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 $\aleph := \{X_1, \ldots, X_m\}$ 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

$$i(\mathbb{X}) := \begin{cases} 0, & \text{if } \mathbb{X} = \emptyset, \\ I(X_{i_1}; \ldots; X_{i_r}), & \text{if } \mathbb{X} = \{X_{i_1}, \ldots, X_{i_r}\}. \end{cases}$$

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 $X_{i_1}, \ldots, X_{i_r}$ can be rewritten as

$$\begin{aligned} H(p_{\overrightarrow{\mathbb{X}}}) = & \sum_{X_j \in \mathbb{X}} H(p_{X_j}) - \sum_{\{X_j, X_k\} \subseteq \mathbb{X}} I(X_j; X_k) \\ & + \sum_{\{X_j, X_k, X_l\} \subseteq \mathbb{X}} I(X_j; X_k; X_l) - \cdots + (-1)^{r+1} I(X_{i_1}; \ldots; X_{i_r}). \end{aligned}$$

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 $k \in \{1, \ldots, m\}$, 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

$$I^{(2)}(\overrightarrow{\mathbb{X}}; \overrightarrow{\mathbb{Y}}) = \sum_{X \in \mathbb{X}} \sum_{Y \in \mathbb{Y}} I(X;Y) \qquad \text{and}$$

$$I^{(3)}(\overrightarrow{\mathbb{X}}; \overrightarrow{\mathbb{Y}}) = I^{(2)}(\overrightarrow{\mathbb{X}}; \overrightarrow{\mathbb{Y}}) - \sum_{X \in \mathbb{X}} \sum_{\{Y_1, Y_2\} \subseteq \mathbb{Y}} I(X; Y_1; Y_2) - \sum_{\{X_1, X_2\} \subseteq \mathbb{X}} \sum_{Y \in \mathbb{Y}} I(X_1; X_2; Y).$$

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 $\{x_1, \ldots, x_r\}$ and $\{y_1, \ldots, y_s\}$. 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.