Anticipated prediction in discriminant analysis on functional data for binary response

G. Damiana Costanzo[1], Cristian Preda[2], and Gilbert Saporta[3]

[1] Università della Calabria, Italia, dmcosta@unical.it
[2] Faculté de Médecine, Université de Lille 2, France, cpreda@univ-lille2.fr
[3] CNAM Paris, France, saporta@cnam.fr

Summary. Linear discriminant analysis is studied when the predictors are data of functional type and the response is a Bernoulli random variable. The aim of this work is to anticipate the prediction of the response earlier than the end of the observed stochastic process. Due to the infinite dimension of the predictor space, discriminant coefficient functions cannot be derived as in the classical way and partial least squares approach is proposed. Results of a simulation study as well as an application to kneading data are presented.

Key words: functional data, partial least squares, discriminant analysis, stochastic process

1 Introduction

Let us consider the particular case of linear discriminant analysis for binary response when the predictor set is a stochastic process with continuous time index, $\{X_t\}_{t\in[0,T]}$. Such data, known in literature as *functional data* ([RS97]), 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. Thus, generalized linear regression models are developed in [Jam02] and more recently by [CS05]. Different linear approaches based on decomposition of the underlying stochastic process are proposed : principal component regression ([CFS99]), partial least squares regression (PLS)([PS05]), logistic regression ([EAV04], [EAV05]). Non-parametric models for regression on functional data using classical kernel estimators are developed in [FV04] and [Pre06], both for scalar and categorical response.

In this paper we are interested to predict a binary response Y, $Y \in \{0, 1\}$, from a stochastic process $X = \{X_t\}_{t\in[0,T]}$ in the following way. Firstly, we measure the predictive capacity of X by considering the process on the whole interval $[0, T]$. Depending on the quality of prediction, we are interested to determine a time $t^* < T$ such that the process X considered on $[0, t^*]$ gives similar results, in terms of prediction of Y, as considered on $[0, T]$. This second point is very important from a practical point of view. If the process X is related to a control parameter affecting the outcome (Y) then it is useful to anticipate the realization of Y, for example, in order to prevent some critical events (which will occur at time T) or just for economical reasons.

Let us consider the example of the kneading process. For a given flour, during the kneading process one can record the resistance of dough in a interval of time $[0, T]$. The obtained curve (see an example in Fig. 1) can be used ([LACMM04], [PSL]) for predicting the quality of cookies obtained with this dough. If the cookie's quality could be anticipated in a short time, reparation could be done for amelioration or for stopping the production process of bad quality cookies. Several other examples where anticipated prediction is useful for controlling process can be found in [RLH02a], [RLH02b], [KLSK00].

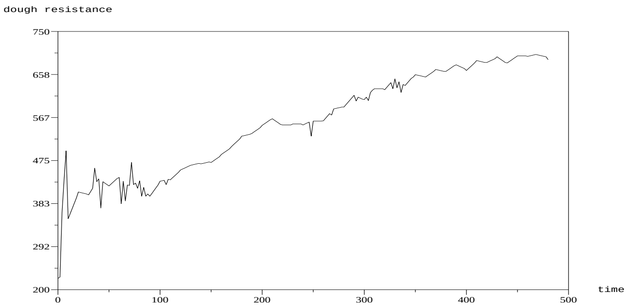

Fig. 1. Example of functional data : dough resistance curve observed during 480 seconds

The paper is organized as follow. In the Section 2 we present some tools for linear discrimination on functional data, in particular the partial least squares (PLS) approach. Section 3 is devoted to anticipated prediction under several criterion of prediction quality. A simulation study as well as an application on kneading data is presented in Section 4.

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 $\{(x_i, y_i)\}_{i=1,\dots,n}$ be n observations of random variables (X, Y) with $x_i = \{x_i(t), t \in [0, T]\}$ and $y_i \in \{0, 1\}$, $i = 1, \dots, n$. Due to infinite dimension of the predictor, the estimation of β is in general an ill–posed problem. In [PSL] it is shown that the optimization problem (1) 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

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

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 (2) has an unique solution under conditions of convergence of series implying the eigenvalues and eigenvectors of the covariance operator of the process X [Sap81]. 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 [PS05] 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 $\mathbf{W}^X_{q-1}\mathbf{W}^Y_{q-1}$, where $\mathbf{W}^X_{q-1}$, respectively $\mathbf{W}^Y_{q-1}$, are the Escoufier's operators ([Sap81]) 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 prediction

Now, 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 first value of s where 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 boosstrap 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.

4 Applications

We use a simulation study for which the anticipated prediction is possible before the end of the process and we evaluate our procedure for this particular case. In the second part, we perform an application of the anticipated prediction on the kneading data ([LACMM04]) provided by Danone Vitapole Research Department (France).

4.1 Simulation study

Let us consider Y be a Bernoulli random variable, $Y \sim \mathcal{B}(0.5)$. The simulated data, $X = \{X_t\}_{t\in[0,2]}$, we consider correspond to the binary response Y for which the predictor has the following form :

$$\text{Class } \{Y = 0\} : X_t = \begin{cases} W(1-t), & 0 \leq t \leq 1 \\ -2\sin(t-1) + W(t-1), & 1 < t \leq 2 \end{cases}$$

$$\text{Class } \{Y = 1\} : X_t = \begin{cases} W(1-t), & 0 \leq t \leq 1 \\ 2\sin(t-1) + W(t-1), & 1 < t \leq 2 \end{cases}$$

where W is the standard brownian motion. Observed data are discretized curves with 201 equidistant points, $t \in \{0, 0.01, 0.02, \ldots, 2\}$. Fig. 2 displays a sample of 100 simulated curves for each class.

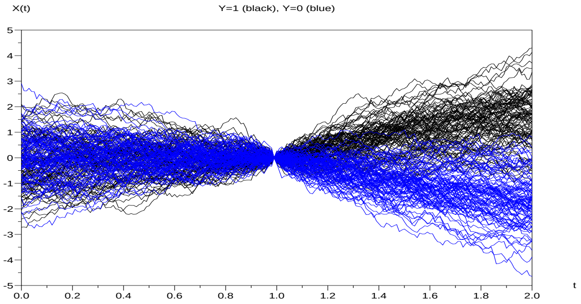

Fig. 2. Sample of size $n = 100$ for each class of Y.
endcenter

Clearly, for $t \leq 1$, the distribution of $X^{(t)} = \{X_s\}_{s\in[0,t]}$ being the same for both classes, any prediction model will fail. For, $t > 1$, $X^{(t)} = \{X_s\}_{s\in[0,t]}$ is intuitively more predictible as t is closer to 2.

For each $s \in \{0, 0.01, 0.02, \ldots, 2\}$ one generates $M = 50$ learning samples of size $n = 100$ for each class. For each learning sample we generate a test sample (same size) which is used to evaluate the model by computing the $AUC(s)$. One obtains in this way $M = 50$ independent realisations of $AUC(s)$.

Using the Wilcoxon test (one-tailed) with the first error type fixed to 0.05, the minimum t^* for which the test is not significant is $t^* = 1.46$. The Wilcoxon statistic is 1.582 and the two averaged AUC corresponding to $t^* = 1.46$ and respectively to $T = 2$ are $\overline{AUC}(t^*) = 0.866$, respectively, $\overline{AUC}(T) = 0.872$. The corresponding averaged ROC curves are presented in Fig. 3.

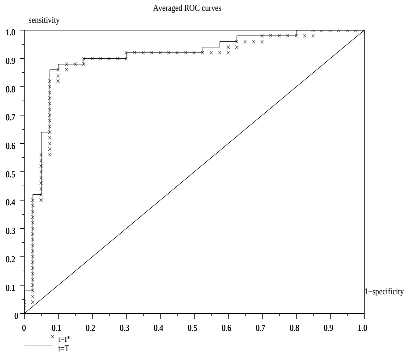

Fig. 3. Averaged ROC curves for $T = 2$ and $t^* = 1.46$

4.2 Application to kneading data

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). After kneading, the dough is processed to obtain cookies. For each flour we have the quality (Y) of cookies which can be *Good*, *Adjustable* or *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 as in [LACMM04] (Fig 4.).

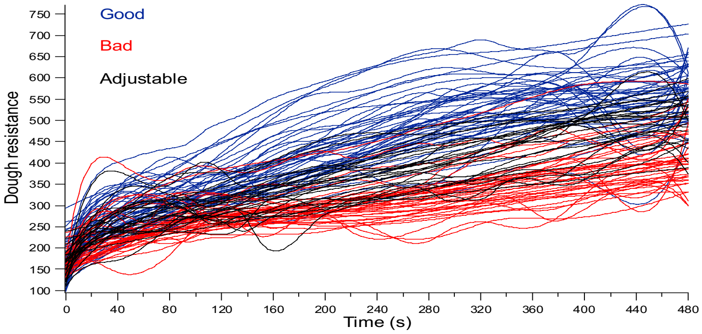

Fig. 4. Sample of kneading curves measuring the resistance of dough during 480 seconds

In the following we consider $Y \in \{Bad, Good\}$. 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. Several discriminant models are fitted in [PSL]. The PLS discriminant analysis gives, for a prediction taking into account the whole interval $[0, 480]$, an average of the test error rate of about 0.112, for an average $\overline{AUC}(T) = 0.746$. The anticipated prediction procedure gives for $M = 50$ and sample size test $n = 30$ (same number of observation in each class), $t^* = 186$. Thus, one can reduce the recording period of the resistance dough to less than half of the current one.

5 Conclusion and perspective

In this paper we addressed the problem of forecasting a random response categorical variable Y, namely a binary one, by predicting on the associated continuous stochastic process $\{X_t\}_{t\in[0,T]}$. Such kind of situation is common to many real applications where a continuous phenomenon evolving in a certain interval of time results in an outcome not observable before the completion of the process itself. We faced the problem by means of the PLS approach for which forecasting of the binary response is drawn as 'anticipated prediction' of the process $\{X_t\}_{t\in[0,T]}$ at $t = T$.

A conceptually different approach would be 'on-line' forecasting Y: instead of using the same anticipated decision time t^* for all data, we adapt t^* to each new trajectory given its incoming measurements. Work in progress comprises the developping of the presented approach by means of PLS-functional logistic model, which would involve detection of an optimal time at which starts the 'best' forecasting and a sequential test procedure to validate the predicted forecasting of Y.

References

[CFS99] Cardot H., Ferraty F., Sarda P. : Functional linear model, Statist. Probab. Lett., **45**, 11-22, (1999).

[CS05] Cardot H. and Sarda P.: Estimation in generalized linear models for functional data via penalized likelihood, Journal of Multivariate Analysis, **92**, 24-41, (2005).

[EAV04] Escabias, M., Aguilera A.M., Valderama M.J. : Principal component estimation of functional logistic regression: discussion of two different approaches Journal of Nonparametric Statistics, **16** (3-4), 365-384, (2004).

[EAV05] Escabias, M., Aguilera A.M., Valderama M.J. : Modelling environmental data by functional principal component logistic regression, Environmetrics, **16** (1), 95-107, (2005).

[FV04] Ferraty F. and Vieu P. : Nonparametric models for functional data with application in regression, time series prediction and curve discrimination, Journal of Nonparametric Statistics, Vol. **16**, No.1-2, 111-125, (2004).

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

[Jam02] James G. M. : Generalized linear models with functional predictors, Journal of the Royal Statistical Society, Series B, **64** (3), 411-432, (2002).

[LACMM04] Lévéder C., Abraham C., Cornillon P. A., Matzner-Lober E., Molinari N. : Discrimination de courbes de pétrissage, Chimiométrie 2004, p. 37 –43, (2004).

[Pre06] Preda C. : Regression models for functional data by reproducing kernel Hilbert space methods, Journal of Statistical Planning and Inference, to appear, (2006).

[PS05] Preda C. and Saporta G. : PLS regression on a stochastic process Computational Statistics and Data Analysis, **48** (1), 149-158, (2005).

[PSL] Preda C. and Saporta G. : PLS classification of functional data, In PLS05, 4th International Symposium on PLS and related methods, Barcelona, 7–9 September 2005 (2005).

[RS97] Ramsay J.O. and Silverman B.W. : Functional Data Analysis, Springer Series in Statistics, Springer-Verlag, New York, (1997).

[RS02] Ramsay J.O. and Silverman B.W. : Applied Functional Data Analysis : Methods and Case Studies, Springer (2002).

[RLH02a] Ratcliffe S. J., Leader L.R., Heller G.Z. : Functional data analysis with application to periodically stimulated foetal heart rate data. I : Functional regression, Statistics in Medicine, **21**, 1103-1114, (2002).

[RLH02b] Ratcliffe S. J., Leader L.R., Heller G.Z., Functional data analysis with application to periodically stimulated foetal heart rate data. II : Functional logistic regression, Statistics in Medicine, **21**, 1115-1127, (2002).

[Sap81] Saporta G. : Méthodes exploratoires d'analyse de données temporelles, Cahiers du B.U.R.O, Université Pierre et Marie Curie,**37-38**, Paris, (1981).

[KLSK00] Kesavan,P., Lee, J.H., Saucedo, V., Krishnagopalan, G.A.: Partial least squares (PLS) based monitoring and control of batch digesters. Journal of Process Control , **10**, 229-236 (2000).