PLS classification of functional data

Cristian Preda[1], Gilbert Saporta[2] and Caroline Lévéder[3]

[1] CERIM - Département de Statistique, Faculté de Médecine, Université de Lille 2, 1, Place de Verdun, 59045 Lille, France (e-mail : cpreda@univ-lille2.fr)
[2] Chaire de Statistique Appliquée, CEDRIC, CNAM, 292, Rue Saint Martin, 75141 Paris Cedex 03, France (e-mail : saporta@cnam.fr)
[3] Danone Vitapole, 128 Route Départementale, 91767 Palaiseau Cedex, France (e-mail : caroline.leveder@danone.com)

Summary

Linear discriminant analysis is studied when the predictors are data of functional type (curves). Due to the infinite dimension of the predictor space, discriminant coefficient functions cannot be derived as in the classical setting. Partial least squares approach is then used in this context. Results of an application to kneading data are presented.

Keywords: PLS regression, Functional data, Linear discriminant analysis.

1 Introduction

Statistical methods for data representing functions or curves have received much attention in recent years and classical tools from the finite multivariate

data analysis are adapted to functional data (see [Ramsay and Silverman, 2002]).

Examples of functional data can be found in several application domains such as medicine, economics, chemometrics and many others (for an overview, see [Ramsay and Silverman, 2002]). 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 H of functions on some set T. For example, a second order stochastic process $X = \{X_t\}_{t\in[0,1]}$, L_2–continuous with sample paths in $L_2([0,1])$ can be used as model for kneading curves describing the resistance of dough during the kneading process (Figure 1).

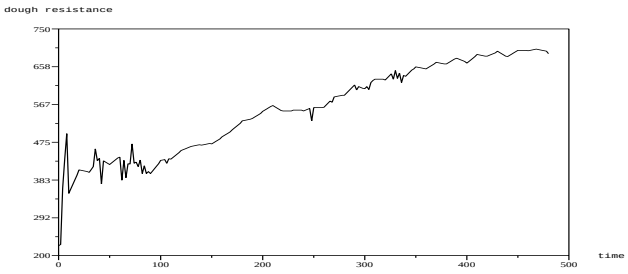

Figure 1: Functional data example: resistance of dough during the kneading process (observed on 480 seconds)

A major interest when dealing with functional data is to develop regression models. One of the difficulties with such data when used as predictor is his multicolinearity feature (the X_t's are highly correlated), the covariance operator being, in general, degenerated. Also, from practical point of view, the number of observations is in general less than the number of predictors (which is theoretically infinite) situation which leads to inconsistency of the estimated model. Depending on the nature of the response variable, several models are developed in literature. Thus, generalized linear regression models are developed in [James, 2002] and more recently by [Cardot and Sarda, 2005]. Different linear approaches based on decomposition of the underlying stochastic process are proposed : principal component regression ([Aguilera *et al.*, 1997], [Cardot *et al.*, 1999]), partial least squares regression (PLS)([Preda and Saporta, 2002]). [Ferraty and Vieu, 2004] and [Preda, 2005] propose non-parametric models for regression on functional data using classical kernel estimators both for scalar and categorical response.

In this paper we are interested in linear discrimination analysis when the predictor X is data of functional type (generally, curves or real functions of time) and the response is a categorical variable Y with K modalities. As an extension of the classical multivariate approach, the aim of linear discriminant analysis (LDA) for functional data is to find linear combinations $\Phi(X) = \int_0^T X_t\beta(t)dt$, $\beta \in L_2([0,T])$ such that the between class variance is

maximized with respect to the total variance, i.e.

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

Let $\{(x_i, y_i)\}_{i=1,\ldots,n}$ be n observations of random variables (X, Y) with $x_i = \{x_i(t), t \in [0,T]\}$ and $y_i \in \{1, \ldots, K\}$, $i = 1, \ldots, n$. Due to infinite dimension of the predictor, the estimation of β is in general an ill–posed problem. For $K = 2$ is well known that the optimization problem (1) is equivalent to find the regression coefficients 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 functional data and propose solutions to overcome this difficulty. Non-parametric 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 is given in [Escabias *et al.*, 2005]. Several applications on real date measure the importance of these models : [Ratcliffe *et al.*, 2002] develops functional regression models for foetal heart data in order to predict the probability of high risk birth outcome, [Ferraty and Vieu, 2003] are interested in classification curves from spectrometric data, [Escabias *et al.*, 2005] develop logistic models for environmental data, etc.

We propose to use PLS regression approach [Preda and Saporta, 2002] in the context of linear discriminant analysis on functional data. The paper is organized as follows. In section 2 we recall some results on PLS regression on functional data. The relationship between linear discriminant analysis (LDA) for binary response and linear regression as well as the PLS methodology in the general case is presented in section 3. An application of PLS classification approach to kneading data is realized in section 4. The aim of this application is to predict, for a given flour, the quality of cookies from a curve recorded during the kneading process (see [Lévéder *et al.*, 2004] for other classification approaches used on the same data).

2 PLS regression on functional data

Let $X = \{X_t\}_{t \in [0,1]}$ be a second order stochastic process L_2-continuous with sample paths in $L_2[0,1]$ and Y a real random variable. Without loss of generality we assume also that $\mathbb{E}(X_t) = 0$, $\forall t \in [0,1]$ and $\mathbb{E}(Y) = 0$.

The partial least squares (PLS) approach offers a good alternative to linear regression on principal components (PCR) 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 $\mathbf{W}_{q-1}^X \mathbf{W}_{q-1}^Y$, where $\mathbf{W}_{q-1}^X$, respectively $\mathbf{W}_{q-1}^Y$, are the Escoufier's operators ([Saporta, 1981]) 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 PLS approximation of Y by $\{X_t\}_{t\in[0,1]}$ at step q, $q \geq 1$, is given by :

$$\hat{Y}_{PLS(q)} = c_1 t_1 + \ldots + c_q t_q = \int_0^1 \hat{\beta}_{PLS(q)}(t) X_t dt. \tag{2}$$

In practice, the number of PLS components used for regression is determined by cross-validation ([Tenenhaus, 2002]).

3 LDA and PLS regression for functional data

3.1 The discriminant coefficient function

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 $f \overset{\mathbf{C}}{\longmapsto} g$, $g(t) = \int_0^1 \mathbb{E}(X_t X_s) f(s) ds$, and by $\mathbf{B}$ the operator on $L_2[0,1]$ defined by $f \overset{\mathbf{B}}{\longmapsto} g$, $g(t) = \int_0^1 B(t,s) f(s) ds$, 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

$$\mathbf{B}\beta = \lambda\mathbf{C}\beta, \quad \text{with } \langle\beta, \mathbf{C}\beta\rangle_{L_2[0,1]} = 1. \tag{3}$$

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

$$\mathbb{E}(Y\, Z_t) = \int_0^1 \mathbb{E}(Z_t\, Z_s)\beta(s)ds, \tag{4}$$

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 (4) 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 PLS approach proposed in Section 2 can be used to compute the discriminant score for a new observation for which one has only the observation of X.

3.2 PLS classification methodology

Let us consider two cases :

Case 1. $K = 2$. Let $Y \in \{0,1\}$. Then, the discriminant function β is the coefficient function of the linear regression of Y on $X = \{X_t\}_{t\in[0,T}$, where Y is encoded by $0 \rightsquigarrow \sqrt{\frac{p_1}{p_0}}$ and $1 \rightsquigarrow -\sqrt{\frac{p_0}{p_1}}$, with (p_0, p_1) the probability distribution of Y. The PLS1 regression of Y on X provides an approximation for the discriminant variable (score) given by $\Phi_{PLS}(X) = \alpha + \int_0^T X_t \hat{\beta}_{PLS}(t)dt$, where $\alpha = -\int_0^T \hat{\beta}_{PLS}(t)\mu(t)dt$ and $\mu(t) = \mathbb{E}(X_t), t \in [0,T]$.

Case 2. $K > 2$. Let $Y \in \{0, \ldots, K\}$ and $\{Y_i\}_{i=1,\ldots,K-1}$ be the dummy variables associated to Y, $Y_i \in \{0,1\}$. We propose to perform linear discriminant analysis of Y and X using the PLS components $\{t_j\}_{j=1,\ldots,q}$ of the PLS2 regression of the vector $\mathbf{Y} = \{Y_1, \ldots, Y_{K-1}\}$ on X, where q is determined by cross validation. The j^{th} PLS component is the eigenvector associated to the largest eigenvalues of the operator $W^X W^{\mathbf{Y}^{j-1}}$, where W^X and $W^{\mathbf{Y}_{j-1}}$ are the Escoufier operators of X, respectively of $\mathbf{Y}^{j-1}$, the residual of $\mathbf{Y}$ at the $j-1$ iteration of the PLS2 regression on X (see [Preda and Saporta, 2002]). One obtains, for each category i of Y, the coefficient discriminant function β^i_{PLS} which allow to compute the associated score, $\Phi^i_{PLS}(X) = \alpha_i + \int_0^T X_t \beta^i_{PLS}(t)dt$, with $\alpha_i \in \mathbb{R}, i = 1, \ldots, K$.

Therefore, given a new observation of X, prediction for Y is made in the classical way. The predictive capacity of the method is then measured on a test sample using either the misclassification rate or the area under the ROC curve (for $K = 2$).

4 Application

PLS approach is applied to predict the quality of cookies from the kneading curve representing the resistance (density) of dough observed during the kneading process. For a given flour, the kneading process is observed during 480 seconds. Since we have 115 different flours we have 115 curves (functions of time), which represent a set of sample paths of the stochastic process $X = \{X_t, t \in [0, 480]\}$. Each curve is observed in a finite number of points corresponding to a discretization of $[0, 480]$ into 240 equispaced instants of time (the same for all flours). Thus, a kneading curve is represented by the set of 241 points $\{(t_i, X_{t_i}), i = 0, \ldots, 240\}$. After kneading, the dough is processed to obtain cookies. For each flour we have the quality (Y) of cookies which can be *Good, Adjustable* and *Bad*. Our sample contains 50 observations for $Y = Good$, 25 observations for $Y = Adjustable$ and 40 for $Y = Bad$.

Due to measuring errors, each curve is smoothed using cubic B-spline functions with the following 16 knots in the interval $[0, 480]$ ([Lévéder *et al.*, 2004]) : $\{10, 42, 84, 88, 108, 134, 148, 200, 216, 284, 286, 328, 334, 380, 388, 478\}$.

Let us consider $Y \in \{Good, Bad\}$. The sample of 90 flours is randomly divided into a learning sample of size 60 and a test sample of size 30. In the test sample the two classes have the same number of observations.

Table 1 presents the error classification rates averaged over 100 test samples using linear discriminant analysis on principal components (LDA_FPCR) respectively the PLS approach (LDA_FPLS) for which we developed software. The number of PLS and PCR components is given by cross validation using the leave-one-out procedure.

Model	**Test error rate**
LDA_FPCR	0.142
LDA_FPLS	0.112

Table 1: Error rate averaged over 100 test samples.

It is of interest to study the set of observations for which $Y = Adjustable$ because, with a small effort, certain of these flours could be *adjusted* to become *Good*. Therefore, it is important to say if a flour with quality *Adjustable* is closer to *Good* flours or to *Bad* ones. For this, we consider the set of *Adjustable* flours as 'test sample' and predict their closeness to *Good* and *Bad* categories using the discriminant coefficient function (Figure 2) given by the PLS approach on the 90 flours (observations form *Good* and *Bad* classes). The discriminant score is $\Phi(X) = -1.46 + \int_0^{480} X_t \hat{\beta}_{PLS}(t)dt$. For a given x, if $\Phi(x) > 0$ then $\hat{Y} = Good$, else $\hat{Y} = Bad$.

One obtains for the 25 *Adjustable* flours, 12 for which the predictions give them closer to class *Good* and 13 closer to class *Bad*.

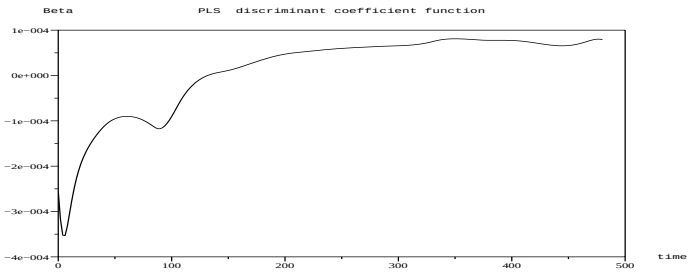

Figure 2: Discriminant coefficient function $\hat{\beta}_{PLS}$ given by PLS approach

For $Y \in \{Good, Adjustable, Bad\}$, the results given by the multivariate PLS discrimination model (see section 3.1, Case 2) are presented in Table 2. We compare these results with those obtained by linear discrimination on the principal components (LDA_FPCR) and on the original smoothed data (LDA) as well as with the K-NN procedure with 13 neighbors ([Lévéder *et al.*, 2004]. Using the 115 flours, the comparison criterion is we averaged error classification rate over 100 test samples of size 35.

Model	LDA_FPLS	K-NN (13)	LDA_FPCR	LDA
Misclassified rate	0.258	0.245	0.262	0.282

Table 2: Misclassified rate for three categories : average over 100 test samples.

5 Conclusion

PLS regression on functional data is used for linear discriminant analysis with binary response. It is a simple and interesting alternative to classical linear methods based on principal components of predictors and his efficiency is confirmed by an example on kneading data.

References

[Aguilera *et al.*, 1997] A. M. Aguilera, F. Ocaña, and M.J. Valderama. An approximated principal component prediction model for continuous-time stochastic process. *Applied Stochastic Models and Data Analysis*, pages 61–72, 1997.

[Biau *et al.*, 2004] G. Biau, F. Bunea, and M. Wegkamp. Function classification in Hilbert spaces. *Submitted*, 2004.

[Cardot and Sarda, 2005] H. Cardot and P. Sarda. Estimation in generalized linear models for functional data via penalized likelihood. *Journal of Multivariate Analysis*, pages 24–41, 2005.

[Cardot *et al.*, 1999] H. Cardot, F. Ferraty, and P. Sarda. Functional linear model. *Statist. Probab. Lett.*, pages 11–22, 1999.

[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.

[Ferraty and Vieu, 2003] F. Ferraty and P. Vieu. Curves discrimination: a nonparametric approach. *Computational Statistics & Data Analysis*, pages 161–173, 2003.

[Ferraty and Vieu, 2004] F. Ferraty and P. Vieu. Nonparametric models for functional data with application in regression, time series prediction and curve discrimination. *Journal of Nonparametric Statistics*, pages 111–125, 2004.

[James, 2002] G. M. James. Generalized linear models with functional predictors. *Journal of the Royal Statistical Society, Series B*, pages 411–432, 2002.

[Lévéder *et al.*, 2004] C. Lévéder, C. Abraham, P. A. Cornillon, E. Matzner-Lober, and N. Molinari. Discrimination de courbes de pétrissage. *Chimiométrie 2004, Paris*, pages 37–43, 2004.

[Preda and Saporta, 2002] C. Preda and G. Saporta. Régression PLS sur un processus stochastique. *Revue de Statistique Appliquée*, pages 27–45, 2002.

[Preda, 2005] C. Preda. Regression models for functional data by reproducing kernel hilbert spaces methods. *Journal of Statistical Planning and Inference*, 2005.

[Ramsay and Silverman, 2002] J. O. Ramsay and B.W. Silverman. *Applied Functional Data Analysis : Methods and Case Studies.* Springer, 2002.

[Ratcliffe *et al.*, 2002] S. J. Ratcliffe, L.R. Leader, and Heller G.Z. Functional data analysis with application to periodically stimulated foetal heart rate data. ii : Functional logistic regression. *Statistics and Medicine*, pages 1115–1127, 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, 2002] M. Tenenhaus. *La régression PLS. Théorie et pratique.* Editions Technip, 2002.