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.

$$\max_{\beta\in L_2[0,1]} \frac{\mathbb{V}(\mathbb{E}(\Phi(X)|Y))}{\mathbb{V}(\Phi(X))}. \tag{1}$$

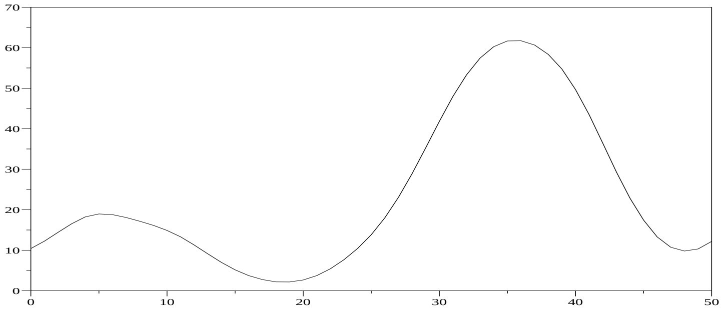


Fig. 1. Knee angular rotation over a complete gait cycle for one subject.

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 $C(t,s) = \text{cov}(X_t, X_s)$, $\forall t, s \in [0,1]$. 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

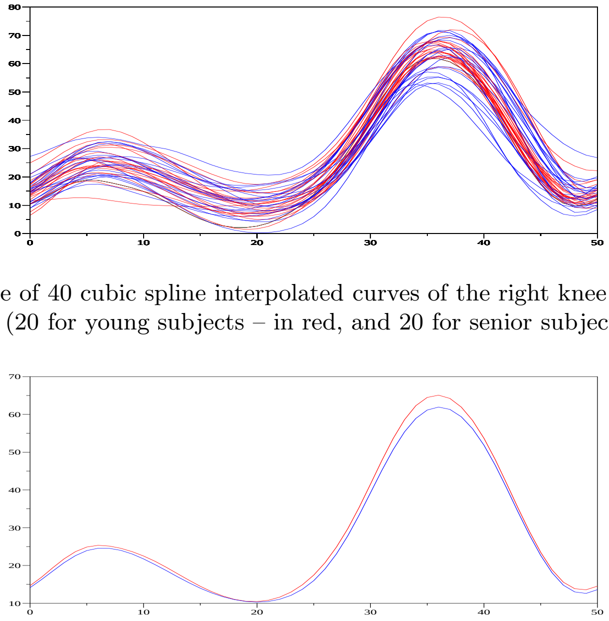
$$\mathbb{E}(YZ_t) = \int_0^1 \mathbb{E}(Z_tZ_s)\beta(s)ds, \tag{10}$$

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.

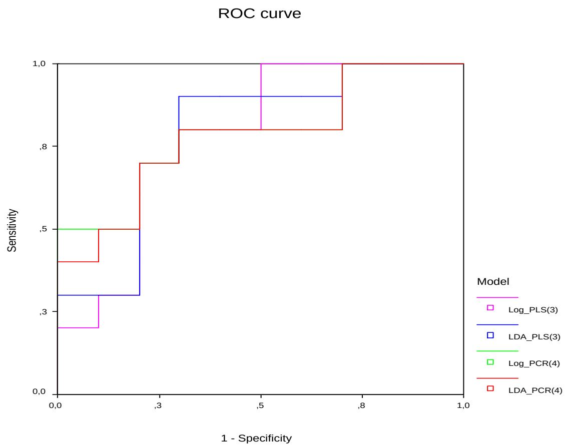


Fig. 3. ROC curves for each discriminant function.

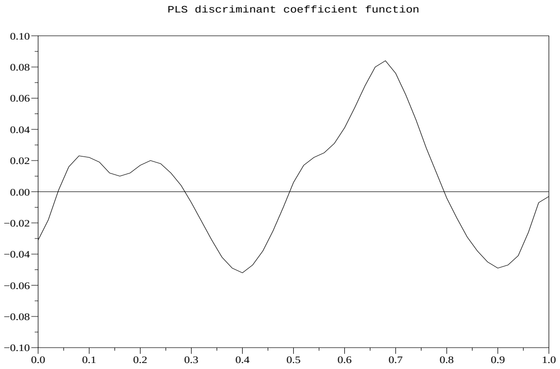


Fig. 4. Discriminant coefficient function $\hat{\beta}_{PLS(3)}$ for LDA_PLS(3)

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.