starting bush, the tool is completely in the hole). Therefore, real-time implementation of this monitoring strategy can be guaranteed.

Acknowledgements

This work has been supported by the Graduate School of Production Engineering and Logistics at the University of Dortmund and the Collaborative Research Centre "Reduction of Complexity in Multivariate Data Structures" (SFB 475) of the German Research Foundation (DFG).

References

[Hackl and Ledolter, 1992]P. Hackl and J. Ledolter. A new nonparametric quality control technique. *Communications in Statistics-Simulation and Computation*, pages 423–443, 1992.

[Messaoud *et al.*, 2004a]A. Messaoud, W. Theis, C. Weihs, and Hering. F. Application and use of multivariate control charts in a bta deep hole drilling process. *to appear in the proceedings of the GFKL 2004*, 2004.

[Messaoud *et al.*, 2004b]A. Messaoud, W. Theis, C. Weihs, and Hering. F. Monitoring the bta deep hole drilling process using residual control charts. *Technical Report 60/2004 of SFB 475, University of Dortmund*, 2004.

[Teräsvirta *et al.*, 1993]T. Teräsvirta, C-F. Lin, and C. Granger. Power of the neural network linearity test. *Journal of Time Series Analysis*, pages 209–220, 1993.

[Theis, 2004]W. Theis. Modelling varying amplitudes. *PhD dissertation, Department of Statistics, University of Dortmund*, 2004.

[Weinert *et al.*, 2002]K. Weinert, O. Webber, M. Hüsken, J. Mehnen, and W. Theis. Analysis and prediction of dynamic disturbances of the bta deep hole drilling process. In *Proceedings of the 3^{rd} CIRP International Seminar on Intelligent Computation in Manufacturing Engineering*, 2002.

Part XVI

Fuzzy Approach

Chaotic Aspects of a GRM1 Innovation Diffusion Model

Christos H. Skiadas[1], Giannis Rompogiannakis[1], Apostolos Apostolou[2], and John Dimotikalis[2]

[1] Technical University of Crete
Department of Production Engineering and Management,
Data Analysis and Forecasting Laboratory,
73100 Chania, Crete, Greece
(e-mail: skiadas@ermes.tuc.gr)
[2] Technological Educational Institute of Crete
Heraclion, Crete, Greece

Abstract. Chaotic behavior of a generalized rational (GRM1) innovation diffusion model is studied. The deterministic continuous version of this model was proposed, analyzed and applied in earlier publications. Here, the chaotic behavior is expressed through the discrete alternative of the continuous GRM1 model. The model shows symmetric and non-symmetric behavior expressed by a parameter σ. In this article it is found that when the diffusion parameter b and the parameter σ verify the relation $b/\sigma \geq 2$ then the chaotic aspects of the model appear. A method is proposed for fitting the model to the data. Time series data expressing the cumulative percentage of steel produced by the oxygen process in various countries are used. Characteristic graphs of the chaotic behavior are given and applications are presented.

Keywords: Chaotic modeling, Diffusion modeling, Speed of diffusion, Innovation diffusion, Non-linear models, Chaotic oscillations.

1 Introduction

It's become a commonplace to call this the information age, but an even more appropriate name might be the information age. In 1997, for example, the U.S. Patent and Trademark Office received 237.000 patent applications, a 15% increase from the year before. Also in 1997, the agency granted 124.127 patents, a record number and an increase of 16% from the volume it recorded at the beginning of the decade in 1991, a year that had also set a record for patent activity. At individual companies, the pace of innovation is even greater. In 1998, IBM Corp. received 2.657 patents for inventions, an increase of 54% from the number it won in 1997, according to a preliminary tally from the patent office. This was not a one-time surge, as IBM has been the leading recipient of U.S. patents for six consecutive years. And IBM was not alone in recording huge increases in U.S. patent activity last year: Sony Corp.'s patent number rose 53%, Eastman Kodak Co.'s 41%, and Motorola Inc.'s 33%, [Maguire and Hagen, 2001]. While not all patents translate into new products or new production methods, these figures clearly demonstrate a tendency, and this explosion of innovation activity presents significant challenges. One of the special challenges firms face in this decade is the challenge of designing, manufacturing, and distributing products in a global marketplace. If customers want new products, and they do, then companies have no choice but to gear up their processes to provide innovative features and the latest designs. This mean that companies must have a proper way to describe the competitive dynamics in a market, [Modis, 1997] and to predict how these new products or production methods will move in the marketplace. One of these methods is described in this paper. A model is proposed and some empirical data are explored. In earlier publications several innovation diffusion models where presented, analyzed and applied to real life data [Bass, 1969], [Mahajan and Schoeman, 1977], [Sharif and Kabir, 1976], [Skiadas, 1985], [Skiadas, 1986], [Skiadas, 1987], [Modis and Debecker, 1992]. A main direction of these applications was focused of the non-symmetric behavior of the models expressed by specific parameters. A relatively simple but very flexible model was proposed in an earlier publication based on a family of Generalized Rational Models, [Skiadas, 1985], [Skiadas, 1986], to express asymmetry during the innovation diffusion process. This model is expressed by the following differential equation:

$$\dot{f} = b\frac{f(F-f)}{F-(1-\sigma)f} \tag{1}$$

Where f is the number of adopters at time t, F is the total number of potential adopters, b is the diffusion parameter, and σ is a dimensionless parameter. This model has a point of infection varying from 0 to F when parameter σ decreases from ∞ to 0. Another interesting property of parameter σ is that it gives a measure of the asymmetry of the model. Perfect symmetry appear for $\sigma = 1$ when equation 1 reduces to equation 2 expressing the popular logistic model:

$$\dot{f} = bf\left(1-\frac{f}{F}\right) \tag{2}$$

For the last model it is easy to show that, by using the transformation:

$$\dot{f} = \frac{df}{dt} \approx \frac{\Delta f}{\Delta t} = \frac{f_{t+1}-f_t}{(t+1)-t} = f_{t+1}-f_t \tag{3}$$

it is expressed by the following difference equation:

$$f_{t+1} = f_t + bf_t\left(1-\frac{f_t}{F}\right) \tag{4}$$

Bifurcation and further chaotic behavior appear when $2 < b \leq 3$. Various applications of the logistic model in several disciplines showed that parameter b of the logistic model lies in very low limits lower that unity. Thus by using the logistic model it is not possible to express chaotic behavior in real situations as the estimated values of parameter b fail to reach the limit at which chaotic behavior appear. On the other hand provided data for various cases show that oscillations and chaotic behavior appear quite frequently and especially when the diffusion process is close to the upper limit F. Moreover when the logistic model is applied in the form:

$$X_{t+1} = bX_t(1 - X_t) \tag{5}$$

Where $X_t = f_t/F$ then, bifurcation and chaotic behavior appears when $3 < b \leq 4$.

Chaos appears for very high values of b, which are not reasonable for real situations. Clearly the model form (4) is more correct as a discrete logistic model expressing behavior similar to the continuous model resulting from differential equation 2. This can be found in an older application done by [Nash, 1976]. The aim of this paper is to show that the model (1) exhibits chaotic behavior for values of parameter b that are quite low and are in accordance to the values estimated in real situations. This is achieved by the help of the flexible parameter σ, which gives a measure of the asymmetry of the model. The chaotic behavior of the model is analyzed and illustrated by using significant graphs. Finally, real life applications are presented.

2 The Generalized Rational Model

The model proposed is a discrete version of the continuous one expressed by equation 1. By introducing the approximation of $\dot{f}$ from equation 3 in the differential equation 1 the following difference equation results:

$$f_{t+1} = f_t + b\frac{f_t(F - f_t)}{F - (1 - \sigma)f_t} \tag{6}$$

Some interesting properties of this model are illustrated in Figures 1to 3.

In Figure 1a the proposed model shows the classical sigmoid form, whereas in Figure 1b the bifurcation appear as a simple oscillation. In Figure 1c a more complicated oscillation with four distinct oscillating levels appears, whereas in Figure 1d - 1f a total chaotic form appears. In all cases presented here the starting value is $f_0 = 1$, the upper limit $F = 100$, $b = 0.3$ and σ takes various values. The value selected for b is within the range 0.1 to 0.5, which is valid in real situations. By varying the dimensionless parameter σ several forms of the model appear.

A very important point is the estimation of the values of parameters b and σ for which bifurcation appear. The presence of the first oscillations and the onset to chaos, which follows, is a very important point when studying innovation diffusion systems. According to the theory of chaotic models, bifurcation for the model (6) starts when:

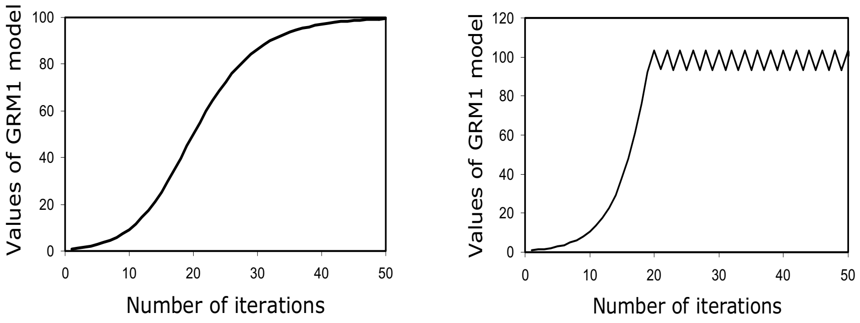


Fig. 1. GRM1 model for a) $b = 0.3$ and $\sigma = 2$ and b) $b = 0.3$ and $\sigma = 0.13$

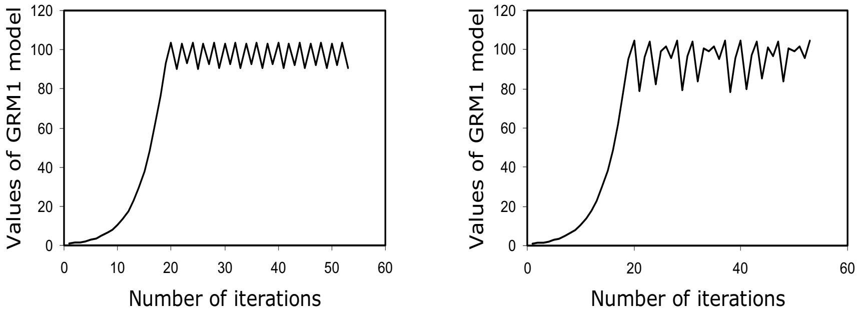


Fig. 2. GRM1 model for a) $b = 0.3$ and $\sigma = 0.12$ and b) $b = 0.3$ and $\sigma = 0.10$

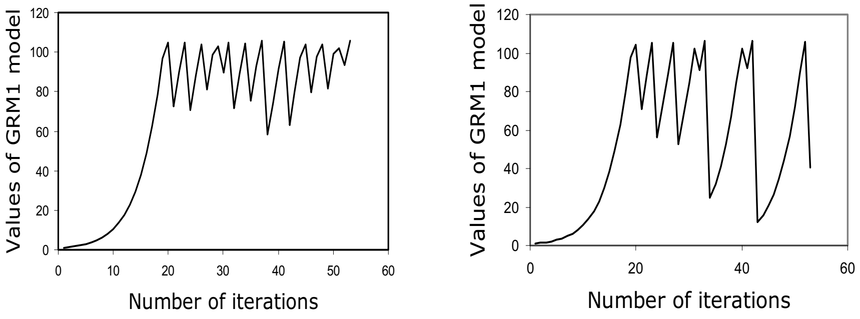


Fig. 3. GRM1 model for a) $b = 0.3$ and $\sigma = 0.09$ and b) $b = 0.3$ and $\sigma = 0.08$

$$f'_{t+1} = -1, \quad f_{t+1} = f_t \tag{7}$$

By applying equations 7 to equation 6 results the following relation for parameters b and σ:

$$\frac{b}{\sigma} = 2 \tag{8}$$

When $b/\sigma > 2$ then oscillation and chaotic behavior appear by gradually augmenting the fraction b/σ . When $\sigma = 1$ which is the case for the logistic model bifurcation appear for values of $b > 2$.

It is also possible to obtain analytic form for the values of f_t after the first bifurcation point and before the second. To achieve this we consider that $f_{t+2} = f_t$. The exact formula is given by:

$$f_t = F\frac{(b+2) \pm \sqrt{\frac{b(b+2)(b-2\sigma)}{(b-2\sigma+2)}}}{2(b+\sigma-1)} \tag{9}$$

For the logistic model $\sigma = 1$ and thus equation 9 reduces to:

$$f_t = F\frac{(b+2) \pm \sqrt{(b+2)(b-2)}}{2b} \tag{10}$$

When $b > 2\sigma$ in equation 9 or $b > 2$ in equation 10 the system oscillates at the values of f_t given by the above formulas respectively. When b is higher of the values expressing the second bifurcation point four distinct oscillating levels appear and later eight and finally 2^n points. For sufficient specifically high values of b, n is very high and the system exhibits chaotic oscillations.

3 Parameters' Estimation of GRM1 Model

The parameters of the discrete GRM1 model are estimated by an Iterative non-linear regression analysis algorithm by minimizing the sum of squared errors ($S = SSE$):

$$S = \sum \epsilon_t^2 = \sum_{t=1}^{n} (y_t - f_t)^2 \tag{11}$$

where ϵ_t is the error term of the stochastic equation:

$$y_t = f_t + \sum_{i=1}^{n} \frac{\vartheta f_t}{\vartheta a_i} \Delta a_i + \epsilon_t \tag{12}$$

y_t denotes provided data and f_t is calculated for every t from equation 6, given a set of initial values of parameters a_i. The estimation of parameters is highly sensitive in the presence of oscillations and chaotic oscillations in the provided data. For a better fitting it was decided to use the non-linear estimation method proposed by Nash for the discrete Logistic model for only three parameters of the model and retaining the dimensionless parameter σ. This parameter is gradualy changed as the iterative procedure proceeds until the sum of squared errors is minimized. The starting values of the partial derivatives need the estimation of the following forms given a set of initial values for the parameters of the model:

$$\frac{\vartheta f_1}{\vartheta b} = \frac{f_0(F-f_0)}{F-(1-\sigma)f_0} \tag{13}$$

$$\frac{\vartheta f_1}{\vartheta f_0} = 1 + b\frac{F^2 - 2Ff_0 + (1-\sigma)f_0^2}{\left(F-(1-\sigma)f_0\right)^2} \tag{14}$$

$$\frac{\vartheta f_1}{\vartheta F} = b\sigma\left(\frac{f_0}{F}\right)^2 \tag{15}$$

After the above estimation of the initial values of the parial derivatives the iterative procedure continues the estimation by using the following formulae:

$$\frac{\vartheta f_{t+1}}{\vartheta b} = \frac{\vartheta f_t}{\vartheta b}(1+bk_t) + \frac{f_t(F-f_t)}{F-(1-\sigma)f_t} \tag{16}$$

$$\frac{\vartheta f_{t+1}}{\vartheta f_0} = \frac{\vartheta f_t}{\vartheta f_0}(1+bk_t) \tag{17}$$

$$\frac{\vartheta f_{t+1}}{\vartheta F} = \frac{\vartheta f_t}{\vartheta F}(1+bk_t) + \frac{b\sigma f_t^2}{\left(F-(1-\sigma)f_t\right)^2} \tag{18}$$

where:

$$k_t = \left(F^2 - 2Ff_t + \frac{(1-\sigma)f_t}{F-(1-\sigma)f_t}\right)^2 \tag{19}$$

4 Illustrations

Time series data expressing the cumulative percentage of steel produced by the oxygen process in various countries are used from an earlier application, [Poznanski, 1983]. Figure 4 illustrates the diffusion of Oxygen steel technology in Spain from 1968 to 1980, for a number of 13 years. The actial data include 18 years but, it is more appropriate to study the last part of the time series data as this part shows the characteristic oscillations that are of special interest in this study. The small cycles indicate the actual data, the dotted line chracterizes the path of the logistic model and the simple line is for the GRM1 model.

Parameter estimates and the sum of squared errors are summarized in Table 1. The parameter b for the Logistic model is relatively high but is far

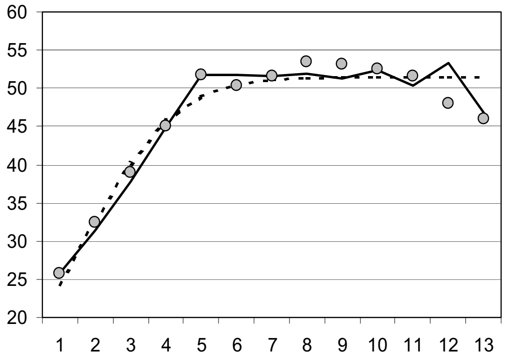


Fig. 4. Spain, Oxygen Steel Process (1968-1980)

Model	b	l	F	$\sigma(b/\sigma)$	SSE
Logistic	0.6309	24.373	51.474	-	72.838
GRM1	0.2331	25.779	51.736	0.084 (2.775)	41.748

Table 1. Parameter Estimates and Sum of Squared Errors (SSE) for Logistic and GRM1 Models in Spain from 1968 to 1980

away from the value needed for the start of bifurcation ($b = 2$). The form of the logistic path presented in the Figure 4 has a smouth form. The model fail to express the oscillating behavior of the actual case studied. Instead the GRM1 model shows a value for the parameter b lower to that of the Logistic model but the extra parameter σ accounts for the presence of oscillating and further of chaotic behavior as the fraction $b/\sigma = 2.775 > 2$. The estimated values for the parameters l and F are very close for both models. The ability of GRM1 model to follow the oscillating behavior of actual data is illustrated in the above Figure and is also expressed by the strong improvement of the Sum of Squared Errors (SSE).

Figure 5 illustrates the diffusion of oxygen steel technology in Italy from 1970 to 1980. The process ends in an oscillating form. The discrete Logistic fails to express these oscillations whereas the discrete GRM1 shows a considerable flexibility to approximate the real data. The sum of the squared errors is very low in the case of GRM1 model compared to that of the Logistic as is demonstrated in Table 2. The fraction $b/\sigma = 3.5292$ for the GRM1 model accounts for the chaotic behavior.

The actual data for the diffusion of the oxygen steel process in Luxemburg are of considerable interest as they cover the scale from 1.5 % during 1962 to that of 100 % in 1980 (Figure 6). The GRM1 model showed a good flexibility as it covers the fast growth process in the first stages of the diffusion process followed by a sudden turn to the high platform of 100%. The small also

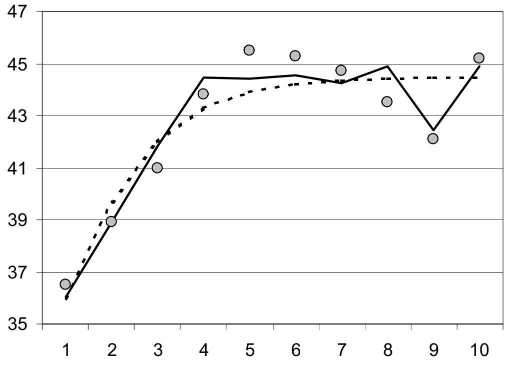


Fig. 5. Italy Oxygen Steel Process (1970-1980)

Model	b	l	F	$\sigma(b/\sigma)$	SSE
Logistic	0.5447	35.957	44.473	-	15.330
GRM1	0.08823	36.0402	44.4614	0.025 (3.5292)	7.431

Table 2. Parameter Estimates and Sum of Squared Errors (SSE) for Logistic and GRM1 Models in Italy from 1970 to 1980

flictuations at the end of the process are also simulated quite well as the fraction $b/\sigma = 3.609$ accounts for the chaotic region of the model. Figure 6 illustrates the case of Luxemburg for the following estimated values for the parameters: $b = 0.1931$, $l = 7.968$, $F = 99.669$ and $\sigma = 0.0535$. The mean squared error is $MSE = 20.872$.

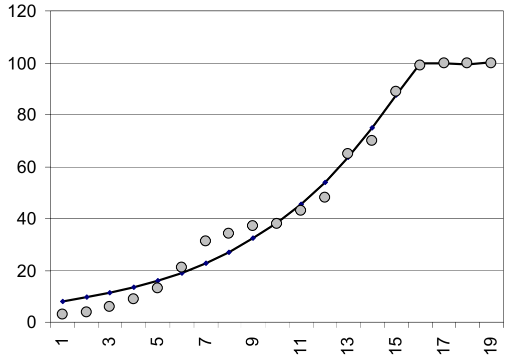


Fig. 6. Luxemburg Oxygen Steel Process (1962-1980)

The flexibility and the ability of GRM1 model to simulate growth processes that show at the end of the process oscillations and also chaotic oscillations is demonstrated in the following case of the diffusion of oxygen steel technology in Bulgaria from 1968 to 1978 (see Figure 7). The estimated parameters have values $b = 0.04046$, $l = 49.2425$, $F = 58.412$ and $\sigma = 0.012$. The sum of squared errors is $SSE = 21.431$ and the fraction $b/\sigma = 3.3718$ indicates that the model behave in the chaotic region.

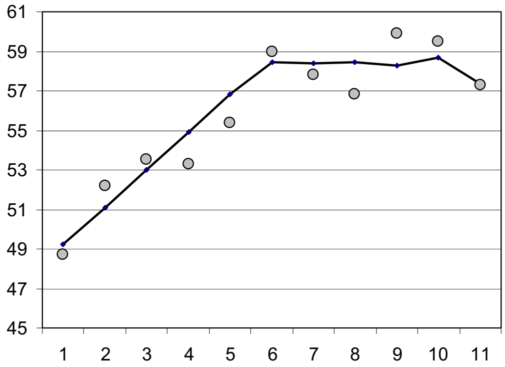


Fig. 7. Bulgaria Oxygen Steel Process (1968-1978)

5 Summary and Conclusions

A nonsymmetric innovation diffusion model is presented and analyzed regarding the chaotic behavior. It is shown that this model exhibits bifurcation and further chaotic behavior for some values of the fraction b/σ of the parameters b and σ. Real time-series data are used and parameters are estimated by an Iterative non-linear algorithm showed that in some cases the model performs oscillations (the fraction b/σ has values higher than 2) whereas in other cases the model showed the classical sigmoid form.

References

[Bass, 1969]F. Bass. A new product growth model for consumer durables. *Management Science*, 15(217–231), 1969.

[Maguire and Hagen, 2001]M. Maguire and M. Hagen. Explosion of new products creates cholleuges. *Quality Progress*, 32(5):29–35, 2001.

[Mahajan and Schoeman, 1977]V. Mahajan and M.E.F. Schoeman. Generalized model for the time pattern of the diffusion process. *IEEE Transactionson Engineering Management*, 24:12–18, 1977.

[Modis and Debecker, 1992]T. Modis and A. Debecker. Chaoslike states can be expected before and after logistic growth. *Technological and Forecasting Social Change*, 41:111–120, 1992.

[Modis, 1997]T. Modis. Genetic re-engineering of corporations. *Technological Forecasting and Social Change*, 56:107–118, 1997.

[Nash, 1976]J.C. Nash. A discrete alternative to the logistic growth function. *Applied Statistics I*, 22:9–14, 1976.

[Poznanski, 1983]K. Z. Poznanski. International diffusion of steel technologies. timelag and the speed of diffusion. *Technol. Forecast. Social Change*, 23:305–323, 1983.

[Sharif and Kabir, 1976]M. N. Sharif and C. Kabir. A generalized model for forecasting technological substitution. *Technol. Forecast. Social Change*, 8:353–364, 1976.

[Skiadas, 1985]C. H. Skiadas. Two generalized models for forecasting innovation diffusion. *Technol. Forecast. Social Change*, 27:39–61, 1985.

[Skiadas, 1986]C. H. Skiadas. Innovation diffusion models expressing asymmetry and/or positively or negatively influencing forces. *Technol. Forecast. Social Change*, 30:313–330, 1986.

[Skiadas, 1987]C. H. Skiadas. Two simple models for early and middle stage prediction of innovation diffusion. *IEEE Trans. Eng. Manag.*, 34:79–84, 1987.

A New Modeling Approach Investigating the Diffusion Speed of Mobile Telecommunication Services in EU-15

Apostolos N. Giovanis[1] and Christos H. Skiadas[1]

Technical University of Crete
Department of Production Engineering and Management,
Data Analysis and Forecasting Laboratory,
73100 Chania, Crete, Greece
(e-mail: skiadas@ermes.tuc.gr)

Abstract. The objective of this paper is to investigate the impact of the time-delay effect on the diffusion of mobile telecommunication services in EU. It has been proved from several studies that the time-delay between the awareness and the adoption phase of mobile services-potential users determines the speed of the mobile telecommunication service diffusion and can be used effectively for ranking or cluster purposes in cases when the diffusion of a new product in different countries is studied. The proposed modeling approach originates from the well-known logistic model where it is assumed that the ordinary contagion process does not take place instantly but after some certain amount of time. A proper modification of the proposed model described by a time lag ordinary differential equation can be solved analytically and its properties for several parameters' combination are investigated. Moreover, a new diffusion speed index is proposed and the correlation between the time-delay index and the proposed diffusion speed index is examined. Finally the model is applied to real data concerning the mobile services diffusion in fifteen counties of EU from 1990 to 2002. Based on the estimated parameters of the model produced for each country a ranking and a clustering of the EU countries based on their derived diffusion speed and time-delay indexes are provided.

Keywords: Time-delay model, Diffusion speed, Innovation diffusion modeling, Technology marketing-management, Modeling Telecommunication services.

1 Introduction

Today, forecasting technology in economic activity is no more avoidable than in forecasting weather in daily life. In fact, voluminous literature has explored different growth-curve models in forecasting the diffusion process of new technologies. Early contributions to this subject are attributed to scientists who noted the analogy between the epidemic process and the social adoption process [Griliches, 1957], [Mansfield, 1961], [Bass, 1969], [Fisher and Pry, 1971], [Blackman, 1972], [Sharif and Kabir, 1976], [Sharif and Ramanathan, 1984]. They came to a general agreement that the proportion of adopters rises at an accelerating rate during the early stages of the diffusion process and then at a declining rate until the population of potential adopters has been exhausted. Later, economists and technologists joined the field to predict the marketability of new products, trajectory of technology process, penetration rate of the advanced manufacturing technologies ([Skiadas, 1985], [Skiadas, 1986], [Skiadas, 1987], [Kumar and Kumar, 1992] and [Mead and Islam, 1998]). Initiated with a simple logistic function, various curves have been empirically derived to investigate the patterns of technological growth process. These curves differ from one another in terms of number of parameters, the point of inflection, the symmetric or non-symmetric shape of their shape, etc.

This paper is using an earlier developed approach ([Poznanski, 1983], [Skiadas, 1986]) originating from the logistic innovation diffusion model which incorporates the time-delay between the awareness and the adoption phase during the classical contagion process between the adopters and the potential adopters of a new technology. It has been proven that the time-delay affects the performance of a new technology launching and speed and it can be used for comparison purposes, in order to study the innovation diffusion among groups of potential adopters with different characteristics. The proposed model can be solved analytically and presents very attractive properties well documented in the field of innovation diffusion representation. Additionally, an expression on the relationship between the time-delay parameter and the diffusion speed is presented. The time-delay innovation diffusion model is applied to the data of mobile telephony services diffusion in EU-15, in order to determine the existence of penetration patterns in relation to the time delay between the awareness and the adoption phase of the potential adopters. Finally, the outcomes are used for a ranking of the investigated countries.

2 A model expressing the time-delay of adoption-diffusion process

The diffusion of an innovation in a stable and homogeneous system with no external influence is traditionally expected to follow a symmetric S-shaped pattern represented by the well known logistic curve [Griliches, 1957]. More specifically, let X_t denote the number of agents that have adopted the new technology in time t. Let X^* denote the total number of potential adopters. Then the following o.d.e expresses the dynamics of the innovation diffusion process through the contagion process between the adopters and the potential adopters:

$$\frac{dX_t}{dt} = \frac{b}{X^*} \cdot X_t \cdot (X^* - X_t) \tag{1}$$

which implies that b represents the growth rate of the numbers of adopters relative to the proportion of agents who have not yet adopted the innovation. The innovation's penetration level follows an S-shaped pattern with maximum diffusion speed reached when half of the total number of potential adopters has adopted the new technology.

This traditional approach in defining the innovation diffusion process assumes that the process takes place in a stable and homogeneous system in which the innovation spreads without any affection of the system's structure. In such cases the diffusion follows a symmetric pattern similar to those provided by eq. (1). The symmetry is also retained in the presence of external influences (e.g. promotional activities) which are not acting directly to the system's structure. However, many studies have proven that the presence of symmetry is not the general rule in innovation diffusion process ([Mahajan *et al.*, 1961], [Skiadas, 1985], [Skiadas, 1986], [Skiadas, 1987]). In the majority of new technology penetration patterns the asymmetry is caused by several factors such as cultural status, economic conditions, demographics (population density, urbanization, and educational level), governmental policy, technology utility, technology familiarity, etc. [Bakalis *et al.*, 1997]. The incorporation of such a critical aspect of the diffusion process into the process representation efforts not only provides more flexible models but can also lead to the revelation of several interesting properties of the innovation diffusion process.

Equation (1) assumes an immediate interaction between the adopters and the potential adopters of a new product leading to a symmetric diffusion pattern. However, this assumption is not always true since there is always a time-delay between the time of interaction occurrence and the adoption time. Thus, the potential adopters $X^* - X_t$ at time t interact with the adopters X_{t-T} at time $t - T$ Taking into account the above consideration, the original logistic model takes the following form:

$$\frac{dX_t}{dt} = \frac{b}{X^*} \cdot X_{t-T} \cdot (X^* - X_t) \tag{2}$$

where T is the mean value of all time-delays occurring between the adopters and the potential adopters of the technology under investigation. Equation (2) cannot be easily handled and therefore an appropriate transformation is needed in order to have an approximate solution. By applying the Taylor series expansion to the expression X_{t-T} we have:

$$X_{t-T} = X_t - T \cdot \frac{dX_t}{dt} + \frac{T^2}{2} \cdot \frac{d^2X_t}{dt^2} - \frac{T^3}{3!} \cdot \frac{d^3X_t}{dt^3} + \cdots \tag{3}$$

Provided that the parameter T is not to large compared to the total time interval, the two first terms of the whs of equation (3) could be retained. Then the equation (3) can be written as:

$$X_{t-T} = X_t - T \cdot \frac{dX_t}{dt} \tag{4}$$

Introducing equation (4) into equation (2) the following delay ordinary differential equation (ODE) results:

$$\frac{dX_t}{dt} = \frac{b}{X^*} \cdot \left[X_t - T \cdot \frac{dX_t}{dt} \right] \cdot (X^* - X_t) \tag{5}$$

The appropriate rearrangements in equation (5) yields:

$$\frac{dX_t}{dt} = \frac{b}{1 + b \cdot T} \cdot \frac{X_t \cdot (X^* - X_t)}{X^* - \frac{b \cdot T}{1 + b \cdot T} \cdot X_t} \tag{6}$$

Setting

$$b^* = \frac{b}{1 + b \cdot T} \tag{7}$$

and then

$$b^* \cdot T = 1 - \sigma \tag{8}$$

equation (6) takes the form:

$$\frac{dX_t}{dt} = b^* \frac{X_t \cdot (X^* - X_t)}{X^* - (1 - \sigma) \cdot X_t} \tag{9}$$

Equation (9), is a special case of a family of generalized innovation diffusion models proposed by [Skiadas, 1985], [Skiadas, 1986] aiming to represent the innovation diffusion process. When $\sigma = 1$ then equation (9) results in the above described logistic model, whereas when $\sigma = 0$ it results in the exponential model. The solution of ODE (9) has given by Skiadas (1985) and has the following form:

$$\ln(X_t) - \sigma \cdot \ln(X^* - X_t) = \ln(X_0) - \sigma \cdot \ln(X^* - X_0) + b \cdot t \tag{10}$$

where X_0 represents the numbers of adopters at time $t = 0$. The inflection point of the above model is given by [Skiadas, 1985] and has the following form

$$X_{inf} = X^* \cdot \frac{1 - \sqrt{\sigma}}{1 - \sigma} \tag{11}$$

The inflection point is considered as measure of asymmetry in every technology diffusion case. Equation (11) reveals that the proposed model is very flexible since the inflection point takes values from 0 to X^* depending on the values of parameter σ. When $\sigma = 1$, then $X_{inf} = X^*/2$ which is the inflection point of the logistic model.

3 Pattern identification in mobile telephony diffusion in EU-15

3.1 Model Identification Results

Mobile Telecommunications has recently developed into a popular innovation of diffusion studies field. In fact, researchers have conducted studies on a national level ([Wright *et al.*, 1997], [Frank, 2003]), a multinational level ([Gruber and Verboven, 2001], [Gruber, 2001])and on a worldwide level [Dekimpe and Sarvary, 1996]. These, multinational or cross-country studies examine the reasons and dynamics behind the differences in the adoption or diffusion processes of a set of countries. The present approach is trying to identify the existence of standardized patterns in mobile telephony diffusion in EU-15 due to the different time-delay effects between adopters and potential adopters during the contagion process.

The available data express the penetration level of mobile telephony in EU-15 from 1990 until 1992 and has been taken from OECD communication outlook (2000, 2001, and 2002). The proposed model is applied to the available data by using an appropriate non-linear regression algorithm [Skiadas, 1987]. The results for the fifteen countries under investigation are summarized in Table 1.

As it can be seen, the model identification performance is very good since it explains for every country more than 99% of the process variance. The parameter σ is statistically significant for every country showing that the assertion of the existence of time-delay between the awareness and adoption phases is true. Based on the outcomes, the time-delay varies from 0.33 yrs to 1.79 yrs. Figure 1 illustrates the time-delay parameters for each country under investigation. Among the countries with the smaller time-delay parameter are Portugal, France and Greece, while the countries with the bigger time-delay parameter are UK, Luxembourg, Germany, and Denmark, Sweden. It is obvious that a catching-up process is present in the diffusion of mobile telecommunications (Gruber and Verboven, 2001), since the countries with high technology level or countries which belong to the originators of the mobile technology present a bigger time-delay parameter than other countries which develop the industry later on. Finally, three countries present almost symmetric diffusion pattern (inflection point around 50% of the saturation level) while all the others not.

3.2 Time-Delay Effect and Speed of Diffusion

It is interesting to examine the relationship between the time-delay effect and the speed of the diffusion process. A frequently utilized measure for the speed is the reciprocal of characteristic duration, a measure expressing the time required to grow from 10% to 90% of the estimated saturation level. Solving Eq. (10) w.r.t. $1/t$ yields:

Country	X_0	b^*	X^*	σ	$V(e_t)$	MSE	Var Expl.	T	X_{inf}
Austria	0,059 (0,025)	0,701 (0,050)	82,459 (1,026)	0,230 (0,073)	1,648	1,141	99,89%	1,10	56%
Belgium	0,067 (0,007)	0,620 (0,0010)	78,790 (0,354)	0,146 (0,015)	0,110	0,076	99,99%	1,38	57%
Denmark	1,878 (0,451)	0,357 (0,041)	95,520 (15,076)	0,416 (0,340)	2,987	2,068	99,72%	1,64	58%
Finland	1,674 (0,390)	0,447 (0,046)	88,994 (4,198)	0,655 (0,253)	2,267	1,569	99,82%	0,77	49%
France	0,016 (0,06)	0,819 (0,049)	67,193 (1,246)	0,699 (0,135)	0,315	0,218	99,96%	0,37	37%
Germany	0,048 (0,023)	0,652 (0,100)	69,997 (1,404)	0,048 (0,015)	3,938	2,726	99,60%	1,46	57%
Greece	0,229 (0,055)	0,768 (0,044)	89,000 (2,242)	0,617 (0,134)	0,477	0,220	99,97%	0,50	50%
Netherlands	0,046 (0,023)	0,693 (0,056)	74,422 (0,964)	0,137 (0,063)	1,756	1,216	99,85%	1,24	54%
Ireland	0,132 (0,028)	0,589 (0,025)	76,599 (0,595)	0,141 (0,040)	0,531	0,368	99,96%	1,46	56%
Italy	0,221 (0,033)	0,579 (0,019)	94,533 (0,998)	0,364 (0,051)	0,356	0,246	99,98%	1,10	60%
Luxembourg	0,223 (0,085)	0,534 (0,041)	99,556 (2,257)	0,101 (0,049)	4,891	3,386	99,74%	1,68	77%
Portugal	0,030 (0,012)	0,801 (0,053)	85,127 (1,656)	0,738 (0,151)	0,613	0,425	99,95%	0,33	46%
Spain	0,027 (0,010)	0,750 (0,083)	83,563 (2,943)	0,473 (0,200)	2,236	1,548	99,82%	0,70	50%
Sweden	3,150 (0,460)	0,330 (0,029)	99,392 (9,747)	0,447 (0,214)	1,881	1,302	99,84%	1,68	60%
UK	0,204 (0,091)	0,538 (0,101)	81,853 (2,486)	0,039 (0,011)	2,358	8,556	99,04%	1,79	69%

Table 1. Parameter Estimates, MSE, and % Variance Explained for the Diffusion of Mobile Telephony in EU-15 (standard errors in parentheses).

$$SPD = \frac{1}{t} = b^* \cdot \left[\ln \frac{X_t}{X_0} + \sigma \cdot \ln \frac{X^* - X_t}{X^* - X_0} \right]^{-1} \tag{12}$$

where for each country X_0 represents the 10% of the saturation level and X_t represents the 90% of the saturation level. Figure 2 illustrates the results concerning the speed of mobile telephony penetration for each country under investigation.

Figure 3 illustrates the mobile telephony speed of EU-15 countries w.r.t. their estimated time-delay effect.

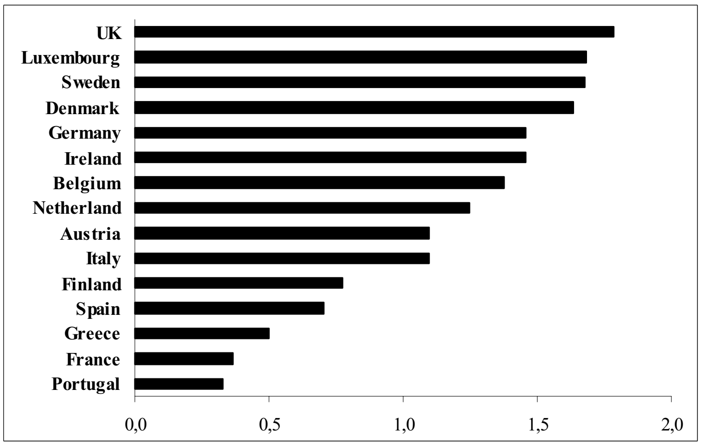


Fig. 1. Time-Delay Parameters for EU-15

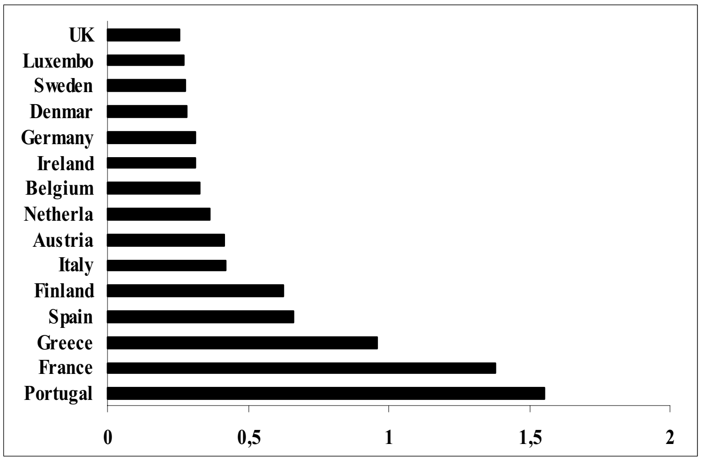


Fig. 2. The Speed of Mobile Telephony Penetration for EU-15

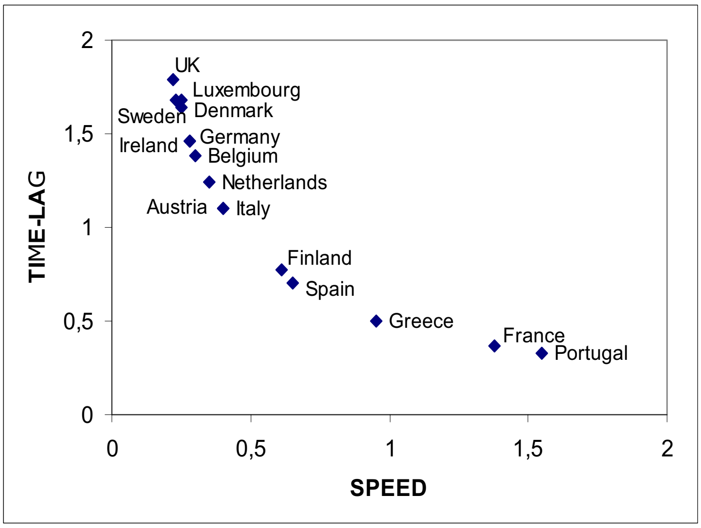


Fig. 3. Time-Delay and Speed of Mobile Telephony Diffusion for EU-15

From the above three figures, it can be seen that the time delay effect in the contagion process has a negative impact on product's penetration rate, at least in the medium phases of the diffusion process.

It could be beneficial in future research efforts to identify and include into the model the market factors affecting the magnitude of the time-delay parameter in order that technology marketing functions to be able to facilitate product's marketability.

4 Conclusions

This paper proposed a new modeling approach for the investigation of the diffusion of mobile telecommunications services in EU-15. It was found that the proposed model which incorporates the notion of the time-delay between the awareness and the adoption phases of a new product plays an important role in studies of new product penetration in different groups of potential agents. The model was applied to the data of mobile telecommunication in EU-15 and the time-delay effect was used for the ranking of the countries under investigation with respect to their ability to adopt and diffuse the new technology. Furthermore, a new speed index was developed aimed to measure the speed of innovation diffusion. The relationship between the speed of the diffusion and the time-delay effect was studied revealing that they are related in a quadratic mode i.e. as the time-delay effect of diffusion increases the speed of diffusions decreases in a quadratic mode.

References

[Bakalis *et al.*, 1997]S. Bakalis, M. Abeln, and E. Mante-Meijer. The adoption and use of mobile telephony in Europe. In L. Haddon, editor, *Communications on the Move: The Experience of Mobile Telephony in the 1990s*, pages 1–10. COST248 Report, 1997.

[Bass, 1969]F. Bass. A new product growth model for consumer durables," ., 15, 217-231. *Management Science*, 15(217–231), 1969.

[Blackman, 1972]A. W. Blackman. A mathematical model for trends forecasts. *Technological Forecasting and Social Change*, 3(441–452), 1972.

[Dekimpe and Sarvary, 1996]P. M. Dekimpe, M. G.and Parker and M. Sarvary. Comparing adoption patterns: A global approach. *INSEAD Working Paper Series*, 37(MKT), 1996.

[Fisher and Pry, 1971]J. C. Fisher and R. H. Pry. A simple substitution model of technical change. *Technological Forecasting and Social Change*, 2:75–88, 1971.

[Frank, 2003]L. Frank. An analysis of the effect of the economic situation on modeling and forecasting the diffusion of wireless communications in Finland. *Technological Forecasting and Social Change*, 2003.

[Griliches, 1957]Z. Griliches. Hybrid corn: An exploration in the economics of technical change. *Econometrica*, 25:501–522, 1957.

[Gruber and Verboven, 2001]H. Gruber and F. Verboven. The diffusion of mobile telecommunications services in the European Union. *European Economic Review*, 45:577–588, 2001.

[Gruber, 2001]H. Gruber. Competition and innovation. the diffusion of mobile telecommunications in Central and Eastern Europe. *Information Economics and Policy*, 13:19–34, 2001.

[Kumar and Kumar, 1992]U. Kumar and V. Kumar. Technological innovation diffusion: The proliferation of substitution models and easing the user's dilemma. *IEEE Trans. Eng. Manag.*, 39:158–168, 1992.

[Mahajan *et al.*, 1961]V. Mahajan, E. Muller, and F. M. Bass. New product diffusion models in marketing: A review and directions for research. *Journal of Marketing*, 54(1):1–26, 1961.

[Mansfield, 1961]E. Mansfield. Technical change and the rate of imitation. *Econometrica*, 29:741–765, 1961.

[Mead and Islam, 1998]N. Mead and T. Islam. Technological forecasting - model selection, model stability, and combining models. *Management Science*, 39(8):1115–1130, 1998.

[Poznanski, 1983]K. Z. Poznanski. International diffusion of steel technologies. timelag and the speed of diffusion. *Technol. Forecast. Social Change*, 23:305–323, 1983.

[Sharif and Kabir, 1976]M. N. Sharif and C. Kabir. A generalized model for forecasting technological substitution. *Technol. Forecast. Social Change*, 8:353–364, 1976.

[Sharif and Ramanathan, 1984]M. N. Sharif and K. Ramanathan. Temporal models of innovation diffusion. *IEEE Trans. Eng. Manag.*, 31(14):76–86, 1984.

[Skiadas, 1985]C. H. Skiadas. Two generalized models for forecasting innovation diffusion. *Technol. Forecast. Social Change*, 27:39–61, 1985.

[Skiadas, 1986]C. H. Skiadas. Innovation diffusion models expressing asymmetry and/or positively or negatively influencing forces. *Technol. Forecast. Social Change*, 30:313–330, 1986.

[Skiadas, 1987]C. H. Skiadas. Two simple models for early and middle stage prediction of innovation diffusion. *IEEE Trans. Eng. Manag.*, 34:79–84, 1987.

[Wright *et al.*, 1997]M. Wright, C. Upritchard, and T. Lewis. A validation of the bass new product diffusion model in new zealand. *Marketing Bulletin*, 8:15–29, 1997.

Time series prediction of the Greek manufacturing index for the non-metallic minerals sector using a Neuro-fuzzy approach (ANFIS)

George Atsalakis[1], Camelia Ucenic[2], and Christos H. Skiadas[1]

[1] Technical University of Crete
Department of Production Engineering and Management,
Data Analysis and Forecasting Laboratory,
73100 Chania, Crete, Greece
(e-mail: atsalakis@ermes.tuc.gr, skiadas@ermes.tuc.gr)

[2] Technical University of Cluj-Napoca
Department of Management and Industrial Systems
Romania
(e-mail: cameliaucenic@yahoo.com)

Abstract. Business units, which work in a competitive economy, are faced with intensifying pressure. Greek businesses have undergone a rapid economic and political development over the last forty years. The relevant sectors constitute an important object of study. The process industries are forced to adopt advanced techniques to improve their global competitiveness due to the increased competition and increasingly environmental regulations. Long range predictive control algorithms are considered by industry to improve the overall plant operability, efficiency and control performance. Many processes have nonlinear and dynamic character. It is difficult to analyze and model using conventional techniques. A new generation of techniques came as an alternative. Soft computing represents one of them. The objectives of this research were threefold: to analyse the economic development of Greek non-metallic sector, to predict its manufacturing index using an Artificial Network with a Fuzzy Inference System (ANFIS) and to compare its forecasting accuracy with various time-series forecasting methods (AR and ARMA). A data set of monthly observations of the manufacturing index is considered. All data is publicly available, and the concerned factors are generally thought to have potentially explaining or predicting capabilities with respect to the industry growth. The data has been generated by the economic market system. The data were available from 1986 to 2002.

Keywords: ANFIS, Neuro-fuzzy, Forecasting, Manufacturing index.

1 General considerations

Business units, which work in a competitive economy, are faced with intensifying pressure. The soft computing approaches are useful due to a high level of uncertainty in dynamic economic processes. The organization's top management is required to not only assess its position, but also to understand the differences inter-firms and inter-industries [Wu *et al.*, 2001].

Artificial intelligence methods have been developed for many business problems. Recent studies have shown that neural networks represent a forecasting technique which is superior to nearly all existing methods such consensus estimates, statistical modelling and simulation [Milam, 1998].

The use of neural networks in economics is still in its relative initial stages. However, in spite of this, a substantial amount of research has been conducted and the number of publications is very extensive. [Moody, 1993] presented empirical results to forecast the U.S. index of industrial production and argued that superior performance can be obtained using state-of-the-art neural network models than using conventional linear time series and regression methods. [Dilli and Wang, 2002] presented an application of the ARIMA model to forecast the production level of the construction industry. [Dilli and Wang, 2003] applied the neural networks to forecast the production level of construction industry. The objective of their paper has been to develop an empirical model for the construction industry in China, which best fits, the data under study and gives better prediction values with minimum errors. The model has to help the planners and the policy makers to formulate proper policies and programs to promote the industry.

The manufacturing index

The index of manufacturing production (IMP) measures changes in the quantity of commodities produced by different producers. It reflects the trends in a constant basket of goods produced by establishments employing a specific number of workers (for example establishments employing 50 or more workers). Nowadays, in Greece, the manufacturing index is compiled from production data covering more than 1 800 factories, selected by branch according to their size (average annual employment of 10 persons and over) (National Statistical Service of Greece (NSSG).

Olga Christodoulaki [Christodoulaki, 1999] presented an analysis of manufacturing output in Greece during the interwar period. The literature usually sees the 1920s as a landmark in the industrialisation of the country and a time when Greek manufacturing achieved an "unprecedented prominence" [Mazower, 1992]. The Supreme Economic Council constructed the first index of industrial production in Greece in the 1930s. This index is described as a weighted volume index, which includes approximately 80% of the total industrial production. It comprised eleven industrial sectors including electricity after 1925. An index constructed by the National Statistical Service was based on up to 61 items. These goods are primarily agricultural products. A few goods from the secondary sector are included, mainly products of food processing industries as well as some imported foodstuffs.

Non-metallic mineral sector

In recent years, a new business environment has been taking shape. The factors, which played an important role, are higher levels of uncertainty, global competition and the European perspective. Strength refers not only to market share but covers the issue of the production cost. The production centres shifted to cheaper areas. The construction industry affects EU's economy to a large extent. It has attained a significant level of competitiveness at the same level as other sectors in the economy at a national or local level. Exports in metallic mineral and non-metallic mineral products within the EU are double than the level of imports from outside EU.

Changes have occurred in the non-metallic minerals market. The Greek cement and building materials market, was expected to be more connected with the technical constructing market as has already happened in other countries. The profits incurred a reduction due to a decline of building activity, recession of the demand regarding products in the European market, and an intensified competition from abroad. Producers were obliged to absorb the biggest part of the incremental production cost when the cost of the production factors increased, in order not to suffer losses in sales and market shares (Federation of Greek Industry, 2000).

2 Methods

Artificial Intelligence forecasting techniques have been receiving much attention lately. They have been cited to have the ability to learn like humans, by accumulating knowledge through repetitive learning activities. Their application in the prediction of economic indicators and financial indices has been demonstrated. [Ranasinghe *et al.*, 1999]

a. Fuzzy logic. Fuzzy logic gives a means of representing uncertainty. It is useful in reasoning with the imprecise data. Fuzzy logic is the convenient way to map the input space to an output space. Fuzzy inference systems (FIS) can express human expert knowledge and experience by using fuzzy inference rules represented in "if-then" statements. The fuzzy inference process has five steps: fuzzify inputs, apply fuzzy operator, apply implication method, aggregate all outputs and defuzzify. In order to obtain a good FIS it is necessary that the researchers possess domain knowledge; the knowledge has to be represented in a symbolic form, be complete, correct and consistent. Unfortunately, fuzzy inference systems tend to become incomplete because experts are reluctant to disclose all the knowledge. In addition it is difficult to express it in a symbolic form. [Nishina and Hagiwara, 1997]

b. Neural networks. Between the biologically inspired computing models there are the artificial neural networks. Artificial NN doesn't approach the complexity of the brain, but have two key similarities: the building blocks are simple computational devices and the connections between neurons determine the function of the network. Layers of neurons form a neural network. A layer includes the weight matrix, the summers, the bias vector b, the transfer function and the output vector [Hagan *et al.*, 1996].

c. Neuro-fuzzy. Neuro-Fuzzy systems use NNs to extract rules and membership functions from input-output data to be used in a Fuzzy Inference System. Using this approach, the black box behaviour of NNs and the problems of finding suitable membership values for FL, are avoided. NFS are suited for applications where user interaction in model design or interpretation is desired. One of the most important NFS is ANFIS.

d. ANFIS. Fuzzy inference systems using neural networks were proposed to avoid the weak points of fuzzy logic. The biggest advantage is that they can use the neural networks' learning capability and can avoid rule-matching time of an inference engine in the traditional fuzzy logic system. Functionally, there are almost no constraints on the node functions of an adaptive network except piecewise differentiability. Structurally, the only limitation of network configuration is that it should be of feedforward type. Due to this minimal restriction, the adaptive network's applications are immediate and immense in various areas. A class of adaptive networks, which are functionally equivalent to fuzzy inference systems, is presented bellow[Jang, 1993]:

We assume the FIS under consideration has two inputs and one output. Suppose that the rule base contains two fuzzy if-then rules of Takagi and Sugeno's type:

Rule1: If x is A_1 and y is B_1 then $f_1 = p_1 \cdot x + q_1 \cdot y + r_1$

Rule2: If x is A_2 and y is B_2 then $f_2 = p_2 \cdot x + q_2 \cdot y + r_2$

The node functions in the same layer are of the same function family as described below:

Layer 1: Every node in this layer is a square node with a node function.

$O_i^1(x) = \mu_{A_i}(x)$ where $x-$ the input to node i, A_i- the linguistic label (small, large, etc.) associated with this node function. In other words, O_i^1 is the membership function of A_i and it specifies the degree to which the given x satisfies the quantifier A_i . Usually is chosen $\mu_{A_i}(x)$ to be bell-shaped with maximum equal to 1 and minimum equal to 0, such as the generalized bell function

$$\mu_{A_i}(x) = \frac{1}{1 + \left[\left(\frac{x-c_i}{a_i}\right)^2\right]^{b_i}}$$

where a_i, b_i, c_i is the parameter set.

As the values of these parameters change, the bell-shaped functions vary accordingly, thus exhibiting various forms of membership function on linguistic label A_i . Parameters in this layer are referred to as premise parameters.

Layer 2: Every node in this layer is a circle node labeled $\prod$, which multiplies the incoming signal and sends the product out.

Layer 3: Every node in this layer is a circle node labeled N. The $i-th$ node calculates the ratio of the $i-th$ rules firing strength to the sum of all rules' firing strengths:

$$\overline{w_i} = \frac{w_i}{w_1 + w_2},\ i = 1, 2$$

For convenience, output of this layer will be called *normalized firing strengths.*
Layer 4: Every node i in this layer is a square node with a node function

$$O_i^4(x) = \overline{w_i} \cdot f_i = \overline{w_i}(p_i \cdot x + q_i \cdot y + r_i)$$

where: $\overline{w_i}$ is the output of layer 3 and $p_i,\ q_i,\ r_i$ is the parameter set. Parameters in this layer will be referred to as consequent parameters.
Layer 5: The single node in this layer is a circle node labeled $\sum$ that computes the overall output as the summation of all incoming signals, i.e.,

$$O_i^5(x) = \sum_i \overline{w}_i \cdot f_i = \frac{\sum\limits_i w_i \cdot f_i}{\sum\limits_i w_i}$$

3 Results and discussions

The object of this research consists of forecasting the manufacturing index for the non-metallic minerals sector from Greece. The forecasting was done using an adaptive neural network with fuzzy inference system. ANFIS uses a hybrid-learning algorithm to identify the membership function parameters of single-output, Sugeno type fuzzy inference systems (FIS). The model was applied for the period 1986-2002. The minimum value of the manufacturing index was 62,1 and the maximum 132,6. The index was less than 100 in 87 months and between 100 and 120 in 43 months. The sector established significant development, reflected in an index greater than 120 only in 26 months in the analyzed interval.

It worked with different numbers of membership functions: two, three, four, five and seven. Different types of membership functions were also chosen - gbellmf, gaussmf, trimf and trapmf The model has one input - the previous value of the manufacturing index of the analyzed sector. The linguistic expressions (small, big, low, high) were transformed into fuzzy sets using membership functions. However, a weak point was the volume of data. Had it been possible to obtain more data, the results could have been more accurate. The model almost always predicted the correct trend of the manufacturing index.

The final model was chosen according to the smallest value of errors. The best results were obtained working with four triangular membership functions. The characteristics of the model for the case of the minimum errors are 20 nodes, 8 linear parameters, 12 nonlinear parameters, 155 training data pairs, 59 checking data pairs and 4 fuzzy rules (their number being obtained with the formula $2^2 = 4 -$ number of membership functions$^{\text{number of imputs}}$).

The scatter plot (figure 1) is a powerful tool, which allows viewing entire data set at once. It displays the relationships between the input and output and identifies the outliers.

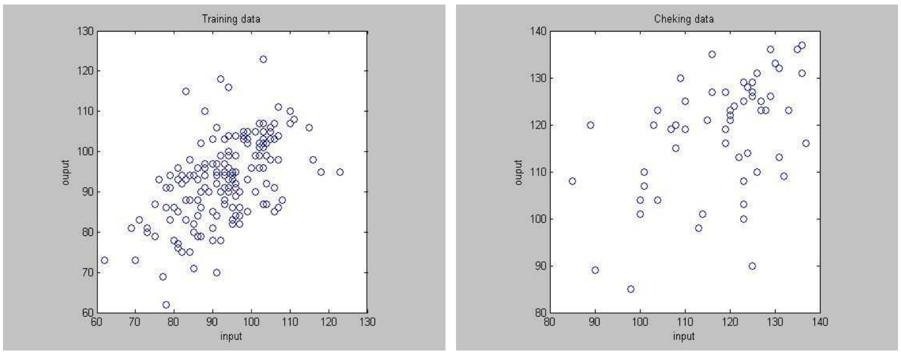


Fig. 1. Scatter plot of input data

The values predicted by the adaptive neural network with fuzzy inference system were compared with the data set. The forecasting accuracy was evaluated by undertaking the comparison with the AR and ARMA methods. The graphical representation of the errors and the comparison between the actual values and the ANFIS predicted values are presented in figure 2

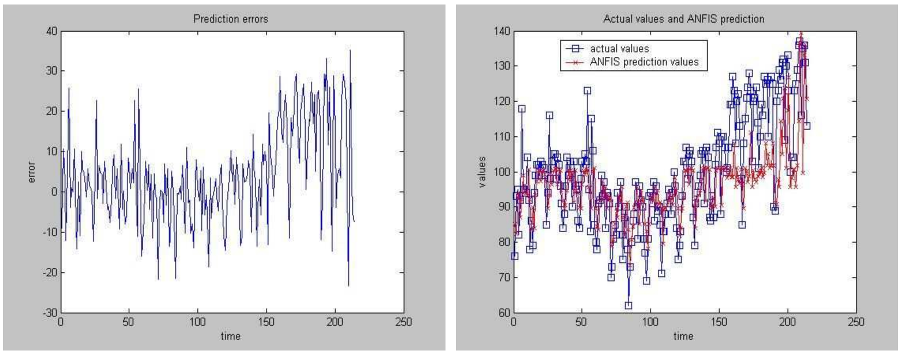


Fig. 2. Error evolution (a) and the comparison between the actual values and the ANFIS values (b) non-metallic

The model displayed a high degree of prediction of the correct trend. The results are better in the first part. Another observation is that at the beginning, the model more accurately forecasted the decrease of the manufacturing index, but afterwards it was able to predict its growth.

The training error, checking error and the step-size are illustrated in figure 3. The training errors are comprised in the interval (8,55; 8,6). The checking errors decreased after 5 epochs by increasing the number of epochs and remained constant after 20 epochs.

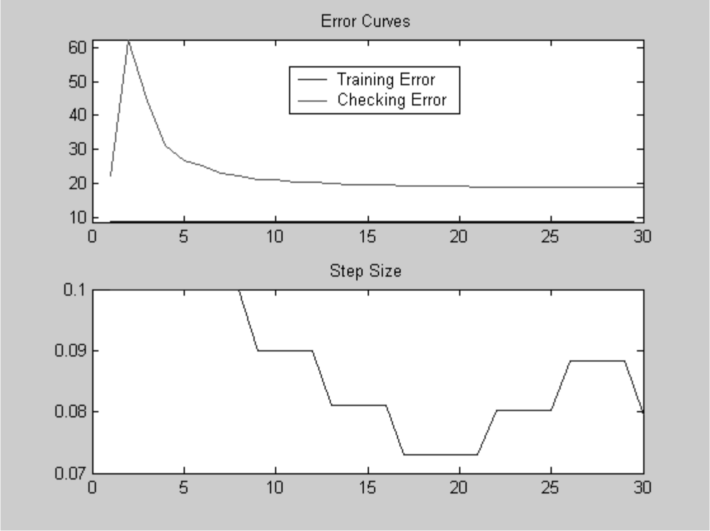


Fig. 3. The error curves and the step-size

The results derived from the application of the adaptive network with fuzzy inference system were compared with those obtained with the use of traditional methods. The comparison of the ANFIS model with the AR and ARMA models is presented in table 1. The comparison was done using relative measures of forecasting accuracy dealing with errors. The measures used in the comparative study are the root mean square error (RMSE), mean absolute error (MAE) and mean absolute percentage error (MAPE). These measures and their application to forecasting have been discussed by many authors [Makridakis *et al.*, 1983][Goh, 1996].

The following ideas can be drawn from the above analysis:

- the application of ARMA took on the smallest RMSE. AR ranked second, followed by the adaptive neural network with fuzzy inference system;
- from the point of view of MAE, the situation was the same. ARMA was ranked first. The second MAE value was obtained for AR, while ANFIS gave the worst value;

Errors	ANFIS	AR	ARMA
RMSE	12.3264	10.6575	10.4546
MAE	9.3513	8.1903	8.1755
MAPE	9.0386	8.4242	8.4588

Table 1. Comparison ANFIS - AR - ARMA for the non-metallic sector

- from the point of view of MAPE, the order of ranking remained unchanged comparing with RMSE and MAE.

4 Conclusions

This research aimed to prove that a neuro-fuzzy approach could be used to forecast the manufacturing index. The weak aspects of other forecasting methodologies for time series could be overcome with the proposed adaptive network with fuzzy inference system (ANFIS). The data available in the form of input output pairs could be used in the ANFIS with relative ease.

Finally, it goes without saying that one of the major limitations of this study is that the practical implementation of the aforementioned approach requires further study and experimentation.

References

[Christodoulaki, 1999]O. Christodoulaki. Industrial growth revisited: Manufacturing output in Greece during the interwar period. *London School of Economics, Working Paper*, (50/99), 1999.

[Dilli and Wang, 2002]R. Dilli and Y. W. Wang. An application of the ARIMA model for forecasting the production level of construction industry. *Journal of Harbin Institute of Technology (New series)*, 9:39–45, 2002.

[Dilli and Wang, 2003]R. Dilli and Y. W. Wang. Neural network forecasting of the production level of chinese construction industry. *Journal of Comparative International Management*, 6, 2003.

[Goh, 1996]B. H. Goh. Resdential construction demand forecasting using economic indicators: A comparative study of artificial neural networks and multiple regression. *Construction Management and Economics*, 14(1), 1996.

[Hagan *et al.*, 1996]M. Hagan, H. Demuth, and M. Beal. *Neural Network Design.* PWS Publishing, Boston MA, 1996.

[Jang, 1993]J.S. Jang. ANFIS: Adaptive-network-based fuzzy inference systems. *IEEE Transactions on Systems, Man and Cybernetics*, 23(3):665–685, 1993.

[Makridakis *et al.*, 1983]S. Makridakis, S.C. Weelwright, and V.E. McGee. *Forecasting: Methods and Applications.* Wiley, New York, 2nd edition, 1983.

[Mazower, 1992]M. Mazower. *Greece and the Interwar Economic Crisis.* Clarendon Press, Oxford, 1992.

[Milam, 1998]A. Milam. Neural networks improve business forecast. *Missisippi Business Journal*, 20(10):34, 1998.

[Moody, 1993]J. et al. Moody. Predicting the U. S. index of industrial production. *Neural Network World*, 3(6):791–94, 1993.

[Nishina and Hagiwara, 1997]T. Nishina and M. Hagiwara. Fuzzy inference neural network. *Neurocomputing*, (14):223–239, 1997.

[Ranasinghe *et al.*, 1999]M. Ranasinghe, B.H. Goh, and T. Barathithasan. A comparative study of artificial neural networks and multiple regression analysis in estimating willingness to pay for urban water supply. *Working Paper, Department of Civil Engineering, University of Moratuwa, Sri Lanka*, 1999.

[Wu *et al.*, 2001]X. Wu, M. Fung, and A. Flitman. Forecasting stock market performance using hybrid intelligence system. In V. N. Alexandrov et al, editor, *ICCS 2001*, pages 447–456. Springer -Verlag Berlin, Heidelberg, 2001.

Solving fuzzy systems of linear equations by a nonlinear programming method

H. Reynaerts[1] and S. Muzzioli[2]

[1] Department of Applied Mathematics and Computer Science,
Ghent University,
Krijgslaan 281, building S9, 9000 Gent, Belgium
(e-mail: huguette.reynaerts@UGent.be)
[2] Department of Economics,
University of Modena and Reggia Emilia,
V.le Berengario 51, 41100 Modena, Italy
(e-mail: muzzioli.silvia@unimore.it)

Abstract. Linear systems of equations, with uncertainty on the parameters, play a major role in various problems in economics and finance. In this paper fuzzy linear systems of the general form $A_1x + b_1 = A_2x + b_2$, with A_1, A_2, b_1 and b_2 matrices with fuzzy elements, are solved by means of a nonlinear programming method. The relation between this methodology and the algorithm proposed in [Muzzioli and Reynaerts, 2004] is highlighted. The methodology is finally applied to an economic and a financial problem.

Keywords: Fuzzy linear systems, fuzzy vector, nonlinear programming.

1 Introduction

Several problems in economics and finance boil down to the solution of a system of linear equations. When we only have some vague knowledge about the actual value of the parameters, it may be convenient to represent some or all of them with a fuzzy number. For such a fuzzy linear system $Ax = b$, where the elements a_{ij} of the $n*n$ matrix A and the elements b_i of the n-vector b are fuzzy numbers, the following solutions have been proposed: the classical solution X_C, the vector solution X_J and the marginal solutions X_E and X_I (see [Buckley and Qu, 1991]). In [Muzzioli and Reynaerts, 2004] this method is extended to the more general fuzzy system of equations $A_1x+b_1 = A_2x+b_2$, with A_1, A_2, b_1 and b_2 matrices with fuzzy elements. Further it is proved that the systems $Ax = b$ and $A_1x + b_1 = A_2x + b_2$ have the same vector solution if $A_1 - A_2 = A$ and $b_2 - b_1$=b. Finally an algorithm to find the vector solution, is introduced.

The aim of this paper is to investigate the solution of the fuzzy linear system by means of a nonlinear programming method and to highlight the relation between this methodology and the algorithm proposed in
[Muzzioli and Reynaerts, 2004].

The plan of the paper is the following: in section 2 we recall the vector solution X_J and the algorithm in order to get this solution. In section 3 we show that the algorithm boils down to a nonlinear programming problem and we work out the Kuhn-Tucker conditions. In section 4 we apply the method to several examples. The last section concludes.

2 The vector solution of the fuzzy system $\boldsymbol{A_1x + b_1 = A_2x + b_2}$

A (triangular) fuzzy number f is defined by three numbers (f_1, f_2, f_3). An α-cut, $\alpha \in [0,1]$, of f is the interval $[\underline{f}(\alpha), \overline{f}(\alpha)]$, with:

$$\underline{f}(\alpha) = (1-\alpha)f_1 + \alpha f_2 \qquad \overline{f}(\alpha) = (1-\alpha)f_2 + \alpha f_3$$

In [Muzzioli and Reynaerts, 2004] we prove that the system $Ax = b$ (where A is a $n * n$- matrix of fuzzy numbers and b a n-vector of fuzzy numbers) and all linear systems $A_1x + b_1 = A_2x + b_2$, where $A_1 - A_2 = A$ and $b_2 - b_1 = b$, have the same vector solution X_J, as defined by [Buckley and Qu, 1991] if all matrices $A(0)$ with $A(0)_{ij} \in a_{ij}(0)$ are nonsingular.

The α-cuts of X_J are the following sets:

$$X_J(\alpha) = \{x \in \mathbb{R}^n \mid A(\alpha)x = b(\alpha), A(\alpha)_{ij} \in a_{ij}(\alpha), b(\alpha)_i \in b_i(\alpha)\}$$

The marginals of $X_{Jj}, j = 1, 2, \ldots, n$, are obtained by projecting X_J on the coordinate axes. In the same paper we consider the following simple algorithm which finds directly the marginals of the vector solution X_J for each unknown. One solves $2^{n(n+1)}$ systems, for each α-cut, where each element of the extended coefficient matrix of those systems is either the lower or the upper bound of the α-cut of the corresponding element of the original fuzzy extended coefficient matrix. The final solution for each unknown, is investigated by taking the minimum and the maximum of the solutions found in each system for this unknown. Since for all parameters a_{ij}, b_i, $[\underline{a}_{ij}(\alpha_1), \underline{a}_{ij}(\alpha_1)] \subset [\underline{a}_{ij}(\alpha_1), \underline{a}_{ij}(\alpha_1)]$ and $[\underline{b}_i(\alpha_1), \underline{b}_i(\alpha_1)] \subset [\underline{b}_i(\alpha_1), \underline{b}_i(\alpha_1)]$ if $\alpha_1 > \alpha_2$, the minimal (resp. maximal) value of $x_k^*(\alpha_1)$ is always greater (resp. smaller) then the minimal value of $x_k^*(\alpha_2)$ and thus $[\underline{x}_k^*(\alpha_1), \overline{x}_k(\alpha_1)^*] \subset [\underline{x}_k^*(\alpha_2), \overline{x}_k(\alpha_2)^*]$.

This has as concequence that $x_k^*(1) \in [\underline{x}_k^*(0), \overline{x}_k^*(0)]$, for all k and thus the solution of the algorithm is always a fuzzy number. This procedure ensures that all possible solutions, consistent with the parameters of the system, are taken. A simplification of the previous method is to find the solutions for $\alpha = 1$ and $\alpha = 0$ and impose ex post a triangular form on the solution, whenever $x_j(1) \in [\underline{x}_j(0), \overline{x}_j(0)]$, for all j. In order to find $x_j(1)$, for all j, one just solves the crisp system, substituting $\alpha = 1$ in the fuzzy system. In order to find $[\underline{x}_j(0), \overline{x}_j(0)]$, for all j, one applies the algorithm for $\alpha = 0$. If $x_j(1) \in [\underline{x}_j(0), \overline{x}_j(0)]$, for all j, then one takes as solution the triangular fuzzy numbers $(\underline{x}_j(0), x_j(1), \overline{x}_j(0))$.

3 The nonlinear programming method

The algorithm can be considered as n nonlinear programming problems where:

- the object functions, $x_k^*(b_1, \ldots, b_n, a_{1,1}, \ldots, a_{nn}), k = 1, 2, \ldots, n$ are the solutions of the system of equations considered as functions of the coefficients,
- with constraints:

$$\begin{array}{l} \underline{b}_1(\alpha) \leq b_1 \leq \overline{b}_1(\alpha) \ldots \underline{b}_n(\alpha) \leq b_n \leq \overline{b}_n(\alpha) \\ \underline{a}_{1,1}(\alpha) \leq b_1 \leq \overline{a}_{1,1}(\alpha) \ldots \underline{a}_{n,n}(\alpha) \leq a_{n,n} \leq \overline{a}_{n,n}(\alpha) \end{array}$$

The object functions should as well be minimized as maximized to find the extremes of the α-cuts of the solution.

The Kuhn-Tucker conditions should be verified for extrema. The Lagrange functions are the following for all $k = 1, 2, \ldots n$:

$$\begin{aligned} L_k(b_1, \ldots, b_n, a_{1,1}, \ldots, a_{nn}) = x_k^*(b_1, \ldots, b_n, a_{1,1}, \ldots, a_{nn} \\ - \lambda_1(b_1 - \overline{b}_1(\alpha)) - \ldots - \lambda_n(b_n - \overline{b}_n(\alpha)) \\ - \lambda_{n+1}(a_{1,1} - \overline{a}_{1,1}(\alpha)) - \ldots - \lambda_{n(n+1)}(a_{nn} - \overline{a}_{nn}(\alpha)) \end{aligned}$$

The (necessary) Kuhn-Tucker conditions for a maximum (resp. minimum) are:

$$\begin{array}{l} \underline{b}_i(\alpha) \leq b_i \leq \overline{b}_i(\alpha) \quad (b_i - \underline{b}_i(\alpha))\dfrac{\partial L_k}{\partial b_i} = 0, \qquad \dfrac{\partial L_k}{\partial b_i} \leq 0 \quad (resp. \geq 0), \\ \lambda_i \dfrac{\partial L_k}{\partial \lambda_i} = 0, \qquad \lambda_i \geq 0 \qquad (resp. \leq 0), \quad \forall i = 1, 2, \ldots, n \end{array}$$

$$\begin{array}{l} \underline{a}_{ij}(\alpha) \leq a_{ij} \leq \overline{a}_{ij}(\alpha) \quad (a_{ij} - \underline{a}_{ij}(\alpha))\dfrac{\partial L_k}{\partial a_{ij}} = 0, \qquad \dfrac{\partial L_k}{\partial a_{ij}} \leq 0 \qquad (resp. \geq 0), \\ \lambda_{i*n+j} \dfrac{\partial L_k}{\partial \lambda_{i*n+j}} = 0, \qquad \lambda_{i*n+j} \geq 0 \quad (resp. \leq 0), \qquad \forall i, j = 1, 2, \ldots, n \end{array}$$

Since the partial derivatives are:

$$\frac{\partial L_k}{\partial b_i} = \frac{\partial x_k^*}{\partial b_i} - \lambda_i \quad \forall i, \qquad \frac{\partial L_k}{\partial a_{ij}} = \frac{\partial x_k^*}{\partial a_{ij}} - \lambda_{i*n+j} \quad \forall i, j$$

$$\frac{\partial L_k}{\partial \lambda_i} = -(b_i - \overline{b}_i(\alpha)) \quad \forall i, \qquad \frac{\partial L_k}{\partial \lambda_{i*n+j}} = -(a_{ij} - \overline{a}_{ij}(\alpha)) \quad \forall i, j$$

the Kuhn-Tucker conditions for a maximum (resp. minimum) are:

$$\underline{b}_i(\alpha) \leq b_i \leq \overline{b}_i(\alpha) \quad \lambda_i \geq 0 \quad (resp. \leq 0)$$

$$(b_i - \underline{b}_i(\alpha))(\frac{\partial x_k^*}{\partial b_i} - \lambda_i) = 0 \tag{1}$$

$$\frac{\partial x_k^*}{\partial b_i} - \lambda_i \leq 0 \quad (resp. \geq 0) \tag{2}$$

$$\lambda_i(b_i - \overline{b}_i(\alpha)) = 0, \forall i = 1, 2, \ldots, n \tag{3}$$

$$\underline{a}_{ij}(\alpha) \leq a_{ij} \leq \overline{a}_{ij}(\alpha) \quad \lambda_{i*n+j} \geq 0 \quad (resp. \leq 0)$$

$$(a_{ij} - \underline{a}_{ij}(\alpha))(\frac{\partial x_k^*}{\partial a_{ij}} - \lambda_{i*n+j}) = 0 \tag{4}$$

$$\frac{\partial x_k^*}{\partial a_{ij}} - \lambda_{i*n+j} \leq 0 \quad (resp. \geq 0) \tag{5}$$

$$\lambda_{i*n+j}(a_{ij} - \overline{a}_{ij}(\alpha)) = 0, \forall i, j = 1, 2, \ldots, n \tag{6}$$

For a maximum (resp. minimum) the following cases can occur:

- Suppose that $\frac{\partial x_k}{\partial b_i} > 0$ (resp. < 0) then from (2) it follows that $\frac{\partial x_k}{\partial b_i} \leq$ (resp. $\geq$)λ_i and thus $\lambda_i \neq 0$. Then from (3) one concludes that $b_i^* = \overline{b}_i(\alpha)$.
- Suppose that $\frac{\partial x_k}{\partial b_i} < 0$ (resp. > 0) then, since $\lambda_i \geq$ (resp. $\leq$) 0 it follows that $\frac{\partial x_k}{\partial b_i} \neq \lambda_i$ and thus from (1) one concludes that $b_i^* = \underline{b}_i(\alpha)$.
- Suppose that $\frac{\partial x_k}{\partial b_i} = 0$ then the Kuhn-Tucker conditions are the following:

  $$(b_i - \underline{b}_i(\alpha))\lambda_i = 0, \qquad \lambda_i \geq 0, \qquad \lambda_i(b_i - \overline{b}_i(\alpha)) = 0$$

  and thus the necessary conditions hold for all $b_i \in [\underline{b}_i(\alpha), \overline{b}_i(\alpha)]$.

The same cases, with analogous conclusions, occur for the coefficients a_{ij}.

4 Economic examples

(1) The market price of a good and the quantity produced are determined by the equality between supply and demand. Demand is the amount of a good that consumers are willing and able to buy at a given price. Supply is the amount of a good producers are willing and able to sell at a given price. Suppose that demand and supply are linear functions of the price:

$$\begin{cases} q_d = a * p + b \\ q_s = c * p + d \end{cases},$$

where q_s is the quantity supplied, that is required to be equal to q_d, the quantity demanded, p is the price and a, b, c and d are coefficients to be estimated. Suppose that we have only some imprecise data on the relation between the quantity supplied and demanded at a given price, then we can naturally describe the parameters by fuzzy numbers. Due to the equilibrium conditions, the following fuzzy linear system should be solved:

$$\begin{cases} x_1 = a * x_2 + b \\ x_1 = c * x_2 + d \end{cases}$$

This corresponds (see [Muzzioli and Reynaerts, 2004]) to find the vector solution of the fuzzy system:

$$\begin{cases} x_1 - a * x_2 = b \\ x_1 - c * x_2 = d \end{cases}$$

If one applies the nonlinear programming method, the following object functions should be maximized (resp. minimized):

$$x_1(a, b, c, d) = \frac{bc - ad}{c - a} \qquad x_2(a, b, c, d) = \frac{b - d}{c - a}$$

with constraints:

$$\begin{array}{cc} \underline{a} \leq a \leq \overline{a}(< 0) & (0 <)\underline{b} \leq b \leq \overline{b} \\ (0 <)\underline{c} \leq c \leq \overline{c} & \underline{d} \leq d \leq \overline{d}(< 0) \end{array}$$

First of all we calculate the partial derivatives of the object functions:

$$\frac{\partial x_1}{\partial a} = \frac{c(b-d)}{(c-a)^2} \qquad \frac{\partial x_1}{\partial b} = \frac{c}{(c-a)}$$

$$\frac{\partial x_1}{\partial c} = \frac{-a(b-d)}{(c-a)^2} \qquad \frac{\partial x_1}{\partial d} = \frac{-a}{(c-a)}$$

$$\frac{\partial x_2}{\partial a} = \frac{(b-d)}{(c-a)^2} \qquad \frac{\partial x_2}{\partial b} = \frac{1}{(c-a)}$$

$$\frac{\partial x_2}{\partial c} = \frac{-(b-d)}{(c-a)^2} \qquad \frac{\partial x_2}{\partial d} = \frac{-1}{(c-a)}$$

Since $\frac{\partial x_1}{\partial a} > 0$ one obtains the maximum of x_1 for $a^{max} = \overline{a}$ and the minimum for $a^{min} = \underline{a}$.
Since $\frac{\partial x_1}{\partial b} > 0$ one obtains the maximum of x_1 for $b^{max} = \overline{b}$ and the minimum for $b^{min} = \underline{b}$.
Since $\frac{\partial x_1}{\partial c} > 0$ one obtains the maximum of x_1 for $c^{max} = \overline{c}$ and the minimum for $c^{min} = \underline{c}$.
Since $\frac{\partial x_1}{\partial d} > 0$ one obtains the maximum of x_1 for $d^{max} = \overline{d}$ and the minimum

for $d^{min} = \underline{d}$.
Since $\frac{\partial x_2}{\partial a} > 0$ one obtains the maximum of x_1 for $a^{max} = \overline{a}$ and the minimum for $a^{min} = \underline{a}$.
Since $\frac{\partial x_2}{\partial b} > 0$ one obtains the maximum of x_1 for $b^{max} = \overline{b}$ and the minimum for $b^{min} = \underline{b}$.
Since $\frac{\partial x_2}{\partial c} < 0$ one obtains the maximum of x_1 for $c^{max} = \underline{c}$ and the minimum for $c^{min} = \overline{c}$.
Since $\frac{\partial x_2}{\partial d} < 0$ one obtains the maximum of x_1 for $d^{max} = \underline{d}$ and the minimum for $d^{min} = \overline{d}$.
The solution to the system is:

$$([\frac{\underline{bc} - \underline{ad}}{\underline{c} - \underline{a}}, \frac{\overline{b}\overline{c} - \overline{a}\overline{d}}{\overline{c} - \overline{a}}], [\frac{\underline{b} - \overline{d}}{\overline{c} - \underline{a}}, \frac{\overline{b} - \underline{d}}{\underline{c} - \overline{a}}])$$

(2) The binary tree model of Cox *et al.* (1979) is used to price options and other derivative securities. A European call option is a financial security that provides its holder, in exchange for the payment of a premium, the right but not the obligation to buy a certain underlying asset at a certain date in the future for a specified price K. In the binary tree model of [Cox *et al.*, 1979] the following assumptions are made: (A1) the markets have no transaction costs, no taxes, no restrictions on short sales, and assets are infinitely divisible; (A2) the lifetime T of the option is divided into N time steps of length T/N; (A3) the market is complete; (A4) no arbitrage opportunities are allowed, which implies for the risk-free interest factor, $1 + r$, over one step of length T/N, that $d < 1 + r < u$, where u is the up and d the down factor. The European call option price at time zero, has a well-known formula in this model,

$$EC(K, T) = \frac{1}{(1+r)^N} \sum_{j=0}^{N} \binom{N}{j} p_u^j p_d^{N-j} \left(S(0)u^j d^{N-j} - K\right)_+ ,$$

where K is the exercise price, $S(0)$ is the price of the underlying asset at time the contract begins, p_u and p_d are the resp. up and down risk-neutral transition probabilities. Fundamental for the option valuation is the derivation of the risk neutral probabilities, which are obtained from the following system:

$$\begin{cases} p_u + p_d = 1 \\ up_u + dp_d = 1 + r. \end{cases} \tag{7}$$

The solution is given by:

$$p_u = \frac{(1+r) - d}{u - d} \qquad p_d = \frac{u - (1+r)}{u - d}.$$

In order to estimate the up and down jump factors from market data, the standard methodology (see Cox *et al.* (1979)) leads to set:

$$u = e^{\sigma\sqrt{T/N}}, d = e^{-\sigma\sqrt{T/N}},$$

where σ is the volatility of the underlying asset.

If there is some uncertainty about the value of the volatility, then it is also impossible to precisely estimate the up and down factors.
[Muzzioli and Reynaerts, 2004] suggest to model the up and down jump factors by triangular fuzzy numbers.

A fuzzy version of the two equations of the system (7) should now be introduced. This can be done (for each equation) in two different ways, since for an arbitrary fuzzy number f there exists no fuzzy number g such that $f + g = 0$ and for all non-crisp fuzzy numbers $f + (-f) \neq 0$:

$$\begin{aligned} p_u + p_d &= (1,1,1) \\ p_u &= (1,1,1) - p_d \end{aligned}$$

respectively

$$\begin{aligned} up_u + dp_d &= (1+r, 1+r, 1+r) \\ up_u &= (1+r, 1+r, 1+r) - dp_d \end{aligned}$$

where p_u and p_d are the fuzzy up and down probabilities and u and d are triangular fuzzy numbers.

Therefore the linear system (7) can be rewritten in four different ways:

$$\begin{cases} p_u + p_d = 1 \\ up_u + dp_d = 1 + r, \end{cases} \tag{8}$$

$$\begin{cases} p_u = 1 - p_d \\ up_u + dp_d = 1 + r, \end{cases} \tag{9}$$

$$\begin{cases} p_u = 1 - p_d \\ dp_d = (1 + r) - up_u, \end{cases} \tag{10}$$

and

$$\begin{cases} p_u + p_d = 1 \\ dp_d = (1 + r) - up_u. \end{cases} \tag{11}$$

Different solutions to the same fuzzy linear system have been found in Muzzioli and Torricelli (2001), and in [Reynaerts and Vanmaele, 2003], by solving system (8) and system (9), respectively.

It is easy to see that the four systems have no classical solution, therefore we investigate the vector solution.

If one applies this algorithm to the financial example, one should solve the following systems:

$$\begin{cases} p_u + p_d = 1 \\ \underline{u}p_u + \underline{d}p_d = 1 + r. \end{cases}$$

$$\begin{cases} p_u + p_d = 1 \\ \overline{u}p_u + \underline{d}p_d = 1 + r. \end{cases}$$

$$\begin{cases} p_u + p_d = 1 \\ \underline{u}p_u + \overline{d}p_d = 1 + r. \end{cases}$$

$$\begin{cases} p_u + p_d = 1 \\ \overline{u}p_u + \overline{d}p_d = 1 + r. \end{cases}$$

The solutions to those systems are resp.:

$$\begin{cases} p_u = \frac{(1+r)-\underline{d}}{\underline{u}-\underline{d}} \\ p_d = \frac{\underline{u}-(1+r)}{\underline{u}-\underline{d}}. \end{cases}$$

$$\begin{cases} p_u = \frac{(1+r)-\underline{d}}{\overline{u}-\underline{d}} \\ p_d = \frac{\overline{u}-(1+r)}{\overline{u}-\underline{d}}. \end{cases}$$

$$\begin{cases} p_u = \frac{(1+r)-\overline{d}}{\underline{u}-\overline{d}} \\ p_d = \frac{\underline{u}-(1+r)}{\underline{u}-\overline{d}}. \end{cases}$$

$$\begin{cases} p_u = \frac{(1+r)-\overline{d}}{\overline{u}-\overline{d}} \\ p_d = \frac{\overline{u}-(1+r)}{\overline{u}-\overline{d}}. \end{cases}$$

The final solution is obtained by taking the minimum and maximum for each unknown:

$$\begin{cases} \underline{p_u} = min(\frac{(1+r)-\underline{d}}{\underline{u}-\underline{d}}, \frac{(1+r)-\underline{d}}{\overline{u}-\underline{d}}, \frac{(1+r)-\overline{d}}{\underline{u}-\overline{d}}, \frac{(1+r)-\overline{d}}{\overline{u}-\overline{d}}) \\ \overline{p_u} = max(\frac{(1+r)-\underline{d}}{\underline{u}-\underline{d}}, \frac{(1+r)-\underline{d}}{\overline{u}-\underline{d}}, \frac{(1+r)-\overline{d}}{\underline{u}-\overline{d}}, \frac{(1+r)-\overline{d}}{\overline{u}-\overline{d}}) \\ \underline{p_d} = min(\frac{\underline{u}-(1+r)}{\underline{u}-\underline{d}}, \frac{\overline{u}-(1+r)}{\overline{u}-\underline{d}}, \frac{\underline{u}-(1+r)}{\underline{u}-\overline{d}}, \frac{\overline{u}-(1+r)}{\overline{u}-\overline{d}}) \\ \overline{p_d} = max(\frac{\underline{u}-(1+r)}{\underline{u}-\underline{d}}, \frac{\overline{u}-(1+r)}{\overline{u}-\underline{d}}, \frac{\underline{u}-(1+r)}{\underline{u}-\overline{d}}, \frac{\overline{u}-(1+r)}{\overline{u}-\overline{d}}). \end{cases}$$

Therefore, the vector of fuzzy numbers:

$$\begin{pmatrix} [\frac{(1+r)-\overline{d}}{\overline{u}-\overline{d}}, \frac{(1+r)-\underline{d}}{\underline{u}-\underline{d}}] \\ [\frac{\underline{u}-(1+r)}{\underline{u}-\underline{d}}, \frac{\overline{u}-(1+r)}{\overline{u}-\overline{d}}] \end{pmatrix},$$

is a solution to the system.

Note that the algorithm boils down to the following nonlinear programming problems (for each α):

$$\begin{aligned} &max_{u,d} \quad (\text{resp.}min_{u,d}) \quad \frac{1+r-d}{u-d} \\ &\text{where} \quad (1+r \leq)\underline{u} \leq u \leq \overline{u} \\ &\qquad \text{and} \quad \underline{d} \leq d \leq \underline{d}(\leq 1+r) \end{aligned}$$

$$
\begin{aligned}
&max_{u,d}(\text{resp.}min_{u,d})\frac{u-(1+r)}{u-d}\\
&\text{where}\quad (1+r\leq)\underline{u}\leq u\leq \overline{u}\\
&\quad\ \ \text{and}\quad \underline{d}\leq d\leq \overline{d}(\leq 1+r)
\end{aligned}
$$

Since $\frac{\partial p_u}{\partial u}=\frac{d-(1+r)}{(u-d)^2}<0$ the maximum of p_u is obtained for $u^{max}=\underline{u}$ and the minimum for $u^{min}=\overline{u}$.
Since $\frac{\partial p_u}{\partial d}=\frac{(1+r)-u}{(u-d)^2}<0$ the maximum of p_u is obtained for $d^{max}=\underline{d}$ and the minimum for $d^{min}=\overline{d}$.
Since $\frac{\partial p_d}{\partial u}=\frac{(1+r)-d}{(u-d)^2}>0$ the maximum of p_d is obtained for $u^{max}=\overline{u}$ and the minimum for $u^{min}=\underline{u}$.
Since $\frac{\partial p_d}{\partial d}=\frac{u-(1+r)}{(u-d)^2}>0$ the maximum of p_d is obtained for $d^{max}=\overline{d}$ and the minimum for $d^{min}=\underline{d}$

5 CONCLUSIONS

In this paper we have investigated the solution of a fuzzy linear system of equations by resorting to a non-linear programming methodology.

We have applied the methodology proposed to two important economic applications.

Acknowledgements

H. Reynaerts would like to acknowledge the financial support by the BOF-project 001104599 of the Ghent University.

References

[Buckley and Qu, 1991]J.J. Buckley and Y. Qu. Solving systems of linear fuzzy equations. *Fuzzy sets and systems*, pages 33–43, 1991.

[Cox *et al.*, 1979]J.C. Cox, S.A. Ross, and S. Rubinstein. Option pricing, a simplified approach. *Journal of Financial Economics*, pages 229–263, 1979.

[Muzzioli and Reynaerts, 2004]S. Muzzioli and H. Reynaerts. Fuzzy binary tree model for european-style vanilla options. In I. Batyrshin, J. Kacprzyk, and L. Sheremetov, editors, *Fuzzy Sets and Soft Computing in Economics and Finance*, pages 222–229, 2004.

[Reynaerts and Vanmaele, 2003]H. Reynaerts and M. Vanmaele. A sensitivity analysis for the pricing of european call options in a binary tree model. In J. Bernard, T. Seidenfeld, and M. Zaffalon, editors, *Imprecise Probabilities and their Applications*, pages 467–481, 2003.

Input Control in Fuzzy Non-Homogeneous Markov Systems

Maria Symeonaki[1] and Giogros Stamou[2]

[1] Panteion University, Department of Social Politics,
136 Syngrou Av., 176 71,
Athens, Greece
(e-mail: msimeon@panteion.gr, msymeon@unipi.gr)

[2] National Technical University of Athens,
Department of Electrical and Computer Engineering,
Image, Video and Multimedia Laboratory,
157 73, Athens, Greece
(e-mail: gstam@softlab.ntua.gr)

Abstract. Certain aspects of input control of a non-homogeneous Markov System (NHMS) using fuzzy set theory and fuzzy reasoning are presented in this paper. This is an effort to provide strategies that direct the changes that take place in the population structures of a Fuzzy Non-homogeneous Markov System (F-NHMS) towards a desirable direction. Our goal is to maintain the population structure of the system, $\mathbf{N}(t)$, between two given population structures, $\mathbf{N}_1$ and $\mathbf{N}_2$, which is a very important issue in practical applications. More specifically, we study the aspect of attainability in a F-NHMS and give the input probability vector that achieves our aim. Maintainability is also studied by providing a necessary and sufficient condition such that $\mathbf{N}(t)$ lies between the two population structures, for each t. Finally, an illustrative example is provided.

Keywords: Markov systems, Fuzzy system models, Control theory.

1 Introduction - Problem statement

Let us first give a short description of a NHMS [Vassiliou, 1982]. Consider a population, which is stratified into classes according to different characteristics and let $S = \{1, 2, ..., n\}$ be the *set of states* of the system, which are assumed to be exclusive and exhaustive. Let also $\mathbf{N}(t) = [N_1(t), N_2(t), ..., N_n(t)]$ be the *expected population structure of the system* at time t, where $N_i(t)$ is the expected number of members in state i at time t. Let $T(t)$ denote the total number of members in the system and $\Delta T(t) = T(t+1) - T(t)$. Let us assume that the individual transitions between the states occur according to the sequence of matrices $\{\mathbf{P}(t)\}_{t=0}^{\infty}$ and that $\{\mathbf{p}_o(t)\}_{t=0}^{\infty}$ is *the sequence of input probability vectors.* Suppose, moreover, that the members that leave the system are transferred in a state $n+1$ denoting the external environment of the system and let $\{\mathbf{p}_{n+1}(t)\}_{t=0}^{\infty}$ be the *sequence of loss probability vectors.* Also assume that $q_{ij}(t) = p_{ij}(t) + p_{i,n+1}(t)p_{oj}(t)$, then we define the sequence of matrices

$\mathbf{Q}(t) = \mathbf{P}(t) + \mathbf{p}'_{n+1}(t)\mathbf{p}_o(t) = \{q_{ij}(t)\}_{i,j\in S}$, where $(\cdot)'$ denotes the transpose of the respective vector. $\{\mathbf{Q}(t)\}_{t=0}^{\infty}$ defines uniquely a non-homogeneous Markov chain, which is called the *embedded non-homogeneous Markov chain*. The (i,j) -element of $\mathbf{Q}(t)$ represents the total transition probability from state i to state j, in the time interval $(t-1,t]$. The expected number of members in the various states at time t is given by:

$$\mathbf{N}(t) = \mathbf{N}(t-1)\mathbf{Q}(t-1) + \mathit{\Delta} T(t-1)\mathbf{p}_o(t-1), or \tag{1}$$

$$\mathbf{N}(t) = \mathbf{N}(t-1)\mathbf{P}(t-1) + R(t-1)\mathbf{p}_o(t-1), \tag{2}$$

where $R(t)$ denotes the *expected number of new members in the system* at time t. In order to apply the model of a NHMS, $q_{ij}(t)$ (or $p_{ij}(t)$) and $p_{oi}(t)$ must be determined, $\forall\ i,j = 1,2,...,n$ and $\forall t$. This estimation obviously depends on statistical data analysis, it can be accomplished whenever enough data is provided and obviously introduces uncertainty due to measurement errors and lack of data. This is the main reason for considering fuzzy logic and fuzzy reasoning in Markov systems. In [Symeonaki *et al.*, 2000], [Symeonaki *et al.*, 2002] the concept of a F-NHMS was introduced. The asymptotic behaviour and variability of the system was provided, but this is only the initial step. We need to proceed in the opposite direction, since the projected structures will seldom coincide with what is desired. In this paper the goal is given and the objective is to provide the input probability vector that achieves the desired goal and the conditions under which the goal is maintained. More specifically, the objective here is to develop a useful methodology for obtaining the transition and input probabilities and provide thereafter the input probability vector such that the population structure of the system lies between two given population structures. A different approach to a similar end can be found in [Hartfiel, 1994]. In this paper the problem is expanded to population systems and more specifically to NHMS, where the transition, input and loss probabilities depend on time t. The present paper is organized as follows. In Section 2, a description of a F-NHMS is provided and the necessary parameters of the system are given. More specifically, attainability and maintainability in a F-NHMS is discussed. Section 3 provides an illustrative example of the conclusions of Section 2.

2 Input Control of a F-NHMS

In this section the central problem of input control of a F-NHMS with $S = \{1,2,...,n\}$ is discussed. It is assumed that the transition probability $p_{ij}(t)$ is a function of the population parameters (e.g. Longevity, Mortality, Fecundity, etc) of the system, i.e. $p_{ij}(t) = f_{ij}(pp_1, pp_2, ..., pp_l)$, where: $\sum_{j\in S} f_{ij}(pp_1, pp_2, ..., pp_l) \leq 1$, for any value of the population parameters $pp_1, pp_2, ..., pp_l$. The idea of the *population parameters* of the system was firstly presented in [Symeonaki *et al.*, 2000] and [Symeonaki *et al.*, 2002].

Each population parameter depends on the values of the basic parameters of the system. In order to determine the population parameters from the basic parameters of the system a Fuzzy Inference System (FIS) is used. The structure of a F-NHMS is illustrated in Figure 1.

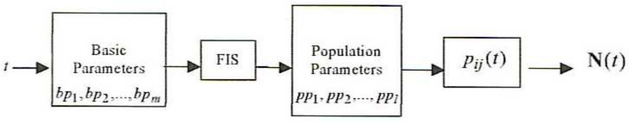


Fig. 1. The structure of the F-NHMS

Assume that the values of the $i-th$ basic parameter of the system range between two values α_i and b_i, i.e. the values of the $i-th$ basic parameter belong to the closed interval $[\alpha_i, b_i]$. A fuzzy partition $A^{(i)}$ of order d_i on the domain $[\alpha_i, b_i]$ is defined and a fuzzy partition $B^{(j)}$ of order r_j is also defined on the universe of discourse of the $j-th$ population parameter. The fuzzy partitions $A^{(i)}$ and $B^{(j)}$ are linguistic representations of their universe of discourses, therefore their elements are linguistic terms like "LOW", "HIGH", etc. The relationship of the crisp universe of discourses is represented using linguistic rules, that derive from the symbolic knowledge that the experts of the system possess and define a mapping of the fuzzy partitions $A^{(i)}$ to the fuzzy partitions $B^{(j)}$. This mapping is said to be a *fuzzy association* and represents the empirical, linguistic rules. As long as the elements of $A^{(i)}$ and the elements of $B^{(j)}$ have a linguistic meaning, heuristic or empirical linguistic rules can be used in order to describe the input-output relationship. We assume that all fuzzy partitions are complete [Stamou and Tzafestas, 1999]. The number of all different rules in the system is denoted by k and we can see that $k = d_1 d_2 \cdots d_m$. We denote by $w_i(t)$ the degree in which the rule i fires at time t. Each rule corresponds to a matrix $\mathbf{P}_i$ and it can easily be proved by induction that if we use as $t-$norm the *product*, then $\sum_{i=1}^{k} w_i(t) = 1$. Therefore, for each t, the transition matrix $\mathbf{P}(t)$ is of the form:

$$\mathbf{P}(t) = \sum_{i=1}^{k} w_i(t)\mathbf{P}_i, \tag{3}$$

with $\mathbf{P}_i\mathbf{1}' \leq \mathbf{1}'$ and $\sum_{i=1}^{k} w_i(t) = 1$, for each $t = 0, 1, 2, ...$, and $\mathbf{1}' = [1, 1, ..., 1]'$. Following the same reasoning for the sequence of input probability vectors, the vector $\mathbf{p}_o(t)$ is of the following form, for each t:

$$\mathbf{p}_o(t) = \sum_{i=1}^{m} u_i(t)\mathbf{p}_{o_i}, \tag{4}$$

$\mathbf{p}_{o_i}\mathbf{1}' = \mathbf{1}'$, $\sum_{i=1}^{m} u_i(t) = 1$, for each $t = 0, 1, 2, ...$, and $u_i(t)$ is the degree in which the rule i for the input probability vector $\mathbf{p}_o(t)$, fires. Therefore, from (2) the expected number of members in the various states of the system at time t, is given by:

$$\mathbf{N}(t) = \mathbf{N}(t-1)\sum_{i=1}^{k} w_i(t)\mathbf{P}_i + R(t-1)\sum_{i=1}^{m} u_i(t-1)\mathbf{p}_{o_i}, or \tag{5}$$

$$\mathbf{N}(t) = \mathbf{N}(0)\prod_{\tau=0}^{t-1}\sum_{i=1}^{k} w_i(\tau)\mathbf{P}_i + \sum_{\tau=1}^{t} R(\tau-1)\sum_{i=1}^{m} u_i(\tau-1)\mathbf{p}_{o_i}\prod_{j=\tau}^{t-1}\sum_{i=1}^{k} w_i(j)\mathbf{P}_i. \tag{6}$$

Let $M_{n,m}(F)$ define the set of all $n \times m$ matrices with elements from the field F.

Definition 1 *[Hartfiel, 1994]: Let two vectors $\boldsymbol{p}, \boldsymbol{q} \in M_{1,k}(R)$ for which it is $\boldsymbol{p} \leq \boldsymbol{q}$. The set of all vectors $\boldsymbol{x} \in M_{1,k}(R)$, which are such that $\boldsymbol{p} \leq \boldsymbol{x} \leq \boldsymbol{q}$, is called box$(\boldsymbol{p}, \boldsymbol{q})$, i.e. box$(\boldsymbol{p}, \boldsymbol{q}) = \{\boldsymbol{x} : \boldsymbol{p} \leq \boldsymbol{x} \leq \ \boldsymbol{q}\}$.*

Now let $\mathbf{N}_1, \mathbf{N}_2$ be two population structures such that $\mathbf{N}_1 \leq \mathbf{N}_2$. Then a NHMS is said to be *stably controllable* if we can maintain the population structure of the system between the desired structures $\mathbf{N}_1$ and $\mathbf{N}_2$ i.e. if: $\mathbf{N}_1 \leq \mathbf{N}(t) \leq \mathbf{N}_2$, $\forall t = 0, 1, 2, ...$. More specifically:

Definition 2 *[Symeonaki, 1998]: If $\forall t = 0, 1, 2, ...$ there exists an input vector $\boldsymbol{p}_o(t)$, such that for each $\boldsymbol{N}(t) \in$ box$(\boldsymbol{N}_1, \boldsymbol{N}_2)$, there exists a $R(t)$ such that:*

$$\boldsymbol{N}(t)\boldsymbol{P}(t) + R(t)\boldsymbol{p}_o(t) \in box(\boldsymbol{N}_1, \boldsymbol{N}_2), \tag{7}$$

then the NHMS is called stably controllable.

Definition 3 *[Hartfiel, 1994]: A vector $\boldsymbol{x} \in M_{1,k}(R)$ is called $(\alpha - \boldsymbol{Q})-$ feasible, if $\boldsymbol{xQ} \leq \alpha\boldsymbol{Q}, \alpha \in R_+$.*

Assume now that:

$$\mathbf{P}_{min} \leq \mathbf{P}(t) \leq \mathbf{P}_{max}, \forall t = 0, 1, 2, ..., \tag{8}$$

where: $\mathbf{P}_{min} = \sum_{i=1}^{k} w_{min_i}(t)\mathbf{P}_i$ and $\mathbf{P}_{max} = \sum_{i=1}^{k} w_{max_i}(t)\mathbf{P}_i$. Notice that this condition is not restrictive since in practice arbitrary movement would be highly undesirable if not impossible. Moreover, the condition applies to real applications where the exact transition probabilities cannot possibly be estimated. We assume now that $\mathbf{P}(t) = \mathbf{P}_{max}$, for some t and that the population structure $\mathbf{N}_1$ is $(1 - \mathbf{P}_{min})-$ feasible. The following theorem is now proved.

Theorem 1 *(attainability): Let a F-NHMS, which satisfies the above conditions. If:*

$$\boldsymbol{p}_o(t) = \frac{1}{R(t)}\boldsymbol{uB}$$

where:

$$\boldsymbol{u} = (\alpha_i), \forall i = 1, 2, ..., 2^k,$$

$$\boldsymbol{B} = [\boldsymbol{N}_1\boldsymbol{P}_{min}, \boldsymbol{N}_2\boldsymbol{P}_{max}] = [b_i], \forall i = 1, 2, ..., 2^k, b_i = X_i - Z_i,$$

and $\boldsymbol{N}_2$ *is* $(1 - \boldsymbol{P}_{max})-$ *feasible, then* $\boldsymbol{N}(t) \in box(\boldsymbol{N}_1, \boldsymbol{N}_2)$.

Proof. Let us assume that the structure $\mathbf{N}_2$ is $(1-\mathbf{P}_{max})-$ feasible. Therefore $\mathbf{N}_2\mathbf{P}_{max} \leq \mathbf{N}_2$. Moreover, from the hypothesis we have that $\mathbf{N}_1\mathbf{P}_{min} \leq \mathbf{N}_1$. Let $\mathbf{N}(t) \in box(\mathbf{N}_1, \mathbf{N}_2)$. Thus, from (8) we conclude that:

$$\mathbf{N}(t)\mathbf{P}(t) \in box(\mathbf{N}_1\mathbf{P}_{min}, \mathbf{N}_2\mathbf{P}_{max}). \tag{9}$$

Let X_i be the vertices of $box(\mathbf{N}_1, \mathbf{N}_2)$ and Z_i the vertices of $box(\mathbf{N}_1\mathbf{P}_{min}, \mathbf{N}_2\mathbf{P}_{max})$. It is assumed that the vertices are being numbered respectively, i.e.

$$(Z_i)_j = \begin{cases} (\mathbf{N}_1\mathbf{P}_{min})_j, \text{ iff } (X_i)_j = (\mathbf{N}_1)_j \\ (\mathbf{N}_2\mathbf{P}_{max})_j, \text{ iff } (X_i)_j = (\mathbf{N}_2)_j. \end{cases} \tag{10}$$

Given that $\mathbf{N}(t)\mathbf{P}(t) \in box(\mathbf{N}_1\mathbf{P}_{min}, \mathbf{N}_2\mathbf{P}_{max})$, the vector $\mathbf{N}(t)\mathbf{P}(t) \in box(\mathbf{N}_1\mathbf{P}_{min}, \mathbf{N}_2\mathbf{P}_{max})$ can be written as:

$$\mathbf{N}(t)\mathbf{P}(t) = \sum_{i=1}^{2^k} \alpha_i Z_i.$$

Let $\mathbf{u} = (\alpha_i)$ for $i = 1, 2, ..., 2^k$, $\mathbf{B} = [\mathbf{N}_1\mathbf{P}_{min}, \mathbf{N}_2\mathbf{P}_{max}, \mathbf{N}_1, \mathbf{N}_2] = [b_i]$ for $i = 1, 2, ..., 2^k$, where $b_i = X_i - Z_i$ and let s be the sum of the elements of the $(1 \times k)-$vector $\mathbf{uB}$. Therefore, if $\frac{1}{R(t)}\mathbf{uB}$, we have that:

$$\mathbf{N}(t)\mathbf{P}(t) + R(t)\mathbf{p}_o(t) = \sum_{i=1}^{2^k} \alpha_i Z_i + \sum_{i=1}^{2^k} \alpha_i (X_i - Z_i) = \sum_{i=1}^{2^k} \alpha_i (X_i) \tag{11}$$

i.e. $\mathbf{N}(t)\mathbf{P}(t) \in box(\mathbf{N}_1, \mathbf{N}_2)$. Therefore, $\mathbf{N}(t+1) \in box(\mathbf{N}_1, \mathbf{N}_2)$.

A necessary and sufficient condition that the system is stably controllable is given in the following theorem.

Theorem 2 *(maintainability): A F-NHMS is stably controllable if and only if the population structure* $\boldsymbol{N}_2$ *is* $(1 - \boldsymbol{P}_{max})-$ *feasible, where* $\boldsymbol{P}_{max} = \sum_{i=1}^{k} w_{max_i}(t)\boldsymbol{P}_i$.

Proof. Let us first assume that the system is stably controllable. Then, since: $\mathbf{N}_2 \in box(\mathbf{N}_1, \mathbf{N}_2)$ and $\mathbf{P}(t) = \mathbf{P}_{max}$, for some t, there exists an input vector $\mathbf{p}_o(t)$ and an $R(t)$ such that:

$$\mathbf{N}_2\mathbf{P}_{max} + R(t)\mathbf{p}_o(t) \in box(\mathbf{N}_1, \mathbf{N}_2),$$

i.e. $\mathbf{N}_2\mathbf{P}_{max} + R(t)\mathbf{p}_o(t) \leq \mathbf{N}_2$. Thus, $\mathbf{N}_2\mathbf{P}_{max} \leq \mathbf{N}_2$. Consequently, the structure $\mathbf{N}_2$ is $(1 - \mathbf{P}_{max})-$ feasible.

It is now assumed that the structure $\mathbf{N}_2$ is $(1 - \mathbf{P}_{max})-$ feasible. From Theorem 1 it follows that $\mathbf{N}(t+1) \in box(\mathbf{N}_1, \mathbf{N}_2)$. Therefore, the system is strongly controllable.

Putting the above results together, we conclude that in a F-NHMS if the structure $\mathbf{N}_2$ is $(1 - \mathbf{P}_{max})-$ feasible, then the limiting population structure given in [Symeonaki *et al.*, 2000] and [Symeonaki *et al.*, 2002] also lies between the two desired structures $\mathbf{N}_1$ and $\mathbf{N}_2$, i.e.:

$$\lim_{t\to\infty} \mathbf{N}(t) = \mathbf{N}(\infty) = T\mathbf{e}_i[\mathbf{I} - (\mathbf{I} - \sum_{i=1}^{s} v_i\mathbf{Q}_i)(\mathbf{I} - \sum_{i=1}^{s} v_i\mathbf{Q}_i)^{\sharp}] \in box(\mathbf{N}_1, \mathbf{N}_2),$$

where $(\cdot)^{\sharp}$ represents the generalized group inverse introduced in [Meyer, 1975], and $\mathbf{Q}_i = \mathbf{P}_j + \mathbf{p}'_{n+1_j}\mathbf{p}_{o_l}$ where j and l depend on i.

3 A numerical example

Let a NHMS with $S = \{1, 2, 3\}$ and let that a number of transition probabilities cannot be estimated due to lack of data. Suppose moreover that we have two factors that influence the transition probabilities. Furthermore, it is assumed that these population parameters depend upon two basic parameters. Combining the rules of the system with the generalized modus ponens (GMP) rule of inference [Klir and Yuan, 1995], [Stamou and Tzafestas, 1999] the multi-conditional approximate reasoning schema (system rules) is formulated. The system rule for the population parameter pp_1, for example, is described as follows:

1^{st} RULE: IF (x_1, x_2) IS (SMALL, LITTLE), THEN y_1 IS LOW
2^{nd} RULE: IF (x_1, x_2) IS (SMALL, AVER), THEN y_1 IS LOW
3^{rd} RULE: IF (x_1, x_2) IS (SMALL, PLENTY), THEN y_1 IS LOW
5^{th} RULE: IF (x_1, x_2) IS (MED, AVER), THEN y_1 IS AVER
6^{th} RULE: IF (x_1, x_2) IS (MED, PLENTY), THEN y_1 IS HIGH
7^{th} RULE: IF (x_1, x_2) IS (LARGE, LITTLE), THEN y_1 IS AVER
8^{th} RULE: IF (x_1, x_2) IS (LARGE, AVER), THEN y_1 IS AVER
9^{th} RULE: IF (x_1, x_2) IS (LARGE, PLENTY), THEN y_1 IS HIGH

Now let that: $\mathbf{P} = \begin{pmatrix} 0.6 & 0.1 & 0 \\ 0.1 & 0.5 & 0.1 \\ 0 & 0.1 & 0.5 \end{pmatrix}$ and $\mathbf{P}_{min} = \begin{pmatrix} 0.7 & 0.15 & 0.1 \\ 0.2 & 0.5 & 0.2 \\ 0 & 0.1 & 0.7 \end{pmatrix}$, and $\mathbf{P}_{min} \leq \mathbf{P}(t) \leq \mathbf{P}_{max}$, $\forall t = 0, 1, 2, ...$ Assume that we want to maintain the population structure of the system between the structures, $\mathbf{N}_1$ and $\mathbf{N}_2$, where: $\mathbf{N}_1 = (100\ 300\ 600)$, and $\mathbf{N}_2 = (100\ 350\ 650)$. $\mathbf{N}_1$ and $\mathbf{N}_2$ are $(1 - \mathbf{P}_{min})-$feasible and $(1 - \mathbf{P}_{max})-$feasible, respectively, since $\mathbf{N}_1\mathbf{P}_{min} = \begin{pmatrix} 90\ 220\ 330 \end{pmatrix} \leq \mathbf{N}_1$, $\mathbf{N}_2\mathbf{P}_{max} = \begin{pmatrix} 280\ 285\ 555 \end{pmatrix} \leq \mathbf{N}_2$. At time t, let:

$$\mathbf{N}(t)\mathbf{P}(t) = \begin{pmatrix} 185\ 261\ 442.5 \end{pmatrix} \in box(\mathbf{N}_1\mathbf{P}_{min}, \mathbf{N}_2\mathbf{P}_{max})$$

where $\mathbf{N}(t)\mathbf{P}(t) = \sum_{i=1}^{8} \alpha_i Z_i$, $\alpha_i = 0.125$, and X_i, Z_i are the vertices of $box(\mathbf{N}_1, \mathbf{N}_2)$ and $box(\mathbf{N}_1\mathbf{P}_{min}, \mathbf{N}_2\mathbf{P}_{max})$, respectively, as numbered in (10). Thus:

$$X_1 = \begin{pmatrix} 100\ 300\ 600 \end{pmatrix}, Z_1 = \begin{pmatrix} 90\ 220\ 330 \end{pmatrix}$$
$$X_2 = \begin{pmatrix} 100\ 300\ 650 \end{pmatrix}, Z_2 = \begin{pmatrix} 90\ 220\ 555 \end{pmatrix}$$
$$X_3 = \begin{pmatrix} 100\ 350\ 600 \end{pmatrix}, Z_3 = \begin{pmatrix} 90\ 285\ 330 \end{pmatrix}$$
$$X_4 = \begin{pmatrix} 100\ 350\ 650 \end{pmatrix}, Z_1 = \begin{pmatrix} 90\ 285\ 555 \end{pmatrix}$$
$$X_5 = \begin{pmatrix} 300\ 300\ 600 \end{pmatrix}, Z_5 = \begin{pmatrix} 280\ 220\ 555 \end{pmatrix}$$
$$X_6 = \begin{pmatrix} 100\ 300\ 650 \end{pmatrix}, Z_6 = \begin{pmatrix} 280\ 220\ 555 \end{pmatrix}$$
$$X_7 = \begin{pmatrix} 300\ 350\ 600 \end{pmatrix}, Z_7 = \begin{pmatrix} 280\ 285\ 330 \end{pmatrix}$$
$$X_8 = \begin{pmatrix} 300\ 350\ 650 \end{pmatrix}, Z_1 = \begin{pmatrix} 280\ 285\ 555 \end{pmatrix}$$

and matrix $\mathbf{B}^{'} = [b_i]^{'} = [X_i - Z_i]^{'}$ is

$$\mathbf{B}^{'} = \begin{pmatrix} 10 & 10 & 10 & 10 & 20 & 20 & 20 & 20 \\ 80 & 80 & 65 & 65 & 80 & 80 & 65 & 65 \\ 270 & 95 & 270 & 95 & 270 & 95 & 270 & 95 \end{pmatrix}$$

According to Theorems 1 and 2, if $R(t) = 291.875$ and $\mathbf{p}_o(t) = \begin{pmatrix} 0.052\ 0.284\ 0.7 \end{pmatrix}$, we have that: $\mathbf{N}(t+1) = \mathbf{N}(t)\mathbf{P}(t) + R(t)\mathbf{p}_o(t) = \begin{pmatrix} 200\ 333.5\ 646.875 \end{pmatrix} \in box(\mathbf{N}_1, \mathbf{N}_2)$ where $box(\mathbf{N}_1, \mathbf{N}_2)$ is shown in Figure 2.

References

[Hartfiel, 1994]D. J. Hartfiel. Input control in nonnegative matrix equations. *Linear Multilinear Algebra*, pages 293–304, 1994.

[Klir and Yuan, 1995]G. J. Klir and B. Yuan. *Fuzzy sets and fuzzy logic: theory and applications.* Prentice Hall PTR, USA, 1995.

[Meyer, 1975]C. D. Meyer. The role of the group generalized inverse in the theory of finite markov chains. *Siam Review*, pages 443–464, 1975.

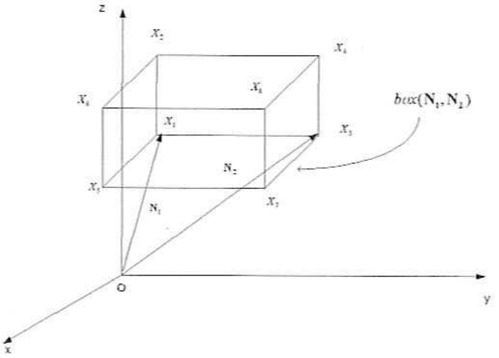


Fig. 2. $box(\mathbf{N}_1, \mathbf{N}_2)$

[Stamou and Tzafestas, 1999]G. B. Stamou and S. G. Tzafestas. Fuzzy relation equations and fuzzy inference systems: an inside approach. *IEEE Trans. on Systems, Man and Cybernetics*, pages 694–702, 1999.

[Symeonaki *et al.*, 2000]M. A. Symeonaki, G. B. Stamou, and S. G. Tzafestas. Fuzzy markov systems for the description and control of population dynamics. In S. G. Tzafestas, editor, *Computational Intelligence in Systems and Control Design and Applications*, pages 301–310, 2000.

[Symeonaki *et al.*, 2002]M. A. Symeonaki, G. B. Stamou, and S. G. Tzafestas. Fuzzy non-homogeneous markov systems. *Applied Intelligence*, pages 203–214, 2002.

[Symeonaki, 1998]M. A. Symeonaki. *Perturbations and Theory of Non-Homogeneous Markov systems.* Aristotle University of Thessaloniki, 1998.

[Vassiliou, 1982]P.-C. G. Vassiliou. Asymptotic behaviour of markov systems. *Journal of Applied Probability*, pages 851–857, 1982.

Part XVII

Internet Modelling

Modelling and analysis of Internet Pricing: introduction and challenges

Yezekael Hayel[1], Patrick Maillé[2], and Bruno Tuffin[1]

[1] IRISA/INRIA
Campus Universitaire de Beaulieu
35042 Rennes Cedex, France
(e-mail: yhayel@irisa.fr, btuffin@irisa.fr)
[2] GET/ENST Bretagne
2 rue de la Châtaigneraie
35576 Cesson Sévigné Cedex, France
(e-mail: patrick.maille@enst-bretagne.fr)

Abstract. This paper/presentation aims at introducing the reasons why switching from the current flat-rate Internet pricing to another scheme is required, at briefly classifying the existing propositions, and at highlighting the challenges that still have to be tackled in the area. Pricing has indeed become a hot topic in the networking literature in order to control congestion, differentiate services among users and somehow fairly share the resource, but is still the subject of debate about how, and even if, it should be implemented.

Keywords: Pricing, Game theory, Modelling, Optimisation.

1 Introduction: why changing?

The Internet has experienced a tremendous success during the last decade. Starting from an academic (and somewhat free) communication network, it has been expanded to commercial purposes. The way customers are currently charged is based on a so-called *flat-rate* price: they pay a fixed subscription fee to an Internet Service Provider (ISP) and have an unlimited access to the network.

Due to the success of this expansion, the amount of Internet traffic has soared in an exponential way, from the increase in the number of subscribers, but also from the more and more demanding applications used by customers, in terms of bandwidth, but also in terms of quality of service (QoS) requirements. Indeed, the proportion of telephony, video and multimedia traffic for instance is increasing with respect to data file transfer and email.

This traffic growth and diversity has highlighted the following problems of the flat-rate pricing scheme, which may therefore have become irrelevant:

i) congestion is observed, resulting in erratic QoS: longer delay, larger jitter and increase of losses. Some people argue that increasing capacity can solve the problem and that, thanks to optical fiber especially, we are safe for a while [Anania and Solomon, 1997]. This is actually the topic of the lasting debate around pricing in the networking community. We indeed believe that this argument may be true for the backbone network, but it seems unlikely to switch from copper lines to optical fiber in access networks, due to a high cost, issue also known as the *last mile* problem [Bernstein, 1997]. Moreover, in wireless networks, capacity (the radio spectrum) is and will probably remain limited.

ii) Next, a flat-rate pricing is an incentive to overuse the network: any selfish user has interest in consuming as much as possible, whatever the loss of QoS imposed on other users is. The charge is thus unfair since small users pay as much as big ones.

iii) Finally, a flat-rate pricing does not allow service differentation among users (and applications), since everybody is served at the same level, with therefore the same QoS.

As a consequence, designing a new pricing scheme is probably the most simple and natural way to cope with congestion, control demand, fairly share resources and differentiate services among users and/or applications [Courcoubetis and Weber, 2003]. The following issues have to be addressed in the design process: which families of new pricing schemes could be used (Section 2)? What are the externalities imposed on other users that have to be dealt with (Section 3)? What modelling tools and properties need to be verified (Section 4)? How do users react to prices (Section 5)? What is the trade-off between mathematical efficiency and engineering feasability (Section 6)? Section 7 also briefly addresses a new challenge in the pricing community: how do independent ISP will exchange traffic and how will they charge each other?

2 Changing to what?

Changing the simple flat-rate pricing scheme to a usage-based or congestion-based one appears thus preferable to us. Some may be worried about the acceptance of such a move, due to the current strong public preference for flat-rate, but it is likely that people will eventually get accustomed to it. Note that sophisticated pricing schemes already exist in other areas such as airfare rate [Odlyzko, 2000] or the new London city toll pricing for instance.

There is a broad range of new schemes proposed in the literature. We can sort them into:

i) pricing schemes for guaranteed services through resource reservation (using RSVP protocol for instance) and admission control (the reader may see [Paschalidis and Tsitsiklis, 2000] or [Songhurst (ed.), 1999] for instance[1]).

[1] Note that the references throughout the paper are not exhaustive but try to be as representative as possible.

ii) A promising proposal, called Paris Metro Pricing [Odlyzko, 1999], consists in partitioning the network into several logical subnetworks, each subnetwork working as the current one, but with different access charges, so that the most expensive ones would likely be less congested. Unfortunately, this proposal has been shown to be inefficient in a competitive context [Gibbens *et al.*, 2000].

iii) Another quite simple scheme is the so-called Cumulus pricing analysed in [Reichl and Stiller, 2001, Hayel and Tuffin, 2005a] where positive or negative points are awarded depending on the respect of the predefined contract, and contract renegociation (with penalties) is periodically applied.

iv) Priority pricing [Cocchi *et al.*, 1991] among different classes (at the packet level) is probably the scheme which fits the most directly the proposed DiffServ architecture. This scheduling policy has nevertheless been compared with other policies such as generalized or discriminatory processor sharing [Hayel *et al.*, 2004] [Hayel and Tuffin, 2005b] when corresponding optimal prices are used. Also, priority for rejection at buffers implementing active queue management has been studied in [Altman *et al.*, 2004].

v) Auctioning, either for priority [Marbach, 2001] or for a proportion of bandwidth [Semret, 1999] [Maillé and Tuffin, 2004] has also recently received a lot of attention.

vi) Finally, a last main group is dealing with pricing based on transfer rates and shadow prices, following. the tremendous work of Kelly *et al* [Kelly *et al.*, 1998].

3 Technologies and externalities: what to price for?

In communication networks, selfish behaviours lead to unsatisfactory outcomes because of *externalities*: the value a user gets from the network depends on the other users. As an example, in a problem of bandwidth sharing on a communication link, a user that is allocated an amount of bandwidth prevents the others from obtaining that resource, and some requests may be rejected. The externality can thus be defined as the loss of valuation a user's presence imposes on the others.

In order to drive users to behave in a more efficient way, externalities have to be taken into account when designing a pricing mechanism. Notice that externalities are often negative, but can also be positive in some cases: the most classical example in resource allocation problems is the case of multicast communications, where several users interested in the same flow have a common interest and therefore an incentive to cooperate. However they still compete against users interested in other flows.

Externalities may take different forms depending on the *technologies* used and the *performance criteria* users are sensitive to: a user willing to transfer a file will be sensitive to the entire transfer duration (losses inducing retransmissions), whereas for some real-time applications delay is the most important constraint, few losses being permitted. Considering wireless ad hoc networks transmission rates and battery consumption [Crowcroft *et al.*, 2003] are additionally critical.

To analyse properly the externalities, the mechanism designer has first to identify the limited resource, that can be bandwidth or computing capacity for wired networks; spectrum, battery and/or transmission power for wireless networks. Then the correlations between the relevant performance criteria can be studied as a function of the limited resource usage. To that extent, the technological specificities of the systems and protocols should be considered. In wireless networks for instance, the way multiple access is provided (Code, Time and/or Frequency Division Multiple Access) has an influence on the externality impact, since it determines the form of the interference that affects the performance criteria through the signal to noise ratio. For real-time applications, the scheduling policies implemented in the different nodes of the network are critical, since they highly influence the overall transmission delay.

For each performance criterion the designer focuses on, and for each communication system, externalities may have a different form, and modelling and studying them raises different problems. One important stage of mechanism design is to carefully study those problems, in order to build the right incentives (through prices) to drive the user behaviour to the desired direction.

4 Mathematical tools and properties involved

A pricing mechanism can be justified by its properties in terms of some economical criteria, such as *efficiency* (maximization of social welfare), *fairness* [Kelly *et al.*, 1998], maximization of network revenue or of the number of accepted clients... Such results need the outcome of the game to be foreseen, which implies that the user behaviour has to be predicted.

Actually, the study of users reactions to a pricing mechanism usually relies on selfishness: the users are expected to act so as to obtain the highest utility, regardless of the consequences on the others. The theoretical framework to study such problems is *game theory* [Fudenberg and Tirole, 1991], and more precisely noncooperative game theory[2]. When the mechanism is well designed, there exists a unique Nash equilibrium that predicts the outcome of the pricing game.

Game theory often implies the use of optimization. Indeed, optimization occurs at different levels:

[2] Game theory also includes the study of cooperative games, however in the context of communication networks it is not likely that users know each other and have an interest in cooperating.

- users try to maximize their utility at the outcome of the game. Depending on the problem considered, that optimization may rely on queueing theory (when delays and losses at the network nodes are the externalities), signal processing (especially in wireless networks) or other mathematical modelling tools adapted to the considered network. An important and interesting property in many pricing schemes, called *incentive compatibility*, states that a user cannot do better than following the designer point of view, that is revealing his real willingness-to-pay for quality of service or choosing the proper class in multiclass systems for instance.
- At the mechanism designer level, since the optimization from the user point of view can be predicted (from what is said in the previous item), the Nash equilibrium can be oriented to a point optimizing some desired criteria.

5 User behavior modelling

As introduced in the previous section, modelling the users' valuation of service is required and is one of the main issues of Internet pricing. Users' preferences (or levels of satisfaction) are expressed by functions called *utility* functions [Fudenberg and Tirole, 1991]. In most Internet pricing papers, the inputs of these functions are the throughput or used bandwidth, the average delay or loss ratio, more generaly the considered externality, and may depend on the type of application. In the literature, the utility functions are selected to model the real user behaviour as closely as possible, but also to verify interesting mathematical properties. Those properties are usually the continuity, differentiability and concavity, to make sure that optimal points exist and are unique [Kunniyur and Srikant, 2003].

Nevertheless, one main challenge is to determine a realistic expression of the utility function (or its distribution over a population). For real-time applications for instance, one would expect non-continuous functions, with thresholds under which the utility of being served becomes null. Very few attempts have been published to solve this question. The only cases we are aware of are as follows. In [Beckert, 2000], the utility function is modelled by a Cobb-Douglas function, which requires the determination of several parameters. These parameters have been estimated using a large-scale experiment testing user behaviour which has been performed at UC Berkeley, called the INDEX project [Edell and Varaiya, 1999]. Another worth-mentioning paper is [Gupta *et al.*, 1998], where a quasi-bayesian update algorithm is developped, aiming at estimating the users' waiting cost per unit of time. This approach can be used to estimate the demand elasticity with respect to prices.

To sum up the section, the choice of utility function has a major impact on the pricing scheme analysis, and should be based not only on mathematical interest, but on practical reality (a usual trade-off in modelling). We now deal with another important trade-off between mathematical and engineering efficiencies.

6 Trade-off between mathematical and engineering efficiencies

Indeed, to obtain a more efficient model, it is often required that prices react dynamically and instantaneously to an externality evolution, so that the system can be continuously kept at its optimal point. Nevertheless, this requires an important signalling overhead, and is difficult to implement from an engineering point of view (at this point, it is important to emphasize that a main reason of the Internet success is its simplicity, which has to be preserved). It is also important to note that, following the previous section, even users are skeptical with respect to a dynamic pricing, as highlighted by the INDEX project [Edell and Varaiya, 1999]. Again, those trade-offs are important issues a designer has to cope with.

It is interesting to note that a good approximation to dynamic pricing is *time-of-day* pricing. It has been shown in [Paschalidis and Tsitsiklis, 2000] that it leads to an asymptotically efficient scheme, while being simple from an engineering and user point of view. Time-of-day pricing is popular in many areas such as telephony, airfare (where it is rather time-of-year)...

The efficiency problem can also be placed at other levels:

- from a mathematical point of view, the efficiency is more easy to reach if charges are applied at each node of the network (or at least for the whole path). This again induces a signalling overhead in terms of accounting (for total charges have to be computed before being billed to the users), but also requires to inform the user in order to make him accept the transaction. A simpler trend is to charge users at the edges of the network, even if it seems difficult to abstract the network status at the edges in an efficient way (especially if the considered traffic does not pass through the existing bottlenecks).
- Also, applying resource reservation (that is making sure that when your session is accepted, you will get a given QoS for the whole duration of your connection) is appealling mathematically and from the user side point of view, but is intricate to apply to a large network of the Internet size. Scalability is thus the reason why the IntServ architecture initially proposed for Internet QoS has gradually taken place to the DiffServ proposal, where no strict reservation is applied.

7 A new challenge: inter-providers pricing

A pricing game that even the opponents of Internet pricing admit to be mandatory is pricing among ISPs in order to deliver their own traffic. Indeed, concurrent ISPs are in competition in the Internet and have to meet traffic forwarding agreements in order to convey their messages to destination if it is not in their network.

A natural way to apply this is to implement auctions between providers and to use, and extend, the Border Gateway Protocol (BGP) usually applied for routing [Feigenbaum *et al.*, 2002]. The goal is then to find lowest-cost routing for sending traffic from an ISP to another that is not directly attainable thanks to BGP. By using VCG auctions, incentive compatibillity can be obtained.

Similarly, pricing for transiting traffic between ISPs and pricing for customers has been jointly studied in [Shakkottai and Srikant, 2005]. Repeated games are used, and, with threat strategies, optimality is shown in the sense that deviating from the equilibrium makes you suffer the worst possible penalty.

References

[Altman *et al.*, 2004]E. Altman, D. Barman, R. El Azouzi, D. Ros, and B. Tuffin. Pricing Differentiated Services: A Game-Theoretic Approach. In *Proceedings of IFIP/TC6 Networking Conference*, Athens, Greece, May 2004.

[Anania and Solomon, 1997]L. Anania and R.J. Solomon. Flat- The Minimalist Price. In Lee W. McKnight and Joseph P. Bailey, editors, *Internet Economics*, pages 91–118. MIT Press, 1997.

[Beckert, 2000]W. Beckert. *Stochastic Demand Analysis.* PhD thesis, UC Berkeley, 2000.

[Bernstein, 1997]L. Bernstein. Managing the last mile. *IEEE Communications Magazine*, 35(10):72–76, Oct 1997.

[Cocchi *et al.*, 1991]R. Cocchi, D. Estrin, S. Shenker, and L. Zhang. A Study of Priority Pricing in Multiple Service Class Networks. In *Proceedings of SIGCOMM'91*, pages 123–130, 1991.

[Courcoubetis and Weber, 2003]C. Courcoubetis and R. Weber. *Pricing Communication Networks—Economics, Technology and Modelling.* Wiley, 2003.

[Crowcroft *et al.*, 2003]J. Crowcroft, R. Gibbens, F. Kelly, and S. Östring. Modelling incentives for collaboration in mobile ad hoc networks. In *Proceedings of WiOpt'03*, Mar 2003.

[Edell and Varaiya, 1999]R. Edell and P. Varaiya. Providing Internet Access: What We Learn From INDEX. In *Proceedings of INFOCOM*, 1999.

[Feigenbaum *et al.*, 2002]J. Feigenbaum, C. Papadimitriou, R. Sami, and S. Shenker. A BGP-based mechanism for lowest-cost routing. In *Proceedings of the twenty-first annual symposium on Principles of distributed computing (PODC'02)*, pages 173–182. ACM Press, 2002.

[Fudenberg and Tirole, 1991]D. Fudenberg and J. Tirole. *Game Theory.* MIT Press, Cambridge, Massachusetts, 1991.

[Gibbens *et al.*, 2000]R. Gibbens, R. Mason, and R. Steinberg. Internet service classes under competition. *IEEE Journal on Selected Areas in Communications*, 18(12):2490–2498, 2000.

[Gupta *et al.*, 1998]A. Gupta, B. Jukic, M. Li, D. Stahl, and A. Whinston. Estimating Internet User's Demand Characteristics. Technical report, National Science Foundation, 1998.

[Hayel and Tuffin, 2005a]Y. Hayel and B. Tuffin. A Mathematical Analysis of the Cumulus Pricing Scheme. *Computer Networks (To appear)*, 2005.

[Hayel and Tuffin, 2005b]Y. Hayel and B. Tuffin. Pricing for Heterogeneous Services at a Discriminatory Processor Sharing Queue. In *Proceedings of IFIP/TC6 Networking Conference*, Waterloo, Canada, May 2005.

[Hayel *et al.*, 2004]Y. Hayel, D. Ros, and B. Tuffin. Less-than-Best-Effort Services: Pricing and Scheduling. In *proceedings of IEEE INFOCOM*, 2004.

[Kelly *et al.*, 1998]F.P. Kelly, A.K. Mauloo, and D.K.H. Tan. Rate control in communication networks: shadow prices, proportional fairness and stability. *Journal of the Operational Research Society*, 49:237–252, 1998.

[Kunniyur and Srikant, 2003]S. Kunniyur and R. Srikant. End-to-End Congestion Control Schemes: Utility Functions, Random Losses and ECN Marks. *IEEE Transactions on Networking*, 11(5), 2003.

[Maillé and Tuffin, 2004]P. Maillé and B. Tuffin. Multi-bid auctions for bandwidth allocation in communication networks. In *IEEE INFOCOM 2004*, Hong-Kong, China, March 2004.

[Marbach, 2001]P. Marbach. Pricing Differentiated Services Networks: Bursty Traffic. In *Proceedings of IEEE INFOCOM 2001*, 2001.

[Odlyzko, 1999]A. Odlyzko. Paris Metro Pricing for the Internet. In *ACM Conference on Electronic Commerce (EC'99)*, pages 140–147, 1999.

[Odlyzko, 2000]A. Odlyzko. The history of communications and its implications for the Internet. Technical report, AT&T Labs, 2000.

[Paschalidis and Tsitsiklis, 2000]I.Ch. Paschalidis and J.N. Tsitsiklis. Congestion-Dependent Pricing of Network Services. *IEEE/ACM Transactions on Networking*, 8(2):171–184, 2000.

[Reichl and Stiller, 2001]P. Reichl and B. Stiller. Edge pricing in space and time: Theoretical and practical aspects of the cumulus pricing scheme. In *Proceedings of the 17th International Teletraffic Congress*, 2001.

[Semret, 1999]N. Semret. *Market Mechanisms for Network Resource Sharing*. PhD thesis, Columbia University, 1999.

[Shakkottai and Srikant, 2005]S. Shakkottai and R. Srikant. Economics of network pricing with multiple ISPs. In *Proceedings of IEEE INFOCOM 2005*, Miami, FL, USA, 2005.

[Songhurst (ed.), 1999]D. Songhurst (ed.). *Charging Communication Networks: from Theory to practice*. Elsevier, Amsterdam, 1999.

Asymptotic behavior of a GPRS/EDGE network with several cells controlled by a global capacity limit

Georges Nogueira[1], Bruno Baynat[1], and Pierre Eisenmann[2]

[1] Laboratoire d'informatique de Paris 6
LIP6 - CNRS
Paris, FRANCE
(e-mail: Georges.Nogueira@lip6.fr, Bruno.Baynat@lip6.fr)

[2] Nortel Networks
Wireless Network Engineering
Chateaufort, FRANCE
(e-mail: pierree@nortelnetworks.com)

Abstract. This paper is a contribution to the generic problem of having simple and accurate models to dimension radio cells with data traffic on a GPRS or EDGE network. It addresses the issue of capacity limitation in a given cell due to coupling with other cells because of a central equipment or transmission link of limited capacity. A mobile can't access the cell although it is alone, because the capacity limit is reached due to traffic on other cells. Our purpose is to extend our previously published Erlang-like law for data traffic to the constrained multiple-cell system and to derive asymptotic developments for all the average performance parameters that are necessary for the dimensioning of the multiple-cell system.

Keywords: GPRS, EDGE, modeling, performance evaluation, dimensionning, discrete-time Markov chain, Erlang, group of cells.

1 Introduction

GPRS is an overlay on GSM networks that allows end-to-end IP-based packet traffic from the terminal to e.g. the Internet. EDGE is an improvement over GPRS whereby the modulation scheme on radio is modified to allow higher throughputs thanks to advanced power amplifier and signal processing technologies. In a GPRS (or EDGE) cell, traffic is split between voice (on circuit) and data (on packet). Data uses a few dedicated circuits which are decomposed into 20 ms ÒblocksÓ carrying elementary packet traffic. The packet-based traffic is managed by the PCU (Packet Control Unit), a standardized network element in charge of the MAC layer (multiplexing of mobiles) and RLC layer (decomposition into elementary blocks and retransmission when radio errors occur). The PCU is connected to the SGSN (Serving GPRS Support Node) which manages the end-user mobility and hides it to the external world. It is linked to the edge router, called GGSN, by an IP tunnel in which traffic is encapsulated. The GGSN is the fixed anchor point to the Internet or service platforms, and a user may change SGSN while going from cell to cell. In this end-to-end chain, it is possible to have traffic limitation in an element in charge of managing several cells (typically a PCU or a SGSN module or a transmission link).

Several research works have investigated the analytical modeling of GPRS systems. Most of the studies develop complex Markovian models that require a numerical resolution in order to evaluate the system performance (see e.g. [Fang and Ghosal, 2003], [Lindemann and Thümmler, 2003], [Vornefeld, 2002], [Foh *et al.*, 2001]). Some of them use approximations to derive closed-form expressions (see e.g. [Ni and Häggman, 1999] and [Pedraza *et al.*, 2002]). However, these studies always focus on a single cell. In [Baynat and Eisenmann, 2004] we have developed a discrete-time Markov chain model for single-cell GPRS/EDGE network engineering. The model captures the main features of the GPRS/EDGE radio resource allocation and assumes an ON/OFF traffic (with infinite sessions) performed by a finite number of users over the cell. The Markov chain is simplified by Taylor series expansion and a simple and accurate Erlang-like law is obtained. Extensions to finite-length sessions traffic has been developed in [Baynat *et al.*, 2004]. In [Nogueira *et al.*, 2005] we study the impact of a capacity limitation imposed upon a group of cells. We assume it can be expressed by a maximum number of concurrent downlink transfers that are allowed in the group of cells. When this limit is reached, any transfer request on any of the cells will be rejected. However, even if the performance parameters can be obtained almost instantaneously with a very good accuracy, [Nogueira *et al.*, 2005] does not provide closed-form expressions. The goal of this paper is to further simplify the performance evaluation of a multiple-cell system, by deriving very simple asymptotic developments. Once again our objective is to obtain closed-form ÒErlang-likeÓ expressions to efficiently help network engineering.

Section 2 addresses single-cell systems, the basic hypotheses and the main results of the Erlang-like model developed in [Baynat and Eisenmann, 2004] are recalled. Then asymptotic developments are given for both low and high load cases. Section 3 deals with multiple-cell systems. In this section we quickly recall the principal steps developed in [Nogueira *et al.*, 2005] for multiple-cell systems. Asymptotic developments are then made to obtain closed-form expressions for the performance parameters. Section 4 presents numerical results.

2 Single cell system

2.1 System description

Our study is focused on the analysis of the bottleneck i.e. the radio link, studied in a particular cell. We are focused on the downlink, assumed to be the critical resource because of traffic asymmetry. We assume that the allocator fairly shares bandwidth between all active mobiles (no QoS is modeled so far).

We also make the following assumptions: there is a fixed number T of time-slots in the cell that are dedicated to GPRS; these time-slots are using a single TDMA. All mobiles have the same reception capability; they are "$(d+u)$" where d is the number of time-slots that can simultaneously be used for the downlink traffic and u is the number of time-slots that can simultaneously be used for the uplink traffic. Note that, as we are only interested in the modeling of the downlink traffic, only the parameter d is relevant for the model. This assumption is presently realistic as most of the time, less than four time-slots are reserved for GPRS ($T \leq 4$). Extensions to the case where $T > d$ are currently under investigation. Note that with this assumption, the parameter $n_0 = \lfloor \frac{T}{d} \rfloor$ used in [Baynat and Eisenmann, 2004] is such that $n_0 \leq 1$.

Our GPRS system is characterized by the following parameters:

- t_B: the system elementary time interval equal to the radio block duration, i.e. $t_B = 20$ ms;
- x_B: number of data bytes that are transferred during t_B over one time-slot. x_B/t_B is the throughput offered by the RLC/MAC layer to the LLC layer. The value of x_B depends on the radio coding scheme [Baynat and Eisenmann, 2004]. As an example, for GPRS CS2, $x_B = 30$ bytes;
- tbf_{max}: maximum number of mobiles that can simultaneously have an active downlink TBF (Temporary Block Flow). This limitation is due to the system hardware characteristics and ensures a minimum throughput per mobile (a TDMA can't be indefinitely shared).

2.2 Markovian analysis

Traffic is modeled as follows. There is a fixed number N of GPRS/EDGE mobiles in the cell, each of them doing an ON/OFF traffic with an infinite number of pages:

- ON periods correspond to the download of an element (a WAP, a WEB page, an email, a file, etc.). Its size is characterized by a discrete random variable X_{on}, with an average value of x_{on} bytes;
- OFF periods correspond to the reading time, which is modeled as a continuous random variable T_{off}, with an average value of t_{off} seconds.

Let us emphasize that there is a limitation $n_{max} = \min(tbf_{max}, N)$ on the number of mobiles that can simultaneously be on active transfer. It involves both the system constraint tbf_{max} and the total mobile population.

[Baynat and Eisenmann, 2004] develops an analytical model for the performance evaluation of a single-cell GPRS system. The smallest time-scale of the system, namely the radio block duration $t_B = 20$ ms, has been accounted for in the modeling process, by developing a discrete-time Markovian model of equal elementary time interval. The model assumes that both the size of ON periods and the duration of OFF periods have memoryless distributions.

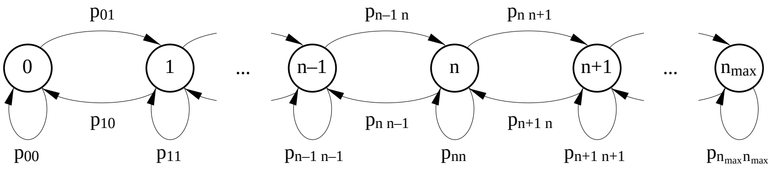


Fig. 1. Linear model

Several models with several levels of approximation have been developed. The simpler one makes the assumption that more than one mobile switching from one state (ON or OFF) to the other one during t_B is negligible, which transforms the model into the discrete-time birth-death process given in Fig. 1. As shown in [Baynat and Eisenmann, 2004], the stationary probabilities of having n mobiles in active transfer in the cell can easily be derived from this linear Markov chain. By further using Taylor series expansions, these probabilities can be expressed as a function of a single dimensionless parameter x as:

$$p(n) = \frac{N!}{T^n(N-n)!} x^n p(0) \qquad 0 \leq n \leq n_{max} \tag{1}$$

where x is given by

$$x = \frac{t_B\, x_{on}}{x_B\, t_{off}} \tag{2}$$

and $p(0)$ is obtained by normalization. Note that relation (1) is a simplification of the expression given in [Baynat and Eisenmann, 2004] that takes into account the fact that $n_0 \leq 1$.

The performance parameters of the cell can be derived from the stationary probabilities as follows [Baynat and Eisenmann, 2004]. The Ònormalized utilizationÓ $\tilde{U}$ of the cell, i.e. the mean number of time-slots used for GPRS, is directly obtained as:

$$\tilde{U} = T \sum_{n=1}^{n_{max}} p(n) = T\,(1 - p(0)) \tag{3}$$

The Ònormalized throughputÓ $\tilde{X}$, i.e. the average number of time-slots given to a mobile for its transfers is given by:

$$\tilde{X} = T \frac{\sum_{n=1}^{n_{max}} p(n)}{\sum_{n=1}^{n_{max}} n\, p(n)} \tag{4}$$

Finally, the ÒblockingÓ (or ÒrejectÓ) probability P_r, i.e. the probability that a mobile that wants to start the download of a new page cannot do it because the limit of n_{max} mobiles in the cell is reached, is obtained as:

$$P_r = 1 - \frac{T}{x} \frac{\sum_{n=1}^{n_{max}} p(n)}{\sum_{n=0}^{n_{max}} p(n)(N-n)} \tag{5}$$

2.3 Asymptotic analysis

We first define the quantity $v = x/T$. We then rewrite the expression of the steady-state probabilities and the performance parameters (relations (1) to (5)) as a function of v:

$$p(n) = \frac{v^n N(N-1)(N-n+1)}{1 + vN + v^2 N(N-1) + ... + v^{n_{max}} N(N-1)...(N-n_{max}+1)} \tag{6}$$

$$\tilde{U} = T \frac{vN + v^2 N(N-1) + ... + v^{n_{max}} N(N-1)...(N-n_{max}+1)}{1 + vN + v^2 N(N-1) + ... + v^{n_{max}} N(N-1)...(N-n_{max}+1)} \tag{7}$$

$$\tilde{X} = T \frac{1 + v(N-1) + ... + v^{n_{max}-1}(N-1)...(N-n_{max}+1)}{1 + 2v(N-1) + ... + n_{max} v^{n_{max}-1}(N-1)...(N-n_{max}+1)} \tag{8}$$

$$P_r = \frac{v^{n_{max}}(N-1)...(N-n_{max})}{1 + v(N-1) + v^2(N-1)(N-2) + ... + v^{n_{max}}(N-1)...(N-n_{max})} \tag{9}$$

From these expressions, we can easily obtain the asymptotes for low and high load. Note that the quantity vN characterizes in an aggregate way the load of the system, as it increases when the size of the downloaded pages or the number of mobiles increase, and decreases when the number of time-slots dedicated to GPRS or the reading time of a page increase. It is in fact the equivalent to the Erlang load factor for a finite number of users doing ON/OFF sessions in data traffic.

When $vN \ll 1$ (low load), we directly obtain from the previous performance expressions the following asymptotic developments:

$$\tilde{U} \approx TvN = xN; \qquad \tilde{X} \approx T(1 - v(N-1)); \qquad P_r \approx 0 \tag{10}$$

When $vN \gg 1$ (high load), we get the following asymptotic developments:

$$\tilde{U} \approx T; \qquad \tilde{X} \approx \frac{T}{n_{max}}; \qquad P_r \approx 1 - \frac{1}{v(N - n_{max})} \tag{11}$$

3 Multiple cell system

3.1 System description

We now assume, as described in Section 1 and [Nogueira *et al.*, 2005], that traffic may be limited because of a capacity constraint in a network element that controls traffic over a group of P cells. Let M_{max} be the total number of mobiles that can currently be in active transfer in the P cells. In order to simplify the derivation of the asymptotes, we assume that the P cells are identical. In each cell, there is a fixed number N of GPRS mobiles generating an ON/OFF traffic as described in Section 3.2 and having the same characteristics (the average page size is x_{on} and the average reading time is t_{off} for all mobiles in all cells). Note however that non identical cells can be handled by the model [Nogueira *et al.*, 2005]. Of course, if $\sum_{i=1}^{P} n_{max}^{i} \leq M_{max}$, the limit does not generate any additional constraint over the system, and each cell can thus be analyzed using the single-cell model described in Section 2. As a consequence, we only consider here the case where $\sum_{i=1}^{P} n_{max}^{i} > M_{max}$. In such a system, a mobile in a given cell i will not be able to start a new transfer, not only because the cell capacity (n_{max}^{i}) is reached, but also because the global system capacity (M_{max}) is reached.

3.2 Model description

We consider a particular cell of the system. As all cells are identical, without loss of generality we can focus on the last one, cell P. When needed, a superscript i will be added to the notations to refer to the parameters of a cell i. The first step of the analysis consists in developing the so-called Òaggregate Markov chainÓ [Nogueira *et al.*, 2005] associated with the considered cell (cell P). As shown in Fig. 2 This aggregate model has the same structure as the single-cell linear Markov chain model (Fig. 1), but the transition between any state n and state $n+1$ is now multiplied by a factor $(1 - r(n))$. $r(n)$ is the probability that an inactive mobile in cell P that wants to start a new transfer cannot do it because the system limit M_{max} is reached, assuming that there are n mobiles currently in active transfer in the cell.

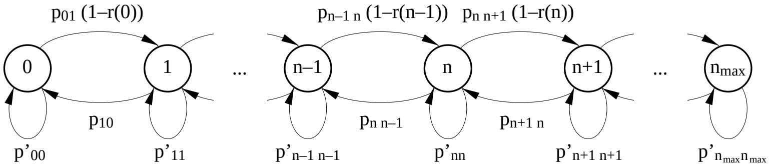


Fig. 2. Aggregate model

As a consequence, $r(n)$ is the probability that the system is full when there are n mobiles in the considered cell, and can thus be estimated by the probability that the $P-1$ other cells (cells 1 to $P-1$) contain $M_{max} - n$ mobiles. It is shown in [Nogueira *et al.*, 2005] that the probabilities $r(n)$ can be estimated by the following expression:

$$r(n) = \frac{p_{uc}^{\{1,\ldots,P-1\}}(M_{max} - n)}{\sum\limits_{k=0}^{M_{max}-n} p_{uc}^{\{1,\ldots,P-1\}}(k)} \tag{12}$$

where the probabilities $p_{uc}^{\{1,...,P-1\}}(k)$ are obtained by convolution over the steady-state probabilities of the $P-1$ other "*unconstrained*" cells, where unconstrained cell i is a virtual cell having the same characteristics as cell i but that is not subjected to the overall constraint M_{max} (see [Nogueira *et al.*, 2005] for derivations):

$$p_{uc}^{\{1,...,P-1\}}(k) = \sum_{\substack{n^1 + ... + n^{P-1} = k \\ n^j \leq n_{max} \\ \forall j = 1,...,P-1}} \left(\prod_{j \in \{1,...,P-1\}} p_{uc}^j(n^j) \right) \tag{13}$$

We can then inject the $r(n)$ parameters in the aggregate model and analyze it. The resulting steady-state probabilities of the aggregate model are thus:

$$p_{agg}(n) = \frac{N!}{T^n (N-n)!} x^n \left(\prod_{k=0}^{n-1} (1 - r(k)) \right) p_{agg}(0) \qquad 0 \leq n \leq n_{max} \tag{14}$$

where x is given by (2) and $p_{agg}(0)$ is obtained by normalization.

Finally, we can derive the normalized utilization $\tilde{U}$ of any cell as well as the normalized throughput $\tilde{X}$ offered to a mobile for its transfers and the blocking probability P_r, from relations (3), (4) and (5), by replacing the probabilities $p(n)$ by the aggregate probabilities $p_{agg}(n)$ obtained from relation (14).

3.3 Asymptotic analysis

In this section we are only interested in the high load case ($vN \gg 1$ and $P \gg 1$). Indeed, in the low load case ($vN \ll 1$), the system constraint does not affect the behavior of the system, and the asymptotes are those of the single-cell case developed in Section 2.3.
We first develop the expression of the probability $p_{uc}^{\{1,...,P-1\}}(k)$:

$$p_{uc}^{\{1,...,P-1\}}(k) = \Psi(k) \left(p_{uc}^i(0) \right)^{P-1} \tag{15}$$

where $\Psi(k)$ comes from relation (13):

$$\begin{aligned}\Psi(k) &= \sum_{\substack{n^1 + ... + n^{P-1} = k \\ n^j \leq n_{max} \\ \forall j = 1,...,P-1}} \left(v^{n^j} N...(N - n^j + 1) \right) ... \left(v^{n^{P-1}} N...(N - n^{P-1} + 1) \right) \\ &= \sum_{\substack{n^1 + ... + n^{P-1} = k \\ n^j \leq n_{max} \\ \forall j = 1,...,P-1}} v^k \left(N...(N - n^j + 1) \right) ... \left(N...(N - n^{P-1} + 1) \right) \qquad (16)\end{aligned}$$

When $vN \gg 1$, we can give the following first order approximation for $p_{uc}^{\{1,\dots,P-1\}}(k)$:

$$p_{uc}^{\{1,\dots,P-1\}}(k) \approx \mathrm{C}_{k+P-2}^{k}(vN)^{k}\left(p_{uc}^{i}(0)\right)^{P-1} \tag{17}$$

Even if it is not exact, we have empirically checked that when this expression is replaced into the $r(n)$ probabilities, it results in a very good approximation:

$$r(n) = \frac{p_{uc}^{\{1,\dots,P-1\}}(M_{max}-n)}{\displaystyle\sum_{j=0}^{M_{max}-n} p_{uc}^{\{1,\dots,P-1\}}(j)} \approx \frac{\mathrm{C}_{M_{max}-n+P-2}^{M_{max}-n}(vN)^{M_{max}-n}}{\displaystyle\sum_{j=0}^{M_{max}-n} \mathrm{C}_{j+P-2}^{j}(vN)^{j}} \tag{18}$$

By replacing the development of the $r(n)$ probabilities into the aggregate model, we get the aggregate steady-state probabilities. Here we only give the expression of $p_{agg}(1)$ and $p_{agg}(2)$:

$$p_{agg}(1) \approx \frac{M_{max}}{M_{max}+P-2}p_{agg}(0); \qquad p_{agg}(2) \approx \frac{N-1}{N}\frac{(M_{max}-1)}{(M_{max}+P-3)}p_{agg}(1) \tag{19}$$

When $P \gg 1$, it appears that the probabilities $p_{agg}(n)$ decrease very fast with n: $p_{agg}(0) \gg p_{agg}(1) \gg p_{agg}(2) \gg \dots$. As a consequence we can obtain the asymptotic developments for the performance parameters by only taking into account the preponderant values of the aggregate probabilities in the summation of expressions (3), (4) and (5). By doing that, we obtain the following expressions for the performance parameters of any cell i:

$$\tilde{U}^{i} \approx T p_{agg}(1) \approx T\frac{M_{max}}{M_{max}+P-2} \tag{20}$$

$$\tilde{X}^{i} \approx T\frac{1+\frac{p_{agg}(2)}{p_{agg}(1)}}{1+2\frac{p_{agg}(2)}{p_{agg}(1)}} \approx T\left(1-\frac{N-1}{N}\frac{(M_{max}-1)}{(M_{max}+P-3)}\right) \tag{21}$$

$$P_{r}^{i} \approx 1-\frac{p_{agg}(1)}{vNp_{agg}(0)} \approx 1-\frac{M_{max}}{vN(M_{max}+P-2)} \tag{22}$$

4 Numerical Results

In this section, we compare the asymptotic developments to the results obtained with the analytical models developed in [Baynat and Eisenmann, 2004] and [Nogueira *et al.*, 2005]. The constraint M_{max} is set to 40. The number T of time-slots reserved to GPRS traffic in each cell is chosen equal to 2. We have tested different values for T (1 to 4) and M_{max} (20 to 100), and the results obtained were very similar to those shown here. All mobiles are assumed to be able to use the T time-slots of the TDMA ($d \geq T$) and generate the same traffic load.

4.1 Influence of the mobile population per cell

First, we investigate the influence of the GPRS mobile population. We set the number of cells to $P = 30$, the dimensionless parameter to $x = 0.268$ (corresponding e.g. to $x_{on} = 4000$ bytes, $t_{off} = 10$ s and $x_B = 30$ bytes), and vary the number N of GPRS mobiles in each cell. We compare the normalized utilization $\tilde{U}$ of a cell, the normalized throughput $\tilde{X}$ offered to a mobile and the blocking probability P_r obtained by the analytical model to those derived from the asymptotic developments. As shown in Fig. 3, 4 and 5, the asymptotic curves are made of two parts. First, when N is low, the system is almost empty, i.e. the bottleneck is not reached. Every cell has the same behavior as if it were alone in the system. Thus, we use the asymptotic expressions (10) developed for low traffic load in single cells. Second, when N is high, the system is saturated because of the important traffic load. Thus, we use the asymptotic expressions (20), (21) and (22) developed for high traffic load in a multiple-cell system. Fig. 3, 4 and 5 show the very good fit between the asymptotic and the analytical curves, for both low and high traffic load. In both cases, the cells performance reach the asymptotic limit quickly. These expressions given by two simple functions are very useful to quickly analyze the qualitative behavior and quantitative bounds on the system performance.

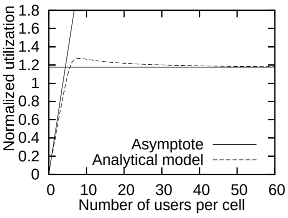


Fig. 3. $\tilde{U}$ function of N

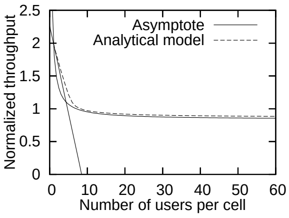


Fig. 4. $\tilde{X}$ function of N

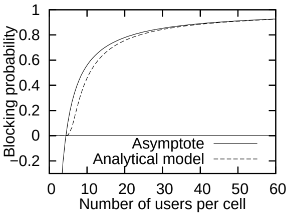


Fig. 5. P_r function of N

4.2 Influence of the number of cells

We now focus on the influence of the number P of cells in the system. We study the system performance evolution for different traffic load profiles ($vN = 0.13, 0.27, 0.54, 1.07$). The asymptotic curves are obtained from relations (20), (21) and (22). We notice a systematic behavior on analytical utilization and throughput curves (Fig. 6 and 7). They nearly follow a horizontal line until they reach the asymptotic curve. The horizontal line corresponds to the performance parameter of the cell that is not subject to the capacity constraint and that can thus be analyzed with the single-cell model of Section 2. Performance remains unchanged when P increases until the capacity constraint starts influencing the cell. We can thus cut out the construction of the performance curves as follows:

i) compute the reference performance parameter for a single-cell system;

ii) draw the asymptote for high traffic load in multiple-cell system;
iii) bind the reference point to the asymptote with a horizontal line.

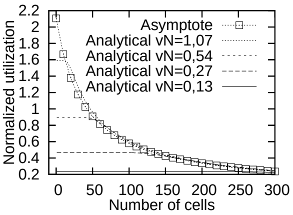


Fig. 6. $\tilde{U}$ function of P

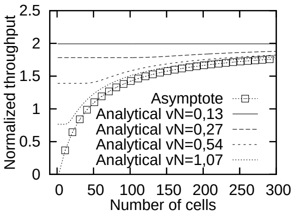


Fig. 7. $\tilde{X}$ function of P

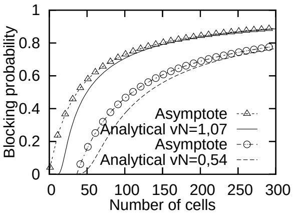


Fig. 8. P_r function of P

5 Conclusion

We have first been able to provide a computationally simple model of the a priori complex system made of a group of cells in a cellular network coupled by capacity limitation in a centralized equipment handlink packet traffic. This model has been further simplified by developing asymptotic expansions for low and high load traffic. The resulting close-form Erlang-like expressions that have been derived allow the construction of even simpler dimensioning models. Indeed, in a dimensioning situation, the problem consists in finding the optimal input system parameter that fulfill a given performance criterion. Our proposal offers simple functions that can easily be inverted in order to obtain directly the required solution without any iteration process. The complexity of such problems is thus drastically reduced.

References

[Baynat and Eisenmann, 2004]B. Baynat and P. Eisenmann. "Towards an Erlang-Like formula for GPRS/EDGE network engineering". In *IEEE International Conference on Communications (ICC)*, 2004.

[Baynat *et al.*, 2004]B. Baynat, K Boussetta, P. Eisenmann, and N. Ben Rached. "Towards an Erlang-Like formula for the performance evaluation of GPRS/EDGE networks with finite-length sessions". In *3rd IFIP-TC6 Networking Conference*, 2004.

[Fang and Ghosal, 2003]Fang and D. Ghosal. "Performance Modeling and QoS Evaluation of MAC/RLC Layer in GSM/GPRS Networks". In *Proc. of IEEE International Conference on Communications*, 2003.

[Foh *et al.*, 2001]C. H. Foh, B. Meini, B. Wydrowski, and M. Zukerman. "Modeling and Performance Evaluation of GPRS". In *Proc. of IEEE VTC*, 2001.

[Lindemann and Thümmler, 2003]Ch. Lindemann and A. Thümmler. "Performance Analysis of the General Packet Radio Service". In *Computer Networks*, pages 1–17, 2003.

[Ni and Häggman, 1999]S. Ni and S. Häggman. "GPRS performance estimation in GSM voice and GPRS shared resource system". In *Proc. of IEEE Wireless Communication and Networking Conference(WCNC)*, pages 1417–1421, 1999.

[Nogueira *et al.*, 2005]G. Nogueira, B. Baynat, and P. Eisenmann. "An analytical model for the dimensioning of a GPRS/EDGE network with a capacity constraint on group of cells". *Unpublished, Technical report in progress*, 2005.

[Pedraza *et al.*, 2002]S. Pedraza, J. Romero, and J. Muoz. "(E)GPRS Hardware Dimensioning Rules with Minimum Quality Criteria". In *Vehicular Technology Conference*, pages 391–395, 2002.

[Vornefeld, 2002]U. Vornefeld. "Analytical Performance Evaluation of Mobile Internet Access via GPRS Networks". In *European Wireless*, 2002.

Estimation of the Memory index of transmission rate measurements using an Infinite Source Poisson model

Gilles Faÿ[1], François Roueff[2], and Philippe Soulier[3]

[1] UST de Lille - CNRS UMR8524
U.F.R. de Mathématiques - Bât. M2
59655 Villeneuve d'Ascq Cedex, France
(e-mail: Gilles.Fay@univ-lille1.fr)

[2] GET - Télécom Paris - CNRS LTCI
46 rue Barrault,
75634 Paris Cedex, France
(e-mail: roueff@tsi.enst.fr)

[3] Université de Paris X - Equipe ModalX - UFR SEGMI
200 avenue de la République,
92001 Nanterre Cedex, France
(e-mail: philippe.soulier@u-paris10.fr)

Abstract. We present long memory processes related to some point processes, give their main properties, asymptotic behaviour and discuss some statistical issues with a view on Internet traffic measurements. The Infinite Source Poisson model is a generalisation of the M/G/∞ queue. Arrivals are driven by a homogeneous Poisson process, durations of active periods are independent and identically distributed (iid) and independent of the arrivals. Each active periods (say dowload sessions) is assumed to have a constant transmission rate and the available bandwidth to be unlimited. Theses rates are iid, independent of the arrivals but possibly depending on the durations. In a traffic modelling context, the obtained process $X(t)$ can serve for modelling the bandwith occupation, often called the *workload*. The stability of the model depends on the tail behavior of the duration distribution. Both in the stable and unstable cases, the tail behavior of the durations can be recovered from the dependence structure of $X(t)$. In particular, heavy-tails durations will result in long range dependence (LRD) in $X(t)$ and the corresponding tail and Hurst indices α and H satisfy $H = (3-\alpha)/2$ for all $\alpha \in (0, 2)$. In practical situations, the process $X(t)$ is observed through passive measurements, by counting packets going trough a point of the network, and then by evaluating the instantaneous workload. Such measurements are much simpler than collectiong complete characterizations of flows. However, from a queuing point of view, as mentionned above about the stability, the important parameter is α. The object of this paper is to rely on the relationship between α and H for estimating α from measurements on $X(t)$.

Keywords: Infinite Source Poisson Model, Heavy tails and long range dependence, Traffic modelling.

1 Modelling transmission rates

We consider the Infinite Source Poisson model with random transmission rate defined by

$$X(t) = \sum_{j \in \mathbb{N}} U_j \mathbf{1}_{\{t_j \leq t < t_j + \eta_j\}} \ . \tag{1}$$

The transmissions are generated at birth times $\{t_j\}$ which are the points of a unit rate homogeneous Poisson process on the positive half-line and have rates given by $\{U_j\}$. The transmissions have positive durations $\{\eta_j\}$. We assume that the vectors $\{(\eta_j, U_j)\}$ are i.i.d. and independent of the arrivals process. The workload at time t is the sum of rates of all surviving present and past transmission. This model was considered by [Resnick and Rootzén, 2000], [Mikosch *et al.*, 2002] among others.

In the following, we consider that the path of the process is observed along continuous time. From a numerical point of view, since the path of X is piece-wise constant, this means that one observes all the jump times and the workload at these times. In practical situations, the transmission rate is measured by counting the packets going through some point of the network link. From the packet counts, one may compute the overall average rate of transmission over equi-spaced time slots $[k\delta, (k+1)\delta]$ $k \in \mathbb{Z}$. From now on, we take $\delta = 1$ without loss of generality. The process X is not aimed to model the traffic at packets level since the transmission rate at the packets level cannot be assumed to be constant. Nevertheless

$$Y_k = \int_k^{k+1} X(s) \, ds$$

is a reasonable model for the overall transmission rate averaged on $[k, k+1]$ because, by locally averaging the instantaneous rate, one eliminates local variations of it. The estimator we will consider is computed from the wavelet coefficients of X. In the case of Haar wavelet these coefficients can be computed exactely from the discrete sequence $\{Y_k, \, k \in \mathbb{Z}\}$. Otherwise, some adaptations are needed but we will not pursue in this direction here and thus will assume either that the continuous time path of X is observed or that the wavelet ψ used below is the Haar wavelet $\psi = \frac{1}{2}(\mathbf{1}_{[0,1)} - \mathbf{1}_{[1,0)})$.

We now introduce the assumption on the joint distribution of the transmissions rates and durations.

Assumption 1 *The random vectors* $\{(\eta, U), \ (\eta_n, U_n), n = 0, \pm 1, \pm 2, \dots\}$ *are i.i.d. with distribution* ν *on* $\mathbb{R}_+ \times \mathbb{R}$ *and independent of the homogeneous Poisson Point Process on the real line with points* $\{t_j\}_{j \in \mathbb{Z}}$*; there exist a real number* $\alpha \in (0, 2)$ *and a positive integer* k^* *such that* $\mathbb{E}[|U|^{k^*}] < \infty$ *and for each integer* $k = 0, 1, \dots, k^*$

$$\mathbb{E}\left[U^k \mathbf{1}_{\{\eta > t\}}\right] = L_k(t) t^{-\alpha} \ . \tag{2}$$

where L_k *are slowly varying as* $t \to +\infty$.

Defining, for each $k \leq k^*$, the signed measure on $\mathbb{R}_+$

$$\nu_k(\mathrm{d}v) := \int u^k \nu(\mathrm{d}v, \mathrm{d}u),$$

and the function

$$H_k(t) = \nu_k(t, \infty) = \mathbb{E}\left[U^k \mathbf{1}_{\{\eta > t\}}\right], \quad t \geq 0,$$

Condition (2) is equivalent to saying that H_k, $k = 0, 1, \ldots, k^*$, are regularly varying with index α.

Assumption 1 implies in particular that the tails of the distribution of η is regularly varying with index α. This in turns implies Assumption 1 if U and η are independent, in which case the functions L_k differ by a multiplicative constant. A more realistic situation for network traffic modelling is the case where the transmission rate U is independent of the amount of transmitted data during the download session (which is equal to $W := U\eta$), given that the rate is above some threshold. Below this threshold, the accessible amount of data is supposed to have light tails, and above this threshold, W is supposed to have heavy tails. In practice this threshold separate high rate connections (say, xDSL/LAN/Cable connection) from low rate connections (say, RTC connection), which are not suitable for downlaoding large data. In this case, it can be shown that the measure ν_k inherits the heavy tails of W for all k such that $\mathbb{E}|U|^k < \infty$.

2 Stationary version and asymptotic behavior

If $\mathbb{E}[\eta] < \infty$, a stationary version of this process is defined by

$$X_S(t) = \sum_{j \in \mathbb{Z}} U_j \mathbf{1}_{\{t_j \leq t < t_j + \eta_j\}} \; t \in \mathbb{R}, \tag{3}$$

where, in the sequel, $\{t_j\}$ are the points of a unit rate homogeneous Poisson process on the line such that $t_k < t_{k+1}$ for all k and $t_{-1} < 0 \leq t_0$.

By Karamata's Theorem, for all such k, we easily obtained the asymptotic equivalences of standard tail behaviors of ν_k. For instance, if $\alpha > 1$,

$$\mathbb{E}\left[U^k \{\eta - t\}_+\right] \mathrm{sim} \frac{1}{\alpha - 1} L_k(t) t^{1-\alpha} \;. \tag{4}$$

Proposition 1 *Let Assumption 1 hold. The process X_S is well defined and strictly stationary if and only if $\mathbb{E}[\eta] < \infty$. If moreover $k^* \geq 2$, then X_S is weakly stationary with expectation and autocovariance function given by*

$$\mathbb{E}[X_S(t)] = \mathbb{E}[U\eta] \;,$$

$$\mathrm{cov}(X_S(0), X_S(t)) = \mathbb{E}[U^2(\eta - t)_+] \mathrm{sim} \frac{1}{1 - \alpha} L_2(t) t^{1-\alpha} \; \textit{if } \alpha > 1 \;,$$

where the equivalence holds as $t \to +\infty$.
The process X *is nonstationary with expectation* $\mathbb{E}[X(t)] = \mathbb{E}[U(\eta \wedge t)]$ *and autocovariance function given, for* $s \leq t$ *by*

$$\mathrm{cov}(X(s), X(t)) = \mathbb{E}[U^2\{s - (t-\eta)_+\}_+] = \int_{t-s}^{t} H_2(v)\mathrm{d}v \ .$$

If $\alpha \in (0,1)$ and if t and s tend to infinity at the same rate, the following asymptotic equivalent of $\mathrm{cov}(X(s), X(t))$ holds. For all $t, s > 0$, as $T \to \infty$,

$$\mathrm{cov}(X(Ts), X(Tt)) \mathrm{sim} \frac{1}{1-\alpha} L_2(T) T^{1-\alpha} \{(s \vee t)^{1-\alpha} - |t-s|^{1-\alpha}\} \ . \qquad (5)$$

The proof of Proposition 1 is a straightforward application of well known properties of Poisson point processes.

If $\mathbb{E}[\eta] < \infty$, the non-stationary process X converges to X_S. By definition, the difference between X and X_S is given by

$$X_S(t) - X(t) = \sum_{k<0} U_k \mathbf{1}_{\{t_k \leq t < t_k + \eta_k\}}, \quad t \geq 0 \ .$$

Since $\mathbb{E}[\eta] < \infty$ and since the η_k are i.i.d and independent of the birth times t_k, a Borel-Cantelli argument yields that this sum has almost surely a finite number of terms, which is at most the number of indices $k < 0$ such that $t_k + \eta_k \geq 0$. Hence, almost surely, $\lim_{t\to\infty}\{X_S(t) - X(t)\} = 0$. This limit also holds in the mean $\mathbb{E}[|X_S(t) - X(t)|] \leq \mathbb{E}[|U|(\eta - t)_+] \to 0$. The asymptotic behavior of the cumulative workload is now investigated.

If we are not in the stable case, that is, for $\mathbb{E}[\eta] = \infty$, the process X_S is not defined (see Proposition 1). We may still consider the weak limit of the cumulative workload but this limit will be very different in the two cases as shown by the next proposition.

For $\alpha < 1$ (implying $\mathbb{E}[\eta] = \infty$), the next proposition gives a straightforward extension of the results of [Resnick and Rootzén, 2000] to the case of random transmission rate U_j. In the case $\alpha > 1$ (implying $\mathbb{E}[\eta] < \infty$), it has been proved under slightly different assumptions by [Mikosch *et al.*, 2002], [Maulik *et al.*, 2002] or [Mikosch and Resnick, 2004].

Proposition 2 *Denote* $H = (3-\alpha)/2$. *If* $0 < \alpha < 1$, i.e. $1 < H < 3/2$, *and if Assumption 1 holds with* $k^* = \infty$, *then the sequence of processes* $\{L_2^{-1/2}(T) T^{-H} \int_0^{Tt} (X(s) - \mathbb{E}[X(s)])\, \mathrm{d}s, t \geq 0\}$ *converges weakly to the Gaussian process* W *with autocovariance function*

$$\mathrm{cov}(W(s), W(t)) = \frac{1}{1-\alpha} \int_0^t \int_0^s \{(u \vee v)^{1-\alpha} - |u-v|^{1-\alpha}\}\, \mathrm{d}u\, \mathrm{d}v \ .$$

If $1 < \alpha < 2$, i.e. $1/2 < H < 1$, *then* $T^{-H} \int_0^{Tt} X(s)\mathrm{d}s$ *converges in probability to 0, and the sequence* $\{T^{-1/\alpha} \int_0^{Tt} (X(s) - \mathbb{E}[X(s)])\, \mathrm{d}s, t \geq 0\}$ *converges weakly to an* α*-stable Levy process.*

This proposition illustrates a change of behavior between the stationary and non-stationary cases.

3 Estimation

3.1 Terminology

The most important parameter for this process is thus the parameter α. In accordance with the notation in use in the context of long memory processes, we define the Hurst index of the process X as $H = (3-\alpha)/2$, because the variance of partial sums scales as T^{2H}. We can also define $d = H - 1/2 = 1 - \alpha/2$, in relation to fractionally integrated processes, such as ARFIMA processes, but this would be quite arbitrary in this context where no fractional integration is involved.

3.2 Methods

The parameter α is a tail index, so traditional methods to estimate a tail index could be used. But it is well known that these methods are not very efficient in the case of dependent data (cf. [Resnick and Stărică, 1995] for instance). Moreover, in the model under consideration here, α is not the tail index of the marginal distribution of the observed process, which has finite variance whereas $\alpha < 2$. Thus it is not at all clear how to use these methods.

But as shown by Proposition 1, the coefficient α is related to the second order properties of the process: the coefficient $H = (3-\alpha)/2$ can be viewed as its Hurst index, *i.e.* H governs the rate of decay of the autocovariance function of the process. Therefore it seems natural to use an estimator of the Hurst index.

3.3 The (wavelet) coefficients

Let ψ be a bounded $\mathbb{R} \to \mathbb{R}$ function with compact support included in $[0, M]$ and such that

$$\int \psi(s)\ \mathrm{d}s = 0\ . \tag{6}$$

For integers $j \geq 0$ and $k \in \mathbb{Z}$, define

$$\psi_{j,k}(s) = 2^{-j/2}\psi(2^{-j}s - k). \tag{7}$$

The wavelet coefficients of the path are defined as

$$d_{j,k} = \int \psi_{j,k}(s)X(s)\ \mathrm{d}s\ , \quad d^S_{j,k} = \int \psi_{j,k}(s)X_S(s)\ \mathrm{d}s\ . \tag{8}$$

Asume that a path is observed between time 0 and T. Since $\psi_{j,k}$ has support in $[k2^j, (k+M)2^j]$, the above coefficients can be computed for all (j,k) such that $T2^{-j} \geq L$ and $k = 0, 1, \ldots, T2^{-j} - M$.

Lemma 1 *Define*

$$\mathcal{L}(z) = z^{\alpha} \int_0^{\infty} \left[\int_{-\infty}^{\infty} \left\{ \int_t^{t+zv} \psi(u)\,\mathrm{d}u \right\}^2 \mathrm{d}t \right] \nu_2(\mathrm{d}v) \; . \tag{9}$$

Then $\mathcal{L}$ is slowly varying at infinity and

$$\mathbb{E}[d^S_{j,k}] = 0 \; , \quad \mathrm{var}(d^S_{j,k}) = \mathcal{L}(2^j)\, 2^{(2-\alpha)j} \; , \tag{10}$$

$$\mathbb{E}[d_{j,k}] = O\left(L_1(k2^j)\, 2^{(3/2-\alpha)j} k^{-\alpha}\right) \; , \tag{11}$$

$$\mathrm{var}(d_{j,k} - d^S_{j,k}) = O\left(L_2(k2^{-j})\, 2^{(2-\alpha)j}\, k^{-\alpha}\right) \; . \tag{12}$$

Remark 31 *The coefficients $d_{j,k}$ are centered in the case where U and η are independent and U is centered, even in the nonstationary case.*

3.4 The estimator

Lemma 1 provides the rationale for the following minimum contrast estimator of α which is related to the local Whittle estimator, cf. [Künsch, 1987], [Robinson, 1995b]. The obtained estimator has been introduced by Moulines, Roueff and Taqqu (2004) and is called the wavelet Whittle estimator. For positive integers $J_0 < J$, define

$$\varDelta = \{(j,k)\;, J_0 < j \leq J \;, \quad 0 \leq k \leq 2^{J-j} - 1\} \;\text{ and }\; \delta = \frac{1}{\#\varDelta} \sum_{(j,k)\in\varDelta} j \; .$$

The scale index J is the maximal scale index available from the data while J_0 is a cut-off tuned by the user. The local Whittle estimator of α is then defined as:

$$\hat{\alpha} = \arg \min_{\alpha' \in (0,2)} \log \left(\sum_{(j,k)\in\varDelta} \frac{d^2_{j,k}}{2^{(2-\alpha')j}} \right) + \delta \log(2)(2-\alpha') \; .$$

Equivalently, we could have defined $\hat{H} = (3-\hat{\alpha})/2$ or $\hat{d} = 1 - \hat{\alpha}/2$.

Theorem 31 *Let $\alpha \in (1,2)$ and let Assumption 1 hold. Suppose that $\mathcal{L}(x) \to 1$ as $x \to \infty$. If $J_0 \to \infty$, $J \to \infty$ and $J_0 < J/\alpha$ then $\hat{\alpha}$ is a consistent estimator of α.*

A corresponding results hold in the case where $\alpha \in (0,1]$ but some adaptations are needed in the definition of the estimator and a second vanishing moment is needed on ψ.

4 Simulations

We have simulated M/G/∞ processes, which correspond to the process X with $U_k = 1$ for all k's, and estimated α via different classical estimators of long range dependence. The obtain paths are represented in Figure 1 and Figure 2, respectively in non-stable ($\alpha = 0.7 < 1$) and stable ($\alpha = 1.5 > 1$) situation. Monte-carlo simulations provided the boxplots and MSE estimates for the several estimators, also represented on these figures. In those graphs, the X-coordinates $V1$, $V2$, ... $V10$ correspond to the scale cut-off VJ_0.

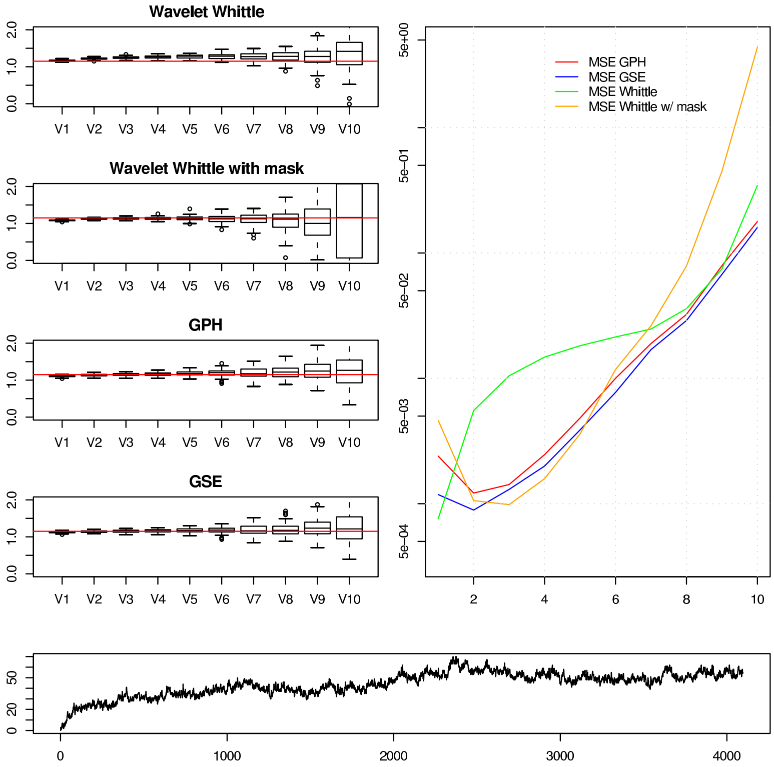


Fig. 1. $\alpha = 0.7$: the process do not converge to a stationary. Its cumulative load is approximately gaussian.

References

[Künsch, 1987]H. R. Künsch. Statistical aspects of self-similar processes. In Yu.A. Prohorov and V.V. Sazonov (eds), *Proceedings of the first World Congres of the Bernoulli Society*, volume 1, pages 67–74. Utrecht, VNU Science Press, 1987.

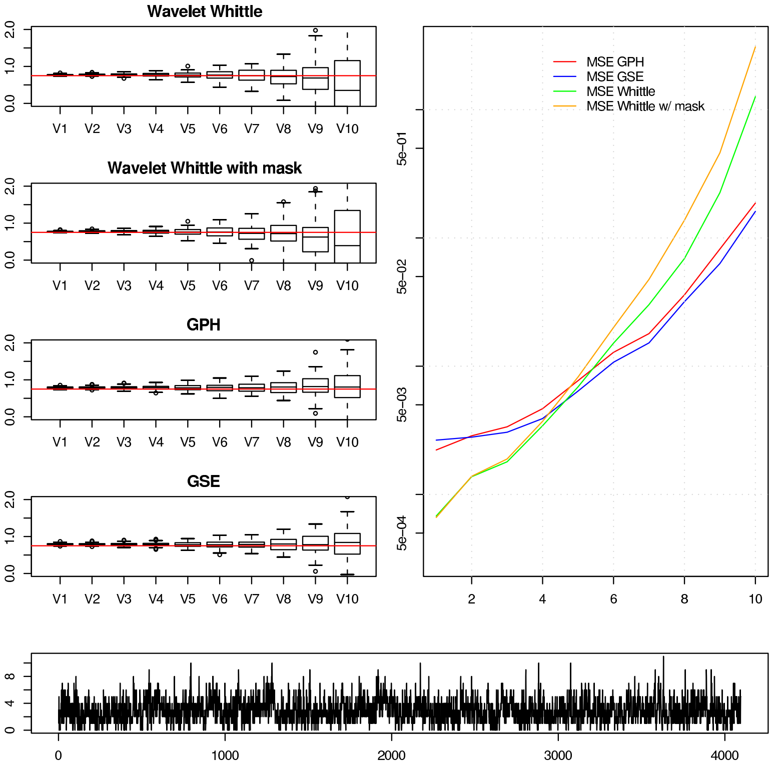


Fig. 2. $\alpha = 1.5$

[Maulik *et al.*, 2002]Krishanu Maulik, Sidney Resnick, and Holger Rootzén. Asymptotic independence and a network traffic model. *Journal of Applied Probability*, 39(4):671–699, 2002.

[Mikosch and Resnick, 2004]Thomas Mikosch and Sidney Resnick. Activity rates with very heavy tails. Technical Report 1411, Cornell University, 2004.

[Mikosch *et al.*, 2002]T. Mikosch, S.I. Resnick, H. Rootzen, and A. Stegeman. Is network traffic approximated by stable Levy motion or fractional Brownian motion? *Annals of Applied Probability*, 12:23–68, 2002.

[Moulines *et al.*,]E. Moulines, F. Roueff, and M. S. Taqqu. A wavelet Whittle estimator. working paper.

[Resnick and Rootzén, 2000]Sidney Resnick and Holger Rootzén. Self-similar communication models and very heavy tails. *The Annals of Applied Probability*, 10(3):753–778, 2000.

[Resnick and Stărică, 1995]Sidney Resnick and Cătălin Stărică. Consistency of Hill's estimator for dependent data. *Journal of Applied Probability*, 32(1):139–167, 1995.

[Robinson, 1995b]P.M. Robinson. Gaussian semiparametric estimation of long range dependence. *Annals of Statistics*, 24(5):1630–1661, 1995b.

Part XVIII

Posters

Discrimination between deterministic trend and stochastic trend processes

Jorge Caiado[1] and Nuno Crato[2]

[1] Escola Superior de Ciências Empresariais,
Instituto Politécnico de Setúbal and CEMAPRE,
Rua do Quelhas 6, 1200-781 Lisboa, Portugal,
(e-mail: jcaiado@esce.ips.pt)

[2] Instituto Superior de Economia e Gestão,
Universidade Técnica de Lisboa and CEMAPRE,
Rua do Quelhas 6, 1200-781 Lisboa, Portugal,
(e-mail: ncrato@iseg.utl.pt)

Abstract. Most of economic and financial time series have a nonstationary behavior. There are different types of nonstationary processes, such as those with stochastic trend and those with deterministic trend. In practice, it can be quite difficult to distinguish between the two processes. In this paper, we compare random walk and determinist trend processes using sample autocorrelation, sample partial autocorrelation and periodogram based metrics.

Keywords: autocorrelation, classification, determinist trend, Kullback-Leibler, periodogram, stochastic trend, time series.

1 Introduction

There are different types of nonstationarity processes. One can consider a deterministic linear trend process $y_t = a + bt + \varepsilon_t$ (with ε_t a white noise term), that can be transformed into a stationary process by subtracting the trend $a + bt$, and a stochastic linear trend process such as the so-called random walk model $(1 - B)y_t = \varepsilon_t$ or $y_t = y_{t-1} + \varepsilon_t$. An interesting, but some times difficult problem is to determine whether a linear process contains a trend, and whether a linear process exhibits a deterministic or a stochastic trend. In particular, it is useful to distinguish between a random walk plus drift $y_t = \mu + y_{t-1} + \varepsilon_t$ and a deterministic trend in the form $y_t = a + \mu t + \varepsilon_t$.

The problem of classifying and clustering time series has been studied by Piccolo (1990), Tong and Dabas (1990), Shaw and King (1992), Kakizawa, Shumway and Taniguchi (1998), Maharaj (2000, 2002), Caiado, Crato and Peña (2005), Xiong and Yeung (2004), among others. In this paper, we use sample autocorrelation, sample partial autocorrelation and periodogram ordinate based metrics to compare deterministic trend and stochastic trend processes.

2 Classification Methods

A fundamental problem in clustering and classification analysis is the choice of a relevant metric. We know that the Euclidean distance is not a good metric for classifying time series since it is invariant to permutation of the coordinates and so it does not take into account the information about the autocorrelations.

Let $X = (x_{1,t}, \dots, x_{k,t})'$ be a vector time series and $\widehat{\rho}_i = (\widehat{\rho}_{i,1}, \dots, \widehat{\rho}_{i,m})$ be a vector of the sample autocorrelations of the time series i for some m such that $\widehat{\rho}_k \overset{\text{sim}}{=} 0$ for $k > m$. A distance between two time series x and y can be defined by $d(x,y) = \sqrt{(\widehat{\rho}_x - \widehat{\rho}_y)'\Omega(\widehat{\rho}_x - \widehat{\rho}_y)}$, where Ω is some matrix of weights (see Galeano and Peña, 2000) . Caiado, Crato e Peña (2004) proposed three possible ways of computing a distance by using the sample autocorrelation function (ACF). The first uses the Euclidean distance between the sample autocorrelations coefficient vectors with uniform weights (ACFU metric),

$$d_{ACFU}(x,y) = \sqrt{\sum_{j=1}^{L} (\widehat{\rho}_{j,x} - \widehat{\rho}_{j,y})^2}, \tag{1}$$

where L is the number of autocorrelations. The second uses the Euclidean distance with geometric weights decaying with the lag (ACFG metric),

$$d_{ACFG}(x,y) = \sqrt{\sum_{j=1}^{L} f_j (\widehat{\rho}_{j,x} - \widehat{\rho}_{j,y})^2}, \tag{2}$$

where $f_j = pq^j$ for $i = 1, 2, ..., L$, $p = 1 - q$ and $0 < p < 1$. The third uses the Mahalanobis distance between the autocorrelations (ACFM metric),

$$d_{ACFM}(x,y) = \sqrt{(\widehat{\rho}_x - \widehat{\rho}_y)'\Omega^{-1}(\widehat{\rho}_x - \widehat{\rho}_y)}, \tag{3}$$

where Ω is the sample covariance matrix of the autocorrelation coefficients given by Bartlett's formula (see Brockwell and Davis, 1991, p. 221-222). A metric based on the sample partial autocorrelation function (PACF) is defined by

$$d_{PACF}(x,y) = \sqrt{(\widehat{\phi}_x - \widehat{\phi}_y)'\Omega(\widehat{\phi}_x - \widehat{\phi}_y)}, \tag{4}$$

where $\widehat{\phi}_{ii}$ are the sample partial autocorrelations and Ω is also some matrix of weights.

A measure based on the Kullback-Leibler (KL) information for time series classification can be defined by

$$d_{KL}(x,y) = tr(R_x R_y^{-1}) - \log \frac{|R_x|}{|R_y|} - n, \tag{5}$$

where R_x and R_y are the sample autocorrelation matrices of time series x and y. Since $d_{KL}(x, y) \neq d_{KL}(y, x)$, one can define a symmetric distance or quase-distance (KLJ metric), known as the *J divergence*, as,

$$d_{KLJ}(x, y) = \frac{1}{2} d_{KL}(x, y) + \frac{1}{2} d_{KL}(y, x), \tag{6}$$

which satisfies all the usual properties of a metric except the triangle inequality.

Caiado, Crato and Peña (2004) introduced also a periodogram-based metric. Let x and y be observed time series with periodograms, $P_x(w_j) = n^{-1}|\sum_{t=1}^{n} x_t e^{-itw_j}|^2$ and $P_y(w_j) = n^{-1}|\sum_{t=1}^{n} y_t e^{-itw_j}|^2$ at frequencies $w_j = 2\pi_j/n$, $j = 1, ..., m$ (with $m = [(n-1)/2]$) in the range 0 to π, and let $NP(w_j) = P(w_j)/\widehat{\gamma}_0$ be the normalized periodogram (with $\widehat{\gamma}_0$ the sample variance). Since the variance of periodogram ordinates is proportional to the spectrum value at the corresponding frequencies, Caiado, Crato and Peña (2004) proposed a metric based on the logarithm of the normalized periodograms (LNPER metric),

$$d_{LNPER}(x, y) = \sqrt{\sum_{j=1}^{m} \left[\log NP_x(w_j) - \log NP_y(w_j)\right]^2}. \tag{7}$$

3 Monte Carlo Simulations

For the Monte Carlo simulations we chose the determinist trend and random walk plus drift models studied by Enders (1995, p. 252),

$$y_t = 1 + 0.02t + \varepsilon_t$$

and

$$y_t = 0.02 + y_{t-1} + \varepsilon_t/3,$$

with ε_t a zero mean and unit variance white noise. These processes were discussed by Enders since it is quite difficult to distinguish between them, as we can see in Figure 1. We performed 250 replicated simulations of five deterministic trend models and five random walk models with those specifications, with sample sizes of 50, 100, 200, 500 and 1000 observations. We used the previously discussed metrics to compute the distance matrices among the 10 time series and to aggregate them into two clusters (determinist trend and stochastic trend) using an hierarchical clustering algorithm (complete linkage method).

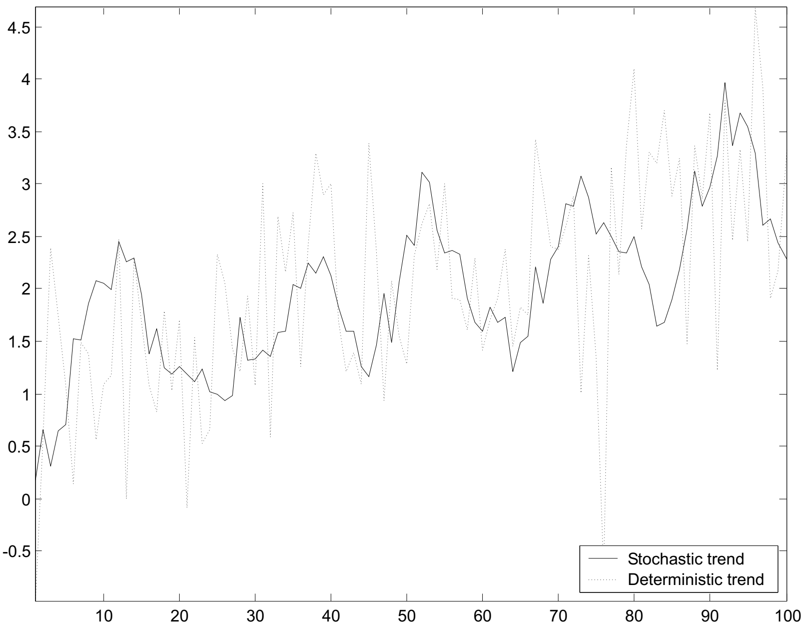


Fig. 1. Simulated stochastic trend and deterministic trend processes.

Table 1 presents the percentage of sucesses obtained in the comparison between the two processes, where n is the sample size, L is the autocorrelation lenght, the sample autocorrelation and sample partial autocorrelation metrics (ACFG and PACFG metrics) uses a geometric decay of $p = 0.05$, in the LNPER metric F for low frequencies corresponds to periodogram ordinates from 1 to $\sqrt{n}$ and F for high frequencies corresponds to periodogram ordinates from $\sqrt{n+1}$ to $n/2$.

The ACF based metrics can discriminate quite well between the deterministic trend models and random walk models. This is particularly evident for the first few autocorrelations, since the ACF of the random walk process is close to unity and the ACF of the deterministic trend tends to approach to zero. Because the PACF of the random walk has a very large first lag and cuts off after lag 1, while the PACF of the deterministic trend exhibits a pattern of a white noise process, the discrimination between the two models based on the first partial autocorrelations is striking. The KLJ metric perform quite well for all data sample sizes and the LNPER metric seems to perform better for periodogram ordinates dominated by high frequencies, which concerns the short-term information of the processes.

n	L	ACFU	ACFG	ACFM	PACFG	KLJ	F	LNPER
50	5	97.28	97.60	99.31	99.27	97.87	low	85.04
	10	92.12	94.88	99.56	99.46	98.53	high	95.24
	25	92.12	91.52	98.01	64.00	97.33	all	94.48
100	5	99.28	98.92	100.0	100.0	98.47		
	10	95.68	97.28	99.73	100.0	98.93	low	92.48
	25	88.16	89.84	96.44	100.0	99.47	high	99.04
	50	85.08	91.80	94.67	70.73	98.53	all	98.72
200	5	99.56	99.36	100.0	100.0	99.72		
	10	95.40	97.36	96.55	100.0	99.49	low	96.08
	20	87.80	91.20	92.22	100.0	99.60	high	99.28
	50	72.76	81.80	87.79	100.0	99.47	all	99.20
	100	70.56	82.56	na	94.37	99.20		
500	5	97.68	97.64	100.0	100.0	98.13		
	10	89.52	92.12	99.28	100.0	99.15	low	94.32
	20	78.00	81.28	96.32	100.0	98.81	high	98.56
	50	68.24	70.32	82.58	100.0	98.13	all	98.16
	125	68.72	70.04	80.97	100.0	99.20		
	250	67.92	70.12	na	na	na		
1000	5	94.48	94.60	100.0	100.0	98.31		
	10	83.04	83.56	95.26	100.0	98.81	low	90.40
	20	72.52	73.92	94.21	100.0	99.32	high	96.72
	50	67.36	68.65	72.97	100.0	97.12	all	93.92
	100	67.52	67.86	70.18	100.0	96.27		
	500	65.12	67.36	na	na	na		

Table 1. Percentage of sucess in the comparison between random walk plus drift and deterministic trend processes.

4 Discussion

In this paper we use different dependence metrics for comparison of a particular type of nonstationary time series models. Simulation results show that the metrics based on the sample autocorrelations, the sample partial autocorrelations, the Kullback-Leibler information measure and the normalized periodogram can distinguish quite well between deterministic trend and stochastic trend processes. In particularly, we point out the performance of the sample partial autocorrelation metric in this type of comparison. For the autocorrelation-based metrics we note that short lags L provide better results. This can be explained by the structure of these models, since the main differences arise for the first ACF and PACF values. Contrarily to what could be expected, the performance of ACF methods decreases with sample size. This does not happen with the PACF method. Kullback-Leibler method shows a remarkable good performance and stability across sample sizes and ACF orders considered. The periodogram-based metric compares well to Kullback-Leibler and is computationally simpler.

Acknowledgment: We thank Daniel Peña of Universidad Carlos III de Madrid for contributions in the development of the methods used in this paper. This research was supported by a grant from the Fundação para a Ciência e Tecnologia (POCTI/FCT).

References

[Brockwell and Davis, 1991]Brockwell, P. J. and Davis, R. A. (1991). *Time Series: Theory and Methods*, Springer, New York.

[Caiado *et al.*, 2005]Caiado, J., Crato, N. and Peña, D. (2005). "A periodogram-based metric for time series classification", *Computational Statistics* & *Data Analysis* (forthcoming).

[Enders, 1995]Enders, W. (1995). *Applied Econometric Time Series*, Wiley, New York.

[Galeano and Peña, 2000]Galeano, P. and Peña, D. (2000). "Multivariate analysis in vector time series", *Resenhas*, 4, 383-404.

[Kakizawa *et al.*, 1998]Kakizawa, Y., Shumway, R. H. and Taniguchi, M. (1998). "Discrimination and clustering for multivariate time series", *Journal of the American Statistical Association*, 93, 328-340.

[Maharaj, 2000]Maharaj, E. A. (2000). "Clusters of time series", *Journal of Classification*, 17, 297-314.

[Maharaj, 2002]Maharaj, E. A. (2002). "Comparison of non-stationary time series in the frequency domain", *Computational Statistics* & *Data Analysis*, 40, 131-141.

[Piccolo, 1990]Piccolo, D. (1990). "A distance measure for classifying ARIMA models", *Journal of Time Series Analysis*, 11, 152-164.

[Shaw and King, 1992]Shaw, C. T. e King, G. P. (1992). "Using cluster analysis to classify time series", *Physica D*, 58, 288-298.

[Tong and Dabas, 1990]Tong, H e Dabas, P. (1990). "Cluster of time series models: an example", *Journal of Applied Statistics*, 17, 187-198.

[Xiong and Yeung, 2004]Xiong, Y. e Yeung, D. (2004). "Time series clustering with ARMA mixtures", *Pattern Recognition* (in press).

Nonparametric laws in the performance evaluation of the reliable, unreliable and renewal systems

Smail Adjabi[1], Karima Lagha[1], Nabil Zougab[2], and Hillal Touati[2]

[1] Laboratoire LAMOS
Université de Béjaia,
06000 Bejaia, Algérie
(e-mail: adjabi@hotmail.com, karima_lagha@yahoo.com)

[2] Departement de Recherche Opérationnelle
Université de Béjaia
06000 Béjaia, Algérie

Abstract. In this paper we consider the performance evaluation problem in a queuing system GI/GI/1, a system with two units of reparable elements and a queuing system with an unreliable server and service repetition, using nonparametric distributions (consider IFR, NBU, DFR and NWU classes). We considered the qualitative properties of the inter-arrival, the repair and the service time, respectively. And presente bounds for the mean stationary waiting time, the mean time of life system and the blocking time in the system, respectively. These bounds are programmed and the characteristics are simulated in order to supplement the work carried out in [Adjabi *et al.*, 2004] and [Lagha and Adjabi, 2004].

Keywords: Nonparametric distribution, performance evaluation, queuing system, bounds, simulation.

1 Introduction

In this paper we are interested to the use of the qualitative properties of distributions for the characterstics evaluation in three systems, it acts a queuing system GI/GI/1, a renewal system and a queuing system with an unreliable server. The distributions considered are those of the inter-arrival, the reparation and the service time, respectively.

The aim of this paper is to calculate the characteristics by simulation in order to verify there membership to the interval delimited by the bounds established in [Adjabi *et al.*, 2004] and [Lagha and Adjabi, 2004]. These bounds are presented in the section 2, 3 and 4, respectively to the considered systems. Whereas the characteristics are simulated in the section 5. The results are interpreated in the section 6.

2 Queuing system GI/GI/1

Consider two systems $A_1/B_1/1$ (known as original) and $A_2/B_2/1$ (known as approximation system). We consider the following notations:

m_{W_i} : means of waiting time in $A_i/B_i/1$ system, $i = 1, 2$.
$A_i, i = 1, 2$: inter-arrival time distribution.
$B_i, i = 1, 2$: service time distribution.
m_B, m_A : means of service and inter-arrival times.
EB^2 : second moment of service time.
σ_B^2, σ_A^2 : variances of service and inter-arrival times.
$C_a = \sigma_A/m_A$: coefficient of variation of the inter-arrival times.
$\rho_i = m_B/m_{A_i}$: intensity of the traffic in the system $i, i = 1, 2$.

Under the external monotonicity property (theorem 5.2.1 in [Stoyan, 1983]),

$$A_2 \leq_{cv} A_1 \quad \text{and} \quad B_1 \leq_c B_2 \ , \tag{1}$$

we obtain the comparison $m_{W_1} \leq m_{W_2}$. Where $\leq_c$ ($\leq_{cv}$) indicates the convex (concave) ordering.
Suppose that $B_1 = B_2 = B$ a general service distribution and A_1 being a nonparametric inter-arrival distribution (IFR or NBU), its lower bound is given from the following table [Sengupta, 1994]:

Class	upper Bound	lower Bound
IFR	$\overline{F}(x) \leq \begin{cases} 1 & \text{if } x < m_r^{1/r} \\ \delta_x & \text{if } x > m_r^{1/r} \end{cases}$ where $\int_0^1 ry^{r-1}\delta_x^y dy = \frac{m_r}{x^r}$	$\overline{F}(x) \geq \begin{cases} \underset{0\leq\beta\leq x}{inf}\ e^{-\alpha} & \text{if } x < m_r^{1/r} \\ 0 & \text{if } x > m_r^{1/r} \end{cases}$ where $\int_0^\infty (\beta + \frac{x-\beta}{\alpha} z)^r e^{-\alpha} dz = m_r$
NBU	$\overline{F}(x) \geq \begin{cases} 1 & \text{if } x < m_r^{1/r} \\ \delta_x & \text{if } x \geq m_r^{1/r} \end{cases}$ where $\int_0^1 ry^{r-1}\delta_x^y dy = \frac{m_r}{x^r}$	$\overline{F}(x) \geq \begin{cases} \delta_x & \text{if } x < m_r^{1/r} \\ 0 & \text{if } x \geq m_r^{1/r} \end{cases}$ where $\sum_{j=0}^\infty \delta_x^j [(j+1)^r - j^r] = \frac{m_r}{x^r}$
DFR	$\overline{F}(x) \leq \begin{cases} e^{\frac{-rx}{x_0}} & \text{if } x < m_r^{1/r} \\ (\frac{x_0}{x})^r e^{-r} & \text{if } x \geq m_r^{1/r} \end{cases}$ where $x_0 = r[\frac{m_r}{\Gamma(r+1)}]^{1/r}$	$\overline{F}(x) \geq 0$
NWU	$\overline{F}(x) \leq \delta_x$ where $\sum_{j=1}^\infty \delta_j^x [j^r - (j-1)^r] = \frac{m_r}{x^r}$	$\overline{F}(x) \geq 0$

Table 1: Bounds on $\overline{F}(x)$ (based on r moment m_r) in various cases.

Using this property (for A_1 being IFR or NBU distribution, see Table 1.) and the relation (1), we presente the upper bounds for the mean stationary waiting time in two cases.

- **In the IFR case**, the upper bound associated is given by the following relation :

$$\frac{\sigma^2_{A_1}+\sigma^2_B}{2m_{A_1}(1-\rho_1)} - 1/2m_{A_1}(\rho_1 + C^2_{a1}) \leq m_{W_1} \leq \frac{EB^2}{\frac{2}{\alpha}(1-e^{-1}-\rho_1)}$$

here $\alpha = \left[\frac{\Gamma(r+1)}{m_r}\right]^{1/r}$.

• **In the NBU case**, the upper bound is given by the following relation :

$$\frac{\sigma^2_{A_1}+\sigma^2_B}{2m_{A_1}(1-\rho_1)} - 1/2m_{A_1}(\rho_1 + 1) \leq m_{W1} \leq \frac{EB^2}{m_{A_1}[1-e^{-1}-2\rho_1]}$$

Remark 1

• *The lower bounds in the IFR and NBU cases are proposed by Stoyan [Stoyan, 1983].*

• *See [Adjabi* et al.*, 2004] for demonstrations.*

3 Renewal theory

Consider a system with two units of reparable elements ξ_1 and ξ_2. At the moment $t = 0$ the element ξ_1 function until there failure at the date $t = X_0$ where repair starts with to be carried out on this element and takes a time equalize with Y_1 whereas ξ_2 starts to work until its failure with the date $t = X_1$. If $X_1 \leq Y_1$,the system stops with date $X_0 + X_1$. If not ξ_1 still function at the date X_0+X_1 whereas the repair of ξ_2 is started. The operating time $X_0, X_1, \ldots$, are supposed iid and independent of times of successive repairs Y_1, Y_2, $\ldots$, which are too iid and with the mean m_r of order, $r > 1$. We defined $N = inf\{n : X_n < Y_n\}$ life time for the system is there r.v. $T = X_0 + X_1 + \cdots + X_N$. So now them X_i is exponentially distributed of parameter λ and arbitrary repair time function R (cumulative distribution function of Y) is a nonparametric distribution (IFR, NBU, DFR or NWU).

Proposition 1 *Consider two systems as described above having for function of repair time distribution R_i, $i = 1, 2$ and λ_i, $i = 1, 2$, indicating parameters of the operating time, respectively. Given the following condition [Stoyan, 1983] :*

$$\lambda_1 \leq \lambda_2 \quad \textit{and} \quad R_1 <_L R_2, \tag{2}$$

the comparison between the mean time of life systems is as $ET_1 \geq ET_2$, where $<_L$ indicates the Laplacien order.

Using the lower bound of R_1 (see the Table 1.) and the relation (2), the upper bounds of the mean time of life system are given by the following relation (see [Adjabi *et al.*, 2004] for demonstrations):

• **IFR case**

$$1 + (1-e^{-\beta})^{-1} \leq ET_1 \leq \frac{1}{\lambda}\Big[1 + \frac{1+\beta^{-1}}{1-e^{-(1+\beta)(\Gamma(r+1))^{1/r}}}\Big]$$

with $\alpha = \left[\frac{\Gamma(r+1)}{m_r}\right]^{1/r}$, $\beta = \frac{\lambda}{\alpha}$ and $r > 0$.

• **NBU case**

$$1 + (1 - e^{-\beta})^{-1} \leq \lambda ET_1 \leq 1 + \frac{\beta}{\beta + e^{-\beta} - 1}, \quad \beta = \lambda m_1$$

Using the upper bound for R_1 (see the Table 1.)and the relation (2), the lower bounds of the mean time of life system are given by the following relation (see [Adjabi *et al.*, 2004] for demonstrations):

• **DFR case**

$$1 + \frac{\theta + r}{\theta(1 - e^{-(r+\theta)}) + (r + \theta)e^{-r}\theta^r \int_\theta^\infty x^{-r} e^{-x} dx} \leq \lambda ET_1 \leq 1 + (1 - e^{-\lambda m_1})^{-1}$$

where $\theta = \lambda x_0$, $x_0 = r\left[\frac{m_r}{\Gamma(r+1)}\right]^{1/r}$ and $r > 0$.

• **NWU case**

$$1 + \frac{1}{1 - \theta e^\theta \int_\theta^\infty x^{-2} e^{-x} dx} \leq \lambda ET_1 \leq 1 + (1 - e^{-\lambda m_1})^{-1}, \quad \theta = \lambda m_1$$

Remark 2

• *The lower bound presented in IFR and NBU cases is proposed by Stoyan [Stoyan, 1983].*

• *This bound became the upper one in the NWU and DFR cases.*

4 Unreliable queuing system

Consider a single-server queuing system with an unreliable server and service repetition. The total time taken by a customer from the instant he enters for service to the instant when he ends his service is called the blocking time which can be represented by:

$$Z_\lambda = X.1_{\{X \leq Y\}} + (Y + Z_\lambda^*).1_{\{X > Y\}}, \quad \lambda > 0. \tag{3}$$

Where X, Y and Z_λ are independent non-negative random variables, with cumulative distribution functions (cdf) denoted by $G(t), R(t)$ and $F(t)$, respectively. X is the service time with free interruption, Y is the server failure time and is assumed to have exponential distribution with mean $1/\lambda$. So $\overline{R}(t) = 1 - R(t) = e^{-\lambda t}$, $0 \leq t \leq \infty$.

Z_λ^* has the same distribution as Z_λ (denoted by $Z_\lambda^* \overset{d}{=} Z_\lambda$), and $1_{\{X \leq Y\}}$ is the indicator function of the event $\{X \leq Y\}$.

Consider the cdf G of X being a nonparametric repair distribution (IFR, NBU, DFR or NWU), its lower or upper bounds are given from the Table 1. To gather with the following Lemma, the bounds of the mean blocking time in the system EZ_λ are established (see [Lagha and Adjabi, 2004] for demonstrations). Let EX^r denote the r *th* moment of the r.v. X.

Lemma 1 *Suppose that X is not degenerate at point zero and Z defined as (3), so*

$$EZ = E(min(X,Y))/p \ ,$$

where $p = P(X \leq Y) = \int_0^{\infty} G(t)dR(t)$ so $E(min(X,Y)) = \int \overline{G}(t)e^{-\lambda t}\, dt$.

Using the corresponding lower bounds for G (see the Table 1.) and the above Lemma the lower bounds of the mean blocking time in the system are given in the following cases :

- **IFR case**

$$EZ_\lambda \geqslant x_0 \Big[\frac{1-e^{-1-\lambda x_0}}{1+x_0\lambda e^{-1-\lambda x_0}}\Big], \ x_0 = EX$$

- **NBU case**

$$EZ_\lambda \geqslant \frac{\beta + e^{-\beta} - 1}{\lambda(1-e^{-\beta})}, \ \beta = \lambda x_0$$

The upper one are given in the remainder cases considered :

- **DFR case**

$$EZ_\lambda \leq \frac{x_0(e^r - e^{-\lambda x_0}) + (r+\lambda x_0)x_0^r I_r}{re^r + \lambda x_0 e^{-\lambda X_0} - \lambda(r+\lambda x_0)x_0^{-r} I_r}$$

where $I_r = \int_{x_0}^{+\infty} t^{-r} e^{-\lambda t}\, dt$, $x_0 = r\Big[\frac{EX^r}{\Gamma(r+1)}\Big]^{1/r}$ and $r > 0$.

- **NWU case**

$$EZ_\lambda \leq \frac{x_0 e^{\beta} I_1}{1 - \beta e^{\beta} I_1}, \ \ \beta = \lambda x_0 \ \text{ and } \ I_1 = \int_{x_0}^{+\infty} t^{-1} e^{-\lambda t} dt$$

Remark 3

- *The complex integral I_r is convergent and simulated (in the following section) to calculate the bounds.*

5 Bounds Computation

Consider in this section some parametric distributions to calculate the bounds given above (for three systems) and simulate characteristics. The application is worked in MATLAB environment.

The results are given in the following tables for three considered problems, respectively.

- **Queuing System GI/GI/1**

System	lower Bound	upper Bound	Simulation
$E_{(4,2)}/E_{(1,3)}/1$	0	0.11936	0.0097321
$E_{(4,3)}/E_{(2,5)}/1$	0	0.27099	0.021166

$E_{(2,3.5)}/W_{(1,4)}/1$	0.083333	0.56199	0.11258
$W_{(3,1.5)}/W_{(1,4)}/1$	0	0.4948	0.051066
$E_{(3,1)}/M_{\mu\ =2}/1$	0	0.17904	0.018002
$M_{\lambda=0.5}/E_{(1,2)}/1$	0.16667	0.32712	0.18617
$M_{\lambda=0.7}/W_{(1,4)}/1$	0.05003	0.095708	0.050896
$M_{\lambda=1.5}/M_{\mu\ =2}/1$	0.3	0.37359	0.3
$IFR/M_{\mu\ =1.2}/1$	0.12987	2.5541	0.32971
$IFR/IFR/1$	0	0.3069	0

Table 2: Bounds and simulation of the waiting average time

- **Renewal system**

Consider for application, the rth moment of the r.v. Y ($r = 1, 5$ and 10).

$exp(\lambda)/R(t)$ model	r order	lower bound	upper bound	simulation
$\lambda = 2/E_{(2,3)}$	1	1.179	1.469	1.2839
	5		1.5305	
	10		3.1114	
$\lambda = 1.2/Exp(1.1)$	1	2.0882	2.7609	2.4071
	5		2.5091	
	10		2.5002	
$\lambda = 1.5/W_{(2,4)}$	1	1.9302	4.2973	3.1272
	5		5.183	
	10		6.355	
$\lambda = 3/W_{(0.5,3)}$	1	0.83127	1.1805	1.0549
	5	0.70901		
	10	0.68856		
$\lambda = 2/IFR$	1	1.0034	1.1015	1.0088
	5		1.2045	
	10		1.3233	
$\lambda = 1.2/W_{(0.8,1.5)}$	1	2.5339	2.5962	2.5132
	5	2.3289		
	10	2.2295		

Table 3: Bounds and simulation of life average time

- **Unreliable system**

Failure rate	$\lambda = 2$	$\lambda = 3$	$\lambda = 3.5$	$\lambda = 1.6$
Service time	$E_{(2,4)}$	$Exp(4)$	$W_{(2,3)}$	IFR
Lower bound	0.3808	0.18274	0.63459	0.58863
Upper bound	x	x	x	x
Simulation	1.5040	0.8947	0.6611	2.4588

Table 4: Bounds and simulation of the blocking average time

6 Results interpretation

In the first system, we remark that the characteristic value obtained by the simulation belongs to the interval delimited by the lower and upper bounds presented in the section 2 and prooved in [Adjabi *et al.*, 2004]. This let us think that these bounds are accepted. Moreover the characteristic value seems to be much closer to the lower bound than the upper one.

Remark in the second system that the characteristic value obtained by simulation belongs to the proposed interval délimited by the bounds presented in the section 3. In the models where the repair time distribution is IFR, the upper bound is an increased function of r but the lower one does not depend on r. In the models where the repair time distribution is DFR, the lower bound is an increased function of r but the upper one does not depend on r.
We remark in addition that, the computed value by simulation turns around 1 when $\lambda \geq 2$ whereas it largely exceeds 1 when λ turns around 1.

In the last system, we considered the IFR and UBU cases for application. So we have only the lower bound to calculate and the values obtained by simulation are finite and higher then those of lower bounds. The "x" means no upper bound is calculated.

7 Conclusion

In this work we considered the performance evaluation problem in the queuing system GI/GI/1 (section 2), a renewal system (section 3) and an unreliable system (section 4), using nonparametric properties of distributions (IFR, NBU, DFR or NWU class). By comparison between distributions with stochastic orders ($<_c, <_{cv}$ and $<_L$), bounds are presented for considered systems. The characteristics bounds obtained are for: mean waiting time, the mean life time and the mean blocking time, respectively.
These systems are simulated in order to supplement the works of [Adjabi *et*

al., 2004] and [Lagha and Adjabi, 2004] to verify the acceptance of the proposed bounds (section 5). This verification is established by the application worked in MATLAB environment.
The bounds presented in this paper can be used for other distributions.
Example: using the following relations

$$IFR \rightarrow IFRA \rightarrow NBU \text{ and } DFR \rightarrow DFRA \rightarrow NWU$$

for IFRA, DFRA,...

References

[Adjabi *et al.*, 2004]S. Adjabi, K. Lagha, and A. Aissani. Application des lois non paramétriques dans les sytèmes d'attente et théorie de renouvellement. *An International Journal on Operation Research: RAIRO-Operations Research* Vol. 38, N^{o}3, pages 243–254, 2004.

[Lagha and Adjabi, 2004]K. Lagha and S. Adjabi. Bounds for blocking time in queueing system with an unreliable server. In Proceedings of The Alamos National Laboratory, editor, *Fourth International Conference on Mathematical methods in Reliability, Methodology and Pratice, June 21-25, Santa Fe, New Mexico, USA*, 2004.

[Sengupta, 1994]D. Sengupta. Another look at the moment bounds on reliability. *J. Appl. Prob. 31*, pages 777–787, 1994.

[Stoyan, 1983]D. Stoyan. *Comparison Methods for Queueing and Other Stochastic Models.* John Wiley, 1983.

Empirical comparison of Arcing algorithms

Riadh Khanchel and Mohamed Limam

I . S . G
41, Avenue de la Liberté
Le Bardo,2000 Tunis, Tunisie
(e-mail: riadh.khanchel@isg.rnu.tn)

Abstract. Adaboost and Arc-x(h) are two ensemble algorithms that belong to Arcing family of algorithms. They use different weight updating rules and combine classifiers using different voting scheme. For $h = 4$, Arc-x(h) performs equally well as Adaboost but higher values of h were not tested. Previous methods used to compare algorithms are based on the performance over test sets. A different approach presented by [Nadeau and Bengio, 2003] takes into account variability in training and test sets. Using this approach, an empirical study is conducted to compare Adaboost and Arc-x(h) for different values of h. Results show that increasing h does not affect the performance of Arc-x(h) whichis comparable to Adaboost.

Keywords: Boosting, Arcing, Adaboost.

1 Introduction

Different classification algorithms have been proposed and used in fields like medicine, business and finance. However, the accuracy of these algorithms may be moderate when applied to complex classification tasks. Ensemble learning is a technique for improving their performance: a collection of moderately accurate and diverse classifiers are constructed then they are combined in order to output highly accurate ones. Different ensemble learning algorithms have been proposed: Adaboost [Freund and Schapire, 1997], Bagging [Breiman, 1996], and Arcing [Breiman, 1998].

Ensemble learning method is developed within the framework of probably approximately correct (P. A. C) model of learning where learning algorithm and hypothesis are used to refer to, respectively, classification algorithm and classifier. This model of learning is specified by a set of measurement space, a label space, an error parameter ϵ, a confidence parameter δ and other parameters that specify the size of the measurement space and the label space. After running for a polynomial time, the learning algorithm outputs a hypothesis which error is less than ϵ with probability $1 - \delta$: this is a P.A.C. hypothesis.

Two extensions to this learning model are strong and weak learning algorithms. Both algorithms run in a polynomial time. The strong learning algorithm outputs a hypothesis that is P.A.C while the weak one outputs a hypothesis with accuracy better than 0.5. The question of whether these two notions are equivalent is referred to as the "hypothesis boosting problem" since in order to show this equivalence we must boost the accuracy of the weak learning algorithm.

The proof that these notions are equivalent is provided by [Schapire, 1990], who presents the first algorithm for converting a weak learning algorithm into a strong one. Based on this idea, Boost-By-Majority, a simpler and more efficient boosting algorithm, is developed by [Freund, 1995]. This algorithm suffers from practical problem for estimating parameters. Adaboost, the first adaptive boosting algorithm, is developed by [Freund and Schapire, 1997]. This algorithm does not require parameter estimation.

Arcing family of algorithms denotes algorithms that **A**daptively **R**esample data and **C**ombine classifiers [Breiman, 1998]. Adaboost belongs to this family. Arc-x(h) is an ad-hoc algorithm developed by [Breiman, 1998] to better understand the behaviour od Adaboost. This algorithm uses a simple weight updating rule and a different method for combining hypotheses. For $h = 4$ Arc-x(h) performs better than $h = 1$ or 2.When compared to Adaboost, Arc-x4 performs equally well. However Breiman argues that higher values for h are not tested so further improvement is possible [Breiman, 1998].

In this paper, different values for the parameter h of Arc-x(h) algorithm are tested and their performance are compared to Adaboost and Arc-x4 in the reweighting framework using C4.5 [Quinlan, 1993] as classification algorithm. In section two, Adaboost and Arc-x(h) are introduced then results of previous empirical study are reviewed. In section three, the general framework of this empirical study is presented: classification and boosting algorithms, datasets and performance measure. Results are presented in section four. Finally, section five provides a conclusion to this work.

2 Arcing Algorithms

Adaboost and Arc-x(h) belongs to the Arcing family of algorithms. In this section, these algorithms are presented then results of previous empirical studies are reviewed.

2.1 Adaboost

Adaboost applies a classification algorithm to a dataset composed with labelled instances for a fixed number of iterations T. In each iteration t, $t = 1, \ldots, T$, a weight, $w_t(Z_i)$, is assigned to each instance $Z_i = (x_i, y_i), i = 1, \ldots, n$ in the dataset. It represents instance's importance. Based on this weight distribution, a classifier is outputted which predicts the class of each instance. Adaboost requires that the weighted error is less than 0.5. A parameter α_t is used to update the weights and to measure classifier's performance. The weight of misclassified instances is increased in order to force the algorithm to concentrate on them in the next iteration.

At the end of the process, a final classifier is obtained by combining classifiers from previous iterations using weighted majority vote. The parameter α_t represents the weight of classifier h_t generated in iteration t. The pseudocode of Adaboost for binary classification is presented in table 1.

```
Algorithm:Adaboost algorithm
Given: ${Z_1 = (x_1, y_1), \ldots, Z_n = (x_n, y_n)}$ where $x_i \in X, y_i \in Y = \{-1; +1\}$
1-Initialize $w_1(Z_i) = 1/n$ for $i = 1, \ldots, n$.
2-For t = 1 to T:
    •Train the algorithm using $w_t$ and get a classifier
     $h_t : X \mapsto \{-1; +1\}$
    • Compute        $\epsilon_t = \sum_{i:h_t(x_i) \neq y_i} w_t(Z_i)$

    • If             $\epsilon_t \geq 0.5$ stop.

    • Choose:        $\alpha_t = \frac{1}{2} \ln(\frac{1-\epsilon_t}{\epsilon_t})$

    • Update:        $w_{t+1}(Z_i) = \frac{w_t(Z_i) \exp(-\alpha_t y_i h_t(x_i))}{N_t}$

     where $N_t$ is a normalization factor

3-output the final hypothesis: $H(x) = sign(\sum_{t=1}^{T} \alpha_t h_t(x))$
```

Table 1. Adaboost algorithm for binary classification

2.2 Arc-x(j)

Arc-x(h) algorithm is developed by [Breiman, 1998] to study the behaviour of Adaboost. It is different from Adaboost in the following:

- it uses a simpler weight updating rule:

$$w_{t+1}(Z_i) = \frac{1 + m(Z_i)^h}{\sum(1 + m(Z_i)^h)}, \tag{1}$$

  where $m(Z_i)$ is the number of misclassifications of instance Z_i by classifiers generated in iterations $1, \ldots, t$ and h is an integer.
- classifiers are combined using simple majority vote.

2.3 Previous results

Empirical results show that Adaboost improves the performance of various classification algorithms, often by dramatic amount. Adaboost decreases the average error rate by 55.2% when applied to decision stump, a weak learning algorithm [Freund and Schapire, 1996]. Boosted decision stump performs equally well as C4.5 [Quinlan, 1993], a strong learning algorithm: Adaboost converts a weak learning algorithm into a strong one.

The ability of Adaboost to improve strong learning algorithm is investigated by [Freund and Schapire, 1996] and [Quinlan, 1996]. Experimental results show that Adaboost improves the average error rate.

Arc-x4 is tested on moderate and large data sets by [Breiman, 1998]. Results show that it improves the performance of CART [Breiman *et al.*, 1984] learning algorithm for all data sets.

Two different frameworks are considered by [Bauer and Kohavi, 1999] to test the performance Arc-x4: reweighting and subsampling. Subsampling uses the weight of instances to generate a different training set in each iteration while reweighting uses a fixed training set for all iterations. Arc-x4 produces a higher error reduction in the subsampling framework than in the reweighting framework.

Adaboost and Arc-x4 are compared in different framework and using different collections of datasets. Arc-x4 has an accuracy comparable to Adaboost without using the weighting scheme to construct the final classifier [Breiman, 1998] and [Bauer and Kohavi, 1999].

Arc-x(h) is tested for $h = 1, 2, 4$ by [Breiman, 1998]. However higher values of h are not tested so improvement is possible. Based on the performance measure used by [Bauer and Kohavi, 1999], increasing the factor h does not improve the performance of Arc-x(h) [Khanchel and Limam, to appear].

3 Empirical Study

In this section, the general framework of our empirical study in presented: classification algorithm, Arcing algorithms and data sets. The performance measure criterion is presented. Then performance of the different algorithms is compared.

3.1 General framework

C4.5 [Quinlan, 1993] is used as subroutine for the different boosting algorithms. In order to compare different boosting algorithms, a collection of data sets from UCI Machine learning Repository [Keogh and Merz, 1998] is used. Details of these data sets are presented in table 2.

Different values of the parameter h, $h \in \{5, 6, 8, 12\}$, are tested for the algorithm Arc-x(h). Results are compared to Adaboost and Arc-x4 in the reweighting framework . Boosting algorithms are applied for 25 iterations.

Data set	number of instances	number of attributes	number of classes
Liver disorders	345	7	2
Heart	270	13	2
Australian	690	14	2
Pima	760	8	2

Table 2. Data sets used in the empirical study

3.2 Performance measure

The performance boosting algorithms is usually evaluated using test error. This criteria takes into account only variability due to the choice of the test set. Comparison is often made without using rigorous statistical framework [Nadeau and Bengio, 2003]. Often it uses liberal estimators and leads to incorrect claims. A new method which takes into account variability due to the choice of training sets and test sets is presented by [Nadeau and Bengio, 2003]. The goal is to estimate the generalization error using the training data.

Given a data set Z^n of size n, a training set of size n_1 is generated from this data set. Using Z^{n_1} a classifier is generated. The loss incurred by this classifier on a new example Z_{n+1} can be expressed by $L(Z^{n_1}; Z_{n+1})$. We are interested in estimating ${}_{n_1}\mu = E[L(Z^{n_1}, Z_{n+1}]$.

To achieve this, we proceed as follows: suppose that the data set Z^n is composed with n labelled instances $Z^n = \{Z_1, \ldots, Z_n\}$. For $m = 1, \ldots, M$, Z^n is randomly splitted into 2 distinct subsets D_m and D_m^c each of size $n/2$. For each subset, we repeat the following process for $j = 1, \ldots, J$:

- Let S_j be a set of random index of size $n_1, n_1 = 4n/10$, chosen from $\{1, \ldots, n/2\}$ and let S_j^c of size $n_2 = n/10$ denote its complement.
- Let $Z_j = \{Z_i / i \in S_j\}$ be the training set and $Z_j^c = \{Z_i / i \in S_j^c\}$ be the test set.
- For $j = 1, \ldots, J$, use Z_j to generate a classifier, and let $L(j, i)$ be:

$$L(j, i) = Q_A(Z_j, i) - Q_B(Z_j, i) \qquad (2)$$

  where Q_A (Q_B) is the loss observed on instance i when the algorithm A (B) uses Z_j to generate classifiers.

  For classification problem, this loss can be expressed as:

$$Q_A(Z_j, i) = \begin{cases} 1 \text{ if instance } i \text{ is incorrectly classified,} \\ 0 \text{ otherwise.} \end{cases} \qquad (3)$$

- First we average over the test set Z_j^c of size n_2 to obtain:

$$\hat{\mu}_j = \frac{1}{n_2} \sum_{i \in S_j^c} L(j, i). \qquad (4)$$

- Then we average over J to obtain:

$$_{n_1}^{n_2}\hat{\mu}_J = \frac{1}{J}\sum_{j=1}^{J}\hat{\mu}_j \tag{5}$$

This process is repeated for D_m^c and for $m = 1, \ldots, M$. Each of the M split yields a pair $(_{n_1}^{n_2}\hat{\mu}_J, _{n_1}^{n_2}\hat{\mu}_J^c)$ which can be denote as $(\hat{\mu}_m, \hat{\mu}_m^c)$.
The generalization error $_{n_1}\mu$ is estimated using $_{n_1}^{n_2}\hat{\mu}_J$ and its variance is estimated using:

$$_{n_1}^{n_2}\hat{\sigma}_J^2 = \frac{1}{2M}\sum_{m=1}^{M}(\hat{\mu}_m - \hat{\mu}_m^c)^2. \tag{6}$$

Since $_{n_1}^{n_2}\hat{\mu}_J$ is the mean of Jn_2 loss $L(j,i)$, its distribution can be approximated by the normal distribution:

$$\frac{_{n_1}^{n_2}\hat{\mu}_J - _{n_1}\mu}{\sqrt{_{n_1}^{n_2}\hat{\sigma}_J^2}}. \tag{7}$$

Using this assumption we can perform inference about the performance of boosting algorithms using confidence interval. A confidence interval for $_{n_1}\mu$ at confidence level $1 - \alpha$ has the following form:

$$_{n_1}\mu \in [_{n_1}^{n_2}\hat{\mu}_J - c\sqrt{_{n_1}^{n_2}\hat{\sigma}_J^2} \quad , \quad _{n_1}^{n_2}\hat{\mu}_J + c\sqrt{_{n_1}^{n_2}\hat{\sigma}_J^2}] \tag{8}$$

where c is a percentile from $N(0,1)$ distribution.

4 Results

For each pair of algorithms and for each dataset we construct a confidence interval at confidence level 95%. If this interval includes zero, we conclude that both algorithms have comparable performance. Confidence interval are presented in table 3. Algorithms producing the same error rate are omitted.
The important observations for this empirical comparison are:

- For 3 data sets: Pima, Heart and Australian, Arc-x(h) outputs the same test error in all iterations and for different values of the parameter h. When compared to Adaboost, the confidence interval is:
  - $[-0.1197, 0.0371]$ for the Australian data and $[-0.0689, 0.1775]$ for the heart data. For these 2 data sets, we conclude that all algorithms have comparable performance.
  - $[-0.0365, -0.0073]$ for Pima data. Adaboost performs slightly better that Arc-x(h) algorithms

data sets	algorithms compared	confidence intervals
Australian	Arc-x(h) - Adaboost,$h \in \{4, 5, 6, 8, 12\}$	[-0.1197, 0.0371]
Heart	Arc-x(h) - Adaboost, $h \in \{4, 5, 6, 8, 12\}$	[-0.0688, 0.1775]
Pima (diabetes)	Arc-x(h) - Adaboost, $h \in \{4, 5, 6, 8, 12\}$	[-0.0365, -0.0073]
Puba (liver disorder)	Arc-x(h)- Arc-x4, $h \in \{5, 6, 8, 12\}$	[-0.0440, 0.0990]
	Arc-x4 - Adaboost	[-0.0976, 0.0214]
	Arc-x(h) - Adaboost	[-0.0227, 0.0014]

Table 3. Confidence intervals for difference of generalization error for different Arcing Algorithms

- For Bupa data, Arc-x(h) outputs the same test error in all iterations for $h = 5, 6, 8$ and12. Arc-x4 outputs a slightly lower test error. This difference is not significant because the confidence interval at 95% confidence level is $[-0.044, 0.099]$. Adaboost has a comparable performance to the different arc-x(h) algorithm: the confidence interval when compared to Arc-x4 is [-0.0976, 0.0214] and [-0.0227, 0.0014] when compared to the other Arc-x(h) algorithms.

5 conclusion

This empirical study is an extension to Breiman's study [Breiman, 1998] of the family of Arcing algorithms. Different values of the parameter h used by Arc-x(h) algorithm in the weight updating rule are tested and compared to Adaboost in the reweighting framework. The approach proposed by [Nadeau and Bengio, 2003] is adopted: performance measures take into account variability due to the training sets and test sets and comparisons are made using confidence intervals.

Based on this empirical study, increasing the factor h used by Arc-x(h) in the weight updating rule does not improve performance. Arc-x(h) performs equally as Adaboost for different values of h. Adaboost performs slightly better for only one data set.

Comparable performance is obtained using two different methods for combining classifiers. This agree with Breiman's claim that the error reduction is due to the weight updating rule.

The size of the data sets used in this empirical study is moderate. The framework proposed by [Nadeau and Bengio, 2003] uses small fractions of these data sets as training and test sets. Also the process generates many training data then averages the performance. This can explain the comparable performance of the different boosting algorithm considered. It will be interesting to test these algorithms on large data sets where large training and test sets can be generated.

References

[Bauer and Kohavi, 1999]E. Bauer and R. Kohavi. An empirical comparison of voting classification algorithm: Bagging, boosting and variants. *Machine Learning*, pages 105–142, 1999.

[Breiman *et al.*, 1984]L. Breiman, J.H. Friedman, R.A. Olshen, and C.J. Stone. *Classification and Regression Trees*. Chapman anf Hall, London, 1984.

[Breiman, 1996]L. Breiman. Bagging predictors. *Machine Learning*, 26(2):123–140, 1996.

[Breiman, 1998]L. Breiman. Arcing classifiers. *The annals of statistics*, 26(3):801–849, 1998.

[Freund and Schapire, 1996]Y. Freund and R.E. Schapire. Experiments with a new boosting algorithm. In *machine learning: Proceedings of the thirteenth international conference*, pages 148–156. Morgan Kaufmann San Francisco, 1996.

[Freund and Schapire, 1997]Y. Freund and R.E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55(1):119–139, 1997.

[Freund, 1995]Y. Freund. Boosting a weak learning algorithm by majority. *Information and Computation*, 12(2):256–285, 1995.

[Keogh and Merz, 1998]C. Blakes E. Keogh and C.J. Merz. Uci repository of machine learning databases http://www.ics.uci.edu/ mlearn/ mlrepository.html. 1998.

[Khanchel and Limam, to appear]R. Khanchel and M. Limam. Empirical comparison of boosting algorithms. Springer-Verlag, to appear.

[Nadeau and Bengio, 2003]C. Nadeau and Y. Bengio. Inference for the generalization error. *Machine Learning*, 52(2):239–281, 2003.

[Quinlan, 1993]J.R. Quinlan. *C4.5: Programs for machine learning*. Morgan Kaufmann, 1993.

[Quinlan, 1996]J.R. Quinlan. Bagging, boosting, and C4.5. In *Proceedings of the Thirteenth National Conference on Artificial Intelligence and the Eighth Innovative Applications of Artificial Intelligence Conference*, pages 725–730, Menlo Park, August4–8 1996. AAAI Press / MIT Press.

[Schapire, 1990]R.E. Schapire. The strength of weak learnability. *Machine Learning*, 5(2):197–227, 1990.

Estimation Statistique de fonctions d'appartenance d'ensembles flous : le cas des fonctions trapézoïdales croissantes

Florence Dupuis[1,2] and Alain Hillion[1]

[1] GET - ENST Bretagne
Département Image et Traitement de l'Information
CNRS UMR TAMCIC
Technopôle de Brest Iroise, CS 83818,
29238 BREST Cedex, France
(e-mail: florence.dupuis@enst-bretagne.fr,
alain.hillion@enst-bretagne.fr)

[2] ENIB
Technopôle de Brest Iroise, CS 73862,
29238 BREST Cedex 3, France
(e-mail: florence.dupuis@enib.fr)

Abstract. Nous présentons ici une méthode nouvelle de détermination des fonctions d'appartenance floue appliquée au cas des fonctions trapézoïdales croissantes. Pour l'estimation, nous considérons un modèle statistique reposant essentiellement sur la notion de processus cohérent. Nous donnons par cette méthode des estimateurs, ponctuels et par région de confiance, des paramètres d'une fonction d'appartenance trapézoïdale croissante. Les résultats obtenus sont illustrés par des simulations.

Keywords: Fonction d'appartenance floue, Processus "expert" cohérent, Estimation ponctuelle, Estimation par région de confiance.

1 Introduction

En logique floue, la description d'un ensemble flou réel A passe par la connaissance de sa fonction $A(x)$. Notre but est d'estimer des fonctions d'appartenance de type trapézoïdal croissant à partir d'informations concernant des points de l'ensemble $\mathbb{R}$, appelés points de contrôle. Dans le domaine de la reconstruction de fonction d'appartenance flou à partir d'observations, les méthodes principalement utilisées relèvent de l'analyse numérique ou de méthodes probabilistes [Shen *et al.*, 2000]; [Tamaki *et al.*, 1998]; [Cheng and Chen, 1997]; [Civanlar and Trussell, 1986]; [Devi and Sarma, 1985]. Nous utilisons ici une méthode statistique nouvelle présentée dans [Dupuis and Hillion, 2004].

2 Processus cohérent

On considère A un sous ensemble flou de $\mathbb{R}$. Pour l'étude nous interrogeons des experts sur un ensemble fini ordonné de points $\mathfrak{X} = \{x_1, \ldots, x_n / \forall i, 1 \leq i \leq n, x_i \in \mathbb{R}, x_1 < x_2 < \ldots x_n\}$. Les x_i représentent les points de contrôle dont l'expert pense qu'ils appartiennent totalement ou pas du tout à A. Pour tout $x \in \mathbb{R}$, on note $X(x)$ la réponse binaire à cette question.

Définition 1 *On définit le processus cohérent $\{X(x)\}$ associé à l'ensemble flou A par (cf [Dupuis and Hillion, 2004]) un processus "expert" $\{X(x)\}_{x \in \mathfrak{X}}$ discret à valeurs dans $\{0,1\}$, auquel on impose pour tout x,*

- $\mathrm{E}[X(x)] = A(x)$.
- *Si $X(x) = 1$, alors pour tout y tel que $A(y) \geq A(x)$, $X(y) = 1$.*
- *Si $X(x) = 0$, alors pour tout y tel que $A(y) \leq A(x)$, $X(y) = 0$.*

Dans le cas où la fonction d'appartenance est croissante, le processus cohérent associé sera nécessairement croissant. On peut alors calculer la loi de la variable aléatoire $X(x_{(n)}) = (X(x_1), X(x_2), \ldots, X(x_n))$ à valeurs dans $\Omega = \cup_{0 \leq k \leq n} \left(\{0\}^k \times \{1\}^{n-k}\right)$. On définit les éléments de Ω par une suite $\alpha_{(n)} = (\alpha_1, \alpha_2, \ldots, \alpha_n)$ où pour tout i, $1 \leq i \leq n$, $\alpha_i = 0$ ou 1, est la valeur réponse à la question: "x_i appartient-il à l'ensemble flou A?" (par convention on pose $\alpha_0 = 0$ et $\alpha_{n+1} = 1$ avec $x_0 = -\infty$ et $x_{n+1} = +\infty$, la fonction $A(x)$ n'étant pas constante). On constate que seul l'instant de saut (Fig.1) du processus "expert" intervient dans l'expression de la probabilité.

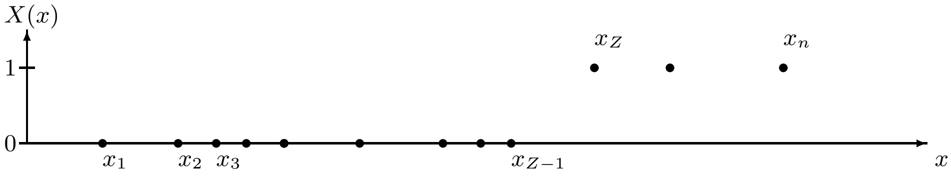


Fig. 1. Exemple de trajectoire "expert". On visualise l'instant de saut "Z".

Théorème 1 *Soit $Z = inf_{1 \leq i \leq n+1}\{i / X(x_i) = 1\}$, $\forall \alpha_{(n)} \in \Omega$, si $r(\alpha) = inf_{1 \leq i \leq n+1}\{i / \alpha_i = 1\}$,*

$$\mathrm{P}[X(x_{(n)}) = \alpha_{(n)}] = (A(x_{r(\alpha)}) - A(x_{r(\alpha)-1})) = \mathrm{P}[Z = r(\alpha)]. \quad (1)$$

*Démonstration.*La suite $\{A(x_i)\}_{1 \leq i \leq n}$ est croissante et par définition du processus cohérent la suite $\{X(x_i)\}_{1 \leq i \leq n}$ est également croissante. Soit $Z = inf_{1 \leq i \leq n+1}\{i / X(x_i) = 1\}$. Par définition de l'inf, pour tout $\forall k$, $1 \leq k \leq n+1$, $\{Z = k\} = \{0 = \ldots = X(x_{k-1}), X(x_k) = \ldots = 1\}$. Donc, $\forall \alpha_{(n)} \in \{0,1\}^n$, si $r(\alpha) = inf_{1 \leq i \leq n+1}\{i / \alpha_i = 1\}$, $\mathrm{P}[X(x_{(n)}) = \alpha_{(n)}] = \mathrm{P}[Z = r(\alpha)] 1\!\!1_{\alpha_{(n)} \in \Omega}$.

Puis d'après la définition du processus cohérent, et étant donné que $X(x_{(n)})$ est une suite croissante, on a la relation suivante pour tout k, $1 \leq k \leq n+1$,

$$\begin{aligned} &\mathrm{P}[X(x_0) = 0, \ldots, X(x_{k-1}) = 0, X(x_k) = 1, \ldots, X(x_{n+1}) = 1] \\ &= \mathrm{P}[X(x_{k-1}) = 0, X(x_k) = 1] \\ &= A(x_k) - A(x_{k-1}). \end{aligned}$$

■

3 Estimation des paramètres

Une fonction de type trapézoïdale croissante (Fig.2) est entièrement définie par la donnée du paramètre $\theta = (\theta_1, \theta_2) \in \mathbb{R}^2$ avec $\theta_1 < \theta_2$, tel que

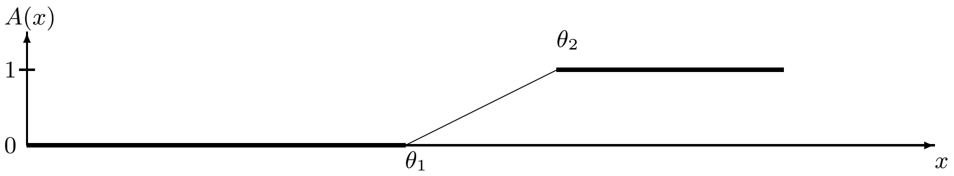


Fig. 2. Graphe d'une fonction trapézoïdale croissante.

$$\begin{aligned} \text{Si} \;\; x \leq \theta_1 \;\; & A_\theta(x) = 0, \\ \text{Si} \;\; \theta_1 < x \leq \theta_2 \;\; & A_\theta(x) = \frac{x - \theta_1}{\theta_2 - \theta_1}, \\ \text{Si} \;\; \theta_2 < x \;\; & A_\theta(x) = 1. \end{aligned} \qquad (2)$$

On se propose d'estimer θ_1 et θ_2 à partir d'un m_échantillon $X_1, X_2, \ldots, X_m$ du processus "expert", $m \geq 2$. Si pour tout i, $1 \leq i \leq m$, Z_i est le saut associé au processus X_i, on note $Z'_1, Z'_2, \ldots, Z'_m$ ces instants de saut réordonnés, i.e tels que $min_{1 \leq j \leq m}(Z_j) = Z'_1 \leq Z'_2 \leq \ldots \leq Z'_m = sup_{1 \leq j \leq m}(Z_j)$, $\forall (\alpha_{j(n)})_{1 \leq j \leq m} \in \Omega^m$, on déduit du théorème 1 l'expression de la vraisemblance $L_\theta(\alpha_{1(n)}, \alpha_{2(n)}, \ldots, \alpha_{m(n)})$

$$\frac{\mathrm{P}_\theta[Z'_1 = r(\alpha'_1)]\mathrm{P}_\theta[Z'_m = r(\alpha'_m)]}{(\theta_2 - \theta_1)^{m-2}} \prod_{j=2}^{m-1} (x_{r(\alpha'_j)} - x_{r(\alpha'_j)-1}) \qquad (3)$$

pour $\theta_1 < x_{r(\alpha'_2)} \leq x_{r(\alpha'_{m-1})-1} < \theta_2$, où les $\alpha'_{1(n)}, \alpha'_{2(n)}, \ldots, \alpha'_{m(n)}$ sont les $\alpha_{1(n)}, \alpha_{2(n)}, \ldots, \alpha_{m(n)}$ réordonnés tels que $r(\alpha'_1) \leq r(\alpha'_2) \leq \ldots \leq r(\alpha'_m)$.

Proposition 1 *Le paramètre $\theta = (\theta_1, \theta_2)$ est identifiable si et seulement s'il y a plus de deux points de contrôle compris entre θ_1 et θ_2.*

En effet, si on note F_{A_θ} la loi du n_uple $X(x_{(n)})$ où X est un processus cohérent de A paramétrée par θ. Soient $\theta = (\theta_1, \theta_2)$ et $\varphi = (\varphi_1, \varphi_2) \in \mathbb{R}^2$, $F_{A_\theta} = F_{A_\varphi} \Longleftrightarrow \forall k, 1 \leq k \leq n,\ A_\theta(x_k) = A_\varphi(x_k)$.

i) Si il y a moins d'un point de contrôle entre θ_1 et θ_2 alors on peut trouver plusieurs fonctions trapézoïdales croissantes ayant les mêmes valeurs aux points de contrôle.

ii) Si il y a plus de deux points de contrôle compris entre θ_1 et θ_2 : $\exists J$ tels que $x_{J-1} \leq \theta_1 < x_J < x_{J+1} < \theta_2$.
De plus, $\begin{cases} A_\theta(x_J) = A_\varphi(x_J) \\ A_\theta(x_{J+1}) = A_\varphi(x_{J+1}) \end{cases}$ est un système de Cramer ($\theta_2 \neq \theta_1$, $x_{J+1} \neq x_J$) de solution unique $\varphi_1 = \theta_1$, $\varphi_2 = \theta_2$. ■

Proposition 2 *La statistique* $(min_{1\leq j\leq m}(Z_j), sup_{1\leq j\leq m}(Z_j))$ *est exhaustive minimale complète.*

D'après (3) et le théorème de factorisation, (Z'_1, Z'_m) est exhaustive. Soit h une fonction définie sur $H = \{(x,y) \in \mathbb{N}^2, x < y\}$ telle que, pour tout $\theta = (\theta_1, \theta_2) \in \mathbb{R}^2$, $\theta_1 < \theta_2$, $\mathrm{E}_\theta[h(Z'_1, Z'_m)] = 0$. On montre par récurrence sur n que $\forall n \in \mathbb{N}^*, \forall k \in \mathbb{N}, h(k, k+n) = 0$. On en déduit que (Z'_1, Z'_m) est exhaustive, complète donc minimale. ■

On définit les indices J et M des points de contrôle encadrant les paramètres, par $x_{J-1} \leq \theta_1 < x_J$ et $x_{M-1} < \theta_2 \leq x_M$.

Proposition 3 *L'estimateur du maximum de vraisemblance* $\left(\hat\theta_1^{(m)}, \hat\theta_2^{(m)}\right) = \left(\underset{1\leq j\leq m}{min}\ (x_{Z_j-1}), \underset{1\leq i\leq m}{max}(x_{Z_j})\right)$ *converge p.s vers* (x_{J-1}, x_M).

Pour maximiser la vraisemblance (3), on étudie la quantité $\frac{\mathrm{P}_\theta[Z'_1=r(\alpha'_1)]\mathrm{P}_\theta[Z'_m=r(\alpha'_m)]}{(\theta_2-\theta_1)^{m-2}} = \frac{\varphi(\theta_1,\theta_2,x_{r(\alpha'_1)},x_{r(\alpha'_m)})}{(\theta_2-\theta_1)^m}$ où $\varphi(\theta_1, \theta_2, x_{r(\alpha'_1)}, x_{r(\alpha'_m)})$ vaut

$$\begin{cases} \left(x_{r(\alpha'_1)} - x_{r(\alpha'_1)-1}\right)\left(x_{r(\alpha'_m)} - x_{r(\alpha'_m)-1}\right) & \text{si } \begin{cases} \theta_1 < x_{r(\alpha'_1)-1} < x_{r(\alpha'_1)} \\ x_{r(\alpha'_m)-1} < x_{r(\alpha'_m)} < \theta_2 \end{cases} \\ \left(x_{r(\alpha'_1)} - \theta_1\right)\left(x_{r(\alpha'_m)} - x_{r(\alpha'_m)-1}\right) & \text{si } \begin{cases} x_{r(\alpha'_1)-1} \leq \theta_1 < x_{r(\alpha'_1)} \\ x_{r(\alpha'_m)-1} < x_{r(\alpha'_m)} < \theta_2 \end{cases} \\ \left(x_{r(\alpha'_1)} - x_{r(\alpha'_1)-1}\right)\left(\theta_2 - x_{r(\alpha'_m)-1}\right) & \text{si } \begin{cases} \theta_1 < x_{r(\alpha'_1)-1} < x_{r(\alpha'_1)} \\ x_{r(\alpha'_m)-1} < \theta_2 \leq x_{r(\alpha'_m)} \end{cases} \\ \left(x_{r(\alpha'_1)} - \theta_1\right)\left(\theta_2 - x_{r(\alpha'_m)-1}\right) & \text{si } \begin{cases} x_{r(\alpha'_1)-1} \leq \theta_1 < x_{r(\alpha'_1)} \\ x_{r(\alpha'_m)-1} < \theta_2 \leq x_{r(\alpha'_m)} \end{cases} \end{cases}.$$

Cette quantité est maximum sur le bord de l'ensemble de définition, i.e si et seulement si $\theta_1 = x_{r(\alpha'_1)-1}$ et $\theta_2 = x_{r(\alpha'_m)}$.

La convergence presque sûre de $\left(\hat{\theta}_1^{(m)}, \hat{\theta}_2^{(m)}\right)$ vers les extrémités du support de la fonction de répartition de Z est assurée par les théorèmes de convergence des valeurs extrêmes [Embrechts *et al.*, 1997]. ■

Dans le cas où le pas varie avec la taille de l'échantillon, on peut obtenir un résultat plus précis :

Théorème 1 *Sous l'hypothèse $p_m = o(\frac{1}{m})$, où p_m est le pas entre deux points de contrôle,*

$$m \left(\frac{\hat{\theta}_1^{(m)} - \theta_1}{\theta_2 - \theta_1}, \frac{\theta_2 - \hat{\theta}_2^{(m)}}{\theta_2 - \theta_1} \right) \textit{ converge en loi vers } (T_1, T_2), \tag{4}$$

où T_1 et T_2 sont deux variables aléatoires indépendantes de même loi exponentielle de paramètre 1.

*Démonstration.*Comme dans la théorie asymptotique des variables extrêmes connue dans [Galambos, 1978], les estimateurs $\hat{\theta}_1^{(m)}$ et $\hat{\theta}_2^{(m)}$ sont asymptotiquement indépendants. D'autre part, si on note $\mathcal{E}nt$ la partie entière, pour tout $t, y \in \mathbb{R}$, $\mathrm{F}_Z\left(t + \frac{y}{m}\right) = \mathrm{P}_\theta\left[Z \leq p_m \mathcal{E}nt\left[\frac{t}{p_m} + \frac{y}{mp_m}\right]\right]$. En conséquence, avec l'hypothèse $p_m = o(\frac{1}{m})$, pour tout $y_1 \in \mathbb{R}$, $\left(1 - \mathrm{F}_Z\left(\theta_1 + \frac{y_1}{m}\right)\right)^m \underset{m \to +\infty}{\longrightarrow} 1 - \left(1 - e^{-\frac{y_1}{\theta_2 - \theta_1}}\right)$, d'où le résultat de convergence. ■

Corollaire 1 *On note, $\forall \alpha \in [0,1]$, $k_\alpha = -\ln(1 - \sqrt{1-\alpha})$. Sous l'hypothèse du théoème 1, on en déduit l'intervalle de confiance asymptotiquement minimum en volume, de seuil de confiance $1 - \alpha$, pour (θ_1, θ_2) :*

$$\left[\hat{\theta}_1^{(m)} - \frac{1}{m}(\hat{\theta}_2^{(m)} - \hat{\theta}_1^{(m)})k_\alpha, \hat{\theta}_1^{(m)}\right] \times \left[\hat{\theta}_2^{(m)}, \hat{\theta}_2^{(m)} + \frac{1}{m}(\hat{\theta}_2^{(m)} - \hat{\theta}_1^{(m)})k_\alpha\right] \tag{5}$$

Démonstration.
Soit $\hat{C}_m = \left[\hat{\theta}_1^{(m)} - \frac{1}{m}(\hat{\theta}_2^{(m)} - \hat{\theta}_1^{(m)})k, \hat{\theta}_1^{(m)}\right] \times \left[\hat{\theta}_2^{(m)}, \hat{\theta}_2^{(m)} + \frac{1}{m}(\hat{\theta}_2^{(m)} - \hat{\theta}_1^{(m)})k'\right]$ tel que $\mathrm{P}_\theta\left[(\theta_1, \theta_2) \in \hat{C}_m\right] \geq 1 - \alpha$. On définit C par $\mathrm{P}_\theta[(\theta_1, \theta_2) \in \hat{C}_m] = \mathrm{P}_\theta\left[m\left(\frac{\hat{\theta}_1^{(m)} - \theta_1}{\theta_2 - \theta_1}, \frac{\theta_2 - \hat{\theta}_2^{(m)}}{\theta_2 - \theta_1}\right) \in C\right]$. D'après le théorème 1, on obtient la convergence $\mathrm{P}_\theta[(\theta_1, \theta_2) \in \hat{C}_m] \underset{m \to +\infty}{\longrightarrow} \mathrm{P}\left[(T_1, T_2) \in C\right]$. Le minimum du volume de $\hat{C}_m$, i.e kk' n'est pas atteint dans le domaine $(1 - e^{-k})(1 - e^{-k'}) > 1 - \alpha$. Sur le bord du domaine, la méthode des multiplicateurs de Lagrange donne comme condition $k = k'$ avec k tel que $\mathrm{P}_\theta[(\theta_1, \theta_2) \in \hat{C}_m] = 1 - \alpha$, d'où le résultat. ■

Remarque 1 Si le contrôle était continu, on aurait $T = inf\{x \in \mathbb{R} / X(x) \neq X(0)\}$ le saut "continu". Les variables Z et T sont liées par les relations suivantes : $x_{Z-1} < T \leq x_Z$ et $\forall i, 1 \leq i \leq n+1$, $\mathrm{P}_\theta\left[Z = i\right] = \mathrm{P}_\theta\left[x_{i-1} < T \leq x_i\right]$.

De plus, pour $x > 0$, $\mathrm{P}_\theta\,[T > x] = \mathrm{P}_\theta\,[X(x) = 0] = 1 - A(x)$ et $\mathrm{P}_\theta\,[T \leq x] = \mathrm{P}_\theta\,[X(x) = 1] = A(x)$.

Dans le cas particulier où A est trapézoïdale croissante, T est une variable aléatoire de loi uniforme sur $[\theta_1, \theta_2]$, $T \sim \mathcal{U}\left([\theta_1, \theta_2]\right)$. Dans le cadre statistique d'un m_échantillon, l'estimateur du maximum de vraisemblance $\left(\hat{\theta}_1^{(m)}, \hat{\theta}_2^{(m)}\right) = \left(\underset{1\leq j\leq m}{inf}\ (T_j), \underset{1\leq i\leq m}{sup}(T_j)\right)$ (cf. [Johnson and Kotz, 1970]) correspond à la version asymptotique (quand le pas tend vers 0) du cas discret. De plus, d'après la théorie des variables extrêmes, on retrouve la convergence presque sûre de $\left(\hat{\theta}_1^{(m)}, \hat{\theta}_2^{(m)}\right)$ vers (θ_1, θ_2) et on a la convergence (4) du théorème 1 (cf. [Galambos, 1978]).

4 Simulations

Pour simuler le processus, il est nécessaire d'en connaitre certaines propriétés:

Proposition 4 *Le processus cohérent associé à une fonction d'appartenance croissante, est un processus de markov (en général non homogène).*

En effet, d'aprés la définition 1 du processus cohérent, $\forall i, 1 \leq i \leq n+1$, $\mathrm{P}[X(x_{i+1}) = \alpha_{i+1}/X(x_i) = \alpha_i, \ldots, X(x_0) = \alpha_0] = \mathrm{P}[X(x_{i+1}) = \alpha_{i+1}/X(x_i) = \alpha_i]$. D'autre part, $\mathrm{P}[X(x_1) = \alpha_1] = A(x_1)1\,1_{\alpha_1=1} + (1 - A(x_1))\mathbb{1}_{\alpha_1=0}$ et pour tout i, si $\mathrm{P}[X(x_i) = \alpha_i] \neq 0$, on a

$$\mathrm{P}[X(x_{i+1}) = \alpha_{i+1}/X(x_i) = \alpha_i] = \begin{cases} 1 & \text{si } \alpha_i = \alpha_{i+1} = 1 \\ \frac{1-A(x_{i+1})}{1-A(x_i)} & \text{si } \alpha_i = \alpha_{i+1} = 0 \\ \frac{A(x_{i+1})-A(x_i)}{1-A(x_i)} & \text{si } \alpha_i < \alpha_{i+1} \end{cases}$$

■

Toutes les simulations suivantes seront effectuées avec comme paramètre $\theta = (4, 6)$. Les graphiques présentent pour chaque simulation l'histogramme des sauts puis la fonction d'appartenance réelle, la fonction d'appartenance estimée et sous l'hypothèse du pas petit (cf théorème 1), une région de confiance à 95%, pour l'estimation de la fonction d'appartenance.
Dans le cas où le pas n'est pas négligeable devant l'inverse du nombre d'observations, par exemple pour m=50 et n=10 (fig. 3), on obtient $\left(\hat{\theta}_1^{(m)}, \hat{\theta}_2^{(m)}\right) = (3.6, 6.2)$.
Sous l'hypothèse du pas petit et pour un intervalle de points de contrôle fixé, si m est le nombre d'"experts" interrogés, on suppose pour les simulations suivantes que $p_m = \frac{1}{m^2}$. On présente les simulations pour un nombre d'observations petit $m = 5$ (fig. 4, $p_m \simeq 0.04$ et le nombre de points de contrôle $n \simeq 180$), on a $\left(\hat{\theta}_1^{(m)}, \hat{\theta}_2^{(m)}\right) = (4.3, 5.6)$. Pour un nombre d'observations plus élevé $m = 20$, ($p_m \simeq 0.0025$ et $n \simeq 2800$), on obtient $\left(\hat{\theta}_1^{(m)}, \hat{\theta}_2^{(m)}\right) = (4, 5.9)$.

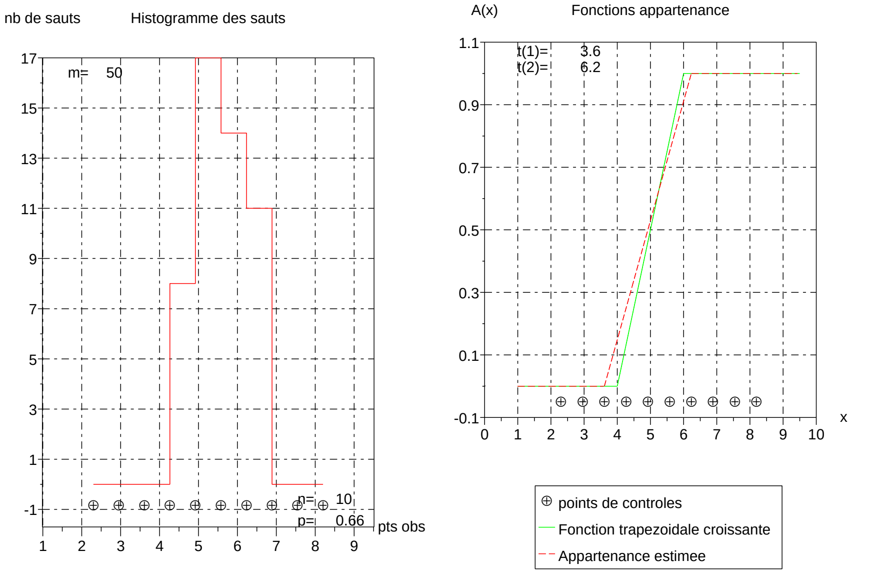


Fig. 3. Exemples de simulation pour n=10, m=50.

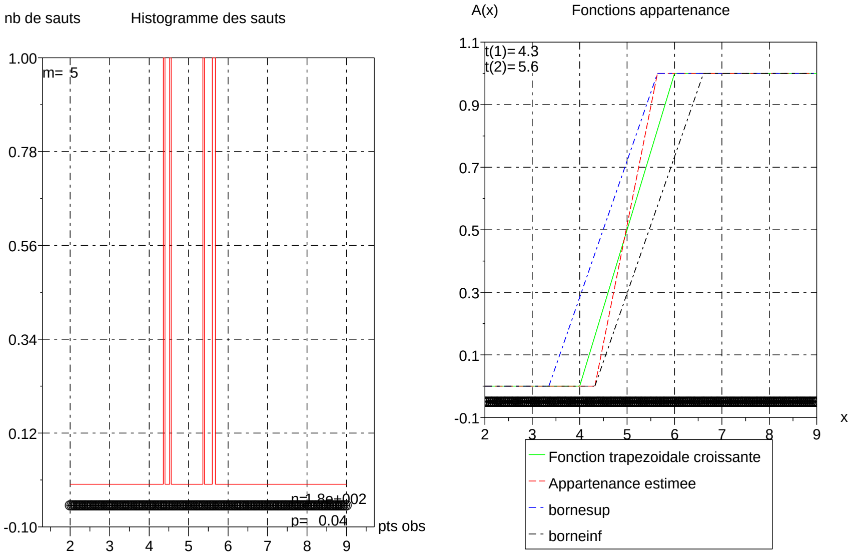


Fig. 4. Simulation sous l'hypothèse du théorème 1, cas m=5.

5 Conclusion

Dans cet article nous avons proposé une estimation des paramètres d'une fonction d'appartenance trapézoïdale croissante. Cette estimation repose sur la méthode du maximum de vraisemblance; la convergence des estimateurs et la loi limite permettent de définir un intervalle de confiance. Des généralisations sont en cours dans plusieurs directions : d'une part nous étendons cette méthode aux fonctions d'appartenance trapézoïdales quelconques, d'autre part nous comparons les résultats d'estimation des paramètres à ceux obtenus par la méthode des moments, enfin nous procédons à l'estimation directe de la fonction d'appartenance par minimisation d'une fonction de coût fondée sur des distances entre ensembles flous.

References

[Arslan and Kaya, 2001]A. Arslan and M. Kaya. Determination of fuzzy logic membership functions using genetic algorithms. *Fuzzy sets and systems*, **118**:297–306, 2001.

[Bloch *et al.*, 1997]I. Bloch, L. Aurdal, D. Bijno, and J. Muller. Estimation of class membership functions for grey-level based image fusion. In *ICIP'97*, volume **III**, pages 268–271, Santa Barbara, CA, October 1997.

[Bloch, 2003]I. Bloch. *Fusion d'informations en traitement du signal et des images.* Hermès science. Lavoisier, 2003.

[Bouchon-Meunier and Marsala, 2003]B. Bouchon-Meunier and C. Marsala. *Logique floue, principes, aide à la décision.* Hermès science. Lavoisier, 2003.

[Caillol *et al.*, 1993]H. Caillol, A. Hillion, and W. Pieczynski. Fuzzy random fields and unsupervised image segmentation. *IEEE Transactions on geoscience and remote sensing*, **31**(4):801–810, 1993.

[Cheng and Chen, 1997]H.D. Cheng and J.R. Chen. Automatically determine the membership function based on the maximum entropy principle. *Information Sciences*, **96**:163–182, 1997.

[Civanlar and Trussell, 1986]M.R. Civanlar and H.J. Trussell. Contructing membership functions using statistical data. *Fuzzy sets and systems*, **18**:1–13, 1986.

[D'Alché-Buc, 2003]F. D'Alché-Buc. Association des systèmes d'inférence floue avec les méthodes connexionnistes et évolutionnistes. In *Traitement de données complexes et commande en logique floue (dir. B. Bouchon-Meunier and C. Marsala)*, Hermès science, chapter 5, pages 199–225. Lavoisier, 2003.

[Devi and Sarma, 1985]B. Bharathi Devi and V.V.S. Sarma. Estimation of fuzzy membership from histograms. *Information Sciences*, **35**:43–59, 1985.

[Dupuis and Hillion, 2004]F. Dupuis and A Hillion. Estimation statistique des fonctions d'appartenance d'ensembles flous. Ecole Nationale Supérieure des Télécommunications de Bretagne, 2004.

[Embrechts *et al.*, 1997]P. Embrechts, C. Klüppelberg, and T. Mikosch. *Modelling Extremal Events.* Springer, 1997.

[Galambos, 1978]J. Galambos. *The asymptotic theory of extreme order statistics.* Wiley, 1978.

[Hummel and Landy, 1988]R.A. Hummel and M.S. Landy. A statistical viewpoint on the theory of evidence. *IEEE Transactions on pattern analysis and machine intelligence*, **10**(2):235–247, 1988.

[Johnson and Kotz, 1970]N L. Johnson and S. Kotz. *Continuous univariate distributions - II.* Wiley, 1970.

[Shen *et al.*, 2000]J. Shen, W. Shen, H.J. Sun, and J.Y. Yang. Fuzzy neural nets with non-symétric π membership functions and applications in signal processing and image analysis. *Signal processing*, **80**(6):965–983, 2000.

[Tamaki *et al.*, 1998]F. Tamaki, A. Kanagawa, and H. Ohta. Identification of membership functions based on fuzzy observation data. *Fuzzy sets and systems*, **93**(2):311–318, 1998.

Parameter optimization for Support Vector Regression

Üstün, B., Melssen, W.J., and Buydens L.M.C.

IMM, Analytical Chemistry,
Radboud University of Nijmegen,
Toernooiveld 1, NL-6525 ED Nijmegen,
The Netherlands

Abstract. Many academic and industrial platforms rely on a statistically sound and robust qualitative (classification) and quantitative (regression) analysis of the data of interest. Recently, Support Vector Machines (SVMs) have emerged as a powerful multivariate modeling technique for classification as well as regression purposes. However, due to the explicit flexibility of the SVM, some vital model parameters need to be selected, which affect the resulting model performance. Therefore, those parameter settings need to be optimized to achieve a good generalization performance. This research focuses on the development of a fast, robust and fully automated method to obtain the optimal parameter settings (that is, kernel type, kernel parameter, the so-called ϵ-insensitive margin and a penalty weight) in case of SVM regression. The optimization of the parameters will be accomplished through the use of Genetic Algorithms (global optimization) in combination with a Simplex optimization (refined local optimization). Preliminary bench-marks on well-known data sets indicate that the optimized SVM outperforms all other methods applied to these data sets. For example, the SVM optimization approach has improved the model performance by approximately 50% on a well-known data set by comparison with the commonly used SVM grid search optimization.

Keywords: SVM.

Bayesian Effetive Sample and Parameter Size

Xiaodong Lin

University of Cincinnati and National Institute of Statistical Sciences

Abstract. Suppose we have a posterior density for a parameter given a sample and we form a second posterior density for the same parameter, based on a different model and a different data set. Then we can evaluate the relative entropy distance between the two posteriors. Minimizing the relative entropy over the second sample gives the virtual sample that would make the second posterior as close as possible to the first in an inferential sense. For instance, if the first posterior is based on a dependent dataset and the second posterior is based on an independence likelihood, the optimization transfers the effective inferential power of the dependent sample into the independent sample. We present further examples of this type of optimization for models with nuisance parameters, finite mixture models and models for correlated data. Finally, we use our approach to choose the effective parameter size in a Bayesian hierarchical model.

Keywords: Bayesian hierarchical model.

Evaluating the Sensitivity of Goodness-of-Fit Indices to Data Perturbation: An Integrated MC-SGR Approach

Massimiliano Pastore[1] and Luigi Lombardi[2]

[1] Department of Psychology
University of Cagliari
Via Is Mirrionis, 1 I-09123 Cagliari, Italy
(e-mail: mpastore@unica.it)

[2] Department of Cognitive Sciences and Education
University of Trento
Via Matteo del Ben, 5 I-38068 Rovereto (TN), Italy
(e-mail: lombardi@form.unitn.it)

Abstract. In this paper, we address the problem of evaluating goodness-of-fit indices in structural equation modeling when corrupted data are considered. Starting from the introduction of a new method, called MC-SGR, we evaluate the sensitivity of four different fit indices (two absolute fit-indices: GFI and AGFI, and two incremental fit-indices: CFI and NNFI) to structured perturbations.

Keywords: Sample Generation by Replacements, Monte Carlo simulations, Goodness-of-fit indices.

1 Introduction

The issue of perturbations in real or simulated data has been substantially neglected in evaluating the adequacy of fit indices used to test covariance structure modeling. Nevertheless, it is certainly legitimate to wonder whether fit indices are reliably sensitive to data corruption. In particular, we would expect that a good index should approach its maximum under correct model specification and uncorrupted data, but also degrade substantially under massive data perturbation. In this paper we provide a possible methodological solution to the problem of evaluating the sensitivity of fit indices in structural equation modeling when perturbed data are considered. In particular, in our study the sensitivity of four different fit indices (two absolute fit-indices: GFI and AGFI, and two incremental fit-indices: CFI and NNFI) to perturbed data is examined in three different factorial models. The sensitivity evaluation is carried out by means of a new integrated approach which combines standard Monte Carlo (MC) simulations and a recent data generating procedure called Sample Generation by Replacements (SGR, [Lombardi *et al.*, 2004]).

The paper is organized as follows. Section 2 outlines the integrated MC-SGR approach. Section 3 describes the simulation study for evaluating the goodness-of-fit indices under perturbed data scenarios. In Section 4 we discuss results of the simulation study. Finally, Section 5 reports some concluding remarks.

2 Integrated approach: MC + SGR

In this section we describe how to integrate the SGR procedure with MC simulations in order to evaluate the sensitivity of fit-indices in structured scenarios of data perturbation.

2.1 Generating data replacements: the SGR method

We think of the full dataset as being represented by an $n \times m$ matrix $\mathbf{D}$ (that is, n observations, each containing m elements), of which a certain portion $\mathbf{D}^c$ is actually represented by corrupted-data (corruption due to possibly fake data points in $\mathbf{D}$). The corrupted-portion $\mathbf{D}^c$ of $\mathbf{D}$ together with the uncorrupted portion $\mathbf{D}^u$ of $\mathbf{D}$, constitutes the full data set, that is to say $\mathbf{D} = \mathbf{D}^c \cup \mathbf{D}^u$. The general idea is the following: under the assumption of $\varrho \leq n \times m$ corrupted data points in $\mathbf{D}$, we replace some portions $\mathbf{D}_1, \ldots, \mathbf{D}_s$ of $\mathbf{D}$, each of which contains exactly ϱ elements, with new components $\mathbf{X}_1^r, \ldots, \mathbf{X}_s^r$ in such a way that for all $h = 1, \ldots, s$, all the corresponding elements in $\mathbf{X}_h^r$ and $\mathbf{D}_h$ are different. The exact uncorrupted portion $\mathbf{D}^u$ is assumed to be unknown and only the value of ϱ is supposed to be known. Moreover, all entries in $\mathbf{D}$ are also assumed to be equally likely in the process of replacements. In the SGR approach the final step consists in analyzing the complete new datasets $\mathbf{X}_1, \ldots, \mathbf{X}_s$ (with $\mathbf{X}_h = \mathbf{X}_h^r \cup \mathbf{D}_h^u$; $h = 1, \ldots, s$).

2.2 Extended MC simulations

Usually, in a Monte Carlo experiment, a hypothesized model is used to generate new data under various conditions. Therefore, the simulated data are used to evaluate some characteristics of the model. This, of course, implies that the distribution of the random component in the assumed model must be known, and it must be possible to generate pseudorandom samples from that distribution under the desired conditions planned by the researcher. In order to evaluate the impact of perturbed data on fit-indices we ought to generate for each MC simulated data $\mathbf{D}_k$ $(k = 1, \ldots, t)$ a family $\mathcal{R}(\mathbf{D}_k, \varrho)$ of SGR perturbed data matrices with exactly ϱ replacements. Therefore, we may think of each new perturbed data $\mathbf{X} \in \mathcal{R}(\mathbf{D}_k, \varrho)$ as an alternative "informative scenario" which is directly derived from the original simulated MC sample $\mathbf{D}_k$. Next, the behavior of a target fit-index can be evaluated with respect to the perturbed samples. In this case, of course, the distributional properties of the fit-index are not those that simply hold under a particular model hypothesis (like for standard Monte Carlo simulation studies); rather they are the properties under a model whose parameters corresponds to values fitted from both the MC generating process and the structured collection of perturbed samples that are generated from the given MC data sets.

3 Simulation study

In this simulation study, four fit-indices were examined with respect to structured perturbation of data. Of the four indices, two were absolute fit-indices (Goodness of Fit Index, GFI, and Adjusted Goodness of Fit Index, AGFI [Jöreskog and Sörbom, 1994]), and two incremental fit-indices (Comparative Fit Index, CFI [Bentler, 1990], and Nonnormed Fit Index, NNFI [Bentler and Bonnett, 1980] or TLI [Tucker and Lewis, 1973]). In this evaluation, three different types of target models were involved.

3.1 Target Models

We selected three target models that [Paxton *et al.*, 2001] considered were commonly encountered in applied research (see Figures 1, and 2). The first model, Model 1, contained nine measured variables and three latent factors. Each variable loaded on a single factor. Further, Factor 2 was regressed on Factor 1, and Factor 3 was regressed on Factor 2. The second model, Model 2, had the same basic structure as Model 1 but contained 15 measured variables, with five indicators per factor. Finally, Model 3 contained 13 measured variables with the same measurement structure as Model 1 (three indicators per factor) but added four observed exogenous variables. Factor 1 depended on all four correlated exogenous variables.

Parameter values were chosen on the basis of effect size (R^2 values) and statistical significance. For Model 1, the primary factor loadings were set to a standardized value of .70 (with $R^2 = .49$). The regression parameters among the latent factors were set to a standardized value of .60 ($R^2 = .36$). For Model 2, all the values were exactly the same as those of Model 1 except for the addition of two measured variables per factor. Finally, for Model 3, we included four exogenous variables. The primary factor loadings were set to .87, .82 and .72 for the first, the second and the third latent factor, respectively.

3.2 Simulation design

The following procedural steps were repeated for each target model M_j ($j = 1, 2, 3$).

i) According to M_j, 1000 raw-data sets $\mathbf{D}_k^j$ with $n = 50$ observations were generated. Next, each $\mathbf{D}_k^j$ ($k = 1, \ldots, 1000$) was discretized on a 5-point scale using the method described by [Jöreskog and Sörbom, 1996].

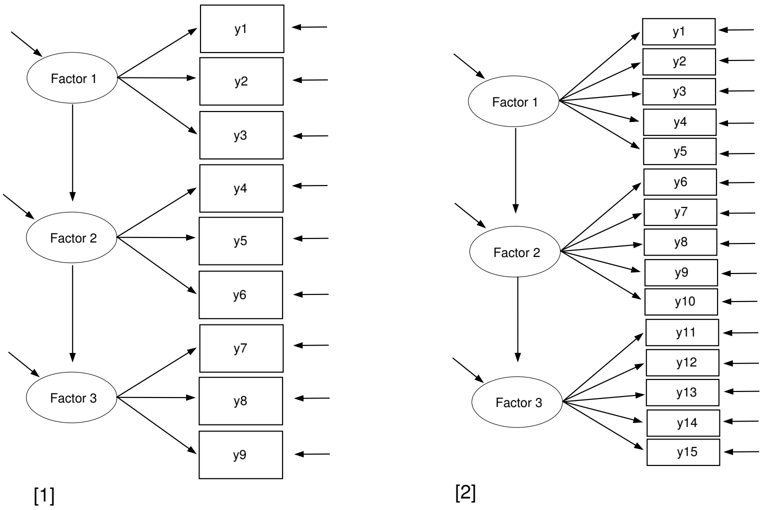


Fig. 1. Model [1]: nine observed variables and three factors. Model [2]: 15 observed variables and three factors.

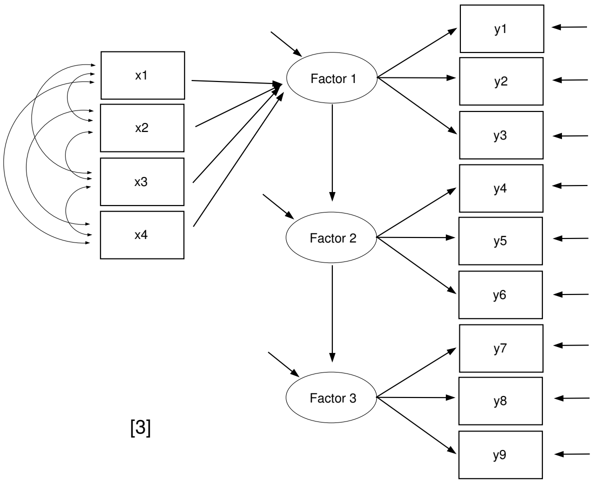


Fig. 2. Model [3]: 13 observed variables (four exogenous and nine endogenous) and three factors.

ii) For each discretized matrix $\underline{\mathbf{D}}_k^j$ we computed its polychoric correlation matrix and, subsequently, used this correlation matrix as input for M_j.

iii) Then the one hundred best fitting discretized matrix were selected by applying the following criteria: Chi-square not significant, Standardized Root-Mean-Square Residual (SRMR) $< .09$, Comparative Fit index (CFI) $> .96$, Nonnormed Fit Index (NNFI) $> .95$ [Hu and Bentler, 1999].

iv) For each best fitting data $\underline{\mathbf{B}}_h^j$ $(h = 1, \ldots, 100)$ we generated a family $\mathcal{R}(\underline{\mathbf{B}}_h^j, \varrho)$ of 50 SGR perturbed data matrices with exactly ϱ replacements. The exact number ϱ of replacements varied as a factor with 10 different levels $l = 1, 2, \ldots, 10$. Each level l denoted the proportion $(l \times 10)/100$ of replacements with respect to the size of the data set.

v) Each perturbed data matrix $\mathbf{X} \in \mathcal{R}(\underline{\mathbf{B}}_h^j, \varrho)$ was subjected to model M_j and the four fit-indices were finally evaluated. The whole procedure generated a total of 50000 new perturbed data matrices $\mathbf{X}$ for each target model.

4 Results

Table 1 reports the percentage of Converging Solutions (CS) and Acceptable Solutions (AS) as a function of percentage of replacements for the three considered models[1]. As expected, the percentage of CS decreased with larger percentage of replaced elements. A similar pattern was also observed for AS.

	Percentage of Replacements										
model	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%	
1	92.40	62.82	28.54	12.16	7.44	6.68	7.00	6.52	6.46	6.80	CS
2	99.68	89.76	50.38	15.24	6.14	3.20	3.52	3.86	3.08	2.72	
3	79.44	52.42	24.94	7.58	2.50	1.92	1.68	1.62	1.56	1.76	
1	85.50	47.46	15.04	4.02	1.76	1.40	1.38	1.26	1.14	1.50	AS
2	99.36	83.18	39.20	9.88	3.06	1.58	1.60	1.98	1.38	1.34	
3	79.44	52.42	24.94	7.58	2.50	1.92	1.68	1.62	1.56	1.76	

Table 1. Percentage of Converging Solutions (CS) (resp. Acceptable Solutions (AS)) as a function of percentage of replacements.

Figure 3 shows the means of GFI and AGFI for the three models. Segments represent standard deviations[2]. Dashed lines represent the cutoff optimal value (.95). Although both indices were constantly less than .95, the GFI (resp. AGFI) mean appeared not to be affected from increasing levels of replacements. Furthermore, very surprisingly, the means of GFI and AGFI increased with larger percentage of replaced elements.

[1] All our analysis were based on the Maximum Likelihood estimation algorithm.

[2] For the evaluation of the fit-indices we considered only AS.

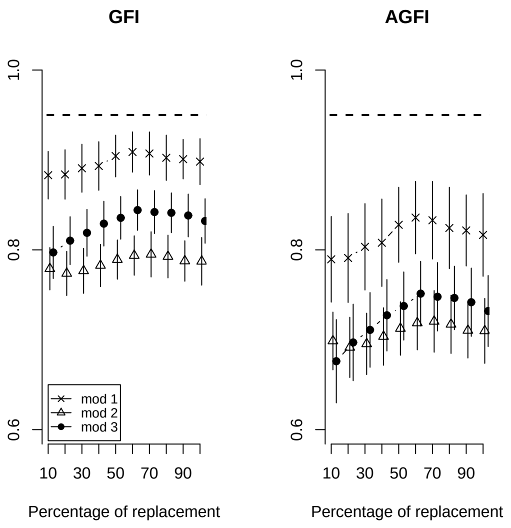


Fig. 3. Means of GFI and AGFI as a function of percentage of replacements. Segments represent standard deviations.

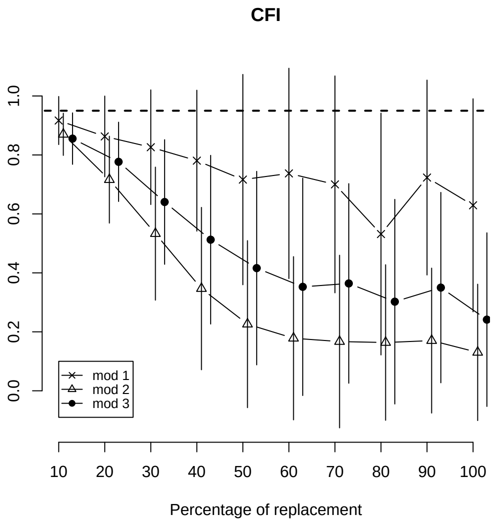


Fig. 4. Means of Comparative Fit Index (CFI) as a function of percentage of replacements. Segments represent standard deviations.

Figure 4 shows the means of CFI as a function of percentage of replacements for the three models. The dashed line indicates the cutoff optimal value (.96). By increasing the percentage of replacements, CFI means decreased and, in general, variability increased. The pattern associated to Model 1 showed that this model was less sensitive to replacements than both Model 3 and Model 2, the latter being the most sensitive to percentage of replacements. Notice that the same patterns were shown also by GFI and AGFI.

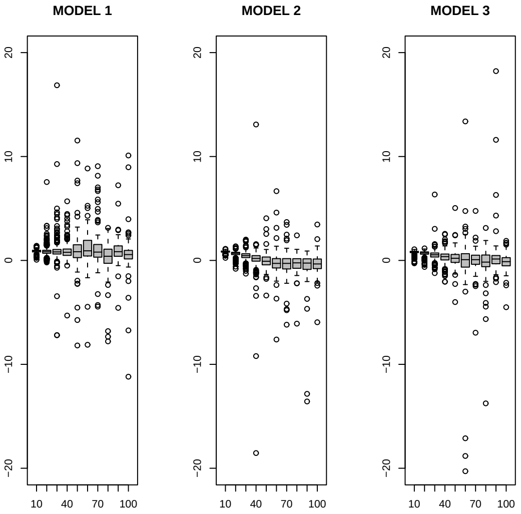


Fig. 5. Distributions of Nonnormed Fit Index (NNFI) as a function of percentage of replacement.

Finally, Figure 5 depicts the distributions of Nonnormed Fit Index (NNFI) for the three models. Remember that a model is a good one, when NNFI ranges between .95 and 1. Unlike both GFI and AGFI, NNFI was very sensitive to increasing levels of replacements. This observation is supported by the fact that a very large proportion of values fell outside the acceptable range [.95-1].
Table 2 reports the proportion of NNFI values within the range [.95-1]. We may notice a strong relationship between replacements and NNFI values. For example, in Model 1, we observed less than 10% of acceptable NNFI values, when 20% of replacements were considered.

5 Concluding remarks

A dominance relation can be read from Figures 3 and 4 as follows $M_2 \succ M_3 \succ M_1$, where $X \succ Y$ denotes that X is more sensitive to perturbations

	Percentage of Replacements									
model	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%
1	14.85	8.72	3.72	2.99	3.41	0.00	0.00	1.59	7.02	4.00
2	6.96	1.64	0.71	0.61	0.65	0.00	1.25	0.00	0.00	0.00
3	4.41	2.56	1.20	0.53	0.80	1.04	1.19	1.23	0.00	0.00

Table 2. Percentage of NNFI in the range [.95-1] as a function of percentage of replacements.

than Y. Overall our results suggested that the performance of the models were sensitive to perturbed data sets. This effect was stronger in the second model as it showed a clear replacement effect. In general, we recommend to choose more sensitive criteria (like NNFI) in order to better evaluate the effect in the model of eventual fake data.

Future applications of this methodology may be used in evaluating the robustness of goodness-of-fit criteria in empirical data set. However, more reasonable replacement scenarios based on external knowledge about process corruption should limit the upper bound of replacements. For example, in a personnel selection context the maximal number of fake answers in a personality questionnaire could be limited to 30%.

References

[Bentler and Bonnett, 1980]P.M. Bentler and D.G. Bonnett. Significance tests and goodness of fit in the analysis of covariance structures. *Psychological Bulletin*, pages 588–606, 1980.

[Bentler, 1990]P.M. Bentler. Comparative fit indexes in structural models. *Psychological Bulletin*, pages 238–246, 1990.

[Hu and Bentler, 1999]L. Hu and P.M. Bentler. Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling*, pages 1–55, 1999.

[Jöreskog and Sörbom, 1994]K.G. Jöreskog and D. Sörbom. *LISREL VI user's guide (3rd ed.).* Scientific Software, Mooresville, IN, 1994.

[Jöreskog and Sörbom, 1996]K.G. Jöreskog and D. Sörbom. *PRELIS 2: User's reference guide.* Scientific Software, Chicago, IL, 1996.

[Lombardi *et al.*, 2004]L. Lombardi, M. Pastore, and M. Nucci. Evaluating uncertainty of model acceptability in empirical applications: A replacement approach. In K. van Montfort, J. Oud, and A. Satorra, editors, *Recent Development on Structural Equation Models*, pages 69–82. Kluwer, Dordrecht, NE, 2004.

[Paxton *et al.*, 2001]P. Paxton, P.J. Curran, K.A. Bollen, J. Kirby, and F. Chen. Monte carlo experiments: Design and implementation. *Structural Equation Modeling*, pages 287–312, 2001.

[Tucker and Lewis, 1973]L.R. Tucker and C. Lewis. A reliability coefficient for maximum likelihood factor analysis. *Psychometrika*, pages 1–10, 1973.

Neural network attempt to nonlinear binary factor analysis of textual data

Dusan Husek[1], Alexander A. Frolov[2], Hana Rezankova[3], Vaclav Snasel[1], Michel Dufosse[4], and Pavel Polyakov[2]

[1] ICS - Acad. of Sci. of Czech Republic
Pod Vododarenskou vezi 2
182 07 Prague, Czech Republic
(e-mail: dusan@cs.cas.cz)

[2] IHNA Russian Acad. of Sci.
Butlerova 5a
117 485 Moscow, Russia
(e-mail: aafrolov@mail.ru)

[3] University of Economics, Prague
W. Churchill Sq. 4
130 67 Prague, Czech Republic
(e-mail: rezanka@vse.cz)

[4] INSERM U483
Universite Pierre and Marie Curie
9 quai Saint Bernard
75 005 Paris, France
(e-mail: michel.dufosse@snv.jussieu.fr)

Abstract. Possible application of a new procedure suitable of binary factorization of signals of large dimension and complexity is discussed. The new procedure is based on the search of attractors in Hoppfield-like associative memory. Starting from random initial state, network activity stabilizes in a attractor which corresponds to one of factors (a true attractor) or one of spurious attractors. Separation of true and spurious attractors is based on calculation of their Lyapunov function. Being applied to textual data the procedure conducted well and even more it showed sensitivity to the context in which the words were used.

Keywords: Neural networks, binary factor analysis, clustering, information retrieval.

1 Introduction

Factor analysis is one of the most efficient method to overcome informational redundancy of high-dimensional data set. Factors extraction is a procedure which maps objects from original space variables into the space of factors. Original signals, factor scores and factor loadings are binary, i.e. possess the values 0 or 1. To avoid computational problems with data large dimensionality we developed a procedure of binary nonlinear factorization based on the search of attractors in Hoppfield-like associative memory. In this case a complex vector signal (pattern) has a form of the Boolean sum of weighted binary factors:

$$\mathbf{X} = \bigvee \mathbf{S}_l \mathbf{f}^l. \tag{1}$$

It was a challenge for us [Frolov *et al.*, 2004] to utilize for binary factorization neural network with parallel dynamics because it has a lot of similarities with the iterative procedure for linear factorization. But there were some peculiarities that we have to solve. First, we have to mention that according our paradigm, the network is learned by signals from original space. During learning phase attractors are created in the energy landscape corresponding to true factors or spurious ones. From this the second problem follows that we have to solve - a procedure development that allows for effective revealing all the learned factors and separation of spurious ones. At the end, we have been successful and developed search procedure effective enough for attractors searching. Starting from random initial state, network activity stabilizes in some attractor which corresponds to one of true factors or one of spurious factors. To separate true and spurious attractors we found procedure based on calculation of their Lyapunov function [Goles-Chacc and Fogelman-Soulie, 1985]. Unlearning of already found factors prevent against their repeated retrieval. Some background on this topic can be found in work [Frolov *et al.*, 2003].

2 Hopfield network

The neural network under consideration consists of N neurons of the McCulloch-Pitts type (integrate-and-fire binary neurons) with gradually ranged synaptic connections between them. Only a fully connected case is considered here.

Network is trained by a set of M patterns of the form $\mathbf{X}^m = \bigvee\limits_{l=1}^{L} \beta_l^m \mathbf{f}^l$, where $\mathbf{f}^l \in B_n^N$ [1] are L factors (N dimensional vectors) and for every m-th pattern $\beta_l^m \in B_C^L$ it is a corresponding factor scores vector. As follows from the definition every factor contains exactly $n = Np$ ones. Every complex pattern $\mathbf{X}^m$ contains in turn exactly the C factors, so it is quite natural to call the *complexity* of the pattern as C. We assumed factors and factor scores to be statistically independent. In a limit case $C = 1$ patterns become pure factors and we obtain an ordinary Hopfield case.

2.1 Learning procedure

The connection matrix $\mathbf{J}$ of this network is a covariation matrix of input signals obtained by using the correlational Hebbian learning rule:

[1] $B_n^N = \{X | X_i \in \{0, 1\}, \sum\limits_{i=1}^{N} X_i = n\}$

$$J_{ij} = \sum_{m=1}^{M} (X_i^m - q^m)(X_j^m - q^m),\ i{\neq}j,\ J_{ii} = 0, \tag{2}$$

where M is the number of patterns in the learning set and $q^m = \sum\limits_{i=1}^{N} X_i^m/N$ is the total activity of the m-th pattern.
Its activity is determined by iterative procedure:

$$X_i(t+1) = \Theta(h_i(t) - T(t)),\ i = 1, \cdots, N \tag{3}$$

where Θ - step function, and $T(t)$ - activation threshold. And third, its activity has following Lyapunov function

$$\Lambda(t+1) = \mathbf{X}^T(t+1)\mathbf{J}\mathbf{X}(\mathbf{t}). \tag{4}$$

Activity of Hopfield-like network with parallel dynamics converges not only to point attractors [Goles-Chacc and Fogelman-Soulie, 1985] but also to cyclic attractors of the length two.

Theoretical analysis and computer simulation performed by Frolov et al. [Frolov *et al.*, 2004] completely confirmed the validity of Hopfield-like network for binary factorization. However, Hopfield-like network has one principal peculiarity. The network dynamics converges to one of the factors (true attractor) only when initial state falls inside its attraction basin. Otherwise it converges to one of the spurious attractors. Thus binary factorization requires special recall procedure to separate true and spurious attractors.

2.2 Recall procedure

To separate true and spurious attractors we developed two-run recall procedure. Its initialization starts by presentation of random initial pattern $\mathbf{X}^{in}$ with $k_{in} = r_{in}N$ active neurons. On presentation of $\mathbf{X}^{in}$, network activity $\mathbf{X}$ evolves to some attractor. The evolution is determined by equation (3). On each time step k_{in} "winners" (neurons with the greatest synaptic excitation) are chosen and only they are active on the next time step. When activity stabilizes at the initial level of activity k_{in}, $k_{in} + 1$ neurons with maximal synaptic excitation are chosen for the next iteration step, and network activity evolves to some attractor at the new level of activity $k_{in} + 1$. Then level of activity increases to $k_{in} + 2$, and so on, until number of active neurons reaches the final level $r_f N$. Thus, the whole procedure (one trial) contains $(r_f - r_{in})N$ iteration steps and several time steps inside each iteration step to reach some attractor for fixed level of activity.

At the end of each iteration step a relative Lyapunov function was calculated by formula: $\lambda = \Lambda/(rN)$ where Λ is given by (**??**). The relative Lyapunov function gives a mean synaptic excitation of active neurons. The time course of the relative Lyapunov function along the recall trajectory provides criterion for separation of true and spurious attractors (see later). Attractors with the highest Lyapunov function would be obviously winners in the most trials of the recall process. Thus, more and more trials are required to obtain new attractor with relatively small value of Lyapunov function. To overcome this problem attractors with high Lyapunov function should be deleted from the network memory. The deletion was performed according to Hebbian unlearning rule by substraction $\Delta J_{ij}, j \neq i$ from synaptic connections J_{ij} where

$$\Delta J_{ij} = \frac{\eta}{2} J(\mathbf{X})[(X_i(t-1) - r)(X_j(t) - r) + (X_j(t-1) - r)(X_i(t) - r), \quad (5)$$

$J(\mathbf{X})$ is the average synaptic connection between active neurons of the attractor, $\mathbf{X}(t-1)$ and $\mathbf{X}(t)$ are patterns of network activity at last time steps of iteration process, r is the level of activity, and η is an unlearning rate. For point attractor $\mathbf{X}(t) = \mathbf{X}(t-1)$ and for cyclic attractor $\mathbf{X}(t-1)$ and $\mathbf{X}(t)$ are two states of attractor.

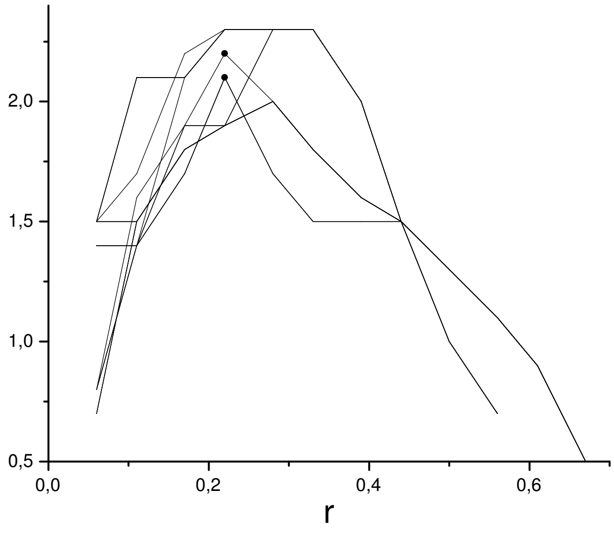


Fig. 1. Relative Lyapunov function λ in dependence on the relative network activity r for 15 titles of medical articles. Circles are points of breaking which were identified as indexes of factors.

3 Computer simulation

We tested our procedure over different examples from literature and text collections. First, we tested binary factorization over the list of titles of 15 medical articles presented in [Berry and Browne, 1999].

The titles were transformed to binary vectors with 18 component. The obtained binary codes of the titles were stored in the network of 18 neurons according to (**??**). Each trial was initiated by activation of one of 18 neurons. Thus the total recall procedure includes only 18 trials. Only two factors were revealed according to the used criterion see Fig.1. The first factor contains words: blood, close, disease and pressure. The second: fast, rats, rise and pressure. It is interesting that the words "culture", "discharge" and "patients" do not create a factor in spite of the fact that they are included into two first titles and, hence, one can expect that they should be tightly connected. However in these titles the word "culture" has different meaning and its banding with words "discharge" and "patients" is not reasonable. Thus we can conclude that our method could be sensitive to the context in which the words are used.

Second we applied our method to the set of 21000 messages of agency Reuters [Reuters, 2004, Rose *et al.*, 2002] as well. The used vocabulary contained 5000 the most often words in the set (consequently network contained

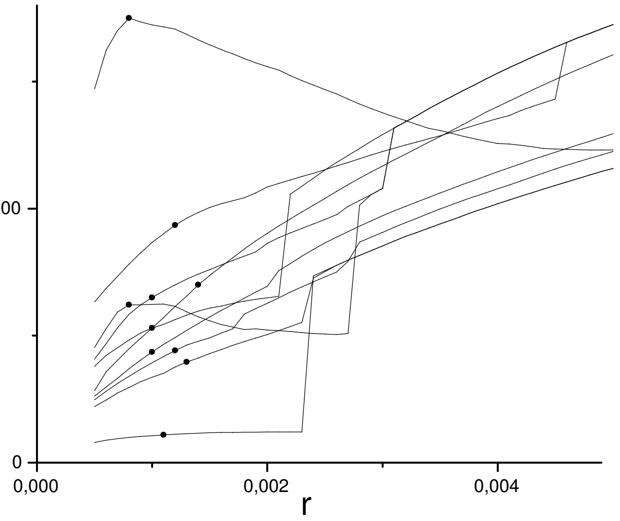


Fig. 2. Relative Lyapunov function λ in dependence on the relative network activity r for 21000 messages of agency Reuters. Circles are points of breaking which were identified as indexes of factors.

5000 neurons). Each message was transformed to binary code dependently on presence or absence of words in the message. Each found factor was deleted from the network memory according to (5) with $\eta = 1$. Fig. 2 demonstrates the first 10 trials which were identified as true. Circles mark the points of curve breaking. All found factors happened to be reasonable and mirror the content of the corresponding messages.

Our method combines words in factors not only according to the frequency of their appearance together at the messages but mainly according to their appearance at the same context. We see that different factors reflect different contexts of word utilization and different topics of news messages, while messages with the same topics are connected with the same factors.

Two messages with highlighted words creating factors are shown below, as an example of the point. These factors may appear in different news messages. But if in several messages the same factors are revealed, then these messages should have the same topic. In particular, the topics of messages from example are *Japanese foreign commerce* and *activity of American administration.* Evidently, factors reflect mutual meaning of the messages quite right.

” Message 1

U.S. ASKS **JAPAN** TO END AGRICULTURE IMPORT CONTROLS **TOKYO**, March 3

The U.S. Wants **Japan**[1] to eliminate import controls on agricultural products within three years, visiting U.S. Under-Secretary of State for **Economic**[1] Affairs Allen Wallis **told**[2] Eishiro Saito, Chairman of the Federation of **Economic**[1] Organisations (Keidanren), a spokesman for Keidanren said. The spokesman quoted Wallis as saying drastic measures would be needed to stave off protectionist legislation by **Congress**[3] .Wallis, who is attending a sub-cabinet-level bilateral **trade**[1] meeting, made the remark yesterday in talks with Saito. Wallis was quoted as saying the **Reagan**[3] **Administration**[3] wants **Japanese**[1] cooperation so the **White House**[3] can ensure any U.S. **Trade bill**[1] is a moderate one, rather than containing retaliatory measures or antagonising any particular country. He was also quoted as saying the U.S. Would be pleased were **Japan**[1] to halve restrictions on agricultural imports within five years if the country cannot cope with abolition within three, the spokesman said. **Japan**[1] currently restricts imports of 22 agricultural products. A ban on rice imports triggered recent U.S. Complaints about **Japan's**[1] agricultural policy.

” Message 2

U.S. COMMERCE SECRETARY QUESTIONS FUJITSU DEAL **WASHINGTON**, March 3

Commerce Secretary Malcolm Baldrige said he felt a proposed takeover by **Japan's**^{1} <Fujitsu Ltd> of U.S.-based Fairchild Semiconductor Corp, a subsidiary of Schlumberger Ltd <SLB>, should be carefully reviewed. He **told**^{2} the Semiconductor Industry Association the deal would soon be discussed by representatives of several different **government**^{3} departments. The **Reagan administration**^{3} has previously expressed concern that the proposed takeover would make Fujitsu a powerful part of the U.S. **market**^{1} for so-called supercomputers at a time when **Japan**^{1} has not bought any American-made supercomputers. In addition, U.S. defense **officials**^{3} have said they were worried semiconductor technology could be transferred out of the United States, eventually giving **Japanese**^{1}-made products an edge in American high-technology markets for defense and other goods. Treasury Secretary James Baker recently **told**^{2} a **Senate**^{3} committee the proposed takeover would be reviewed by the cabinet-level **Economic**^{1} Policy Council.

Here terms marked ^{1} are contained in the first factor, terms marked ^{2} are common words - contained in both factors and terms marked ^{3} are words contained in the second factor. One can see that factorizations is really nonlinear as there is nonempty set of common words.

4 Conclusion

In this work we have shown next step in development of Hopfield based neural network capable of performing binary factorization of the signals of high dimension and complexity. Advantage of our NN attempt should be possibility of incremental learning and capability to analyze large multidimensional data sets. This method is suitable for text collections analysis as shown in example. Being applied to textual messages of agency Reuters [Reuters, 2004], [Rose *et al.*, 2002], result showed not only full applicability of this method but moreover sensitivity to the context in which the words were used. Therefore we see big future potential for this application.

Acknowledgement: This work was supported by grant BARRANDE No. 2005-06-060-1 awarded by the Ministry of Education of the Czech Republic, 1ET100300414 and GA CR No. 201/05/0079.

References

[Berry and Browne, 1999]M.W. Berry and M. Browne. *Understanding Search Engines: Mathematical Modeling and Text Retrieval.* SIAM, NY, 1999.

[Frolov *et al.*, 2003]A.A. Frolov, D. Husek, and P. Muravjev. Informational efficiency of sparsely encoded Hopfield-like autoassociative memory. *Optical Memory and Neural Networks (Information Optics)*, pages 177–198, 2003.

[Frolov *et al.*, 2004]A.A. Frolov, A.M. Sirota, D. Husek, and P. Muravjev. Binary factorization in Hopfield-like neural networks: single-step approximation and computer simulations. *Neural Networks World*, pages 139–152, 2004.

[Goles-Chacc and Fogelman-Soulie, 1985]E. Goles-Chacc and F. Fogelman-Soulie. Decreasing energy functions as a tool for studying threshold networks. *Discrete Mathematics*, pages 261–277, 1985.

[Reuters, 2004]Reuters. *http://about.reuters.com/researchandstandards/corpus/.* Reuters, NY, 2004.

[Rose *et al.*, 2002]T. Rose, M Stevenson, and M. Whitehead. The Reuters corpus volume 1 - from yesterday's news to tomorrow's language resources. In *Proceedings of the Third International Conference on Language Resources and Evaluation*, pages 397–402, 2002.

Computing all-terminal reliability of stochastic networks with Binary Decision Diagrams

Gary Hardy[1], Corinne Lucet[1], and Nikolaos Limnios[2]

[1] LaRIA, FRE 2733, 5 rue du Moulin Neuf 80000 AMIENS
email:(corinne.lucet, gary.hardy)@u-picardie.fr
[2] LMAC, UTC, BP 20529 60205 COMPIEGNE Cedex
email:Nikolaos.Limnios@utc.fr

Abstract. In this paper, we propose an algorithm based on Binary Decision Diagram (BDD) for computing all-terminal reliability. It is defined as the probability that the nodes in the network can communicate to each other, taking into account the possible failures of network links. The effectiveness of this approach is demonstrated by performing experiments on several large networks represented by stochastic graphs. [1]

Keywords: Network reliability, Binary Decision Diagram (BDD), Stochastic graph.

1 Introduction

A stochastic network is modeled by an undirected graph $G = (V, E)$ where V is the vertex set and E is the edge set. Sites correspond to vertices and links to edges. The all-terminal reliability $R(G)$ is the probability that G remains connected assuming all edges can fail independently with known probability and nodes are perfect. Provan [Provan, 1986] showed that even for planar graphs this problem is still NP-hard. In literature, two classes of algorithms for computing the network reliability can be distinguished. The first class deals with the enumeration of all the minimum paths. The *inclusion-exclusion* or *sum of disjoint products* methods have to be applied since this enumeration provides non-disjoint events. The algorithms in the second class are factoring algorithms improved by reductions. It consists in reducing the size of the network while preserving its reliability. When no reduction is allowed, the factoring method is used. The idea is to choose a component and decompose the problem into two sub-problems: the first assumes the component has failed, the second assumes it is functioning. Satyanarayana and Chang [Satyanarayana and Chang, 1983] and Wood [Wood, 1985] have shown that the factoring algorithms with reductions are more efficient than the classical path or cut enumeration method for solving this problem. This was confirmed by the experimental works of Theologou and Carlier [Theologou and Carlier, 1991].

[1] Acknowledgment: This research was supported by the Conseil Regional de Picardie.

This paper is organized as follows. First, we give a brief introduction to BDD in Section 2. Then, in Section 3 we proposed a description of our method for computing network reliability. In Section 4, we introduce an other important reliability measure (Birnbaum importance measure) and its fast computation via BDD. Next, we present experimental results in Section 5. Finally, we draw some conclusions and outline the direction of futur works in Section 6.

2 Binary Decision Diagram (BDD)

Akers [Akers, 1978] first introduced BDD for representing boolean function. Bryant popularized the use of BDD by introducing a set of algorithms for efficient construction and manipulation of the BDD structure [Bryant, 1992]. Nowadays, BDD are used in a wide range of area, including hardware synthesis and verification, model checking and protocol validation. Their use in the reliability analysis framework has been introduced by Madre and Coudert [Coudert and Madre, 1992b] [Coudert and Madre, 1992a] and developped by Rauzy [Rauzy, 1993]. Sekine and Imai were the first to use the BDD structure in network reliability [Sekine and Imai, 1998]. A BDD is a directed acyclic graph (DAG) based on Shannon's decomposition. The Shannon's decomposition is defined as follows:

$$f = xf_{x=1} + \bar{x}f_{x=0}$$

where x is one of decision variables and $f_{x=i}$ is the boolean function f evaluated at $x = i$.
The graph has two sink nodes labeled with 0 and 1 representing the two corresponding constant expressions. Each internal node is labeled with a boolean variable x and has two out-edges called 0-edge and 1-edge. The node linked by 1-edge represents the boolean expression when $x = 1$, i.e $f_{x=1}$ while the node linked by 0-edge represents the boolean expression when $x = 0$, i.e $f_{x=0}$.
An ordered binary decision diagram (OBDD) is a BDD where variables are ordered according to a known total ordering and every path visits variables in an ascending order. Afterwards, BDDs will be considered as ordered. Leaves of the BDD give the value of f for the assignment corresponding to a path from the root to the leaf. The size of a BDD structure depends critically on variable ordering. Finding an ordering that minimizes the size of BDD is also a NP-complete problem [Friedman and Supowit, 1990].

3 Computing all-terminal reliability

Definitions and notations

A graph G is connected if there exists at least one path between any two vertices. Our network model is an undirected stochastic graph $G = (V, E)$.

x_1	x_2	x_3	f
0	0	0	0
0	0	1	0
0	1	0	0
0	1	1	1
1	0	0	0
1	0	1	1
1	1	0	0
1	1	1	1

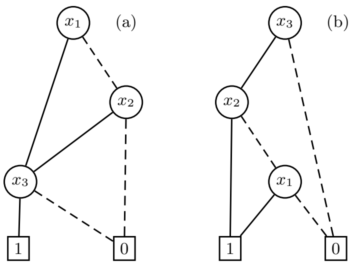


Fig. 1. Function $f(x_1, x_2, x_3) = (x_1 \wedge x_3) \vee (x_2 \wedge x_3)$ represented by its truth table and BDDs with order: (a) $x_1 < x_2 < x_3$ and (b): $x_3 < x_2 < x_1$. A dashed (solid) line represents the value 0 (1).

Each edge e_i of E ($i \in \{1, 2, \ldots, m\}$ where $m = |E|$) can fail independently with known probability q_i ($p_i = 1 - q_i$ is the functioning probability of e_i) and we consider that vertices of G are perfectly reliable. A state $\mathcal{G}$ of the stochastic graph G is denoted by $(x_1, x_2, \ldots, x_m)$ where x_i stands for the state of edge e_i, i.e, $x_i = 0$ when edge e_i fails and $x_i = 1$ when it functions. The associated probability of $\mathcal{G}$ is defined as:

$$Pr(\mathcal{G}) = \prod_{i=1}^{m} (x_i.p_i + (1 - x_i).q_i)$$

At each state $\mathcal{G}$ is associated a partial graph $G(\mathcal{G}) = (V, E')$ such that $e_i \in E'$ if and only if $e_i \in E$ and $x_i = 1$. The all-terminal reliability can be define as follows:

$$R(G) = \sum_{G(\mathcal{G}) \text{ is connected}} Pr(\mathcal{G})$$

We denote by G_{*e} the graph G with contracted edge e and by G_{-e} the graph G with deleted edge e.

Construction of the all-terminal reliability function

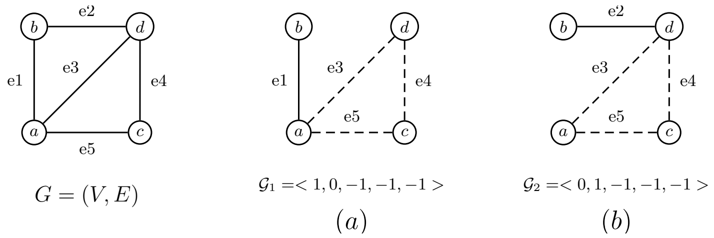
Our algorithm follows three steps:

1 The edges are ordered by using a heuristic.
2 The BDD is generated to encode the network reliability.
3 From this BDD, we obtain the all-terminal reliability.

We apply recursively the factoring algorithm in the order of $e_1, e_2, \ldots, e_m$ in a top-down way. The computation process can be represented as a binary tree such that the root corresponds to the original graph G and children correspond to graphs obtained by deletion /contraction of edges. Nodes in the binary tree correspond to subgraphs of G. We use the method introduced by Carlier [Carlier and Lucet, 1996] for represententing graph by partition. It is an efficient way for representing graph and finding isomorphic graphs during the computation process. By sharing the isomorphic subgraphs an expansion tree is modified as a rooted acyclic graph (therefore a BDD).

Sharing isomorphic graphs

Consider that $E_k = \{e_1, e_1, \ldots, e_k\}$ and $\bar{E_k} = \{e_{k+1}, \ldots, e_m\}$. The graphs in the k-th level of the BDD are sub-graphs of G with the edge set $\bar{E_k}$. For each level k, we define the boundary set F_k as a vertex set such that each vertex of F_k is incident to at least one edge in E_k and one edge in $\bar{E_k}$. Then we gather vertices in blocks according the following rule: two vertices s and t of F_k are in the same block if and only if there exists a path made of functioning edges linking s to t. For instance in figure 3(a), in the first level, the boundary set is equal to $\{a, b\}$. G_{*e_1} can be represented by partition [ab] and G_{-e_1} by partition [a][b]. Now, we order partitions in the same level k in order to identify and stock them in an efficient way. We number the partition from 1 to $Bell(|F_k|)$ where $Bell(|F_k|)$ (known as the Bell number) is the theoretical maximum number of partitions in level k. This number grows exponentially with i, consequently the number of classes grows exponentially with the size of the boundary set. From now on, we only manipulate partitions instead of graphs during the all-terminal reliability computation.



b
d
a
c
e1
e2
e3
e4
e5
$G = (V, E)$
$\mathcal{G}_1 =< 1, 0, -1, -1, -1 >$
(a)
$\mathcal{G}_2 =< 0, 1, -1, -1, -1 >$
(b)

Fig. 2. G($\mathcal{G}_1$) and G($\mathcal{G}_2$) represent sub-graphs in level 2 in the computation process illustred in figure 3(a). G($\mathcal{G}_1$) and G($\mathcal{G}_2$) has the same partition: $[a][d]$ during the computation. $e_i = -1$ means the state of e_i is not yet fixed.

All-terminal reliability computation

In the previous section, BDD of the all-terminal reliability function was constructed. The BDD can be recognized as a graph-based set of disjoint products. Based on the disjoint property of this structure, we can now easily compute the all-terminal reliability of G. Given the non-failure probabilty p_k

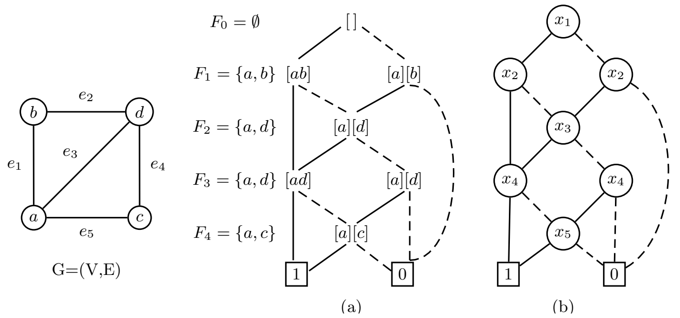


Fig. 3. Graph G and its BDD (b). A dashed (solid) line represents the value 0 (1). (a) illustrates the computation process of the BDD.

($k \in \{1, 2, \ldots, m\}$) of edge e_k, the all-terminal reliability of a BDD-based function f can be recursively obtain by:

$$R(G) = Pr(f = 1) = Pr(x_k.f_{x_k=1} = 1) + Pr(\bar{x_k}.f_{x_k=0} = 1) \ (\textit{disjoint property})$$

$$R(G) = Pr(f = 1) = p_k.Pr(f_{x_k=1} = 1) + q_k.Pr(f_{x_k=0} = 1) \ (\textit{independent property})$$

The reliability is evaluated by traversing the BDD from the root to the leaves.

4 Importance measure

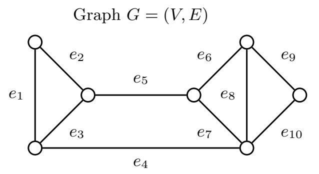
Finding the critical components is also an important issue for reliability analysis and the optimization design of network topology. The aim is to obtain information concerning a component's contribution to the system reliability. The three most used importance measures are: Birnbaum, Critically and Fussell-Vesely. We briefly explain here the Birnbaum importance measure. The Birnbaum importance measure of a component e_k is the probability that a system is in a critical state with respect to e_k and that the failure of component e_k will then cause the system to fail. Here, the Birnbaum importance measure of edge e_k, noted I_k^B, is defined as:

$$I_k^B = Pr(f_{x_k=1} = 1) - Pr(f_{x_k=0} = 1)$$

The figure 4 shows the importance measures for the reliability graph G.

5 Experimental results

Computations are done by using Pentium 4 with 512 MB memory. Our program is written in C language. The experimental results are shown in Tables 1 and 2. The unit of time is in second. The running time includes the BDD generation and the all-terminal reliability computation. The heuristic



ordering: e_5,e_1,e_2,e_9,e_{10},e_4,e_6 ,e_3,e_8,e_7

e_k	q_k	I_k^B
e_1	*0.3*	*0.2849*
e_2	*0.3*	*0.2849*
e_3	*0.3*	*0.1611*
e_4	*0.4*	*0.2285*
e_5	*0.05*	*0.3534*
e_6	*0.4*	*0.1872*
e_7	*0.4*	*0.1497*
e_8	*0.2*	*0.1543*
e_9	*0.3*	*0.2625*
e_{10}	*0.3*	*0.2625*

Fig. 4. Sensibility analysis of graph G. According to the Birnbaum importance measure, e_5 has the highest degree of contribution to the graph reliability.

used for ordering edges (and so variables in BDD) in the experiments is known as a breadth-first-search (BFS) ordering. We give two characteristics of the generated BDD: its *size* (number of nodes) and its *width* (if $|W_i|$ is the number of nodes in the ith level then the bdd width is: $\max_i |W_i|$). $|F_{max}|$ corresponds to the maximal size of the boundary set during the computation process. The computation speed heavily depends on $|F_{max}|$ and so the edge ordering.

6 Conclusion

A method for evaluating the all-terminal reliability via BDD has been proposed in this paper. Based on this approach, our futur works will focus on computing other kinds of reliability and reusing the BDD structure in order to optimize design of network topology.

References

[Akers, 1978]B. Akers. Binary decision diagrams. *IEEE Trans. On Computers*, vol. C-27:509–516, 1978.

[Bryant, 1992]R. E. Bryant. Symbolic Boolean manipulation with ordered binary-decision diagrams. *ACM Computing Surveys*, 24(3):293–318, 1992.

[Carlier and Lucet, 1996]J. Carlier and C. Lucet. A decomposition algorithm for network reliability evaluation. In *Discrete Applied Mathematics*, volume 65, pages 141–156, 1996.

[Coudert and Madre, 1992a]O. Coudert and J. C. Madre. Implicit and incremental computation of primes and essential primes of boolean functions. In *Proceedings of the 29th ACM/IEEE Design Automation Conference (DAC'92)*, pages 36–39. IEEE Computer Society Press, June 1992.

[Coudert and Madre, 1992b]O. Coudert and J. C. Madre. A new method to compute prime and essential prime implicants of boolean functions. In *Advanced Research in VLSI and Parallel Systems*, pages 113–128, March 1992.

type	n	m	*time*	*size*	*width*	$\lvert F_{max}\rvert$
K08	8	28	0.03	2745	405	7
K09	9	36	0.06	10265	1265	8
K10	10	45	0.13	39856	3925	9
K11	11	55	0.52	160793	15105	10
K12	12	66	2.14	673934	652	11
K13	13	78	9.97	2932248	279981	12
K14	14	91	50.00	13227624	1191235	13
K15	15	105	490	61780095	5021561	14

Table 1. Benchmark on complete graphs

type	n	m	*time*	*size*	*width*	$\lvert F_{max}\rvert$
6x6	49	84	0.15	39523	858	8
7x7	64	112	0.7	179410	2860	9
8x8	81	144	3.16	797916	9724	10
9x9	100	180	14.75	3495491	33592	11
10x10	121	220	67.94	15137188	117572	12
15x10	176	325	101.2	33360848	117572	12
20x10	231	430	101.2	24249018	117572	12
11x11	144	264	321.3	64959137	416024	13

Table 2. Benchmark on lattice graphs

[Friedman and Supowit, 1990]S. J. Friedman and K. J. Supowit. Finding an optimal variable ordering for binary decision diagrams. *IEEE Trans. On Computers*, vol. C-39:710–713, 1990.

[Provan, 1986]J.S. Provan. The complexity of reliability computations on planar and acyclic graphs. *SIAM J. Computing*, 15(3):694–702, 1986.

[Rauzy, 1993]A. Rauzy. New algorithms for fault tolerant trees analysis. *Reliability Engineering and System Safety*, pages 203–211, 1993.

[Satyanarayana and Chang, 1983]A. Satyanarayana and M.K. Chang. Network reliability and the factoring theorem. *Networks*, 13:107–120, 1983.

[Sekine and Imai, 1998]K. Sekine and H. Imai. Computation of the network reliability (extended abstract). Technical report, Department of Information Science, University of Tokyo, 1998.

[Theologou and Carlier, 1991]O. Theologou and J. Carlier. Factoring and reductions for networks with imperfect vertices. In *IEEE Trans. on Reliability*, volume 40, pages 210–217, 1991.

[Wood, 1985]R.K. Wood. A factoring algorithm using polygon-to-chain reductions for computing k-terminal network reliability. In *Networks*, volume 15, pages 173–190, 1985.

On the multivariate kernel distribution estimator for distribution functions under association

Cecília Azevedo[1] and Paulo E. Oliveira[2]

[1] CMAT- Centro de Matemática
Universidade do Minho
4710-057 Braga, Portugal
(e-mail: cecilia@math.uminho.pt)

[2] Departamento de Matemática
Universidade de Coimbra
3000 Coimbra, Portugal
(e-mail: paulo@mat.uc.pt)

Abstract. In this note we consider the estimation of the multivariate distribution function $\mathbf{F}_p$ of the p-dimensional marginal of a stationary associated sequence. We show, under certain regularity conditions, the almost sure consistency and characterize the asymptotic behavior of the MSE. We also characterize the asymptotic optimal bandwidth. Under some stronger assumptions on the covariance this bandwidth rate is shown to be the same as for the independent case.

Keywords: Association, Kernel estimator, Optimal bandwidth, Mean squared error.

1 Introduction and assumptions

Estimation of distribution functions has been one of the main problems in statistics. Given a stationary sequence of random variables we will consider the estimator of it's p- dimensional marginal distribution function assuming some kind of positive dependence. The various types of positive dependence have received some interest in the literature since the early 1990's. We will consider associated random variables as introduced in Esary et al (1967). For the one-dimensional marginal, the estimator has been studied by Roussas [Roussas, 1993], [Roussas, 2000] and Cai, Roussas [Cai and Roussas, 1998]. Motivated by the need to approximate covariance functions appearing in the study of empirical processes Azevedo, Oliveira [Azevedo and Oliveira, 2000] and Henriques, Oliveira [Henriques and Oliveira, 2002] studied the two dimensional case. This note extends results in [Azevedo and Oliveira, 2000] for the p-dimensional case. We start by recalling the definition of association, as stated in Esary et al (1967).

Definition 1 *For a finite index set I, the random variables (r.v.'s) $\{X_i\}_{i\in I}$ are said to be associated, if for any real-valued coordinatewise increasing* *functions G and H defined on $\mathbb{R}^I$,* Cov $\{G(X_i, i \in I), H(X_j, j \in I)\} \geq 0$, *provided* $\mathbb{E}\left(G^2(X_i, i \in I)\right) < \infty$ *and* $\mathbb{E}\left(H^2(X_j, j \in I)\right) < \infty$*. A sequence of r.v's is said to be associated if any finite subset of the r.v.'s is associated.*

Definition 2 *A smooth estimate of $\mathbf{F}_p$, d.f. of the random vector $\mathbf{X} = (X_1, \ldots, X_p)$, with $p \geq 2$, $\widehat{F}_{n,p}$ is defined, for each $\mathbf{x} = (x_1, \ldots, x_p) \in \mathbb{R}^p$, by*

$$\widehat{F}_{n,p}(\mathbf{x}) = \frac{1}{n-p} \sum_{i=1}^{n-p} \mathbf{U}\left(\frac{\mathbf{x} - \mathbf{X}_{i,p}}{h_n}\right), \tag{1}$$

where $\mathbf{U}$ is a $p-$variate known d.f., the kernel function and, for each fixed p and $i = 1, \ldots, n-p$, $\mathbf{X}_{i,p} = (X_{i+1}, \ldots, X_{i+p})$. The (bandwidths) h_n are positive numbers tending to 0, as $n \to \infty$.

Jin, Shao [Jin and Shao, 1999] have been shown that, under independence, the optimal bandwidth of the p-dimensional kernel distribution estimator of $\mathbf{F}_p$ has order $n^{-1/3}$, for all dimensions. For associated samples, several properties of the univariate estimate $\widehat{F}_n$ of the marginal d.f. F have been investigated by Cai, Roussas [Cai and Roussas, 1998]. These authors proved that the optimal bandwidth rate is of order n^{-1}. The rate $n^{-1/3}$ becomes optimal under some stronger assumptions on the covariance structure. Azevedo, Oliveira [Azevedo and Oliveira, 2000] studied properties of the bivariate estimate $\widehat{F}_{n,k}$ of the d.f. of (X_1, X_{k+1}) with fixed $k = 1, \ldots n-1$, characterizing the optimal bandwidth rate. The results obtained on [Azevedo and Oliveira, 2000] extended the one-dimensional ones.

The set of conditions bellow are basically the same as in Cai, Roussas [Cai and Roussas, 1998] together with the conditions used by Jin, Shao [Jin and Shao, 1999] under independence.

Assumptions

(A_1) $\{X_n\}_{n \in \mathbb{N}}$ is a strictly stationary sequence of random variables with bounded density function f and continuous marginal distribution function F;

(A_2) The derivative of f exists and is continuous and bounded;

(A_3) The d.f., $\mathbf{F}_p$, of the random vector $\mathbf{X} = (X_1, \ldots, X_p)$ has bounded and continuous partial derivatives of first and second orders;

(A_4) For each positive integer j, the d.f. of $\mathbf{X}_{p,j} = (X_1, \ldots, X_p, X_{j+1}, \ldots, X_{j+p})$, $\mathbf{F}_{p,j}$, has bounded and continuous partial derivatives of first and second order;

(A_5) The kernel function $\mathbf{U}$ is $p-$differentiable and $\mathbf{u} = \frac{\partial^p \mathbf{U}}{\partial x_1 \ldots \partial x_p}$ is such that:

$$(i) \int_{\mathbb{R}^p} \mathbf{u}(\mathbf{x}) d\mathbf{x} = 1; \; (ii) \int_{\mathbb{R}^p} \mathbf{x}\mathbf{u}(\mathbf{x}) d\mathbf{x} = \mathbf{0}; \; (iii) \int_{\mathbb{R}^p} \mathbf{x}\,\mathbf{x}^T\, \mathbf{u}(\mathbf{x}) d\mathbf{x} < \infty;$$

(A_6) The sequence of bandwidths is such that $n\, h_n^2 \to 0$;

(A_7) $\sum_{n=1}^{\infty} n\, Cov^{1/3}(X_1, X_n) < \infty;$ $\qquad (A_7)'$ $\sum_{n=1}^{\infty} Cov^{1/3}(X_1, X_n) < \infty;$

(A_8) $\mathbf{V} = \dfrac{\partial^p \mathbf{U}^2}{\partial x_1 \ldots \partial x_p}$, is such that $\int_{\mathbb{R}^p} \mathbf{x}\,\mathbf{x}^T\, \mathbf{V}(\mathbf{x}) d\mathbf{x} < \infty.$

Remark 1 *Note that* $\int_{\mathbb{R}^p} \mathbf{V}(\mathbf{x}) d\mathbf{x} = \mathbf{U^2}(+\infty, \ldots, +\infty) = 1.$

The conditions $(A_1), (A_2)$ and (A_7) have already been used in Cai and Roussas [Cai and Roussas, 1998] for the treatment of the univariate case. Note further that (A_7) implies $(A_7)'$ which implies the $L^2[0, 1]$ weak convergence of empirical process, as proved in Oliveira and Suquet [Oliveira and Suquet, 1999].

Let us define the auxiliar functions $\mathbf{V_1}, \mathbf{V_2}$, $\mathbf{V_3}$ and $\mathbf{V_4}$ from $\mathbb{R}^p$ to $\mathbb{R}$, such that for each $\mathbf{x} = (x_1, \ldots, x_p)$,

- $\mathbf{V_1}(\mathbf{x}) = \sum_{i=1}^{p} \dfrac{\partial^2 \mathbf{F}_p}{\partial x_i^2}(\mathbf{x}) \int_{\mathbb{R}^p} a_i^2 \mathbf{u}(\mathbf{a}) d\mathbf{a} + 2 \sum_{j=1}^{p-1} \sum_{i=j+1}^{p} \dfrac{\partial^2 \mathbf{F}_p}{\partial x_j\, \partial x_i}(\mathbf{x}) \int_{\mathbb{R}^p} a_i a_j \mathbf{u}(\mathbf{a}) d\mathbf{a};$
- $\mathbf{V_2}(\mathbf{x}) = \sum_{i=1}^{p} \dfrac{\partial \mathbf{F}_p}{\partial x_i}(\mathbf{x}) \int_{\mathbb{R}^p} a_i \mathbf{V}(\mathbf{a}) d\mathbf{a};$
- $\mathbf{V_3}(\mathbf{x}) = \sum_{i=1}^{p} \dfrac{\partial^2 \mathbf{F}_p}{\partial x_i^2}(\mathbf{x}) \int_{\mathbb{R}^p} a_i^2 \mathbf{V}(\mathbf{a}) d\mathbf{a} + 2 \sum_{i=1}^{p-1} \sum_{j=i+1}^{p} \dfrac{\partial^2 \mathbf{F}_p}{\partial x_j\, \partial x_i}(\mathbf{x}) \int_{\mathbb{R}^p} a_i a_j \mathbf{V}(\mathbf{a}) d\mathbf{a};$
- $\mathbf{V_4}(\mathbf{x}) = \sum_{i=1}^{2p} \dfrac{\partial^2 \mathbf{F}_{p,j}}{\partial x_i^2}(\mathbf{x}, \mathbf{x}) \int_{\mathbb{R}^{2p}} a_i^2 \mathbf{u}(\mathbf{a}) d\mathbf{a} + 2 \sum_{i=1}^{2p-1} \sum_{j=i+1}^{2p} \dfrac{\partial^2 \mathbf{F}_{p,j}}{\partial x_j\, \partial x_i}(\mathbf{x}, \mathbf{x}) \int_{\mathbb{R}^{2p}} a_i a_j \mathbf{u}(\mathbf{a}) d\mathbf{a}.$

2 Consistency of the estimator.

In this section we present some results concerning to consistency of the estimator (1). We first show that $\widehat{F}_{n,p}$ is asymptotic unbiased, characterizing also the convergence rate of $\mathbb{E}\left(\widehat{F}_{n,p}(\mathbf{x})\right)$. To derive the asymptotic consistency of $\widehat{F}_{n,p}$, we apply a strong law of large numbers to the random variables $\mathbf{U}\left(\frac{\mathbf{x}-\mathbf{X}_{i,p}}{h_n}\right)$, $i = 1, \ldots, n-p$. To achieve this last step we shall need to characterize the behavior of each entry of the covariance matrix of the random vector whose entries are the preceding variables.

Theorem 1 *Suppose* $\{X_n\}_{n\in\mathbb{N}}$ *satisfy* $(A_1),(A_3)$ *and* (A_5)*. Then, for each* $\mathbf{x}\in\mathbb{R}^p$,

$$\mathbb{E}\left(\widehat{F}_{n,p}(\mathbf{x})\right)=\mathbf{F}_p(\mathbf{x})+\frac{V_1(\mathbf{x})}{2}h_n^2+o(h_n^2).$$

Proof: First note that the kernel estimator (1) can be written as

$$\widehat{F}_{n,p}(\mathbf{x})=\int_{\mathbb{R}^p}\mathbf{U}\left(\frac{\mathbf{x}-\mathbf{s}}{h_n}\right)d\widehat{\phi_n}(\mathbf{s}), \tag{2}$$

where $\widehat{\phi_n}(\mathbf{x})=\dfrac{1}{n-p}\displaystyle\sum_{i=1}^{n-p}\mathbb{1}_{(-\infty,x_1]\times\cdots\times(-\infty,x_p]}(\mathbf{X}_{i,p})$, with $\mathbb{1}_A$ the characteristic function of the set A.
As $\mathbb{E}\left(\widehat{\phi_n}(\mathbf{x})\right)=\mathbf{F}_p(\mathbf{x})$, it follows from (2) applying Fubini's Theorem, that $\mathbb{E}\left(\widehat{F}_{n,p}(\mathbf{x})\right)=\displaystyle\int_{\mathbb{R}^p}\mathbf{U}\left(\frac{\mathbf{x}-\mathbf{s}}{h_n}\right)d\mathbf{F}_p(\mathbf{s})=\int_{\mathbb{R}^p}\mathbf{u}(\mathbf{t})\mathbf{F}_p(\mathbf{x}-\mathbf{t}h_n)d\mathbf{t}$. Now, by using a Taylor expansion of order 2 of $\mathbf{F}_p$ and taking account of (A_3) and (A_5), and of the continuity of the second order partial derivatives of $\mathbf{F}_p$, (A_3), the result follows. ■
Note that (A_3) *and* (A_5) *are only required in order to establish a convergence rate. In fact, the convergence of* $\mathbb{E}\left(\widehat{F}_{n,p}(\mathbf{x})\right)$ *to* $\mathbf{F}_p(\mathbf{x})$ *follows from an application of the Dominated Convergence Theorem.*
In order to establish the almost sure convergence of (1) we need to control some covariances. Define

- $\mathbf{I}_{nj}(\mathbf{x})=\mathrm{Cov}\left(\mathbf{U}\left(\frac{\mathbf{x}-\mathbf{X}_{1,p}}{h_n}\right),\mathbf{U}\left(\frac{\mathbf{x}-\mathbf{X}_{j,p}}{h_n}\right)\right)$
- $\mathbf{I}_j(\mathbf{x})=\mathrm{Cov}\left(\mathbb{1}_{(-\infty,\mathbf{x}]}(\mathbf{X}_{1,p}),\mathbb{1}_{(-\infty,\mathbf{x}]}(\mathbf{X}_{j,p})\right)$.

Lemma 1 *Suppose that* $\{X_n\}_{n\in\mathbb{N}}$ *satisfy* (A_1), (A_3), (A_4) *and* (A_5)*. Then, for each* $j>1$, *and* $\mathbf{x}\in\mathbb{R}^p$,

(i) $\mathbf{I}_{nj}(\mathbf{x})=\mathbf{I}_j(\mathbf{x})+O(h_n^2)=\mathbf{F}_{p,j}(\mathbf{x},\mathbf{x})-\mathbf{F}_p^2(\mathbf{x})+O(h_n^2)$;

(ii) *For* $j>p-1,\mathbf{I}_j(\mathbf{x})\le\displaystyle\sum_{k=1}^{p}(p-k+1)\mathrm{Cov}^{1/3}(X_1,X_{j+k})+$
$+\sum_{k=1}^{p-1}(p-k)\mathrm{Cov}^{1/3}(X_1,X_{j-k+1})$.

Proof: Condition (i) follows from rewriting the covariance
$I_{nj}=\displaystyle\int_{\mathbb{R}^{2p}}\mathbf{U}(\frac{\mathbf{x}-\mathbf{s}}{h_n})\mathbf{U}(\frac{\mathbf{x}-\mathbf{t}}{h_n})d\mathbf{F}_{p,j}(\mathbf{s},\mathbf{t})-\left(\int_{\mathbb{R}^p}\mathbf{U}(\frac{\mathbf{x}-\mathbf{s}}{h_n})d\mathbf{F}_p(s)\right)^2$. For the first term, writing the function $\mathbf{U}$ as an integral and by using Fubini's Theorem, we have $\displaystyle\int_{\mathbb{R}^{2p}}\mathbf{u}(\mathbf{a})\mathbf{u}(\mathbf{b})\mathbf{F}_{p,j}(\mathbf{x}-\mathbf{a})(\mathbf{x}-\mathbf{b})d\mathbf{a}d\mathbf{b}$. So, expanding $\mathbf{F}_{p,j}$ to the second order and using (A_4) and (A_5), this integral is equal to

$\mathbf{F}_{pj}(\mathbf{x},\mathbf{x}) + O(h_n^2)$, which together with the behavior of $\mathbb{E}\left(\widehat{F}_{n,p}(\mathbf{x})\right)$, completes the proof of (i). To prove condition (ii) we need use the inequality,

$$\text{Cov}\ \left(\mathbb{1}_{(-\infty,s]}(Y_1), \mathbb{1}_{(-\infty,t]}(Y_2)\right) \leq M\text{Cov}^{1/3}(Y_1,Y_2), \tag{3}$$

where Y_1, Y_2 are associated random variables with common distribution function with a bounded density and $M > 0$ is constant (see Sadikova [Sadikova, 1966]), and the following lemma (Lebowitz [Lebowitz, 1972]),

Lemma 2 *Let A and B be subsets of $\{1,\ldots,n\}$ and x_i real with $i \in A \cup B$. Let $H_{A,B} = P(X_i > x_i, i \in A \cup B) - P(X_j > x_j, j \in A)P(X_k > x_k, k \in B)$.*
If $(X_1,\ldots,X_n)$ is associated then, $0 \leq H_{A,B} \leq \sum_{i\in A, j\in B} H_{\{i\},\{j\}}$.

In fact, according lemma 2,

$$\text{Cov}\ \left(\mathbb{1}_{(-\infty,\mathbf{x}]}(\mathbf{X}_{1,p}), \mathbb{1}_{(-\infty,\mathbf{x}]}(\mathbf{X}_{j,p})\right) \leq$$

$$\leq \sum_{k=1}^{p}\sum_{i=1}^{p} \text{Cov}\ \left(\mathbb{1}_{(-\infty,x_k]}(X_k), \mathbb{1}_{(-\infty,x_{j+i}]}(X_{j+i})\right).$$

Now applying innequality (3), we have
$\text{Cov}\ \left(\mathbb{1}_{(-\infty,x_k]}(X_k), \mathbb{1}_{(-\infty,x_{j+i}]}(X_{j+i})\right) \leq M\,\text{Cov}^{1/3}(X_k, X_{j+i})$, so
$\mathbf{I}_j(\mathbf{x}) \leq M \sum_{k=1}^{p}\sum_{i=1}^{p} \text{Cov}^{1/3}(X_k, X_{j+i})$. The sequence $\{X_n\}_{n\in\mathbb{N}}$ being stationary,
$\mathbf{I}_j(\mathbf{x}) \leq M\sum_{k=1}^{p}(p-k+1)\text{Cov}^{1/3}(X_1, X_{j+k}) + \sum_{k=1}^{p-1}(p-k)\text{Cov}^{1/3}(X_1, X_{j-k+1})$. ■

Remark 2 *Note that if the covariance sequence*

$$\{\text{Cov}\,(X_1, X_{j+1})\}_{j\in\mathbb{N}} \tag{4}$$

is decreasing, $\mathbf{I}_j(\mathbf{x}) \leq p^2\text{Cov}^{1/3}(X_1, X_{j+1})$.

Theorem 2 *Suppose the variables $X_n, n \geq 1$, satisfy (A_1), (A_2), (A_3), (A_4), (A_5), (A_7) and (A_8). Then, for every $\mathbf{x} \in \mathbb{R}^p$, $\widehat{F}_{n,p}(\mathbf{x}) \to \mathbf{F}_p(\mathbf{x})$ almost surely.*

Proof: As proved in Theorem 1, $\mathbb{E}\left(\widehat{F}_{n,p}(\mathbf{x})\right) \to \mathbf{F}_p(\mathbf{x})$, so it's enough to prove that the variables $\mathbf{U}\left(\frac{\mathbf{x}-\mathbf{X}_{m,p}}{h_n}\right)$, $m \geq 1$ satisfy a strong law of large numbers. These variables are stationary and associated, as $\mathbf{U}$ is coordinatewise nondecreasing. Then, according to Newman [Newman, 1980] they satisfy a strong law of large numbers if

$$\lim_{n\to\infty} \frac{1}{n-p}\sum_{j=1}^{n-p} I_{n,j}(\mathbf{x}) = 0. \tag{5}$$

From conditions (i) and (ii) of the preceding lemma, $I_{n,j}(\mathbf{x}) \leq M\sum_{k=1}^{p}(p-k+1)\mathrm{Cov}^{1/3}(X_1,X_{j+k}) + \sum_{k=1}^{p-1}(p-k)\mathrm{Cov}^{1/3}(X_1,X_{j-k+1}) + O(h_n^2)$. Now condition (5) is a consequence of (A_7) and association, so the theorem follows. ■

3 The behavior of the mean square error.

In this section we study the asymptotics and convergence rate of the mean square error (MSE). This characterization will then be used to derive the optimal bandwidth convergence rate. This convergence rate for the bandwidth is, as will be explained later, of order n^{-1}, thus a different convergence rate than the one in the independent case. But if we consider a decreasing rate on the sequence of the covariances (see Cai, Roussas [Cai and Roussas, 1998]) we obtain a convergence rate of order $n^{-1/3}$, as in the independent case (see Jin, Shao [Jin and Shao, 1999]), for all dimensions p.

As usual write MSE $\left(\widehat{F}_{n,p}(\mathbf{x})\right) = \mathrm{Var}\left(\widehat{F}_{n,p}(\mathbf{x})\right) + \left(\mathbb{E}\left(\widehat{F}_{n,p}(\mathbf{x})\right) - \mathbf{F}_p(\mathbf{x})\right)^2$. The behavior of $\mathbb{E}\left(\widehat{F}_{n,p}(\mathbf{x})\right)$ being known (cf.Theorem 1), we need to describe the asymptotics and convergence rate for the variance term.

Lemma 3 *Suppose the sequence* $\{X_n\}_{n\in\mathbb{N}}$ *satisfy* (A_1), (A_3), (A_4), (A_5) *and* (A_8). *Then for all* $\mathbf{x}$ *in* $\mathbb{R}^p$,

$$\textbf{(i)}\quad \mathbb{E}\big(\mathbf{U}^2(\frac{\mathbf{x}-\mathbf{X}_{i,p}}{h_n})\big) = \mathbf{F}_p(\mathbf{x}) - h_n\mathbf{V}_2(\mathbf{x}) + \frac{h_n^2}{2}\mathbf{V}_3(\mathbf{x}) + o(h_n^2)$$

$$\textbf{(ii)}\quad \left|\mathrm{Var}\left(\mathbf{U}\left(\tfrac{\mathbf{x}-\mathbf{X}_{i,p}}{h_n}\right)\right) - \mathbf{F}_p(\mathbf{x})(1-\mathbf{F}_p(\mathbf{x})) + h_n\mathbf{V}_2(\mathbf{x})\right| = \\ = h_n^2(\mathbf{V}_3(\mathbf{x}) - \mathbf{F}_p(\mathbf{x})\mathbf{V}_1(\mathbf{x})) + o(h_n^2).$$

Proof: In what concerns to (**i**), we have, by definition,
$\mathbb{E}\left(\mathbf{U}^2\left(\frac{\mathbf{x}-\mathbf{X}_{i,p}}{h_n}\right)\right) = \int_{\mathbb{R}^p}\mathbf{U}^2\left(\frac{\mathbf{x}-\mathbf{s}}{h_n}\right)d\mathbf{F}_p(\mathbf{s}) \int_{\mathbb{R}^p}\left(\int_{(-\infty,\mathbf{x}]}\mathbf{V}(\mathbf{a})d\mathbf{a}\right)d\mathbf{F}_p(\mathbf{s})$
By using Fubini Theorem and changing variables,
$\mathbb{E}\left(\mathbf{U}^2\left(\frac{\mathbf{x}-\mathbf{X}_{i,p}}{h_n}\right)\right) = \int_{\mathbb{R}^p}\mathbf{V}(\mathbf{a})\mathbf{F}_p(\mathbf{x}-\mathbf{a}h_n)d\mathbf{a}$. Using a Taylor expansion of order 2 of $\mathbf{F}_p$ and taking account of (A_5) and the definitions of $\mathbf{V}_2$ and $\mathbf{V}_3$, we have (**i**). In order to obtain (**ii**), knowing that
$\mathrm{Var}\left(\mathbf{U}\left(\frac{\mathbf{x}-\mathbf{X}_{i,p}}{h_n}\right)\right) = \mathbb{E}\left(\mathbf{U}^2\left(\frac{\mathbf{x}-\mathbf{X}_{i,p}}{h_n}\right)\right) - \left(\mathbb{E}\left(\mathbf{U}\left(\frac{\mathbf{x}-\mathbf{X}_{i,p}}{h_n}\right)\right)\right)^2$, it is suffices to apply (**i**) and Theorem 1. ■

Definition 3 *Let* $\boldsymbol{\sigma^2}(\mathbf{x}) = \mathbf{F}_p(\mathbf{x}) - \mathbf{F}_p^2(\mathbf{x}) + 2\sum_{j=2}^{\infty}\left(\mathbf{F}_{p,j}(\mathbf{x},\mathbf{x}) - \mathbf{F}_p^2(\mathbf{x})\right)$ *and*
$\mathbf{c}_n(\mathbf{x}) = 2\sum_{j=n-p+1}^{\infty}\left(\mathbf{F}_{p,j}(\mathbf{x},\mathbf{x}) - \mathbf{F}_p^2(\mathbf{x})\right) + \frac{2}{n-p}\sum_{j=2}^{n-p}(j-1)\left(\mathbf{F}_{p,j}(\mathbf{x},\mathbf{x}) - \mathbf{F}_p^2(\mathbf{x})\right)$

Theorem 3 *Suppose that* $\{X_n\}_{n\in\mathbb{N}}$ *satisfy* (A_1), (A_3), (A_4), (A_5), (A_6), (A_7) *and* (A_8). *Then*

$(n\ -\ p)\text{Var}\left(\widehat{F}_{n,p}(\mathbf{x})\right)\ =\ \sigma^2(\mathbf{x})\ -\ h_n\mathbf{V}_2(\mathbf{x})\ +\ (n\ -\ p\ -\ 1)h_n^2\left(\mathbf{V}_4(\mathbf{x})-\mathbf{F}_p(\mathbf{x})\mathbf{V}_1(\mathbf{x})\right)+O(h_n^2)-c_n(\mathbf{x})$.

Proof: $\text{Var}\left(\widehat{F}_{n,p}(\mathbf{x})\right)=\dfrac{1}{(n-p)^2}\displaystyle\sum_{i,j=1}^{n-p}\text{Cov}\left(\mathbf{U}\left(\frac{\mathbf{x}-\mathbf{X}_{i,p}}{h_n}\right),\mathbf{U}\left(\frac{\mathbf{x}-\mathbf{X}_{j,p}}{h_n}\right)\right)$.

By stationarity, $\text{Var}\left(\widehat{F}_{n,p}(\mathbf{x})\right)=\dfrac{1}{n-p}\text{Var}\left(\mathbf{U}\left(\dfrac{\mathbf{x}-\mathbf{X}_{1,p}}{h_n}\right)\right)+\dfrac{2}{(n-p)^2}$

$\displaystyle\sum_{j=2}^{n-p}(n-p-j+1)\mathbf{I}_{n,p}(\mathbf{x})$ By using the preceding lemma and lemma 1,

$$(n-p)\text{Var}\left(\widehat{F}_{n,p}(\mathbf{x})\right)=\mathbf{F}_p(\mathbf{x})-\mathbf{F}_p^2(\mathbf{x})-\mathbf{V}_2(\mathbf{x})h_n+(\mathbf{V}_3(\mathbf{x})-\mathbf{F}_p(\mathbf{x})\mathbf{V}_1(\mathbf{x}))h_n^2+$$
$$+\frac{2}{n-p}\sum_{j=2}^{n-p}(n-p-j+1)\times\left(\mathbf{F}_{p,j}(\mathbf{x},\mathbf{x})-\mathbf{F}_p^2(\mathbf{x})+\frac{h_n^2}{2}(\mathbf{V}_4(\mathbf{x})-\mathbf{F}_p(\mathbf{x})\mathbf{V}_1(\mathbf{x}))\right).$$

We have now,

$$(n-p)\text{Var}\left(\widehat{F}_{n,p}(\mathbf{x})\right)=\mathbf{F}_p(\mathbf{x})-\mathbf{F}_p^2(\mathbf{x})-\mathbf{V}_2(\mathbf{x})h_n+(\mathbf{V}_3(\mathbf{x})-\mathbf{F}_p(\mathbf{x})\mathbf{V}_1(\mathbf{x}))h_n^2+$$
$$+\sum_{j=2}^{n-p}\left(\mathbf{F}_{p,j}(\mathbf{x},\mathbf{x})-\mathbf{F}_p^2(\mathbf{x})\right)+(n-p-j+1)h_n^2\left(\mathbf{V}_4(\mathbf{x})-\mathbf{F}_p(\mathbf{x})\mathbf{V}_1(\mathbf{x})\right)-$$
$$\frac{2}{n-p}\sum_{j=2}^{n-p}(n-p-j+1)\times\left(\mathbf{F}_{p,j}(\mathbf{x},\mathbf{x})-\mathbf{F}_p^2(\mathbf{x})\right)+O(h_n^2).$$

Replacing $\displaystyle\sum_{j=2}^{n-p}\left(\mathbf{F}_{p,j}(\mathbf{x},\mathbf{x})-\mathbf{F}_p^2(\mathbf{x})\right)$ by $\displaystyle\sum_{j=2}^{\infty}\left(\mathbf{F}_{p,j}(\mathbf{x},\mathbf{x})-\mathbf{F}_p^2(\mathbf{x})\right)$ and subtracting to later result $\displaystyle\sum_{j=n-p+1}^{\infty}\left(\mathbf{F}_{p,j}(\mathbf{x},\mathbf{x})-\mathbf{F}_p^2(\mathbf{x})\right)$, we obtain now the expression for the variance of $\widehat{F}_{n,p}(\mathbf{x})$. ■

We may present now the behavior of the MSE.

Theorem 4 *Suppose* $\{X_n\}_{n\in\mathbb{N}}$, *satisfy* (A_1), (A_3), (A_4), (A_5), (A_6), (A_7) *and* (A_8). *Then,*

$(n-p)\text{MSE}\left(\widehat{F}_{n,p}(\mathbf{x})\right)=\boldsymbol{\sigma}^{\mathbf{2}}(\mathbf{x})-h_n\,\mathbf{V_2}(\mathbf{x})+O(n\,h_n^2)+o(h_n+n\,h_n^2)-\mathbf{c}_n(\mathbf{x})$.

Note that $c_n\to 0$, according to the assumptions made, and that c_n is independent of the bandwidth choice. It is now evident that an optimization of the convergence rate of the MSE is achieved by choosing $h_n=O(n^{-1})$ for all dimensions p. In fact, $h_n(\mathbf{x})=\frac{\mathbf{V}_2(\mathbf{x})}{2(n-p-1)(\mathbf{V}_4(\mathbf{x})-\mathbf{F}_p(\mathbf{x})\mathbf{V}_1(\mathbf{x}))}$.

To obtain, as in the independent case, the asymptotic optimal bandwidth of order $n^{-1/3}$, we replace assumptions (A_6) and (A_7) by,

(A_6^*) $nh_n^4\to 0$ (A_7^*) $\sum_{j=1}^{\infty}\left(\text{Cov}\,(X_1,X_{j+1})\right)^{\frac{1-\tau}{3}}<\infty,\ 0<\tau<1,$

as Cai and Roussas, 1998, did in the univariate case and providing that the sequence of covariances (4) is decreasing.

Theorem 5 *Suppose* $\{X_n\}_{n\in\mathbb{N}}$, *satisfy* (A_1), (A_3), (A_4), (A_5), (A_8), (A_6^*) *and* (A_7^*). *Then,*
$(n-p)\mathrm{MSE}\left(\widehat{F}_{n,p}(\mathbf{x})\right) = \sigma^2(\mathbf{x}) - h_n\,\mathbf{V_2}(\mathbf{x}) + O(n\,h_n^4) + o(h_n + nh_n^4) - \mathbf{c}_n(\mathbf{x}).$

Proof: To prove this result we use the identity $\mathbf{I}_{nj}(\mathbf{x}) = \mathbf{I}_j(\mathbf{x}) + O(h_n^2)$ (cf. Lema 1). As we noted in Remark 2, if we obtain an upper bound for $\mathbf{I}_j$ and, consequently, for $\mathbf{I}_{nj}$ we may use the following identity

$$\begin{aligned}|\mathbf{I}_{nj}(\mathbf{x}) - \mathbf{I}_j(\mathbf{x})| &= |\mathbf{I}_{nj}(\mathbf{x}) - \mathbf{I}_j(\mathbf{x})|^\tau\,|\mathbf{I}_{nj}(\mathbf{x}) - \mathbf{I}_j(\mathbf{x})|^{1-\tau} \le c^\tau\, h_n^{2\tau}\, p^{2(1-\tau)} \\ &\cdot\left|\left(\mathrm{Cov}^{1/3}(X_1,X_{j+1})\right)^{1-\tau}\right| = \tilde{c}\; h_n^{2\tau}\left|\left(\mathrm{Cov}^{1/3}(X_1,X_{j+1})\right)^{1-\tau}\right|,\end{aligned}$$

where $\tilde{c} = c^\tau\, p^{2(1-\tau)}$ is constant.

If we consider the following expression for the variance,
$(n-p)\mathrm{Var}\left(\widehat{F}_{n,p}(\mathbf{x})\right) = \mathrm{Var}\left(\mathbf{U}\left(\frac{\mathbf{x}-\mathbf{X}_{1,p}}{h_n}\right)\right) +$
$+\dfrac{2}{n-p}\displaystyle\sum_{j=2}^{n-p}(n-p-j+1)\left|\mathbf{I}_{nj}(\mathbf{x}) - \mathbf{I}_j(\mathbf{x})\right| + \sum_{j=2}^{n-p}(n-p-j+1)\mathbf{I}_j(\mathbf{x})$, then,
$\dfrac{1}{n-p}\displaystyle\sum_{j=2}^{n-p}(n-p-j+1)\left|\mathbf{I}_{nj}(\mathbf{x}) - \mathbf{I}_j(\mathbf{x})\right| \le \sum_{j=2}^{n-p}\left|\mathbf{I}_{nj}(\mathbf{x}) - \mathbf{I}_j(\mathbf{x})\right| \le$
$\tilde{c}h_n^{2\tau}\sum_{j=2}^{\infty}\left(\mathrm{Cov}^{1/3}(X_1,X_{j+1})\right)^{1-\tau} = O(h_n^{2\tau})$, by using (A_7^*). The result now follows readily. ■
Once again, is now evident that an optimization of the convergence rate of the MSE is achieved by choosing $h_n = O(n^{-1/3})$, for all dimensions p.

Corollary 1 *Suppose* $\{X_n\}_{n\in\mathbb{N}}$, *satisfy* $(A_1), (A_3), (A_4), (A_5)$, (A_6^*), (A_7^*) *and* (A_8). *Suppose further that the covariance sequence (4) is decreasing. Then, the asymptotic optimal bandwidth* $\{h_n\}_{n\in\mathbb{N}}$ *of kernel estimator of* $\mathbf{F}_p$ *is, for all dimensions p, in the MSE sense, of order* $O(n^{-1/3})$.

This work has been partially supported by CMAT and FCT under the program POCI 2010.

References

[Azevedo and Oliveira, 2000]C. Azevedo and P.E. Oliveira. Kernel-type estimation of bivariate distribution function for associated random variables. In *New Trends in Probability and Statistics*, pages 17–25, 2000.

[Cai and Roussas, 1998]Z. Cai and G. Roussas. Efficient estimation of a distribution function under quadrant dependence. *Scand. J. Statist.*, pages 211–224, 1998.

[Henriques and Oliveira, 2002]C. Henriques and P. Oliveira. Covariance estimation for associated random variables. *préprint*, 2002.

[Jin and Shao, 1999]Z. Jin and Y. Shao. On kernel estimation of a multivariate distribution function. *Statist. Probab. Lett.*, pages 163–168, 1999.

[Lebowitz, 1972]J. Lebowitz. Bounds on the correlations and analyticity properties of ferromagnetic ising spin systems. *Comm. Math. Phys.*, pages 313–321, 1972.

[Newman, 1980]C.M. Newman. Normal fluctuations and the fkg inequalities. *Comm. Math. Phys.*, pages 119–128, 1980.

[Oliveira and Suquet, 1999]P. Oliveira and Ch. Suquet. An $l^2[0,1]$ invariance principle for lpqd random variables. *Port. Mathematica*, pages 367–379, 1999.

[Roussas, 1993]G. Roussas. Smooth estimation of the distribution function under association: Asymptotic normality. *Tech. Report*, 1993.

[Roussas, 2000]G. Roussas. Asymptotic normality of the function under *Statist. Probab. Lett.*, pages 1–12, 2000.

[Sadikova, 1966]S.M. Sadikova. Two dimensional analogues of an inequality of essén with applications to the central limit theorem. *Theory Probab. Appl.*, pages 325–335, 1966.

Reference curves estimation via Sliced Inverse Regression

Ali Gannoun[1], Stéphane Girard[2], Christiane Guinot[3], and Jérôme Saracco[4]

[1] Equipe de Probabilités et Statistique, CC051
Institut de Mathématiques et de Modélisation de Montpellier,
UMR CNRS 5149,
Université Montpellier II,
Place Eugène Bataillon, 34095 Montpellier Cedex, France
(e-mail: gannoun@math.univ-montp2.fr)

[2] SMS/LMC/IMAG, BP 53
Université Grenoble I,
38041 Grenoble Cedex 9, France
(e-mail: Stephane.Girard@imag.fr)

[3] CE.R.I.E.S,
Biometrics and Epidemiology Department,
20, Rue Victor Noir, 92521 Neuilly sur Seine Cedex, France
(e-mail: christiane.guinot@ceries-lab.com)

[4] Equipe "Applications des Mathématiques",
Institut de Mathématiques de Bourgogne, UMR CNRS 5584,
Université de Bourgogne,
9 avenue Alain Savary, 21078 Dijon Cedex,
(email: Jerome.Saracco@u-bourgogne.fr)

Abstract. In order to obtain reference curves for data sets when the covariate is multidimensional, we propose a new methodology based on dimension-reduction and nonparametric estimation of conditional quantiles. This semiparametric approach combines sliced inverse regression (SIR) and a kernel estimation of conditional quantiles. The convergence of the derived estimator is shown. By a simulation study, we compare this procedure to the classical kernel nonparametric one for different dimensions of the covariate. The semiparametric estimator shows the best performance. The usefulness of this estimation procedure is illustrated on a real data set collected in order to establish reference curves for biophysical properties of the skin of healthy French women.

Keywords: Conditional quantiles, Dimension reduction, Kernel estimation, Semiparametric method.

1 Introduction

The reference intervals are a tool of some importance in clinical medecine. They provide a guideline to clinicians seeking to interpret a measurement obtained from a new patient. Many experiments, in particular in biomedical studies, are conducted to establish the range of values that a variable of interest, say Y whose values are in $\Re$, may normally take in a target population. Here "normally" refers to values that one can expect to see with a given probability under normal conditions and for typical individuals. The conventional definition of a reference interval is a pair of numbers that bind, for example, the central 90% of a set of values obtained from a specified group of subjects (the reference subjects).

The need for reference curves, rather than a simple reference interval, arises when a covariate, say X whose values are in $\Re$, is simultaneously recorded with Y. Norms are then constructed by estimating a set of conditional quantile curves. Conditional quantiles are widely used for screening biometrical measurement (height, weight, circumferences and skinfold) against an appropriate covariate (age, time). For details, the readers may refer, for example, to the work of [Healy *et al.*, 1998].

Let $\alpha \in (0, 1)$, the conditional quantile of Y given $X = x$, denoted by $q_\alpha(x)$, is naturally defined as the the root of the equation

$$F(y|x) = \alpha, \tag{1}$$

where $F(y|x) = P(Y \leq y \mid X = x)$ denotes the conditional distribution function of Y given $X = x$. For $\alpha > 0.5$, the $(2\alpha - 1)\%$ reference curves are defined, when x varies, by

$$I_\alpha(x) = [q_{1-\alpha}(x), q_\alpha(x)].$$

So, estimating reference curves is reduced to estimating conditional quantiles.

In the last decade a nonparametric theory has been developed in order to estimate the conditional quantiles. From (1), an estimator of the conditional distribution induces an estimator of corresponding quantiles. For instance, a *Nadaraya-Watson* estimator, $\hat{F}_n(y|x)$, can be assigned to $F(y|x)$:

$$\hat{F}_n(y|x) = \sum_{i=1}^{n} K\{(x - X_i)/h_n\} I_{\{Y_i \leq y\}} \bigg/ \sum_{i=1}^{n} K\{(x - X_i)/h_n\}\ , \tag{2}$$

where h_n and K are respectively a bandwidth and a bounded (kernel) function. The estimator of $q_\alpha(x)$ is then deduced from $\hat{F}_n(y|x)$ as the root of the equation

$$\hat{F}_n(y|x) = \alpha. \tag{3}$$

Many authors are interested in this estimator, see, for mathematical details, [Samanta, 1989] or [Berlinet *et al.*, 2001]. Note that various other nonparametric methods are explored in order to estimate $q_\alpha(x)$. Among them we can cite the *local polynomial*, the *double kernel*, the *weighted Nadaraya-Watson* methods.

Although, theoretically, the extension of conditional quantiles to higher dimension p of X is obvious, its practical success, while depending on the number of observations, suffers from the so-called *curse of dimensionality.* Further, because reference curves are, in this case, a pair of p-dimensional hyper-surfaces, their visual display is rendered difficult making it less directly useful for exploratory purposes (unlike the one-dimensional case). When $p > 2$, viewing all the data in single $(p + 1)$-dimensional plot may no longer be possible.

Motivated by this, the key is then to reduce the dimension of the predictor vector X without loss of information on the conditional distribution of Y given X and without requiring a prespecified parametric model. Sufficient dimension-reduction leads naturally to the idea of a sufficient summary plot that contains all information on the regression available from the sample. Moreover, it is a very helpful step in nonparametric estimation for circumvening the curse of dimensionality. Methods to reduce the dimension exist in the literature. For instance, [Stone, 1985] or [Stone, 1986] used additive regression models to cope with curse of dimensionality in nonparametric function estimation. [Chaudhuri, 1991] used this technique in order to estimate conditional quantiles. In this paper, we focus on a linear projection method of reducing the dimensionality of the covariates in order to construct a more efficient estimator of conditional quantiles and consequently reference curves. The specific dimension reduction method used is based on Li's well known Sliced inverse regression (SIR), see[Li, 1991] or [Chen and Li, 1998]. From a computational point of view, SIR is very fast. Note that this method is used as a pre-step of the main analysis of the data, in order to get an efficient estimator of conditional quantiles from which we can then deduce reference curves. It is fairly robust, especially against some outliers in the regressor observations.

The rest of the paper is organized as follows. In Section 2, we present the dimension-reduction context and we derive the corresponding semiparametric estimator of conditional quantiles. We also give an asymptotic result. Simulations are conducted in Section 3 to assess the performance of this estimator in finite-sample situation. Numerical example involving real data application is reported in Section 4.

2 Dimension-reduction context and estimation procedure

2.1 Dimension-reduction context

Suppose that there exists a matrix β such that

$$Y \perp X \mid \beta^T X, \tag{4}$$

where the columns of the $p \times d$ matrix β $(d \leq p)$ are linearly independent. Consequently, in the current study, statement (4) is equivalent to

$$F(y|x) = F(y|\beta^T x),$$

for all values of x in the sample space. Straightforwardly, it follows that

$$q_\alpha(x) = q_\alpha(\beta^T x).$$

The SIR method can be used to estimated a basis of the subspace $S(\beta)$ spanned by the columns of β. More details and comments on the SIR estimation procedure can be found in [Li, 1991] or [Chen and Li, 1998].

2.2 Estimation procedure

Let Y_i denote the ith observation on the univariate response and let X_i denote the corresponding $p \times 1$ vector of observed covariate values, $i = 1, \ldots, n$.

• **Step 1: SIR estimation step.** With SIR method, we get $\{\hat{b}_k\}_{k=1}^d$, an estimated basis of $S(\beta)$. In practice, the dimension d is replaced with an estimate $\hat{d}$ equal to the number of singular values that are inferred to be nonzero in the population, see for example, [Li, 1991] or [Ferré, 1998] for testing procedure in order to identify d. Moreover, the eigenvalues scree plot approach used here is a useful explonatory tool in determining the number $\hat{d}$ of EDR directions to keep. From a practical point of view, we look for a visible jump in the scree plot and $\hat{d}$ is then the number of the eigenvalues located before this jump. Note that if no jump is detected, no dimension reduction is possible with SIR approach.

• **Step 2: Conditional quantile estimation step.** For the sake of convenience, we assume that $d = 1$ and we use the notation $\hat{b} = \hat{b}_1$. Using the SIR estimates and following (2), a kernel estimator of $F(y|x)$ is defined, from the data $\{(Y_i, \hat{b}^T X_i)\}_{i=1}^n$, by

$$F_n\left(y \,\middle|\, \hat{b}^T x\right) = \frac{\sum_{i=1}^n K\{(\hat{b}^T x - \hat{b}^T X_i)/h_n\} I_{\{Y_i \leq y\}}}{\sum_{i=1}^n K\{(\hat{b}^T x - \hat{b}^T X_i)/h_n\}}. \tag{5}$$

Then, as in (3), we derive from (5) an estimator of $q_\alpha(x)$ by

$$q_{n,\alpha}\left(\hat{b}^T x\right) = F_n^{-1}(\alpha \mid \hat{b}^T x). \tag{6}$$

As a consequence of the above result, for $\alpha > 0.5$, the corresponding estimated $(2\alpha - 1)\%$ reference curves are given by the following

$$I_{n,\alpha}(x) = [q_{n,1-\alpha}(\hat{b}^T x), q_{n,\alpha}(\hat{b}^T x)], \quad \text{as } x \text{ varies.}$$

2.2.0.2 Remark. The above definitions have been presented in the context of single index. A natural extension is to consider the general multiple indices $(d > 1)$ and to work with $\{\hat{b}_k\}_{k=1}^d$ and $\{(Y_i, \hat{b}_1^T X_i, \ldots, \hat{b}_d^T X_i)\}_{i=1}^n$. Then we use the classical multi-kernel estimation to get $q_{n,\alpha}(\hat{b}_1^T x, \ldots, \hat{b}_d^T x)$ as in (6).

2.3 Asymptotic property.

Under usual assumptions, we obtain the consistency of $q_{n,\alpha}(\hat{b}^T x)$: for a fixed x in $\Re^p$,

$$q_{n,\alpha}(\hat{b}^T x) \longrightarrow q_\alpha(x) \quad \text{in probability, as } n \to +\infty.$$

The proof is given in [Gannoun *et al.*, 2004].

3 Simulation study

We study the numerical performances of the proposed method on simulated data. In particular, we compare our method with the classical nonparametric estimation method. Let us introduce the following estimators of $q_\alpha(x)$:

(a) $q^{(a)}_{n,\alpha}(x) := q_{n,\alpha}(\widehat{b}^T x)$ is the estimator defined in (6).

(b) $q^{(b)}_{n,\alpha}(x) := q_{n,\alpha}(\beta^T x)$ has no practical interest, it is only introduced for the sake of comparison. It is similar to **(a)** except the dimension-reduction direction is not estimated but fixed to the theoretical one.

(c) $q^{(c)}_{n,\alpha}(x) := q_{n,\alpha}(x)$ is the classical conditional nonparametric quantile estimator.

The kernels are the densities of the standard normal or multinormal distribution, and the bandwidth is chosen by a cross-validation technique. The estimated conditional quantiles are computed by numerically inversing the corresponding conditional distribution function.

3.1 Simulated models

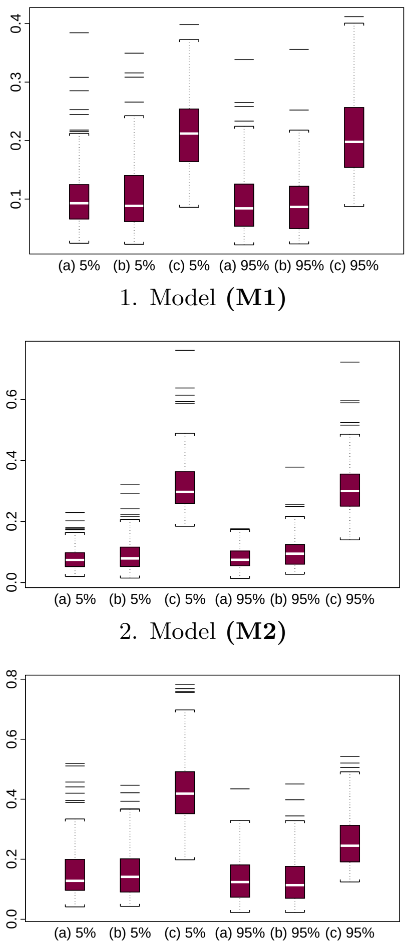
We consider the following regression model $Y = f(\beta^T X) + \varepsilon$, where X follows the standard multinormal distribution $\mathcal{N}_p(0, I_p)$ and where ε is normally distributed $\varepsilon \text{sim} \mathcal{N}(0,1)$ and is independent from X. We examine three situations:

(M1) $p = 3$, $f(t) = 1 + 2t/3$ and $\beta^T = 2^{-1/2}[1, -1, 0]$.

(M2) $p = 10$, $f(t) = 1 + 2t/3$ and $\beta^T = 3^{-1}[1, 1, 1, 1, 1, -1, -1, -1, -1, 0]$.

(M3) $p = 3$, $f(t) = 1 + \exp(2t/3)$ and $\beta^T = 2^{-1/2}[1, -1, 0]$.

Our motivation for considering the pair of models **(M1,M2)** is to investigate the behavior of the estimation methods when the dimension increases. The pair of models **(M1,M3)** is introduced to evaluate the influence of the link function f on the accuracy of the estimation methods. Let us note that $q_\alpha(x) = f(\beta^T x) + N_\alpha$, where N_α is the α-quantile of the standard normal distribution.



1. Model **(M1)**

2. Model **(M2)**

3. Model **(M3)**

Fig. 1. Boxplots obtained on the three different models with the three different estimates.

3.2 Evaluation of the results

Our goal is to compare successively the three estimators **(a)**, **(b)** and **(c)** to the true quantile in the situations **(M1)**, **(M2)** and **(M3)**. To this end, the $N = 100$ data sets with size $n = 200$ are simulated in each of the above situations. The conditional quantiles are estimated for $\alpha = 5\%$ and $\alpha = 95\%$ on a p dimensional grid. This grid is composed of 125 points $\{z_\ell,\ \ell = 1, \ldots, 125\}$ randomly generated with a uniform distribution on $[-3/2, 3/2]^p$. Then, the performance of the estimators can be assessed on each of the N

simulated data sets by a mean square error criterion:

$$E_{n,\alpha}^{(\Theta)} = \frac{1}{125}\sum_{\ell=1}^{125}\left(q_{n,\alpha}^{(\Theta)}(z_\ell) - q_\alpha(z_\ell)\right)^2, \quad \text{where } \Theta \in \{a,b,c\}.$$

The boxplots of the mean square error $E_{n,\alpha}^{(\Theta)}$ for $\Theta \in \{a,b,c\}$ and $\alpha \in \{0.05, 0.95\}$ on each model are represented on Figure 1. Figure 1.1 shows no difference between the distribution of $E_{n,\alpha}^{(a)}$ and $E_{n,\alpha}^{(b)}$. The estimation of the direction β by $\widehat{b}$ has no significant consequence on the accuracy of the estimation of the reference curves. On the contrary, results obtained by the estimators **(a)** and **(c)** are very different. The proposed estimator **(a)** gives better results than the estimator without dimension-reduction **(c)**. Besides, this difference of quality increases with the number p of covariates (see Figure 1.3). In this case, the curse of dimensionality becomes an essential limitation to the use of estimator **(c)**, and thus estimator **(a)** is particularly useful in such situations. Note that the quality of the estimation of β is not severely affected by the covariates number. Finally, in view of Figure 1.2, the nature of the link function f does not seem to have any influence on the relative behaviors of the three estimators.

4 Application to real data

4.1 Data

When studying biophysical skin properties of healthy women, knowledge about the reference "curves" of certain parameters is lacking. The aim is to establish 90% reference "curves" for some of the biophysical properties of the skin (here the conductance of the skin) of healthy Caucasian women, on two facial areas and one forearm area, using the age and a set of covariates. The data collection was conducted from November 1998 to March 1999 on $n = 322$ Caucasian women between 20 and 80 years old with apparently healthy skin, and living in the Ile de France (in around Paris) area. The volunteers were preselected by a subcontractor company. Each healthy volunteer was examined at CE.R.I.E.S ("CEntre de Recherches et d'Investigations Epidermiques et Sensorielles" or Epidermal and Sensory Research and Investigation Centre) in a controlled environment. This evaluation included self-administered questionnaires on skin-related habits, a medical examination and a biophysical evaluation. The age of the volunteer, the temperature and relative humidity of the controlled environment occur in each study as covariates. The other available covariates included are some biophysical properties of the skin (as the the skin temperature or the skin pH).

4.2 Results

We only give here the results for the forearm area. In step 1, the SIR method gives $\hat{d} = 1$ and the corresponding vector $\hat{b}$. Then in step 2, after a simplifi-

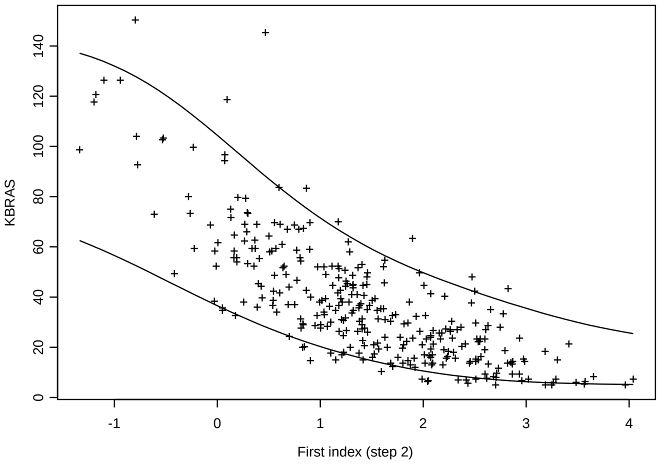


Fig. 2. Estimated 90%-reference curves for the forearm area.

cation of the index $\hat{b}^T X$ (see [Gannoun *et al.*, 2001] or [Gannoun *et al.*, 2004] for details), we construct the 90% reference curves for the conductance of the skin (variable named KBRAS) using this estimated index, see Figure 2. The results of the analysis on the forearm index show that apart from age five covariates enter in the model: two of these represent the environmental conditions of the measurements, which is to be expected, the three other covariates are directly clinically-related with skin hydration: skin pH, capacitance and transepidermal water loss. The studies of the two facial areas can be found in [Gannoun *et al.*, 2001].

References

[Berlinet *et al.*, 2001]A. Berlinet, A. Gannoun, and E. Matzner-Lober. Asymptotic normality of convergent estimates of conditional quantiles. *Statistics*, pages 139–169, 2001.

[Chaudhuri, 1991]P. Chaudhuri. Global nonparametric estimation of conditional quantile functions and their derivative. *Journal of Multivariate Analysis*, pages 246–269, 1991.

[Chen and Li, 1998]C.H. Chen and K.C. Li. Can SIR be as popular as multiple linear regression?. *Statistica Sinica*, pages 289–316, 1998.

[Ferré, 1998]L. Ferré. Determining the dimension in sliced inverse regression and related methods. *Journal of the American Statistical Association*, pages 132–140, 1998.

[Gannoun *et al.*, 2001]A. Gannoun, S. Girard, S. Guinot, and J. Saracco. Dimension-reduction in reference curves estimation. *Technical report, Unité de Biométrie, ENSAM-INRA-UMII*, 2001.

[Gannoun *et al.*, 2004]A. Gannoun, S. Girard, S. Guinot, and J. Saracco. Sliced inverse regression in reference curves estimation. *Computational Statistics and Data Analysis*, pages 103–122, 2004.

[Healy *et al.*, 1998]M.J.R. Healy, J. Rasbash, and M. Yang. Distribution-free estimation of age-related centiles. *Annals of Human Biology*, pages 17–22, 1998.

[Li, 1991]K.C. Li. Sliced inverse regression for dimension reduction (with discussion). *Journal of the American Statistical Association*, pages 316–342, 1991.

[Samanta, 1989]T. Samanta. Non-parametric estimation of conditional quantiles. *Statistics and Probability Letters*, pages 407–412, 1989.

[Stone, 1985]C.J. Stone. Additive regression and other nonparametric models. *The Annals of Statistics*, pages 689–705, 1985.

[Stone, 1986]C.J. Stone. The dimensionality reduction principle for generalized additive models. *The Annals of Statistics*, pages 590–606, 1986.

Acknowledgement: We are grateful to Pr. Denis Malvy, University Bordeaux 2, for his advice and clinical expertise, to Pr. Erwin Tschachler for his encouragement, and to all the CE.R.I.E.S. team for their important contribution to the data.

Estimation of the population mean using auxiliary information when some observations are missing

M del Mar Rueda[1], Silvia González[2], Antonio Arcos[1], Yolanda Román[1], M Dolores Martínez[1], and Juan Fco. Muñoz

[1] Department of Statistics and Operational Research
Faculty of Sciences. University of Granada,
18071 Granada, Spain
(e-mail: mrueda@ugr.es, arcos@ugr.es, yroman@ugr.es, mmiranda@ugr.es)

[2] Department of Statistics and Operational Research
University of Jaén
Jaén, Spain
(e-mail: sgonza@ujaen.es)

Abstract. In this paper we use tools of classical statistical estimation theory in finding a suitable estimator of the population mean using auxiliary information when some observations in the sample are missing. We study model and design properties of the proposed estimator. We also report the results of a wide simulation study on the efficience of the estimator which reveals very promising results.

Keywords: Auxiliary information, missing data, superpopulation model.

1 Introduction

Missing data is a common problem in virtually all surveys. Frequently, survey sampling is conducted to gather complete information on all sampling units but, due to a variety of reasons, for a fraction of the subjects, either no data at all is available or information on one or more variables is missing. Missing data can contribute to bias in the estimates and make the analyses harder to conduct and results harder to present.

The most frequently used method to compensate for item non response is imputation (see [Little and Rubin, 1987]). Some statistics specialists are reluctant to apply this method because it manipulates the original information. Many empirical studies do not follow this approach. They simply discard all the sampling units with missing values and employ the usual inference procedures, which can produce that the actual sample size was less than the planned one, biases in estimations and increases in sampling variance if missing data follows any pattern.

Contending that the deleted observations may contain valuable information, an alternative approach is to try to improve the precision of the estimators by including all cases available for their calculation.

In this paper we propose a prediction approach to deal with the presence of missing data. Specifically, we address the case where only the value of the variable of interest is missing for some subjects, and the value of the auxiliary variable is missing for other distinct subjects. We propose a new estimator for the mean of the variable of interest, using all known data for principal and auxilary variables.

2 Estimation with auxiliary information and missing data

Indirect estimation methods are easily comprehensible techniques for the estimation of total population in survey sampling when an auxiliary characteristic correlated with the study characteristic is available; see, e.g., [Singh, 2003], [Sampath and Chandra, 1990], [Srivastava and Jhajj, 1981]. These methods of estimation assume that the sample data contains no missing observations. This specification may not be tenable in many practical applications; see; e.g., [Rubin, 1977].

Some authors have defined indirect estimators when the sample is drawn by a simple random sampling without replacement when some observations are missing and the population mean of auxiliary characteristic is available (see [Tracy and Osahan, 1994], [Toutenburg and Srivastava, 1998] and [Rueda and González, 2004]).

There appears to be no effort reported in the literature when both the asumptions are violated simultaneously (some observations are missing in both variables and the population mean of the auxiliary variable is not known). We will consider this situation under a general sampling design.

Let be a population,U, of N units from which a random sample, s, of fixed size, n is drawn according to a noninformative sample design $d = (S_d, P_d)$, with first order inclusion probabilities π_i. For this sample we observe the values of two variables, (y_i, x_i), $i = 1, \dots, n$, for the estimation of some parameters of variable y.

We assume that only a set of $(n-p-q)$ complete observations on selected units in the sample are available. In addition to these, observations on the x characteristic on p units in the sample are available but the corresponding observations on the y characteristic are missing. Similarly, we have a set of q observations on the y characteristic in the sample but the associated values on the x characteristic are missing. Further, p and q are assumed to be integer numbers verifying $0 < p, q < n/2$.

For the sake of simplicity, we separate the unit of the sample s into three disjoint sets:

$$\begin{aligned}
s_1 &= \{i \in s/x_i, y_i \text{ are available}\} \\
s_2 &= \{i \in s/x_i \text{ are available, but } y_i \text{ is not}\} \\
s_3 &= \{i \in s/y_i \text{ are available, but } x_i \text{ is not}\}
\end{aligned}$$

Prediction theory for sampling surveys can be considered as a general framework for statistical inference on the characteristics of finite populations. The prediction approach is based on this idea: for any given $s \in S_d$ of size n we can write:

$$\overline{Y} = f_s\overline{y}_s + (1 - f_s)\overline{y}_{\tilde{s}} \tag{1}$$

where $f_s = n/N$ is the sampling rate, $\overline{y}_s = \sum_s Y_k/n$ is the mean for units in the sample, and $\overline{y}_{\tilde{s}} = \sum_{\tilde{s}} Y_k/(N - n)$ is the mean for the nonsample units.

In this representation of the mean, the sample mean $\overline{y}_s$ is known, and then we attempt a post survey prediction of the mean $\overline{y}_{\tilde{s}}$ of the nonsurveyed units.

Now for any given $s \in S_d$ we can write:

$$\overline{Y} = f_{s1}\overline{y}_{s1} + f_{s2}\overline{y}_{s2} + f_{s3}\overline{y}_{s3} + (1 - f_s)\overline{y}_{\tilde{s}} \tag{2}$$

where $f_{s1} = \dfrac{n - p - q}{N}$, $f_{s2} = \dfrac{p}{N}$, $f_{s3} = \dfrac{q}{N}$ and $f_{s4} = 1 - \dfrac{n}{N}$,

In this representation of the mean, the sample means $\overline{y}_{s1}$ and $\overline{y}_{s3}$ are known, thus the problem of predicting $\overline{Y}$ is equivalent to the problem of predicting the means $\overline{y}_{s2}$ and $\overline{y}_{\tilde{s}}$.

We denote by E_ξ the expected value under the model ξ and E_d the expected value under the design d. The minimum $E_\xi MSE_d$ criterium will be considered. We only consider the linear and unbiased under model predictors.

Consider any predictor T of $\overline{Y}$; it can be represented, for any given sample s as:

$$T = f_{s1}\overline{y}_{s1} + f_{s2}U_2 + f_{s3}\overline{y}_{s3} + (1 - f_s)U_4 \tag{3}$$

where U_2 and U_4 are considered as predictors of $\overline{y}_{s2}$ and $\overline{y}_{\tilde{s}}$ respectively. Tools of classical statistical estimation theory will be useful in finding the suitable predictors U_2 and U_4.

Firstly we study the problem of estimation of $\overline{y}_{s2}$. If the predictor T is of the form 3 and it verify: $E_\xi(T) = \mu = \frac{1}{N}\sum_{i\in U} E_\xi(Y_i)$, it is logical to consider the class of linear estimators U_2 with the condition: $E_\xi(U_2) = \mu_{s2} = \frac{1}{p}\sum_{i\in s2} E_\xi(Y_i)$. In the sample s_2 we do not have the values of the study characteristic but we have all the values of the auxiliary charasterictic, x. We now consider the frequently used regression model, where $\eta_i = \beta x_i$, $i = 1, ..., N$, where β is a unknown quantity. By generalized least squares theory, the minimum variance linear unbiased under the model estimator of β is, for a given sample, given by $\widehat{\beta}$ the sample regression coefficient. Then we consider the predictor $U_2^* = \widehat{\beta\overline{x}}_{s2}$ that is linear and unbiased under the model of $\overline{y}_{s2}$.

Regarding the estimation of $\overline{y}_{\tilde{s}}$, there is not any information available in s_4, neither from the study characteristic neither from the auxiliary characteristic, so it is logical to consider the sample mean $U_4^* = \overline{y}_{s1\bigcup s3}$.

We consider the predictor of $\overline{Y}$:

$$T^* = f_{s1}\overline{y}_{s1} + f_{s2}U_2^* + f_{s3}\overline{y}_{s3} + (1 - f_s)U_4^* \tag{4}$$

As $E_d(\overline{y}_{s1}) = E_d(\overline{y}_{s3}) = 0$, T^* is a linear ξ-unbiased predictor of $\overline{Y}$ for any design d, and therefore the random variable obtained from T^* if Y_k is fixed at y_k is ξ-unbiased estimator of population mean $\overline{y}$. The estimator T^* is also asymptotically normal. The proof use the asymptotical normality of U_4^* and U_2^* (see, e.g., Valliant et al., 2000).

Writting
$k_1 = \dfrac{n-p-q(N-p)}{N(n-p)}$, $k_2 = \dfrac{q(N-p)}{N(n-p)}$ and $k_3 = \dfrac{p}{N}$
the proposed estimator can be expressed as follows:

$$T^* = k_1\overline{y}_{s1} + k_2\overline{y}_{s3} + k_3\widehat{\beta}\overline{x}_{s2} \tag{5}$$

2.1 Simple random sampling

Next, we are going to consider a simple random sampling without replacement. We are interested in finding the statistical properties of the estimator with respect to this sampling design.

First, the estimator is unbiased under this design the approximate variance of T^* is

$$AV(T^*) = S_y^2\left[k_1^2 a + k_2^2 b + 2k_1k_2c\right] + \beta^2 k_3^2 S_x^2 d + 2k_3\beta S_{xy}\left[k_1 e + k_2 f\right] \tag{6}$$

where
$a = \frac{1}{n-p-q} - \frac{1}{N}$, $b = \dfrac{1}{q} - \dfrac{1}{N}$, $d = \dfrac{1}{p} - \dfrac{1}{N}$

$$c = \begin{cases} \dfrac{1}{n-p-q} - \dfrac{1}{N} & \text{if} \quad \dfrac{n-p}{2} \geq q \\[2ex] \dfrac{1}{q} - \dfrac{1}{N} & \text{if} \quad \dfrac{n-p}{2} < q \end{cases}$$

$$e = \begin{cases} \dfrac{1}{n-p-q} - \dfrac{1}{N} & \text{if} \quad \dfrac{n-q}{2} \geq p \\[2ex] \dfrac{1}{p} - \dfrac{1}{N} & \text{if} \quad \dfrac{n-q}{2} < p \end{cases}$$

$$f = \begin{cases} \dfrac{1}{p} - \dfrac{1}{N} & \text{if} \quad p \geq q \\[2ex] \dfrac{1}{q} - \dfrac{1}{N} & \text{if} \quad p < q \end{cases}$$

A consistent estimator of $AV(T^*)$ can be simply obtained by substituting S_y^2, S_x^2 and S_{yx} with their sample values s_y^2, s_x^2 and s_{yx}.

Table 1. Relation between lines type and nonresponse rates

Type of line	CASE 1	CASE 2	CASE 3
dotted	p = 0.32n	p = 0.32n	p = 0.4n
	q = 0.4n	q = 0.48n	q = 0.48n
dashed	p = 0.4n	p = 0.48n	p = 0.48n
	q = 0.32n	q = 0.32n	q = 0.4n

3 Simulation study

The next step in our study consists of carrying out a simulation study to reveal the behaviour of the proposed estimator. For this purpose, we examined four populations: Cancer, Co60, Co70 and Hospital (see [Valliant *et al.*, 2000]).

In order to study the properties of the proposed estimator, the following process was repeated 1000 times: a simple random sample was selected, for which in a completely random way the selected proportion of cases for both variables was removed. The values of the proposed estimator T^* and of the estimator of the simple mean were then calculated. The results of this simulation are shown in Figure 1, and Table 1 describes the correspondence between the types of line and the nonresponse rates.

The above Figure represents the log-ratios of the mean squared errors of both estimators. The simulation results shown that for all the populations, sampling sizes and nonresponse rates considerated, the behaviour of the proposed estimator is better than that of the standard one (the sample mean). Moreover, there is an absence of variation in the error of estimation, produced by exchanging the proportion of nonresponders between the main variable and the auxiliary variable. Another interesting feature is that the precision improves in proportion to the increase in the sample size.

After comparing the T^* estimator and the standard estimator of the mean, we considered it useful to study the relation between the efficiency of the proposed estimator and that of the estimator defined by Toutenburg and Srivastava (1988), under the same conditions. We conclude that the behaviour of the T^* estimator is considerably better than that the Toutemburg estimator $\hat{y}_{T4}$.

References

[Little and Rubin, 1987]R. J. A. Little and D. B. Rubin. *Statistical analysis with missing data*. John Wiley, New York, 1987.

[Rubin, 1977]D. B. Rubin. Formalizing subjective notions about the effect of nonrespondents in sample surveys. *Journal of the American Statistical Association*, pages 538–543, 1977.

[Rueda and González, 2004]M. Rueda and S. González. Missing data and auxiliary information in surveys. *Computational Statistics*, pages 555–567, 2004.

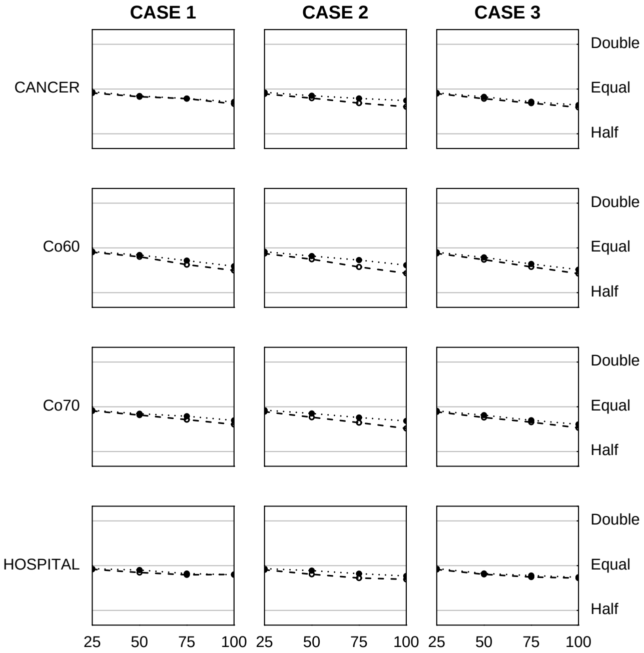


Fig. 1. Log ratios of standar error of the *predictive* estimators over the *simple* estimator.

[Sampath and Chandra, 1990]S. Sampath and S. K. Chandra. General class of estimators for the population total under unequal probability sampling schemes. *Metron*, pages 409–419, 1990.

[Singh, 2003]S. Singh. *Advanced sampling theory with applications. How Michael selected Amy.* Kluwer Academic Press, London, 2003.

[Srivastava and Jhajj, 1981]S. K. Srivastava and H. S. Jhajj. A class of estimators of the population mean in survey sampling using auxiliary information. *Biometrika*, pages 341–343, 1981.

[Toutenburg and Srivastava, 1998]H. Toutenburg and V. K. Srivastava. Estimation of ratio of population means in survey sampling when some observations are missing. *Metrika*, pages 177–187, 1998.

[Tracy and Osahan, 1994]D. S. Tracy and S. S. Osahan. Random non-response on study variable versus on study as well as auxiliary variables. *Statistica*, pages 163–168, 1994.

[Valliant *et al.*, 2000]R. Valliant, A. H. Dorfman, and R. M. Royall. *Finite population sampling and inference.* John Wiley, London, 2000.

A Comparison of Methods for Joint Modelling of Mean and Dispersion

Ilmari Juutilainen and Juha Röning

Computer Engineering Laboratory
PO BOX 4500
90014 University of Oulu, Finland
(e-mail: ilmari.juutilainen@ee.oulu.fi, juha.roning@ee.oulu.fi)

Abstract. We considered the suitability of the methods for joint modelling of mean and dispersion for prediction based on large data sets under the assumption of normally distributed errors. Methods that seemed capable of handling a problem with 25 explanatory variables and 100000 observations were compared in predicting the strength of steel in a real data set collected from the production line of a steel plate mill. A neural network model for mean and dispersion gave the best prediction. The results indicate that neural networks are suitable for joint modelling of mean and dispersion in large data sets.

Keywords: Joint modelling of mean and dispersion, Heteroscedasticy.

1 Introduction

Joint modelling of mean and dispersion is a common problem in statistics. In many real problems, not only mean but also variance and even other moments of the conditional distribution of the response variable depend on the explanatory variables. In these cases, dispersion modelling is needed to predict the conditional distribution realistically. The variance model has often been employed to make mean model estimation more efficient. In many applications, including industrial quality improvement experiments, the variance function itself has been the focus of the interest.

A single observation does not give any information about variance, and many more observations are needed to estimate a model for variance than a model for mean. Although joint modelling of mean and dispersion has been applied in many fields, applications to large data sets seem to be lacking. The different methods for joint modelling of mean and dispersion have not been compared to each other, and their prediction abilities and suitability to large data sets are rather unclear. This paper gives insight into the suitability of different methods proposed for joint prediction of mean dispersion based on large data sets. The models are compared for their accuracy in predicting the mean and variance of the strength of steel plates using a real data set with about 25 explanatory variables and 100000 observations.

2 Joint modelling of mean and dispersion

We denote the observations of the response variable with $Y = (y_1, y_2, \ldots, y_N)^{\mathrm{T}}$ and let $x_i = (x_{i1}, x_{i2}, \ldots, x_{ip})$ denote the values of the p explanatory variables of the ith observation. We assume that y_i are normally, independently distributed $y_i \mathrm{sim} N(\mu_i, \sigma_i^2)$, with both the mean $\mu_i(x_i)$ and the variance $\sigma_i^2(x_i)$ depending on the explanatory variables. Joint modelling of mean and dispersion can be divided into two tasks: estimation of the mean function and estimation of the variance function [Carroll and Ruppert, 1988]. In the iterative estimation method, the mean function is estimated with weighted least squares by keeping the variance model fixed and by using weights proportional to the inverses of the predicted variances. The variance function is then estimated by keeping the mean model fixed [Carroll and Ruppert, 1988]. There has been controversy as to the number of iterations needed. Sometimes good results have been obtained using only one iteration [Yu and Jones, 2004], and two iterations have often been considered best [Carroll and Ruppert, 1988]. Simple models can be estimated without iteration using full maximum likelihood or restricted maximum likelihood (REML).

The selection of the response for dispersion model fitting is not obvious because direct measurements of variance cannot be made without replication. Natural measurement of the variance is provided by the squared residual $\widehat{\varepsilon}_i^2 = (y_i - \widehat{\mu}(x_i))^2$. Fitting of the mean model biases the estimation of the variance function because the fitted model always adapts itself to the estimation data. This bias can be corrected by modifying the response: for example, in a regression context the response $\varepsilon_i^2/(1 - h_{ii})$, where h_{ii} are the diagonal elements of the hat matrix, corresponds to the REML estimation and leads to unbiased fitting [Smyth *et al.*, 2001]. If the fit can be expressed using a smoother matrix, $\widehat{Y} = SY$, the expectation of a squared residual in the estimation data is $E\widehat{\varepsilon}_i^2 = \sigma_i^2 - 2S_{ii}\sigma_i^2 + \sum_{j=1}^{N} S_{ij}^2\sigma_j^2 + (\mu_i - E\widehat{\mu}_i)^2$ [Ruppert *et al.*, 1997]. Defining $\mathit{\Delta} = \mathrm{diag}(2S - SS^{\mathrm{T}})$ and assuming the fit to be conditionally unbiased, the result motivates the $\mathit{\Delta}$-corrected response $r_i = \widehat{\varepsilon}_i^2/(1 - \mathit{\Delta}_i)$.

The learning method, i.e. model type and estimation method, is another major selection problem in dispersion modelling. In principle, most of the learning methods can be used for modelling dispersion. If the residuals are normally distributed, $\varepsilon_i \mathrm{sim} N(0, \sigma_i^2)$, then the squared residuals are gamma distributed, $\varepsilon_i^2 \mathrm{simGamma}(\sigma_i^2, 2)$, and the fitting can be based on gamma log-likelihood. For most other possible responses (e.g. $|e_i|$ or $\log |e_i|$) no such helpful result is available, and the least squares method has been commonly used.

3 Methods

Heteroscedastic regression (HetReg), mean and dispersion additive models (MADAM), local linear regression for mean and dispersion (LLRMD) and neural network modelling of mean and dispersion (NNMMD) were compared in a real data set. The estimation data were collected from an industrial process of steel plate production and consisted of 90 000 observations. Two response variables were measured from finished products; tensile strength and yield strength, both being approximately normally distributed. In the modelling, 27 explanatory variables related to the steel plate production process and likely to have an effect on the responses were used. The explanatory variables were related to the concentrations of alloying elements, the thermomechanical treatments made during the process of production and the size and shape of the plate and the test specimen. In the modelling of variance, 12 of the explanatory variables with a likely effect on the conditional variance were used. [Myllykoski, 1998] has studied the reasons affecting the variance in the strength of thin steel sheets.

The fitting of models was accomplished using the iterative approach. First, the model for mean was fitted, and the variance model was then fitted based on the corrected or uncorrected squared residuals from the mean model fit. In the optional second iteration, the mean model was weighted with the inverses of the predicted variances, and the variance model was fitted again. The parameters of the mean model were estimated with the least squares, and the parameters of the variance model were estimated with the gamma log-likelihood or least squares. For the models MADAM and NNMMD the likelihoods were penalised. A linear link was used for the mean and a square root link or log link for the variance.

The test data set was collected from the production line after the training data set and consisted of 25 000 observations. The prediction accuracies of the models were compared using the negative log-likelihood of the test data set under a gaussian assumption. Variance predictions smaller than 16 (including negative predictions) were transformed to 16; otherwise, single bad predictions could have blurred the results.

Heteroscedastic linear regression is a simple method, which can be easily applied to large data sets [Smyth *et al.*, 2001]. We used a heteroscedastic regression model of the form

$$\begin{aligned} f(\mu_i) &= \tilde{z}_i^{\mathrm{T}}\beta \\ g(\sigma_i^2) &= z_i^{\mathrm{T}}\tau \end{aligned} \tag{1}$$

where the link functions f and g define the relationship between the linear predictors and the mean and variance, respectively. The input vectors $\tilde{z}_i$ and z_i include transformations and product terms of the original explanatory variables to allow non-linear effects and interactions between the explanatory variables. We made the model selection manually based on the prediction accuracy in the validation data set. The selected mean models included about 110 terms and the dispersion models about 25 terms. The model estimation was carried out using the iterative REML of [Smyth *et al.*, 2001].

Generalised additive models are known to be able to handle large data sets pretty well [Hastie *et al.*, 2001]. [Rigby and Stasinopoulos, 1996] proposed mean and dispersion additive models for joint modelling of mean and dispersion. We used an additive model resembling the model of [Yau and Kohn, 2003] and allowing two-way interactions

$$\begin{aligned} f(\mu_i) &= \sum_{j=1}^{p} h_j(x_{ij}) + \sum_{j=1}^{p}\sum_{k=1}^{p} h_{jk}(x_{ik}, x_{ij}) \\ g(\sigma_i^2) &= \sum_{j=1}^{p} k_j(x_{ij}) + \sum_{j=1}^{p}\sum_{k=1}^{p} k_{jk}(x_{ik}, x_{ij}). \end{aligned} \tag{2}$$

The functions $h_j(\cdot)$ and $k_j(\cdot)$ were linear functions or univariate penalised regression splines with 10 knots. The functions $h_{ij}(\cdot)$ and $k_{ij}(\cdot)$ were zero functions or two-dimensional penalised regression splines with 10 knots selected out of 100 candidates. The estimation of the smoothing parameters of the different terms was accomplished using generalised cross-validation criteria. The non-zero terms of the models (about 50 in the mean models and 15 in the variance models) were selected using a simple algorithm, which expands the model by adding terms that improve the model's performance significantly in a validation data set.

In local methods, the whole set of estimation data serves as the model, and prediction is based on the nearest neighbours of the query point. Local linear regression was proposed for joint modelling of mean and dispersion by [Ruppert *et al.*, 1997]. [Yu and Jones, 2004] improved the method by proposing that the variance is estimated by minimising the local gamma likelihood instead of the sum of squares. They also used a link function $g(t) = \log(t)$ for variance in local estimation, leading to

$$\begin{aligned} \widehat{\mu}_i &= \widehat{a} \\ (\widehat{a}, \widehat{\beta}) &= \arg\min_{a,\beta} \sum_{j=1}^{N} (y_j - a - (x_j - x_i)^{\mathrm{T}}\beta)^2 K_1\Big(\frac{||x_j - x_i||}{h_1}\Big) \\ \widehat{\sigma}_i^2 &= g^{-1}(\widehat{c}) \\ (\widehat{c}, \widehat{\tau}) &= \arg\min_{c,\tau} \sum_{j=1}^{N} \Big[\frac{\varepsilon_j^2}{g^{-1}(c + (x_j - x_i)^{\mathrm{T}}\tau)} + \log g^{-1}(c + (x_j - x_i)^{\mathrm{T}}\tau)\Big] \\ &\qquad \cdot K_2\Big(\frac{||x_j - x_i||}{h_2}\Big). \end{aligned} \tag{3}$$

Here, K_1 and K_2 are kernel functions and the bandwidths h_1 and h_2 are chosen independently. The suitability of local methods to high-dimensional problems has been questioned, because the distances between the neighbouring points grow rapidly with the number of dimensions and the local neighbourhood becomes too sparse [Hastie *et al.*, 2001]. We used the local likelihood method of [Yu and Jones, 2004] with the Epanechnikov quadratic kernel $K_\lambda(x_0, x) = \frac{3}{4}(1 - |x - x_0|/\lambda)^2 I(|x - x_0| < 1)$. A simple adaptive bandwidth, which gives positive weights to a constant number (few thousands) of estimation data instances, was used. The model selection task was simplified to the selection of a suitable number of neighbours to be used in prediction, which was decided on the basis of performance in validation data.

Neural networks are known as a flexible modelling method with good predictive performance in large data sets [Hastie *et al.*, 2001]. We fitted neural network models for mean and dispersion. The idea is not completely new, see [Myllykoski, 1998]. We used single-layer perceptron model with skip-layer connections of the form

$$\begin{aligned} f(\mu_i) &= x_i^{\mathrm{T}}\beta + \sum_{j=1}^{h} f_j(x_i^{\mathrm{T}}\beta_j) \\ g(\sigma_i^2) &= x_i^{\mathrm{T}}\tau + \sum_{j=1}^{h} g_j(x_i^{\mathrm{T}}\tau_j) \end{aligned} \tag{4}$$

where the activation functions $f_j(\cdot)$ and $g_j(\cdot)$ are logistic $e^{-t}/(1+e^{-t})$. We fitted the variance model by maximising the penalised gamma log-likelihood related to squared residuals of the mean model. Model selection consisted of selecting the number of hidden neurons h and selecting the smoothing parameter. Different models were tested and the model that worked best in the validation data was selected. We modelled variance using single-layer perceptrons with 10 and 15 hidden neurons.

4 Results

We compared the prediction accuracy of joint modelling of mean and dispersion using the negative log-likelihood in the test data set T

$$\text{-log-lik} = \frac{1}{2}\sum_{i\in T} \ln 2\pi\widehat{\sigma}_i^2 + \frac{1}{2}\sum_{i\in T}\frac{(y_i-\widehat{\mu}_i)^2}{\widehat{\sigma}_i^2}. \tag{5}$$

It can be easily seen that the gamma log-likelihood of the dispersion model is equivalent to the likelihood of the whole model when the mean model is kept fixed. Thus, the comparison of dispersion models by keeping the mean model fixed can be based on the full likelihood. For the comparison of mean models, the root mean squared errors rMSE $= \sqrt{\mathrm{ave}(\widehat{\epsilon}^2)}$ are also presented.

Table 1 shows the achieved prediction accuracies of the different methods for joint modelling of mean and dispersion in the test data set. To compare especially the dispersion models, we fixed the mean models to the fitted neural network models and fitted the dispersion models using the squared residuals. The results are presented in Table 2.

The basic method for fitting the dispersion model was to use the response $\varepsilon_i^2/(1-\mathit{\Delta}_i)$ and the square root link function and to fit the model using gamma likelihood without iterating the mean model and variance model estimation. Some alternatives for the basic setting were tested: effects are

model	Tensile strength		Yield strength	
	rMSE	-log-lik	rMSE	-log-lik
HetReg	9.25	95125	14.39	108399
MADAM	9.67	95837	14.28	108172
LLRMD	9.23	95468	14.09	107800
NNMMD	8.95	94442	13.90	107482

Table 1. Prediction accuracy in the test data set.

model	Tensile S.	Yield S.
HetReg	94410	107646
MADAM	94726	107623
LLRMD	94593	107514
NNMMD	94442	107482

Table 2. The negative log-likelihoods (the smaller, the better) in the test data set when the mean model was kept fixed.

model	Tensile strength				Yield strength			
	ε^2	gaussian	log-link	weighted	ε^2	gaussian	log-link	weighted
HetReg	0	-56	-24	+61	0	-303	-6	+187
MADAM	-36	-2050	-375	+117	+12	-643	+13	-665
LLRMD	-80	-68	·	·	-27	-73	·	·
NNMMD	·	-350	+30	+251	·	-230	-185	-211

Table 3. The differences in test data log-likelihood between the standard fitting method and the alternatives. The plus sign means that the alternative gave better likelihood in the reduced test data set.

presented in Table 3. Using the response e^2 had only a small effect on the results; prediction accuracy usually decreased. If the parameters were estimated under gaussian likelihood instead of gamma likelihood, the likelihood of the test data decreased significantly. The effect of a link function was moderate, in most cases log-link for the variance function gave worse results. The number of iterations in the joint modelling of mean and dispersion had a major but fluctuating effect on the results. Usually, the weighted estimation of the second iteration gave better results when measured using likelihood but worse results when rMSE was used. The differences in rMSE were 0, -0.10 and -0.02 for tensile strength and +0.04, -0.57 and -0.10 for yield strength (in the same order as in Table 3). The third iteration changed the results of the second iteration only slightly, and the differences in log-likelihood were about 10-20. The subsequent iterations had a very small effect on the results, the change in log-likelihood being about 1-4.

In neural network modelling, it was noticeable that a network with skip-layer connections was much better than an ordinary single-layer perceptron without skip-layer connections. For yield strength the difference in log-likelihood was 800 and for tensile strength 1300. The use of log-link for variance with local likelihood fitting caused convergence problems at several prediction points, and log-link could thus not be used. Constant bandwidth seemed to work poorly; the difference in log-likelihood with the adaptive bandwidth was about 2000. We did not try the weighted version of local linear modelling, because too large computations would have been needed.

The computational requirements of modelling methods are a focus of interest when prediction is based on large data sets. We tested the computational needs using R software (`http://www.r-project.org/`) installed on a SunOS unix machine with 15 Gb of memory. The CPU power used in the computation was 900 MHz. R is known to be fast but to use memory inefficiently. The observed need for memory and computation time for fitting the model for strength are shown in Table 4. The time needed to produce 25000 predictions for the test data set is also presented. We used a simple model selection algorithm for each case; the approximate computation times used by the model selection procedures are also presented in Table 4.

	Fitting	Prediction	Model selection	Memory need (Mb)
HetReg	1 min	< 1 min	15 h	800 Mb
MADAM	70 min	< 1 min	12 h	3500 Mb
LLRMD	70 h	20 h	240 h	400 Mb
NNMMD	120 min	< 1 min	10 h	400 Mb

Table 4. The required computational resources for applying different methods to the strength of steel data.

5 Discussion

The results on the predictive performance of the models in predicting the distribution of the strength of steel plates are presented. This is the first extensive comparison of the methods for joint modelling of mean dispersion in a real prediction problem.

Modification of the response in dispersion model fitting with Δ-corrections to take into account the effect of estimating the mean model has a small effect on prediction. In heteroscedastic regression with a large number of observations, Δ-corrections have practically no impact, but the effect increases with the complexity of the model. We suggest that good results are obtained with an uncorrected response, but if the Δ-corrections are easily available, the corrected response should be used.

The traditional log-link ensures the positivity of predicted variance, but it did not perform very well in our case study. Log-link implies that the explanatory variables have a multiplicative effect on variance, which is not necessarily a rational assumption. We suggest that a linear model for variance and a linear model for deviation should be also considered when selecting link function.

Iteration of mean model estimation and variance model estimation increases the computation time needed for model fitting. Our results agree well with the earlier results claiming that two iterations are needed, and the subsequent iterations have only a minor effect on the results. In our data set, the first iteration also gave pretty good results. Our suggestion is to use two iterations. We compared two loss functions in variance function estimation; least squares and gamma log-likelihood. Least squares yielded poor results, which was expected, as the distribution of squared residuals is far from normal.

A wide variety of learning methods can be used for modelling dispersion, and the choice of the model type has a great influence on the accuracy of the prediction. The results suggest that neural networks are included among the methods that provide a suitable model framework for joint prediction of mean and dispersion based on large data sets. The fitting of additive spline models to large data sets requires a huge amount of memory, which makes them difficult to use. Local linear modelling is time-consuming, and it may not be applicable to real-time applications. Heteroscedastic regression models are appropriate when simplicity and interpretability are required.

References

[Carroll and Ruppert, 1988]R.J. Carroll and D. Ruppert. *Transformation and Weighting in Regression*. Chapman and Hall, New York, 1988.

[Hastie *et al.*, 2001]T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. Springer, New York, 2001.

[Myllykoski, 1998]P Myllykoski. A study on the causes of deviation in mechanical properties of thin steel sheets. *Journal of Materials Processing Technology*, pages 9–13, 1998.

[Rigby and Stasinopoulos, 1996]R.A. Rigby and D.M. Stasinopoulos. A semi-parametric additive model for variance heterogeneity. *Statistics and Computing*, pages 57–65, 1996.

[Ruppert *et al.*, 1997]D. Ruppert, M.P. Wand, U. Holst, and O. Hössjer. Local polynomial variance-function estimation. *Technometrics*, pages 262–273, 1997.

[Smyth *et al.*, 2001]G.K. Smyth, A.V. Huele, and A.P. Verbyla. Exact and approximate reml for heteroscedastic regression. *Statistical Modelling*, pages 161–175, 2001.

[Yau and Kohn, 2003]P Yau and R Kohn. Estimation and variable selection in non-parametric heteroscedastic regression. *Statistics and Computing*, pages 191–208, 2003.

[Yu and Jones, 2004]K. Yu and M.C. Jones. Likelihood-based local linear estimation of the conditional variance function. *Journal of the American Statistical Association*, pages 139–144, 2004.

Authors Index