Anticipated and Adaptive Prediction in Functional Discriminant Analysis

Cristian Preda[1], Gilbert Saporta[2], and Mohamed Hadj Mbarek[3]

[1] Ecole Polytehnique Universitaire de Lille& Laboratoire Painlevé, UMR 8524
Université des Sciences et Technologies de Lille, France,
cristian.preda@polytech-lille.fr
[2] Chaire de statistique appliquée & CEDRIC, CNAM
292 rue Saint Martin, Paris, France, *gilbert.saporta@cnam.fr*
[3] Institut Supérieur de Gestion de Sousse, Tunisie, *benmbarekmhedi@yahoo.fr*

Abstract. Linear discriminant analysis with binary response is considered when the predictor is a functional random variable $X = \{X_t, t \in [0,T]\}, T \in \mathbb{R}$. Motivated by a food industry problem, we develop a methodology to anticipate the prediction by determining the smallest T^*, $T^* \leq T$, such that $X^* = \{X_t, t \in [0,T^*]\}$ and X give similar predictions. The adaptive prediction concerns the observation of a new curve ω on $[0,T^*(\omega)]$ instead of $[0,T]$ and answers to the question "How long should we observe ω ($T^*(\omega) =?$) for having the same prediction as on $[0,T]$?". We answer to this question by defining a conservation measure with respect to the class the new curve is predicted.

Keywords: functional data, discriminant analysis, classification, adaptive prediction

1 Introduction

Statistical methods for data represented by curves (or functions) have received much interest in the last years. Random variables taking values into an infinite dimensional function space are called *functional random variables* (Ferraty and Vieu (2006)) and methods dealing with such variables define the *functional data analysis* (FDA) framework (Ramsay and Silverman (1997)). Examples of functional data can be found in several application domains such as medicine (patient evolution over time), economics (stock-exchange data), chemometrics (spectrometric data) and many others (for an overview, see Ramsay and Silverman (2002)).

A well accepted model for univariate functional data is to consider it as paths of a stochastic process $\mathbf{X} = \{X_t\}_{t\in\mathcal{T}}$ taking values into a Hilbert space of real-valued functions defined over some set $\mathcal{T}$. For example, if $\mathcal{T} = [0,T]$ with $T \in \mathbb{R}_+$, a second order stochastic process $\mathbf{X} = \{X_t\}_{t\in[0,T]}$ L_2– continuous with sample paths in $L_2([0,T])$ can be used as model for describing the behavior of some quantitative parameter associated to a process observed on a time interval of length T.

Suppose that for each curve we have a single response variable Y. If Y is categorical we have a classification (or discrimination) problem and a regression one if Y is numerical.

In this paper we assume that Y is a binary response and all trajectories of X are observed continuously on $[0,T]$ and belong to $L_2([0,T])$. The main purpose of discriminant analysis (supervised classification) is to define a discriminant score $\Phi(X)$, $\Phi : L_2([0,T]) \to \mathbb{R}$, such that the prediction of Y using $\Phi(X)$ is as good as possible.

The linear discriminant analysis for functional data considers that

$$\Phi(X) = \int_0^T X_t \beta(t) dt, \quad \beta \in L_2([0,T]),$$

and has been addressed by James and Hastie (2001) and Preda et al. (2007). This problem is not new and comes back to Fisher (1924) who used the expression *integral regression*. It is well known that the estimation of this regression model by least squares criterion yields to an ill-posed problem. Regularization techniques such as principal component regression (PCR) and partial least squares regression (PLS) have been proposed in Preda and Saporta (2005).

An estimating procedure of the functional logistic model is proposed by Escabias et al. (2004, 2005) with environmental applications. Nonparametric models have been proposed by Ferraty and Vieu (2003), Biau et al. (2005) and Preda (2007). More details on nonparametric models for functional data can be found in the recent monograph of Ferraty and Vieu (2006).

In this paper we are firstly interested to the problem of *anticipated prediction* : find minimal T^*, $T^* < T$, such that the predictor X observed on $[0,T^*]$ gives "similar results"', in terms of prediction of Y, as considered on $[0,T]$. By "similar results" we mean, for example, not significantly different results with respect to some statistical test. This problem comes naturally from the following practical example : in Preda et al. (2007) we have developed functional discriminant models to predict the quality of cookies at Danone from curves representing the resistance of dough observed during the first 8 minutes of the kneading process (Lévéder et al. (2004)). The discriminant power of the linear model is satisfactory with a misclassified rate of about 11.2%. Then, the interest of reducing the observation time and take decision keeping the same discriminant power is evident.

Secondly, we address the problem of the prediction of Y from X for a new observation ω in an *adaptive* way. Usually, the new curve X_ω is observed on the whole interval $[0,T]$ and then the prediction is made using the score $\Phi(X_\omega)$. Provided the existence of a good prediction model of Y from X, in adaptive prediction we are interested to determine a time $T^*(\omega)$ such that the prediction of Y from the observation X on $[0,T^*\omega)]$ is similar to the prediction with X on $[0,T]$. In other words, to observe X after $T^*(\omega)$ will no change the prediction.

The paper is organized as follows. In section 2 we present some basics of the PLS approach for linear discriminant analysis with functional data. The

anticipated and adapted prediction are introduced in Section 3. The Section 4 presents the results of the anticipated and adaptive prediction for the quality of cookies to Danone.

2 Linear discriminant analysis on functional data. The PLS approach

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

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

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 $y_i \in \{0,1\}$, $i = 1,\ldots,n$. Due to infinite dimension of the predictor, the estimation of β is in general an ill–posed problem. In Preda and Saporta (2005) it is shown that the optimization problem (13) is equivalent to find the regression coefficients in the linear model which predicts Y (after a convenient encoding) by the stochastic process X under the least-squares criterion.

Without loss of generality, let us recode Y by : $0 \rightsquigarrow \sqrt{\frac{p_1}{p_0}}$ and $1 \rightsquigarrow -\sqrt{\frac{p_0}{p_1}}$, where $p_0 = \mathbb{P}(Y = 0)$ and $p_1 = \mathbb{P}(Y = 1)$. If β is a solution of (1) then β satisfies the Wiener-Hopf equation

$$\mathbb{E}(YX_t) = \int_0^T \mathbb{E}(X_tX_s)\beta(s)ds, \tag{2}$$

which is the equation giving, up to a constant, the regression coefficient function of the linear regression of Y on $X = \{X_t\}_{t\in[0,T]}$. Equation (14) 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.

Using this result, there are several ways to approximate the discriminant score $\Phi(X)$. Thus, $\Phi(X)$ can be approximate using the linear regression on the principal components of X. The choice of principal components used for regression is not easy and should be a trade off between the quality of the model and the quality of the representation of X. The PLS approach proposed

in Preda and Saporta (2005) is an efficient alternative and provides generally better results. It allows to approximate $\Phi(X)$ by $\Phi_{PLS}(X) = \int_0^T \beta_{PLS}(t)X_t dt$ and thus, to compute for a new observation the discriminant score for further prediction.

2.1 The PLS approximation

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 $\{X_t\}_{t\in[0,T]}$ at step q, $q \geq 1$, is given by :

$$\hat{Y}_{PLS(q)} = c_1 t_1 + \cdots + c_q t_q = \int_0^T \hat{\beta}_{PLS(q)}(t) X_t dt. \tag{3}$$

In practice, the number of PLS components used for regression is determined by cross-validation.

2.2 Quality criterion. The ROC curve

Let denote by $d_T = \Phi_{PLS}(X) = \int_0^T \beta_{PLS}(t)X_t dt$ the approximation for the discriminant score given by the PLS regression on the process $X = \{X_t\}_{t\in[0,T]}$. There are several criteria to evaluate the quality of the discriminant model, for example the error rate for a defined threshold, the squared correlation ration $\eta^2(d_T|Y) = \dfrac{\mathbb{V}(\mathbb{E}(d_T|Y))}{\mathbb{V}(d_T)}$, the ROC curve, etc.

For a binary target Y, the ROC curve is generally accepted as the best measure of the discriminating power of a discriminant score.

Let $d_T(x)$ be the score value for some unit x. Given a threshold r, x is classified into $Y = 1$ if $d_T(x) > r$. The true positive rate or "sensitivity" is $P(d_T > r|Y = 1)$ and the false positive rate or $1 -$ "specificity", $P(d_T > r|Y = 0)$. The ROC curve gives the true positive rate as a function of the false positive rate and is invariant under any monotonic increasing transformation of the score. In the case of an inefficient score, both conditional distributions of d_T given $Y = 1$ and $Y = 0$ are identical and the ROC curve is the diagonal line. In case of perfect discrimination, the ROC curve is confounded with the edges of the unit square.

The Area Under ROC Curve or AUC, is then a global measure of discrimination. It can be easily proved that $AUC = P(X_1 > X_0)$, where X_1 is

a random variable distributed as d when $Y = 1$ and X_0 is independently distributed as d for $Y = 0$. Taking all pairs of observations, one in each group, AUC is thus estimated by the percentage of concordant pairs (Wilcoxon-Mann-Whitney statistic).

3 Anticipated and adaptive prediction

3.1 Anticipated prediction

Let denote by d_t the approximation for the discriminant score given by PLS regression on the process X considered on the interval time $[0, t]$, with $t \leq T$. The objective here is to find $T^* < T$ such that the discriminant function d_{T^*} performs quite as well as d_T.

The stochastic process $\{d_t\}_{t\in[0,T]}$ is such that :

- $d_t = Y - \varepsilon_t$, where Y is recoded by $0 \rightsquigarrow \sqrt{\frac{p_1}{p_0}}$ and $1 \rightsquigarrow -\sqrt{\frac{p_0}{p_1}}$. $\mathbb{E}(d_t) = 0$.
- $\mathbb{E}(\varepsilon_t, d_s) = 0$, $\forall s \leq t$,
- $\mathbb{E}(d_t d_s) = \mathbb{E}(d_s Y) = \sqrt{p_0 p_1}(\mathbb{E}(d_s|Y = 0) - \mathbb{E}(d_s|Y = 1))$, $\forall s \leq t$.

Once a quality measure Q_s is defined, a solution could be to define T^* as the smallest value of s such that Q_s is not significantly different from Q_T. Since Q_s and Q_T are dependent random variables, we will use a non parametric paired comparison test.

We will use in the following the AUC criterion for defining the quality of the discriminant model.

Since the distribution of AUC is not known, we will test the equality of $AUC(s)$ with $AUC(T)$, by using booststrap methodology: we resample M times the data, according to a stratified scheme in order to keep invariant the number of observations of each group. Let $AUC_m(s)$ and $AUC_m(T)$ be the resampled values of AUC for $m = 1$ to M, and δ_m their difference. Testing if $AUC(s) = AUC(T)$ is performed by using a paired t-test, or a Wilcoxon paired test, on the M values δ_m.

3.2 Adaptive prediction

Let $\Omega = \{\omega_1, \ldots, \omega_n\}$, $n \geq 1$, be a training sample and $\{(x_1, y_1), \ldots, (x_n, y_n)\}$ be the observation of (X, Y) on Ω, X being considered on $[0, T]$. Let also suppose that one has a good discriminant score d_T for the prediction of Y by $\{X\}_{t\in[0,T]}$ with respect to some criterion (misclassified rate, AUC, R^2, etc).

Let now consider a new data ω.

By *adaptive prediction* for ω we understand to find the smallest time $T^* = T^*(\omega)$ such that the prediction of $Y(\omega)$ on $[0, T^*]$ is similar to that on $[0, T]$. Let observe that T^* is here a random variable, whereas in the anticipated approach T^* is a constant.

Let h be the step of a convenient discretisation of $[0,T]$ and suppose that for ω the process X is observed until the time t, $t < T$. It is clear that the decision to continue the observation $X(\omega)$ at $t+h$ or to stop it ($T^* = t$) depends on the similarity of $X(\omega)$ with $x_1, \ldots, x_n$ with respect to the prediction of Y. We define this similarity at the time t in the following way :

Let d_t be a discriminant score for Y using only the observation interval $[0,t]$ and denote by $\hat{Y}_t(\omega)$ and $\hat{Y}_{t,i}, i = 1, \ldots, n$ the predictions for ω, respectively Ω, with respect to d_t.

Denote by

$$\Omega_\omega(t) = \{\omega_i \in \Omega | \hat{Y}_t(\omega) = \hat{Y}_{t,i}\} \text{ and } \overline{\Omega}_\omega(t) = \Omega - \Omega_\omega(t)$$

the class of elements having the same prediction as ω, respectively its complement with respect to Ω.

Let

$$p_{0|\Omega_\omega(t)} = \frac{\left|\{\omega' \in \Omega | \hat{Y}_T(\omega') = 0\} \cap \Omega_\omega(t)\}\right|}{|\Omega_\omega(t)|} \tag{4}$$

be the observed rate of elements in $\Omega_\omega(t)$ predicted in the class $Y = 0$ at the time T. Similarly, let $p_{1|\Omega_\omega(t)}$, $p_{0|\overline{\Omega}_\omega(t)}$and $p_{1|\overline{\Omega}_\omega(t)}$. Obviously,

$$p_{0|\Omega_\omega(t)} + p_{1|\Omega_\omega(t)} = 1 \text{ and } p_{0|\overline{\Omega}_\omega(t)} + p_{1|\overline{\Omega}_\omega(t)} = 1.$$

Let define $C_{\Omega_\omega(t)} = \max\{p_{0|\Omega_\omega(t)}, p_{1|\Omega_\omega(t)}\}$ and $C_{\overline{\Omega}_\omega(t)} = \max\{p_{0|\overline{\Omega_\omega}(t)}, p_{1|\overline{\Omega}_\omega(t)}\}$ the *conservation* rate of prediction group at the time t with respect to the time T for the elements of $\Omega_\omega(t)$, respectively of $\overline{\Omega}_\omega(t)$. As a global measure of conservation we consider

$$C_\Omega(\omega, t) = \min\{C_{\Omega_\omega(t)}, C_{\overline{\Omega}_\omega(t)}\}. \tag{5}$$

For each $t \in [0,T]$, $C_\Omega(\omega,t)$ is such that $0.5 \leq C_\Omega(\omega,t) \leq 1$ and $C_\Omega(\omega,T) = 1$.

Given a confidence conservation threshold $\gamma \in (0,1)$, e.g. $\gamma = 0.90$, we define the following rule :

Adaptive prediction rule for ω and t :

(1) if $C_\Omega(\omega,t) \geq \gamma$ then the observation of X for ω on the time interval $[0,t]$ is sufficient for the prediction of $Y(\omega)$. $\hat{Y}(\omega)$ is then the same as the prediction at time T of the subgroup of $\Omega_\omega(t)$ corresponding to $C_{\Omega_\omega(t)}$.
(2) if $C_\Omega(\omega,t) < \gamma$ then the observation process of X for ω should continue after t. Put $t = t + h$ and repeat the adaptive prediction procedure.

Then, $T^*(\omega)$ is the smallest t such that the condition (1) of the adaptive prediction rule is satisfied.

An important role in the proposed adaptive prediction methodology is the observation of the discriminant score process d_t, $t \geq 0$ for the new data ω. We propose two approaches to define $d_t(\omega)$.

(M1) by *completion* : Use a functional regression model with functional response as in Preda and Saporta (2005) or Lian (2007) and predict $X(\omega)$ on $[t, T]$. Then, by completion one obtains a trajectory $X(\omega)$ on $[0, T]$ for which the score $d_T(\omega)$ provides a prediction for Y. Put $d_t(\omega) = d_T(\omega)$.

(M2) *sequential* : Construct the discriminant score d_t progressively for each time t and predict Y using d_t.

The first approach uses two regression models : one for the completion step, which is sequently performed for each t, and a second one for discrimination. The discrimination model concerns the training sample Ω for which X is considered on the whole interval $[0, T]$. The second approach, M2, involves the estimation of several discriminant models, one for each considered time t. However, our intuition is that the error associated to M1 is greater than that given by M2 since the first approach cumulates errors from both models. Moreover, even for t close to T, the prediction error given by the regression model used in the completion step of the trajectories in the training sample (PRESS) could be important and thus misleading for the discrimination step. For these reasons, we used the sequential approach in our application on kneading data.

4 Application

We use the anticipated approach for kneading data from Danone. The quality (Y) of cookies produced by a set of 90 flours for which one knows the dough resistance (X) during the first 480 seconds of the kneading process is evaluated. One obtains 50 flours yielding to good quality of cookies and 40 to a bad one. Because of large local variation, the curves are smoothed using cubic B-spline basis. Figure 1 shows the set of the 90 flours before and after smoothing.

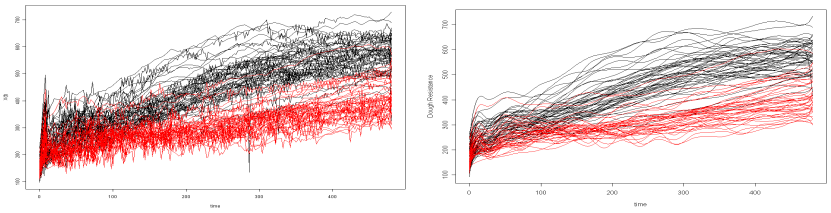

Fig. 1. Good (black) and bad (red) flours. Left : original data. Right : smoothed data

We use for prediction the smoothed curves that we consider as sample paths of a stochastic process $\{X_t\}_{t\in[0,480]}$. Considering $Y \in \{Bad, Good\}$, the PLS approach for discrimination of Y from $\{X_t\}_{t\in[0,480]}$ Preda et al. (2007) yields to a misclassification rate of about 11.2%. For a signification

level of 5% and using the AUC criterion, the anticipated approach provides $T^* = 186$.

Thus, the predictive power of the dough curves for the cookies quality is resumed by the first 186 seconds of the kneading process. In the next paragraph we extend the anticipated prediction methodology by adapting the optimal time T^* to each new trajectory given its incoming measurements, in that sense that observation of the trajectory after the time T^* does not change the prediction of Y.

Adaptive prediction 25 new flours have been tested for adaptive prediction. These flours were classified by Danone as being of quality "adjustable", somewhere between "good" and "bad". 12 of these flours are predicted by the PLS discriminant analysis using the interval time [0, 480] into the "good" class. Using as training sample Ω the set of the 90 flours considered in the anticipated prediction approach, we perform for each one the adaptive prediction starting from $t = 100$.

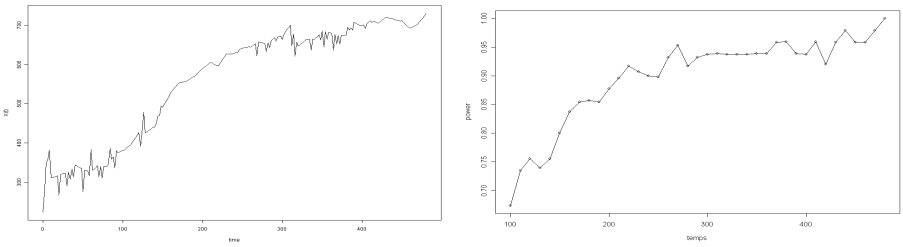

Fig. 2. Left : new flour ω. Right : $C_\Omega(\omega, t)$, $t \in [100, 480]$, $\gamma = 0.90$.

In Figure 3 (left), we present one of these flours (ω) which was observed on the whole interval $[0, 480]$. The conservation rate evolution $C_\Omega(\omega, t)$ is presented in Figure 3 (right) $t \in [100, 480]$. For a conservation rate threshold $\gamma = 0.90$, the adaptive prediction rule provides $T^*(\omega) = 220$ and predicts ω in the "good" class of flours.

The empirical cumulative distribution function of T^* obtained with the 25 flours is presented in Figure 4. Notice that there are 5 time points which are earlier than the optimal time for anticipated prediction ($T^* = 186$). 10 flours are predicted in the "good" class.

5 Conclusions

In this paper we addressed the problem of the prediction of a binary response Y using as predictor data of functional type represented by paths

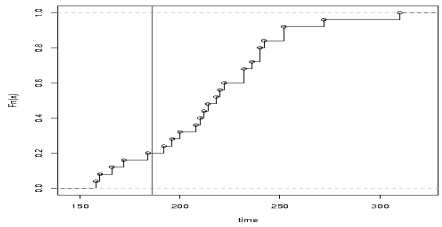

Fig. 3. Empirical cumulative distribution function of T^* (in red, the time point t=186).

of a continuous–time stochastic process $\{X_t\}_{t\in[0,T]}$. We faced the problem by means of the PLS approach for which forecasting the binary response is drawn as 'anticipated prediction' from the process $\{X_t\}_{t\in[0,T^*]}$ with $T^* < T$. Under the hypothesis of existence of an acceptable prediction model, we investigated the possibility of reducing the length of the observation period without loss of the quality prediction. We provided methodologies for anticipated and adaptive prediction for preserving the global quality model as well as the quality prediction of individual curves. An example is provided on kneading data from Danone.

References

BIAU, G., BUNEA, F. and WEGKAMP, M. (2005): Function classification in Hilbert spaces. *IEEE Transactions on Information Theory*, 51, 2162-2172.

ESCABIAS, M., AGUILERA, A.M. and VALDERAMA, M.J. (2004): Principal component estimation of functional logistic regression: discussion of two different approaches. *Journal of Nonparametric Statistics, 16 (3-4), 365-384.*

ESCABIAS, M., AGUILERA, A.M. and VALDERAMA, M.J. (2005): Modelling environmental data by functional principal component logistic regression. *Environmetrics, 16 (1), 95-107.*

FERRATY, F. and VIEU, P. (2006): *Nonparametric functional data analysis. Theory and practice*, Springer.

FERRATY, F. and VIEU, P. (2003): Curves discrimination: a nonparametric approach. *Computational Statistics & Data Analysis, 44, 161-173.*

FISHER, R.A. (1924): The Influence of Rainfall on the Yield of Wheat at Rothamsted. *Philosophical Transactions of the Royal Society, B 213, 89-142.*

FISHER, R.A. (1936):, The use of multiple measurement in taxonomic problems. *Ann. Eugen*, 7, 179-188.

JAMES, G.M. and HASTIE, T.J. (2001): Functional discriminant analysis for irregularly sampled curves. *Journal of the Royal Statistical Society, Series B, 63,533-550.*

LÉVÉDER, C., ABRAHAM, C., CORNILLON P. A., MATZNER-LOBER, E. and MOLINARI N. (2004): Discrimination de courbes de pétrissage. *Chimiométrie, p. 37–43.*

LIAN, H. (2007): Nonlinear functional models for functional responses in reproducing kernel Hilbert spaces. *The Canadian Journal of Statistics, 35, 597-606.*

PREDA, C. (2007): Regression models for functional data by reproducing kernel Hilbert space methods. *Journal of Statistical Planning and Inference*, Vol. 137, 3, p. 829-840.

PREDA, C. and SAPORTA, G. (2005): PLS regression on a stochastic process. *Computational Statistics and Data Analysis 48 (1), 149-158.*

RAMSAY, J.O. and SILVERMAN, B.W. (1997): *Functional Data Analysis*, Springer Series in Statistics, Springer-Verlag, New York.

RAMSAY, J.O. and SILVERMAN, B.W. (2002): *Applied Functional Data Analysis: Methods and Case Studies*, Springer.

SAPORTA, G. (1981): Méthodes exploratoires d'analyse de données temporelles. *Cahiers du B.U.R.O, Université Pierre et Marie Curie*, 37-38, Paris.