Adaptive Forecasting on Functional Data

Gilbert Saporta, Cristian Preda

Adaptive forecasting on functional data

Previsione adattiva per dati funzionali

Gilbert Saporta[1] and Cristian Preda[2]

[1] Chaire de Statistique Appliquée & CEDRIC, CNAM, Paris, France
e-mail: gilbert.saporta@cnam.fr

[2] Faculté de Médecine, Université de Lille 2, France

Riassunto: Discrimination with a functional predictor is performed through PLS regression. In addition to the definition of a global optimal observation time $[0, t^*]$, we present adaptive methods giving a specific t^* for each new data.

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

1. Introduction

Statistical methods for data representing functions or curves have received much attention in recent years. Such data, known in literature as *functional data* (Ramsay and Silverman, 1997), has received in the last years a large interest for research, especially due to the difficulty to deal with infinite dimensional spaces in the context of classical multivariate methods. 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 functional data is to consider it as paths of a stochastic process $\mathbf{X} = \{X_t\}_{t\in T}$ taking values in a Hilbert space of functions on some set T. For example, for $T \in \mathbb{R}_+$, 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. We assume here that all trajectories are observed continuously on a time interval $[0; T]$ and that the variables X_t have zero mean. The functional linear model considers that the prediction may be expressed as an integral sum

$$\hat{Y} = \int_0^T X_t \beta(t) dt \tag{1}$$

This problem is not new and comes back to Fisher (1924) who used the expression *integral regression*. It is well known that this regression model yields to an ill-posed problem and it is necessary to use regularization techniques: Karhunen-Loeve expansion and functional PLS regression have been compared in Preda and Saporta (2005).
In this paper we are interested to predict a binary response $Y \in \{0, 1\}$, from a stochastic process $\mathbf{X} = \{X_t\}_{t\in[0,T]}$: the extension of functional regression to binary classification is easy, since Fisher's linear discriminant function is equivalent to a multiple regression where the response variable Y is coded with two values a and b.

In previous works (Costanzo *et al.* (2006) and Preda *et al.* (2007)) we have applied PLS functional classification to predict the quality of cookies from curves representing the resistance of dough observed during the kneading process (Lévéder *et al.*, 2004) and proposed to determine a optimal time $t^* < T$ such that the process X considered on $[0, t^*]$ gives similar results, in terms of prediction of Y, as considered on $[0, T]$.

We extend here the anticipated prediction problem by adapting the optimal time t^* to each new trajectory given its incoming measurements: for some cases it could be necessary to observe the process during a longer period than $[0, t^*]$ while for others a shorter period could be enough. We will present two sequential approaches for *on-line* or adaptive forecasting.

2. Linear discriminant analysis on functional data and the PLS approach

The aim of linear discriminant analysis (LDA) for functional data is to find linear combinations $\Phi(\mathbf{X}) = \int_0^T X_t\beta(t)dt$, $\beta \in L_2([0, T])$ such that the between class variance is maximized with respect to the total variance, i.e.

$$\max_{\beta \in L_2[0,T]} \frac{\mathbb{V}(\mathbb{E}(\Phi(X)|Y))}{\mathbb{V}(\Phi(X))} \tag{2}$$

$\Phi(\mathbf{X})$ will be called the *score* function. This is the extension to functional data of the formulation given in Fisher (1936) for LDA.

Due to the infinite dimension of the predictor, the estimation of β is an ill–posed problem. For $K = 2$ it is well known that the optimization problem (2) is equivalent to find the regression coefficients of the linear regression of Y (after a convenient encoding) on the stochastic process $\mathbf{X}$. Nonparametric approaches are proposed in Biau et al. (2003) for analyzing spectrometric data. Solutions based on explicit regularization techniques like ridge–regression have been developed as well as solutions based on a projection onto a finite dimensional space. A functional linear discriminant analysis (FLDA) model is developed by James and Hastie (2001) for irregularly sampled curves using natural cubic spline basis to represent data and the corresponding coefficients for classification.

We propose to use the functional PLS regression approach developed in Preda and Saporta (2005) in order to perform LDA on functional data.

2.1. Functional linear discriminant analysis as a particular case of functional regression

Minimizing $E\left(Y - \int_0^T X_t\beta(t)dt\right)^2$ leads to the Wiener–Hopf equation:

$$E(X_tY) = \int_0^T E\left(X_tX_s\right)\beta(s)ds \tag{3}$$

Let us now consider the case of a binary response, $Y \in \{0, 1\}$ and recode Y by $0 \rightsquigarrow \sqrt{\frac{p_1}{p_0}}$ and $1 \rightsquigarrow -\sqrt{\frac{p_0}{p_1}}$. The discriminant coefficient function, $\beta \in L_2[0, T]$, which satisfies the criterion given in (2) is also solution of:

$$E(Z_tY) = \int_0^T E\left(Z_tZ_s\right)\beta(s)ds \tag{4}$$

Equation (4) is equivalent to the least squares criterion for the linear regression of $\mathbf{Y}$ on the process $\mathbf{Z} = \{Z_t\}_{t\in[0,T]}$ where $Z_t = \sqrt{\lambda}X_t$, $t \in [0, T]$. Thus, the discriminant coefficient function β is, up to a constant, the regression coefficient function of the linear regression of Y on $\mathbf{X}$.
But Picard's theorem states that $\beta(t)$ is unique if and only if $\sum_{i=1}^{\infty} \frac{c_i^2}{\lambda_i^2} < \infty$ where:

$$c_i = \mathrm{cov}(Y, \xi_i) = \mathrm{cov}(Y, \int_0^T f_i(t)X_t dt) = \int_0^T E(X_t Y) f_i(t) dt \tag{5}$$

The $f_i(t)$ are the *factors* of the Karhunen-Loeve expansion

$$X_t = \sum_{i=1}^{\infty} f_i(t)\xi_i \tag{6}$$

associated to eigenvalues λ_i and the ξ_i are the principal components of the process,

$$\int_0^T E(X_t X_s) f_i(s) ds = \lambda_i f_i(t). \tag{7}$$

Picard's condition is generally not satisfied, especially when n is finite since $p > n$.

The equivalence between LDA and linear regression shows that, in general, one cannot write the discriminant score $\Phi(\mathbf{X})$ as linear combination of variables $\mathbf{X}$, $\Phi(\mathbf{X}) = \int_0^T X_t\beta(t)dt$, with $\beta \in L_2([0, T])$. The discriminant score is the orthogonal projection of Y (after appropriate encoding) on the linear space spanned by $\{X_t,\ t \in [0, T]\}$. Then, regularized techniques for linear regression can be used in order to derive an approximation for the discriminant score.

2.2. PCR and PLS approach for linear discrimination

The functional version of principal component regression (PCR) consists in using the components derived from the Karhunen-Loeve expansion as predictors. If we use all principal components we have:

$$\hat{Y} = \sum_{i=1}^{\infty} \frac{\mathrm{cov}(Y, \xi_i)}{\lambda_i}\xi_i = \sum_{i=1}^{\infty} \frac{c_i}{\lambda_i}\xi_i \tag{8}$$

and

$$R^2(Y, \hat{Y}) = \sum_{i=1}^{\infty} R^2(Y, \xi_i) = \sum_{i=1}^{\infty} \frac{c_i^2}{\lambda_i} \tag{9}$$

for finite n we have $R^2 = 1$. In practice we need to choose an approximation of order q:

$$\hat{Y}^{(q)} = \sum_{i=1}^{q} \frac{\mathrm{cov}(Y; \xi_i)}{\lambda_i}\xi_i\hat{\beta}^{(q)}(t) = \sum_{i=1}^{q} \frac{\mathrm{cov}(Y; \xi_i)}{\lambda_i} f_i(t) \tag{10}$$

But using principal components for prediction is heuristic because they are computed irrespective of the response: the components corresponding to the q largest eigenvalues are not necessarily the q most predictive, but it is impossible to rank an infinite number of components according to R^2.

The basic idea of PLS approach is to construct a set of uncorrelated random variables $\{t_i\}_{i\geq 1}$ (PLS components) in the linear space spanned by $\mathbf{X}$, taking into account the correlation between Y and $\mathbf{X}$. Replacing the least squares criterion with that of maximal covariance between $\mathbf{X}$ and Y,

$$\max_{w} \operatorname{cov}^2(Y, \int_0^\infty w(t)X_t dt) \tag{11}$$

the PLS regression offers a good alternative to PCR (Preda and Saporta (2005)). The first PLS component is given by $t_1 = \int_0^\infty w(t)X_t dt$ and further PLS components are obtained by maximizing the covariance criterion between the residuals of both Y and X_t with the previous components. The PLS approximation is given by:

$$\hat{Y}_{PLS(q)} = c_1 t_1 + \ldots + c_q t_q = \int_0^T \hat{\beta}_{PLS(q)}(t)X_t \mathrm{dt} \tag{12}$$

For functional data the same property as in finite dimension holds: "PLS fits closer than PCR":

$$R^2(Y; \hat{Y}_{PLS(q)}) \geq R^2(Y; \hat{Y}_{PCR(q)}) \tag{13}$$

For a supervised classification (*ie* discrimination) in 2 classes, one has just to use the encoding of Y defined earlier. Note that Barker and Rayens (2003) presented a connection between PLS and LDA for the finite dimensional case, but their PLS approach is different in that sense that they look for the orthogonality of coefficient functions giving the PLS scores instead of looking for uncorrelated PLS components (scores).

If the response has more than two categories ($K > 2$) then LDA is equivalent to canonical correlation analysis. However, as for the binary response case, the canonical factor for $\mathbf{X}$ (which is also the discriminant coefficient function) cannot be obtained directly. Considered as a penalized canonical correlation analysis, PLS approach provides a regularization method for LDA by using the PLS components as predictors.

3. Anticipated prediction

Let d_t be the approximation for the discriminant score given by PLS regression on the process X considered on the interval time $[0, t]$, $t \leq T$,

$$d_t = \Phi(\{X_t\}_{[0,t]}) = \int_0^t \hat{\beta}_{PLS,t}(s)X_s ds \tag{14}$$

The purpose of the anticipate prediction is to find $t^* < T$ such that the discriminant score d_{t^*} performs quite as well as d_T. Costanzo *et al.* (2006) propose a solution based on the area under the ROC curve, as quality measure for the discriminant score d_t, and a paired sign test based on a bootstrap resampling : t^* is the smallest t such that the $AUC(t)$ is not significantly different of $AUC(T)$ for a fixed type I risk of error.

The prediction of Y on $\{X_t\}_{t\in[0,T]}$ is then of the same quality as that given by $\{X_t\}_{t\in[0,t^*]}$.

We used 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.

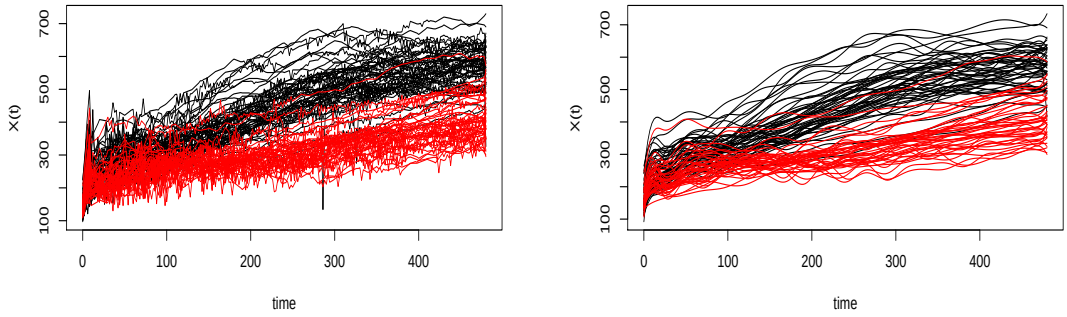

Figure 1: *Good (red) and bad (black) flours. Left : original data. Right : smoothed data*

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 $t^* = 186$. The evolution of the p-value of the sign test is presented in Figure 2.

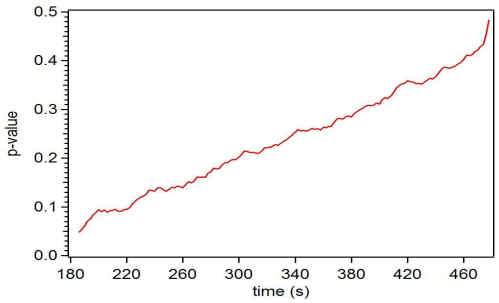

Figure 2: *p-value of the sign test according to time*

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 section 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.

4. 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) which is known for each element of Ω.

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

$$p_{0|\Omega_\omega(t)} = \frac{\left|\{\omega' \in \Omega | \hat{Y}_T(\omega') = 0\} \cap \Omega_\omega(t)\}\right|}{|\Omega_\omega(t)|} \tag{15}$$

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

$$C_\Omega(\omega, t) = \min\{C_{\Omega_\omega(t)}, C_{\overline{\Omega}_\omega(t)}\}. \tag{16}$$

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 $C_{\Omega_\omega(t)}$.

(1) 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 played by 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 (Preda and Saporta (2005), 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, as in Costanzo *et al.* (2006), 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.

Adaptive prediction for kneading data. 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 $t = 100$.

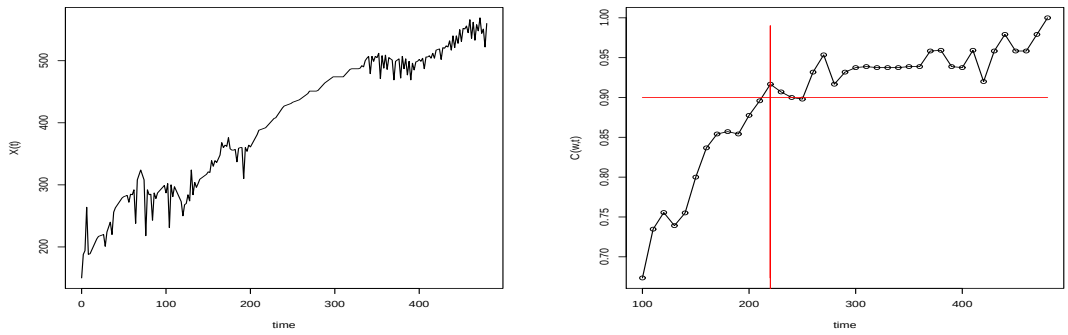

Figure 3: *Left : new flour* ω. *Right :* $C_\Omega(\omega, t)$, $t \in [100, 480]$, $\gamma = 0.90$.

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.

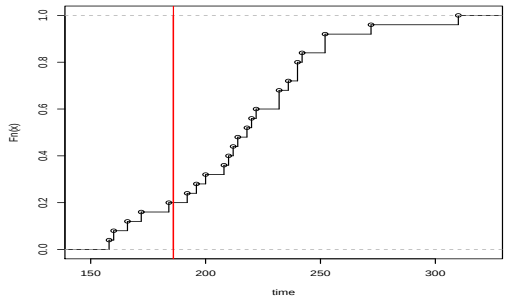

Figure 4: *Empirical cumulative distribution function of t^* (in red, the time point t=186).*

5. Conclusion

Prediction of a binary response has been considered in the context of functional data predictor. 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. Works in progress are devoted to simulation studies and new applications on real data.

References

Barker M., Rayens W. (2003) Partial least squares for discrimination. *Jounal of Chemometrics*, 17:166–173.

Biau G., Bunea F., Wegkamp, M. (2005) Function classification in Hilbert spaces. *IEEE Transactions on Information Theory*, 51:2162–2172.

Costanzo D., Preda C., Saporta G. (2006) *Anticipated prediction in discriminant analysis on functional data for binary response*, In Compstat2006, 17th Symposium on Computational Statistics and Data Analysis, Rome, August 28 - September 1st, 821–828.

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., Hastie T.J. (2001) Functional discrimnant analysis for irregularly sampled curves. *Journal of the Royal Statistical Society, Series B*, 63(3):533–550.

Lévéder C., Abraham C., Cornillon P. A., Matzner-Lober E., Molinari N. (2004) Discrimination de courbes de pétrissage. *Chimiométrie 2004, Paris*, 37–43.

Lian H. (2007) Nonlinear functional models for functional responses in reproducing kernel Hilbert spaces. *The Canadian Journal of Statistics*, Vol. 35, No 4, 597–606.

Preda C., Saporta G. (2005) PLS regression on a stochastic process, *Computational Statistics and data analysis*,48, 149–158.

Preda C., Saporta G., Lévéder C. (2007) PLS classification for functional data, *Computational Statistics*, 22, 223–235.

Ramsay J. O., Silverman B.W. (1997) *Functional Data Analysis*. New York, Springer.

Ramsay J. 0., Silverman B.W. (2002) *Applied Functional Data Analysis : Methods and Case Studies*. Springer.