Discriminant analysis on functional data

Gilbert Saporta
Chaire de Statique Appliquée & CEDRIC, CNAM, Paris - saporta@cnam.fr

Abstract: Discriminant analysis or "supervised" classification for functional data occurs when for each curve or path of a stochastic process we have a single categorical response Y. Linear methods looks for predictors which may be expressed as an integral sum. Fisher's linear discriminant function being equivalent to a multiple regression with a coded response, one can use techniques for the regression problem when Y is continuous. When t takes continuously its values in an interval $[0;T]$, multicollinearity leads to inconsistent estimation of the parameters. Components derived from the Karhunen-Loeve decomposition are, for functional data, the equivalent of principal components regression (PCR). Partial least squares performs better than PCR, since principal components are obtained irrespective of the response (Preda *et al.*, 2007). Functional logistic regression is another approach advocated by Aguilera *et al.* , 2006. We determine an optimal time $t^* < T$ giving a prediction based on $[0;t^*]$ almost as good as the prediction based on $[0;T]$ (Costanzo *et al.*, 2006) by using a bootstrap test for AUC criterion.

Keywords: functional data, regression, discriminant analysis, classification

1 Introduction

Functional data (Ramsay, Silverman 1997) occur when we observe curves or trajectories from a stochastic process X_t. If for each curve we have a single categorical response Y, we have a classification 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.

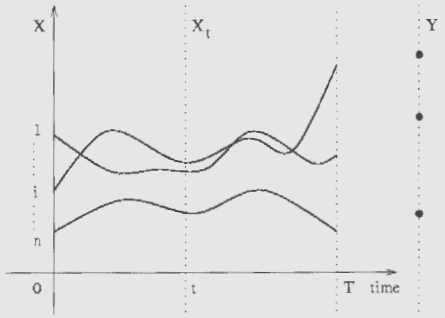

Figure 1: A scheme for prediction with functional data

There are many recent works on this topic; using mainly nonparametric techniques. We focus here on the extension of the classical linear model.

2 The functional linear model

The functional linear model considers a predictor which may be expressed as an integral sum:

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

The problem is not new and comes back to Fisher (1924) who used the expression "integral regression" for solving a problem of predicting the amount of crop with temperature curves. It is well known that this regression model yields to an ill-posed problem: the least squares criterion leads to the Wiener-Hopf equation which in general has not an unique solution.

$$E(X_t Y) = \int_0^T E\left(X_t X_s\right) \beta(s) ds$$

and the problem is even worse when we try to estimate the regression coefficient function $\beta(t)$ with a finite number of observations. This is quite clear by using the Karhunen-Loeve decomposition.

2.1 Functional principal components analysis or Karhunen-Loeve expansion

The generalization of principal components analysis to functional data (Saporta, 1985) relies on the Karhunen-Loeve decomposition. The principal component analysis (PCA) of the stochastic process (X_t) consists in representing X_t as:

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

where the principal components $\xi_i = \int_0^T f_i(t) X_t dt$ are obtained through the eigenfunctions of the covariance operator:

$$\int_0^T C(t,s) f_i(s) ds = \lambda_i f_i(t)$$

This integral equations cannot be solved analytically for empirical data. However for a finite number of observations, there exists an exact solution: If $\mathbf{W}$ is the matrix of all inner products between trajectories $w_{uv} = \int_0^T x_u(t)\, x_v(t) dt \quad u, v = 1, 2, .., n$, then the principal components are its eigenvectors and we have $f(t) = \frac{1}{n}\frac{1}{\lambda} \sum_{u=1}^{n} \xi_u X_u(t)$

Otherwise one has to discretize the trajectories.

2.2 Linear functional regression

For an integral linear predictor $\hat{Y} = \int_0^T X_t \beta(t) dt$ Picard's theorem states that the regression function $\beta(t)$ is unique if and only if $\sum_{i=1}^{\infty} \frac{c_i^2}{\lambda_i^2} < \infty$ where

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

Picard's condition is generally not satisfied especially when n is finite: since $p > n$. We have a perfect fit when minimizing:

$$\frac{1}{n} \sum_{i=1}^{n} \left(y_i - \int_0^T \beta(t) x_i(t) dt \right)^2$$

Many techniques have been applied to solve this kind of problem, mostly by using explicit regularization techniques. High dimensionality and multicollinearity also involves some smoothing. In the functional linear approach, functional data (the predictor) and functional parameter can be modelled as linear combinations of a basis functions from a given functional family. Literature on that subject essentially differs in the choice of the basis and the way parameters are estimated. Basis functions should be chosen to reflect the characteristics of the data: for example, Fourier basis are usually used to model periodic data, while B-spline basis functions are chosen as they have the advantage of finite support. We will focus here on linear methods based on an orthogonal decomposition of the predictors.

2.3 Linear regression on principal components

Using components derived from the Karhunen-Loeve expansion is, for functional data, the equivalent of principal components regression (PCR).

If we use all principal components we have:

$$\hat{Y} = \sum_{i=1}^{\infty} \frac{\operatorname{cov}(Y, \xi_i)}{\lambda_i} \xi_i = \sum_{i=1}^{\infty} \frac{c_i}{\lambda_i} \xi_i \quad \text{and} \quad R^2(Y, \hat{Y}) = \sum_{i=1}^{\infty} R^2(Y, \xi_i) = \sum_{i=1}^{\infty} \frac{c_i^2}{\lambda_i}$$

but for finite n, $R^2 = 1$.

In practice we need to choose an approximation of order q : .

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

But using principal components for prediction is heuristic because they are computed independently 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...

2.4 Functional PLS regression (Preda & Saporta, 2005)

PLS regression offers a good alternative to the PCR method by replacing the least squares criterion with that of maximal covariance between (X_t) and Y .

$$\max_w \operatorname{cov}^2(Y, \int_0^\infty w(t)X_t dt) \quad \text{with} \quad \|w\|^2 = 1$$

The first PLS component is given by $t_1 = \int_0^\infty w(t)X_t dt$.

The PLS regression is iterative 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:

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

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

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

since PCR components are obtained irrespective of the response. Preda & Saporta (2005) showed the convergence of the PLS approximation to the approximation given by the classical linear regression:

$$\lim_{q \to \infty} E(\left\| \hat{Y}_{PLS(q)} - \hat{Y} \right\|^2) = 0$$

In practice, the number of PLS components used for regression is determined by crossvalidation.

3 Supervised classification on functional data by linear methods

3.1 Functional linear discrimination

Regression methods for functional data are easily generalized to binary classification, since Fisher's linear discriminant function is equivalent to a multiple regression where the response variable Y is coded with 2 values a and b : most frequently ± 1, but also conveniently $\sqrt{\frac{p_1}{p_0}}$ and $-\sqrt{\frac{p_0}{p_1}}$ with (p_0, p_1) the probability distribution of Y.

Costanzo et al. (2006) and Preda et al. (2007) have applied PLS functional classification to predict the quality of cookies from curves representing the resistance (density) of dough observed during the kneading process. For a given flour, the resistance of dough is recorded during the first 480 s of the kneading

process. We have 115 curves which can be considered as sample paths of a L^2-continuous stochastic process. Each curve is observed in 240 equispaced time points of the interval time [0, 480]. 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 for Y= Adjustable and 40 for Y= Bad . Due to measuring errors, each curve is smoothed using cubic B-spline functions with 16 knots.

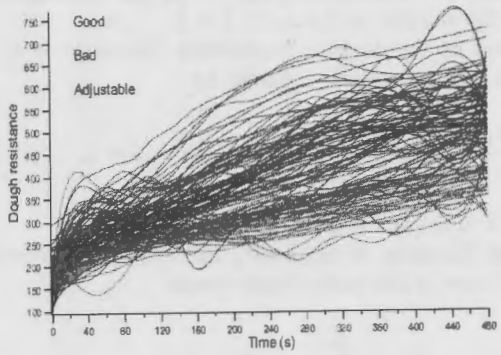

Figure 2: Smoothed kneading curves

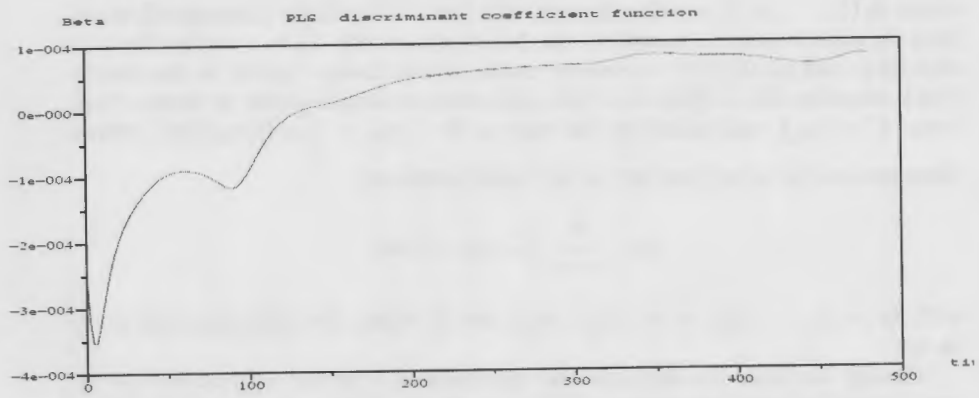

Figure 3: Discriminant coefficient function

Some of these flours could be adjusted to become Good. Therefore, we have considered the set of Adjustable flours as the test sample and predict for each one the group membership, Y = {Good, Bad}, using the discriminant coefficient function (Fig. 2) given by the PLS approach on the 90 flours. PLS functional discriminant analysis gave an average error rate of 11% which is better than discrimination based on principal components.

3.2 Functional logistic regression

Let Y be a binary random variable and $y_1, \ldots, y_n$ the corresponding random sample associated to the sample paths $x_i(t)$, $i = 1, \ldots, n$.

A natural extension of the logistic regression (Ramsay et al., 1997) is to define the functional logistic regression model by :

$$\ln\left(\frac{\pi_i}{1-\pi_i}\right) = \alpha + \int_0^T x_i(t)\beta(t)dt; \quad i = 1, \ldots, n$$

where $\pi_i = P(Y = 1|X = x_i(t); t \in T)$.

It may be assumed (Ramsay et al., 1997) that the parameter function and the sample paths $x_i(t)$ are in the same finite space:

$$\beta(t) = \sum_{q=1}^{p} b_q \psi_q(t) = \mathbf{b}'\psi$$

$$x_i(t) = \sum_{q=1}^{p} c_{iq} \psi_q(t) = \mathbf{c}_i'\psi$$

where $\psi_1(t), \ldots, \psi_q(t)$ are the elements of a basis of the finite dimensional space. Such an approximation transform the functional model (1) in a similar form to standard multiple logistic regression model whose design matrix is the matrix which contains the coefficients of the expansion of sample paths in terms of the basis, $\mathbf{C} = (c_{iq})$, multiplied by the matrix $\Phi = (\phi_{kq} = \int_T \psi_k(t)\psi_q(t)dt)$, whose elements are the inner product of the basis functions

$$\ln\left(\frac{\pi}{1-\pi}\right) = \alpha\mathbf{1} + \mathbf{C}\Phi\mathbf{b}$$

with $\mathbf{b} = (b_1, \ldots, b_p)$, $\pi = (\pi_1, \ldots \pi_p)$ and $\mathbf{1}$ being the p-dimensional unity vector.

Finally, in order to estimate the parameters a further approximation by truncating the basis expansion could be considered. Alternatively, regularization or smoothing may be get by some roughness penalties approach.

In a similar way as we defined earlier functional PCR, Leng and Müller (2006) use functional logistic regression based on functional principal components with

the aim of classifying gene expression curves into known gene groups.With the explicit aim to avoid multicollinarity and reduce dimensionality, Escabias *et al.* (2004) and Aguilera *et al.* (2006) propose an estimation procedure of functional logistic regression, based on taking as covariates a reduced set of functional principal components of the predictor sample curves, whose approximation is get in a finite space of no necessarily orthonormal functions. Two different forms of functional principal components analysis are then considered, and two different criterion for including the covariates in the model are also considered. Müller and Stadtmüller (2005) consider a functional quasi likelihood and an approximation of the predictor process with a truncated Karhunen-Loeve expansion. The latter also developed asymptotic distribution theory using functional principal scores.

Comparisons with functional LDA are in progress, but it is likely that the differences will be small.

4 Anticipated prediction

In many real time applications like industrial process, it is of the highest interest to make anticipated predictions. Let denote d_t the approximation for a discriminant score considered on the interval time $[0, t]$, with $t < T$: $d_t = \int_0^t X_t\hat{\beta}(t)dt$

The aim is to find $t^* < T$ such that the discriminant score $d_t{}^*$ performs quite as well as d_T .

Costanzo *et al.* (2006) proposed a procedure for a binary target Y, based on the ROC curve and the AUC (Area Under Curve) criterion. 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", $P(d_t > r|Y = 0)$. 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.

Define t^* as the first value of s where AUC(s) is not significantly different from AUC(T) Since AUC(s) and AUC(T) are not independent variables, we use a bootstrap test for comparing areas under ROC curves: 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.

The previous methodology has been applied to the kneading data: the sample of 90 flours is randomly split 50 times into a stratified learning sample of size 60 and a stratified test sample of size 30. Functional PLS discriminant analysis gave, with the whole interval [0, 480], an average test error rate of 0.112, for an average AUC(T) = 0.746. We used here a Wilcoxon test and the first time where the p-value was greater than 0.05 was t^* = 186 see figure 4. Thus, one can reduce the recording period to less than half of the current one.

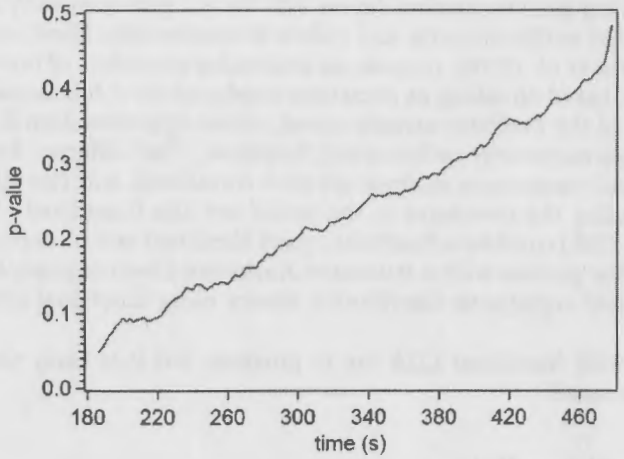

Figure 4: p- value of the Wilcoxon test

5 Conclusions and perspectives

PLS regression is an efficient and simple way to get linear prediction for functional data. Anticipated prediction could be solved by means of a bootstrap procedure

Work in progress concern "on-line" forecasting: instead of using the same anticipated decision time t^* for all data, in a forthcoming paper we adapt t^* to each new trajectory given its incoming measurements.

References

[1] Aguilera, A.M., Escabias, M. and Valderrama M.J. (2006). Using principal components for estimating logistic regression with high-dimensional multicollinear data. *Computational Statistics & Data Analysis*, 50, 1905-1924.

[2] Barker M. and Rayens W. (2003). Partial least squares for discrimination. *J Chemomericst*, 17, 166-173.

[3] Costanzo D., Preda C. and Saporta G. (2006). Anticipated prediction in discriminant analysis on functional data for binary response. *In COMPSTAT2006*, 821-828. Physica-Verlag.

[4] Escabias, M. and Aguilera A.M. & Valderrama M.J. (2004). Principal Component Estimation of Functional Logistic Regression: discussion of two different approaches. *Nonparametric Statistics*, 16, 365-384.

[5] 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.

[6] Leng, X. and Müller, H.G. (2006). Classification using functional data analysis for temporal gene expression data. *Bioinformatics*, 22, 68-76.

[7] Lévéder C., Abraham C., Cornillon P. A., Matzner-Lober E. and Molinari N. (2004). Discrimination de courbes de pétrissage. *Chimiometrie*, 37-43.

[8] Müller and H.G, Stadtmüller, U. (2005). Generalized functional linear models. *The Annals of Statistics*, 33, 774-805.

[9] Preda C. and Saporta G. (2005). PLS regression on a stochastic process. *Computational Statistics and Data Analysis*, 48, 149-158.

[10] Preda C., Saporta G. and Lévéder C., (2007). PLS classification of functional data. *Computational Statistics*, 22(2), 223-235.

[11] Ramsay J.O. and Silverman (1997). *Functional data analysis.* Berlin: Springer.

[12] Saporta G. (1985). Data analysis for numerical and categorical individual time series. *Applied Stochastics Models and Data Analysis*, 2, 109-119.