Functional PLS logit regression model

- Manuel Escabias, Ana M. Aguilera, Mariano J. Valderrama
- Functional PLS logit regression model
- *Computational Statistics & Data Analysis, Volume 51, Issue 10, 2007, Pages 4891-4902*
- DOI: https://doi.org/10.1016/j.csda.2006.08.011

Volume 169
May 2022

ISSN 0167-9473

COMPUTATIONAL
STATISTICS
& DATA ANALYSIS

An Official Publication of

The International Association for Statistical Computing
A Section of The International Statistical Institute

The network Computational and Methodological Statistics

Available online at www.sciencedirect.com

ScienceDirect

Available online at www.sciencedirect.com
ScienceDirect
ELSEVIER
Computational Statistics & Data Analysis 51 (2007) 4891 – 4902
COMPUTATIONAL STATISTICS & DATA ANALYSIS
www.elsevier.com/locate/csda

Functional PLS logit regression model

M. Escabias*, A.M. Aguilera, M.J. Valderrama

Department of Statistics and O.R. University of Granada, Spain

Available online 28 August 2006

Abstract

Functional logistic regression has been developed to forecast a binary response variable from a functional predictor. In order to fit this model, it is usual to assume that the functional observations and the parameter function of the model belong to a same finite space generated by a basis of functions. This consideration turns the functional model into a multiple logit model whose design matrix is the product of the matrix of sample paths basic coefficients and the matrix of the inner products between basic functions. The likelihood estimation of the parameter function of this model is very inaccurate due to the high dependence structure of the so obtained design matrix (multicollinearity). In order to solve this drawback several approaches have been proposed. These employ standard multivariate data analysis methods on the design matrix. This is the case of the functional principal component logistic regression model. As an alternative a functional partial least squares logit regression model is proposed, that has as covariates a set of partial least squares components of the design matrix of the multiple logit model associated to the functional one.

Keywords: Logistic regression; Functional data; Partial least squares (PLS)

1. Introduction

Functional logistic regression (FLR) has received much attention in literature in recent years with different objectives. On the one hand, James (2002) used FLR in the more general framework of functional generalized linear models with the emphasis on the prediction of the response variable (5-year survival on a randomized placebo controlled trial of the drug D-penicillamine on patients with primary biliary cirrhosis of the liver). With the same objective, Ratcliffe et al. (2002) used a FLR model for predicting if human foetal heart rate responds to repeated vibroacoustic stimulation, and Rossi et al. (2002) predicted whether an examinee would answer a certain item of a specific test correctly. In this context, FLR can be used as a curve discrimination method alternative to those developed in literature, as linear discriminant analysis on functional data (Preda and Saporta, 2005b) or nonparametric curves discrimination (Ferraty and Vieu, 2003). On the other hand, Escabias et al. (2004) paid their attention to the accurate estimation of the parameter function of the FLR model without forgetting the importance of an accurate prediction of the response. An interpretation of the parameter function in terms of odds ratios was developed in Escabias et al. (2005) with a climatological application to establish the relationship between the risk of drought and time evolution of temperatures. More recently, Müller and Stadtmüller (2005) have developed asymptotic inference for a class of generalized functional linear regression models based on approximating the predictor process with a truncated Karhunen–Loève expansion.

* Corresponding author. Universidad de Granada, Departamento de Estadística e I.O., Facultad de Farmacia, Campus de Cartuja, 18071-Granada, Spain. Tel.: +34 958240640; fax: +34 958249046.

E-mail address: escabias@ugr.es (M. Escabias).

0167-9473/$ - see front matter © 2006 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2006.08.011

These different objectives of FLR have shown different ways of dealing with the problem with a common point, that is the multiple approach to the functional logit model after considering that both the observations of the functional predictor and the parameter function are expressed in terms of basis functions. James (2002) and Rossi et al. (2002) consider the basis coefficients as unobservable quantities and use the EM algorithm to estimate them. On the other hand, Ratcliffe et al. (2002) used least squares approximation of the basis coefficients from discrete time observations of the sample paths meanwhile Escabias et al. (2005) developed a quasi-natural cubic spline interpolation procedure. All of these methods transform the functional model into a multiple logistic regression model with high multicollinearity that makes the estimation of the parameters of the model inaccurate with high estimated variance in spite of the good prediction of the response variable. In order to solve this problem and to provide an accurate estimation of the parameter function of the model, Escabias et al. (2004) studied two different functional principal component (PC) approaches that take as covariates a reduced set of PCs of the design matrix of the multiple logit model equivalent to the original FLR model.

Many criticisms have been made to principal component analysis (PCA) as a resource to avoid multicollinearity, mainly in multiple regression methods, based on the fact that PCs are calculated without taking into account the response variable. As an alternative to PCA of the predictors in regression methods (also known as principal component regression (PCR)) Wold (1975) introduced the partial least squares (PLS) analysis, that consists of obtaining a set of uncorrelated latent variables (as the PCs) that takes into account the relationship between the response and the predictor variables, and to fit the regression model with the PLS components as covariates. Different algorithms for computing the PLS components have been developed in the literature as NIPALS (see Geladi and Kowalski, 1986), SIMPLS (De Jong, 1993) and more recently powered PLS (Indahl, 2005). Moreover, PLS has been incorporated within the iteratively reweighted least squares (IRWLS) steps for multinomial or binary logistic regression (Marx, 1996). Frank and Friedman (1993) and Garthwaite (1994) compared PLS with different methods for solving multicollinearity in linear regression, i.e. PLS, PCR, ridge regression (RR), variable subset selection (VSS), ordinary least squares (OLS), forward variable selection, PCR and a Stein shrinkage method, and concluded that PLS is suited to models with many variables and large error variances. In addition, Garthwaite (1994) provides an interpretation of the PLS components that can be viewed as weighted averages of predictor variables where each predictor holds the residual information in an explanatory variable that is not contained in earlier components.

Due to the great interest raised by PLS in recent years (see the special issue on PLS published in Computational Statistics and Data Analysis in 2005), this technique has been extended and adapted to be used with other statistical methods such as generalized linear regression models (Bastien et al., 2005). For the special situation of logistic regression in presence of multicollinearity, Aguilera et al. (2006) show that PC logistic regression (PCLR) can provide an estimation of the parameters of the model as accurate as the one given by the PLS logit algorithm of Bastien et al. (2005), with a subtle variation in the selection of PCs, that consists of selecting them by a stepwise method based on conditional likelihood ratio test.

Other statistical field where literature has been wide over the last few years, has been functional data analysis (FDA) where data are functions usually observed at a finite set of time points. Classical statistical techniques as PCA, linear regression or canonical correlation analysis, among others, have been extended to the functional case (see Ramsay and Silverman, 2005 for a detailed study). Functional methods have also been developed to solve the prediction problem in several areas of statistics as for example the PCP models to forecast a process in the future from its recent past (Aguilera et al., 1999a,b) and its adaptation to forecast continuous time series (Aguilera et al., 1999c). On the other hand, PLS regression have recently been generalized to PLS of a stochastic process (Preda and Saporta, 2005a) for solving the problem of estimating the functional linear model.

The objective of this paper is to give a new step in all these generalizations by introducing the functional PLS logistic regression (FPLSLR) model as an alternative to the functional principal component logistic regression (FPCLR) introduced by Escabias et al. (2004) for solving multicollinearity. Our approach is different from the one proposed by Preda and Saporta (2005a) in the linear case where the PLS components of the multiple model that result after selecting a set of knots and considering the average (between knots) of the sample paths, were obtained. They state that the so obtained multiple PLS components were the same as the functional ones. In this paper, we propose to use PLS multiple logit regression on the design matrix of the multiple logit model obtained after considering that the parameter function and the functional predictor belong to a finite space generated by a basis of functions. In order to do this, we apply the PLS generalized linear regression algorithm proposed by Bastien et al. (2005) to the particular case of the logit model.

The paper has been organized in four sections. After the introductory section, we summarize the functional logit regression model, its parameter function interpretation, the fitting problems, the multiple-based solution and the multicollinearity problem. In Section 3, we describe the algorithm to obtain the PLS components for the logit regression model and present our proposed FPLSLR model. Finally, in Section 4 we develop three simulation studies, the two first focused on the accurate estimation of the parameter function, and the last one based on curve discrimination.

2. The functional logistic regression model and the problem of multicollinearity

As we have emphasized in the Introduction, functional data analysis methodologies have received much attention in recent years, generalizing a lot of statistical techniques to the functional field. Among these methods is logistic regression that has been extended to the functional case to forecast a binary response from a functional predictor whose observations are functions instead of vectors as in the classical multivariate setting. As in other functional regression methods, the fitting procedure is usually based on expressing the functional observations of the predictor variable and the parameter function of the model in terms of a basis of functions. Let us formulate the FLR model and the most used method of fitting, showing the problems that arise and how they are usually solved.

Let $\{x_1(t), \ldots, x_n(t) : t \in T\}$ be a random sample of observations (sample paths) of a functional variable $X(t)$, and $\{y_1, \ldots, y_n\}$ a set of observations of a binary response variable Y associated to them (i.e. $y_i \in \{0, 1\},\ \forall i = 1, \ldots, n$). In order to formulate the FLR model, we will assume that the random variables X and Y are defined on the same probability space with X valued in the separable Hilbert space L^2_T, whose usual inner product is defined as

$$\langle \alpha, \beta \rangle = \int_T \alpha(t)\beta(t)\,\mathrm{d}t \quad \forall \alpha, \beta \in L^2_T.$$

Then, the FLR model is given by

$$y_i = \pi_i + \varepsilon_i, \quad i = 1, \ldots, n,$$

where the probability that the binary response variable takes value one given a functional observation, $\pi_i = P\{Y = 1|X(t) = x_i(t)\}$, is modelized as

$$\pi_i = \frac{\exp\{\alpha + \langle x_i, \beta \rangle\}}{1 + \exp\{\alpha + \langle x_i, \beta \rangle\}} = \frac{\exp\{\alpha + \int_T x_i(t)\beta(t)\,\mathrm{d}t\}}{1 + \exp\{\alpha + \int_T x_i(t)\beta(t)\,\mathrm{d}t\}} \tag{1}$$

with α being a real parameter and β the parameter function of the model that belong to the space L^2_T. As in the multiple logistic regression model (Hosmer and Lemeshow, 2000), ε_i are centered independent random errors with variance $Var[\varepsilon_i] = \pi_i(1 - \pi_i)$. Eq. (1) can be expressed alternatively in terms of the logit transformation, $l_i = \log[\pi_i/(1 - \pi_i)]\ \forall i = 1, \ldots, n$, as

$$l_i = \alpha + \int_T x_i(t)\beta(t)\,\mathrm{d}t, \tag{2}$$

so that the FLR model can be seen as a particular case of the functional generalized linear models introduced in James (2002). Let us observe that the conditional distribution of Y given $X(t)$ is a Bernoulli distribution that belongs to the exponential family.

Formula (2) leads us to the interpretation of the parameter function. That is, the integral of the parameter function multiplied by a constant K, can be interpreted as the multiplicative change in the odds of response $Y = 1$ obtained when a functional observation is incremented constantly in K units along T (Escabias et al., 2005).

As in the case of the functional linear regression model, in the FLR model we have to take into account different aspects. Firstly, we cannot observe the functional form of the sample paths continuously, as much we can observe each of them in a finite set of discrete time points. Secondly, it is impossible to estimate the parameter function (that is infinite no numerable) with a finite number of observations n. In order to circumvent the first task which is common to all functional methods, different proposals can be found in literature. In the context of functional PLS, Preda and Saporta (2005a) reconstruct the functional form of sample paths by fixing a set of knots and considering the average between knots of the sample paths observed at a finite set of points. In the more general context of functional generalized linear models, Cardot and Sarda (2005) discuss identifiability of the model and propose an estimation procedure based on

B-splines and penalized likelihood. In that paper, it is proved that in order to ensure the identifiability of the FLR model it is necessary to assume that the functional variable X is of second order and the eigenvalues of its covariance operator are nonzero.

In this paper, we will apply the usual solution in functional regression methods that solves at the same time the two pointed questions by expressing the sample paths and the parameter function in terms of a basis of functions that spans the space where the sample paths belong to (see Ramsay and Silverman (2005) for a general revision). Different basis as trigonometric functions (Aguilera et al., 1995; Ratcliffe et al., 2002), cubic spline functions (Aguilera et al., 1996; Escabias et al., 2005) or wavelets functions (Amato et al., 2006; Ocaña et al., 2006) can be used depending on the nature of the functional predictor sample paths. In the particular case of doubly stochastic Poisson processes, Bouzas et al. (2006) have proposed a new procedure for estimating the mean process based on monotone piecewise cubic interpolation of its sample paths.

If we denote by $\Phi=(\phi_1(t),\ldots,\phi_p(t))'$ a vector of basic functions (the prime denote transpose) that span the space where $x_i(t)$ and $\beta(t)$ belong to, then we have

$$x_i(t)=a_i'\Phi, \quad \beta(t)=\beta'\Phi,$$

where $a_i=(a_{i1},\ldots,a_{ip})'$ is the vector of sample path basis coefficients and $\beta=(\beta_1,\ldots,\beta_p)'$ the parameter function basis coefficients. Therefore, the functional model in terms of the logit transformations (2) is expressed in matrix form as

$$L=\mathbf{1}\alpha+A\Psi\beta, \tag{3}$$

with $L=(l_1,\ldots,l_n)'$, $\mathbf{1}=(1,\ldots,1)'$, A being the $n\times p$ matrix which has as rows the sample path basis coefficients a_i' and $\Psi=(\psi_{jk})$ being the $p\times p$ matrix which has as entries the basic functions inner products

$$\psi_{jk}=\int_T \phi_j(t)\phi_k(t)\,\mathrm{d}t, \quad j,k=1,\ldots,p.$$

The A matrix is usually obtained by least squares approximation from the observations of the sample curves at discrete points that could be different for each sample individual.

The multiple logistic regression model (3) so obtained provides a design matrix with high dependence structure (multicollinearity). The multicollinearity problem has an undesirable effect in regression methods providing inaccurate estimated parameters and increasing their estimated variances (see Jollife (2002) for multicollinearity discussion in linear regression or Aguilera et al. (2006) for logistic regression). In order to solve this problem and to obtain an accurate estimation of the parameter function, Escabias et al. (2004) proposed several FPCLR models based on different functional PCAs of the functional predictors and a stepwise procedure based on conditional likelihood ratio tests for selecting the PCs to be included in the logit model. In this paper, we are going to develop a functional PLS logit model that uses as covariates of the multiple logit model (3) the logit PLS components of its design matrix $A\Psi$. The next section summarizes the algorithm to obtain the logit PLS components, the model in terms of the logit PLS components and the estimation of the parameter function basis coefficients.

3. A solution to multicollinearity: the functional PLS logit model

As an alternative to FPCLR, we propose a new functional regression method based on the PLS logit model introduced by Bastien et al. (2005) that consists of adapting the classical PLS regression algorithm (Wold, 1975) to the logistic regression model.

In any PLS regression method, the PLS components are defined as uncorrelated linear spans of the explicative variables of the regression model that maximize the covariance between linear spans of explicative and response variables, respectively. These linear spans are usually obtained by algorithmic methods. In the logistic regression case the PLS algorithm proposed by Bastien et al. (2005) considers as the linear spans's coefficients, a transformation of the slope parameters of the simple logit fittings of the response Y on each single explicative variable as covariate. The functional PLS logit model that we propose in this work is based on a multiple PLS logit regression model that has $A\Psi$ as design matrix (see Eq. (3)).

The algorithm for computing the FPLSLR model consists of the three following steps:

(1) Computation of a set of PLS components.
Let Y be a binary variable and $H_j,\ \ j=1,\dots,p$ the columns of the design matrix of the functional logit model. Then, the algorithm that computes the logit PLS components is summarized as follows:

Step 1: First logit PLS component

- Logit regression fits Y/H_j $(j=1,\dots,p)$, denoting the estimated slope parameters by $\widehat{\delta}_1=(\widehat{\delta}_{11},\dots,\widehat{\delta}_{1p})'$ and by $V_1=(v_{11},\dots,v_{1p})'$ its normalized vector $(v_{1j}=\widehat{\delta}_{1j}/\|\widehat{\delta}_1\|)$.
- Set equal to zero those coefficients v_{1j} that are not significant in base to the usual Wald test. That means to delete those components of V_1 that verify $|\widehat{\delta}_{1j}/SE(\widehat{\delta}_{1j})|\leqslant z_{\alpha/2}$, with $SE(\widehat{\delta}_{1j})$being the estimated standard deviation of $\widehat{\delta}_{1j}$ and $z_{\alpha/2}$ being a fixed critical value of the standard normal distribution.
- The first logit PLS component is defined as $T_1=v_{11}H_1+\cdots+v_{1p}H_p$.

Step l: Given $T_1,\dots,T_{l-1}$ the first $l-1$ logit PLS components, the lth one is obtained as follows:

- Logit regression fits $Y/(T_1,\dots,T_{l-1},H_j)$ $(j=1,\dots,p)$, denoting by $\widehat{\delta}_l=(\widehat{\delta}_{l1},\dots,\widehat{\delta}_{lp})'$ the estimated slope parameter corresponding to the variables H_j in each logit regression and by $V_l=(v_{l1},\dots,v_{lp})'$ its normalized vector $(v_{lj}=\widehat{\delta}_{lj}/\|\widehat{\delta}_l\|)$.
- Set equal to zero those coefficients v_{lj} that are not significant, $|\widehat{\delta}_{lj}/SE(\widehat{\delta}_{lj})|\leqslant z_{\alpha/2}$, with $SE(\widehat{\delta}_{lj})$ being the estimated standard deviation of $\widehat{\delta}_{lj}$.
- Linear regression fits $H_j/(T_1,\dots,T_{l-1})$ $(j=1,\dots,p)$, denoting by $R_1^{(l-1)},\dots,R_p^{(l-1)}$ their residual vectors.
- The lth logit PLS component is defined by $T_l=v_{l1}R_1^{(l-1)}+\cdots+v_{lp}R_p^{(l-1)}$.

The algorithm stops at the step where the component T_l is not significant because none of the coefficients v_{lj} is significantly different from 0. That means that $|\widehat{\delta}_{lj}/SE(\widehat{\delta}_{lj})|\leqslant z_{\alpha/2}\ \forall j=1,\dots,p$. Other usual method for selecting the number s of PLS components to be retained is cross-validation on the predictive power of the model (see Tenenhaus, 2002 for a detailed study).

(2) Logit regression fitting of the response variable Y on the retained PLS components.
First, all logit PLS components are expressed in terms of the original covariates (columns H_j of the $A\Psi$ matrix) instead of the corresponding residual vectors. If we denote by Γ the matrix of logit PLS components of the design matrix $A\Psi$, then we have $\Gamma=A\Psi V$, with V being the matrix whose columns are the vector of coefficients of the logit PLS components in terms of the original predictors.

Let us consider the multiple expression (3), then the logit model in terms of the PLS components is expressed as

$$\widehat{L}=\mathbf{1}\widehat{\gamma}_0+\Gamma\widehat{\gamma},$$

where $\widehat{\gamma}=(\widehat{\gamma}_1,\dots,\widehat{\gamma}_s)'$ are the maximum likelihood estimators of the coefficients of the logit model in terms of the s logit PLS components obtained in the previous algorithm.

(3) Formulation of the PLS logit regression model in terms of the original predictors

$$\widehat{L}=\mathbf{1}\widehat{\gamma}_0+A\Psi V\widehat{\gamma},$$

so that we can finally obtain the following estimation of the parameter function: $\widehat{\beta}(t)=\widehat{\beta}'\Phi$, in terms of its basis coefficients estimated from the gamma parameters $\widehat{\beta}=V\widehat{\gamma}$.

4. Simulation study

In order to test the performance of functional PLS logit model, we have developed three simulation examples. In the first and the second ones, we follow two different simulation schemes proposed in Escabias et al. (2004), where the effects of multicollinearity on the parameter function estimation were analyzed. In the third example, the simulation process is the one developed in Ferraty and Vieu (2003). The first and second examples are devoted to the accurate

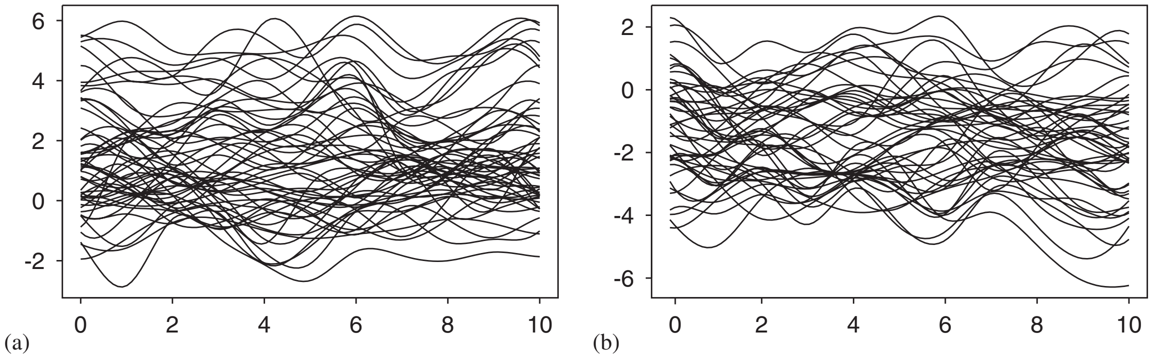

Fig. 1. Simulated sample curves of Example 1: (a) curves with response $Y=1$; (b) curves with response $Y=0$.

estimation of the parameter function of the FLR model meanwhile the third one deals with its curve discrimination ability. In all cases we have calculated the usual goodness of fit measures of the logistic regression model, i.e. the deviance statistic, given by

$$G^2 = 2\sum_{i=1}^{n}\left[y_i \log\frac{y_i}{\widehat{\pi}_i} + (1-y_i)\log\frac{1-y_i}{1-\widehat{\pi}_i}\right],$$

its p-value, the area under the receiver operating curve (ROC), and the correct classification rate (CCR) defined as the rate of correct response classifications by using 0.5 as cut-point taking into account that those observations with predicted probability greater than 0.5 are classified as $Y=1$ and the rest as $Y=0$. In these examples, the results provided by FPLSLR are compared with other existing estimation methods as the nonfunctional multiple PLS logit regression (MPLSLR) (Bastien et al., 2005) that uses as predictors the vectors of the observations of the sample curves at a finite set of time points, and other functional approaches as the FPCLR (Escabias et al., 2004), among others.

4.1. Example 1

In the first example, we considered as functional explicative variable the one whose sample curves are cubic spline functions expressed in terms of the basis of cubic B-spline functions defined by the knots $\{0, 1, 2, \ldots, 10\}$

$$x(t) = \sum_{j=-1}^{11} a_j B_j(t), \tag{4}$$

with $B_j(t)$ being the cubic B-spline functions. In order to simulate a set of sample paths we simulated 100 vectors of dimension 13 of a centered multivariate normal distribution with highly correlated components, which correspond to the coefficients of the B-splines in expression (4) (see Escabias et al., 2004 for a detailed explanation of the simulation method). Fig. 1 shows these sample paths. As parameter function of the logit model we considered the natural cubic spline interpolation of the function $\sin(t-\pi/4)$ on the knots previously defined (see Fig. 3). Finally, each observation of the response variable y_i $(i=1,\ldots,n)$ was simulated by using a Bernoulli distribution with probability π_i calculated by expression (1) with $\alpha=0.5$.

After simulating the data, we fitted the functional logistic regression model given by Eq. (3), obtaining the estimated parameter function given in Fig. 2. Let us observe the great difference from the simulated one. The variance of the estimated parameter function defined in this paper as the sum of the estimated variances of its basis coefficients, is 668.4326. This extremely high value is a symptom of inaccuracy of the parameter function estimation caused by multicollinearity. In order to evaluate the functional logistic regression model we obtained some goodness of fit measures as the CCR giving 88% of correct classifications and the deviance statistic $G^2=56.48$ that provides a p-value of 0.9942. All these measures show that the logistic model is a good choice for these data. As measure of accuracy of the parameter function estimation, we calculated the integrated mean squared error (IMSEB) of the parameter function

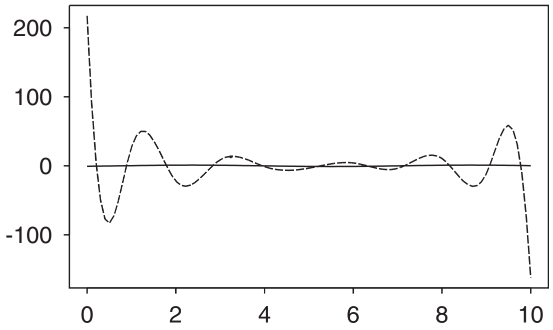

Fig. 2. Simulated parameter function (solid line) and FLR estimation (broken line) for a particular simulation of Example 1.

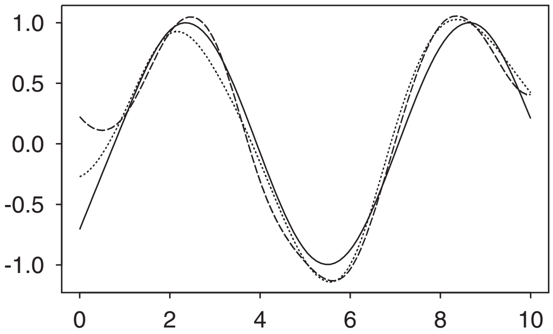

Fig. 3. Simulated parameter function (solid line), FPLSLR estimation (broken line) and best FPCLR estimation (dotted line) for a particular simulation of Example 1.

defined as

$$IMSEB = \frac{1}{T}\int (\beta(t) - \widehat{\beta}(t))^2 \, \mathrm{d}t.$$

This measure showed that the estimation so obtained was very inaccurate providing $IMSEB = 1.16E + 5$.

In order to better estimate the parameter function of the FLR model by avoiding the multicollinearity problem, we obtained the functional logit PLS components by the algorithm proposed in Section 3, finding that there were only three components. After that, we fitted the logit model in terms of these components and reconstructed the parameter function as was shown in the previous section. The estimated parameter function so obtained can be seen in Fig. 3 from which we can state that this estimation is much better than the one seen in Fig. 2. Following this, we obtained the goodness of fit measures of this model with $CCR = 88\%$ and deviance statistic $G^2 = 59.66$ (p-value $= 1.00$) which showed that the model fitted well. Finally the parameter function accuracy measures were calculated, obtaining $IMSEB = 0.03$ and estimated variance of the estimated parameter function $Var = 1.44$, all of which were much lower than the ones obtained without using PLS components. All these measures showed a better estimation of the parameter function by using the functional PLS logit model.

In order to compare the performance of FPLSLR with other competitors as the FPCLR model, we obtained the PCs of the design matrix of Model (3) and fitted the multiple logit model in terms of different numbers of PCs. We found that the model with the first six PCs (the ones that cumulate more than 95% of the total variability) provided the best parameter function estimation with the lowest *IMSEB*.

From Table 1 and Fig. 3, we can observe that the parameter function estimations obtained by FPLSLR and FPCLR are very similar, in both cases they provide a clear improvement in the estimations with respect to the functional logit model. The FPLSLR model may be superior to FPCLR model in a bigger reduction of dimension in the number of components needed to obtain the best possible estimation. Escabias et al. (2004, 2005) proved that in the case of FPCLR it is better to include PCs in the model in the order given by a stepwise method than in the natural order of explained variability, in order to obtain an accurate estimated parameter function.

Table 1
Goodness of fit and accuracy measures of the different logit models fitted to a particular simulation of Example 1

Method	Covariates	IMSEB	ROC area	CCR	G^2	p-value
FLR	13	1.16E+5	0.912	88	56.48	0.990
FPCLR	6	1.78E − 1	0.942	88	59.57	0.997
FPLSLR	3	2.76E − 2	0.944	88	59.66	0.998

Table 2
Mean and standard deviation of goodness of fit and accuracy measures of the different optimum logit models after 350 simulations of Example 1

Measures	FPLSLR		FPCLR	
	Mean	SD	Mean	SD
Covariates	2.72	0.53	4.19	0.91
ROC area	0.945	0.022	0.937	0.019
CCR	87.11	3.42	85.97	2.96
IMSEB	0.52	2.19	0.95	0.76
G^2	57.13	11.66	0.98	0.001

In order to check if the previously proved results are the general rule, we repeated the simulation of the binary response variable 350 times and fitted the FPLSLR and FPCLR models in each one. For the FPCLR model, we considered as optimum model in each repetition the one which had the lowest *IMSEB* after fitting the models in terms of different number of PCs included one by one by explained variability. In the two types of models, we obtained the mean and standard deviation of the goodness of fit and accuracy measures previously defined for the optimum models. The results can be seen in Table 2, and from them we can see that the mean CCR and ROC area are similar in both methods, that the mean number of components needed for the best possible estimation of the parameter function is lower in the FPLSLR model (2.72) than in FPCLR model (4.19) and that the mean *IMSEB* is lower too in FPLSLR model (0.52) than in the FPCLR model (0.95). From these results, we can state that the conclusions seen before are the rule and not a particular case.

4.2. Example 2

In order to compare the results of FPLSLR and FPCLR with MPLSLR regression, we have developed a new example where the sample paths are only observed at a finite set of time points given by {0, 0.1, 1.3, 1.4, 2.6, 3.1, 3.5, 3.6, 3.9, 7.10, 7.1, 7.60, 7.66, 7.9, 8.6, 8.7, 9.8, 10.351, 10.353, 10.4, 11.1, 11.2, 12.3, 12.71, 12.74, 13.5, 13.8, 14.8, 14.9, 15}. We have considered as functional predictor the stochastic process

$$X(t) = Z(t) + t/4 + 5E,$$

where $Z(t)$ is a zero mean Gaussian stochastic process with covariance function given by $C(t, s) = (1/2)^{80|t-s|}$ and E is a Bernoulli random variable with parameter $p = 0.1$. Then we have simulated observations of a sample of 100 curves of this process at the previous set of 30 unequally spaced times of the interval [0,15].

In order to reconstruct the functional form of the original curves we consider that they are cubic splines and can be expressed in terms of the basis of B-splines defined by the knots {0, 1.1, 2.5, 3.7, 5.1, 6.9, 8.2, 9.6, 10.9, 12.1, 13.3, 15}. The basis coefficients of the corresponding expansion of each sample curve is obtained by least squares approximation.

The parameter function considered in this example is the cubic spline interpolation of $\cos(t - \pi/4)$ on the nodes that define the B-spline functions. Each value of the response variable was again simulated by using a Bernoulli distribution whose parameter was computed by using expression (1) with $\alpha = 1.5$.

The following steps for fitting the FPLSLR and FPCLR models are the same as those seen in the first example. In this example, we have also repeated the simulation and the fits 340 times and we have obtained the same resume measures for the optimum models. In addition, we have estimated for each simulation a MPLSLR model based on logit regression of the response on the logit PLS components associated to the data matrix whose rows are the vectors of observations of

Table 3
Mean and standard deviation of goodness of fit and accuracy measures of the different optimum logit models after 340 simulations of Example 2

Measures	FPLSLR		FPCLR		MPLSLR	
	Mean	SD	Mean	SD	Mean	SD
Covariates	1.72	0.82	5.17	2.04	1.96	0.76
ROC area	0.921	0.054	0.926	0.109	0.949	0.047
CCR	94.76	2.13	94.97	1.91	95.75	1.95
IMSEB	0.68	1.51	0.47	0.19	4.20	25.47
G^2	27.62	9.31	27.39	8.54	22.02	9.44

each predictor curve at discrete time points. In order to compute the IMSEB associated to this multiple model, we have computed the parameter function by using least squares approximation of the estimated parameter vector (values of the parameter function at the observation time points) on the same B-splines basis used for approximating the predictor sample paths.

The mean and standard deviation of the goodness of fit and accuracy measures of the three models (FPLSLR, FPCLR and MPLSLR) adjusted to each simulation can be seen in Table 3. From this table, we can corroborate the results of Example 1 and compare them with the ones provided by MPLSLR. Let us observe that the mean CCR and ROC area are similar for the three methods, the mean number of components needed for the best possible estimation of the parameter function is lower with FPLSLR and MPLSLR than with FPCLR, and the mean *IMSEB* is similar for the functional models (FPLSLR and FPCLR) and higher for the MPLSLR approach.

4.3. Example 3

In Examples 1 and 2, we have shown the ability of the FPLSLR model to provide an accurate estimation of the parameter function. Next we have developed another simulated example with the emphasis on its prediction ability. As we have noted in previous sections, logistic regression is a good method for discrimination, so we have used the FPLSLR model for curve discrimination in order to compare it with different methods which can be found in literature, such as the classification and regression tree procedure (CART), functional discriminant analysis (FDA), multivariate partial least-squares regression (MPLSR), penalized discriminant analysis (PDA) or nonparametric curve discrimination (NPCD) (see Ferraty and Vieu, 2003 for a detailed explanation).

In order to follow the simulation scheme developed in Ferraty and Vieu (2003) we simulated 1000 curves of two different classes, for the first class we simulated 500 curves according to the function

$$x(t) = uh_1(t) + (1-u)h_2(t) + \varepsilon(t),$$

and another 500 curves were simulated for the second class according to the function

$$x(t) = uh_1(t) + (1-u)h_3(t) + \varepsilon(t),$$

with u and $\varepsilon(t)$ being uniform and standard normal simulated random values, respectively, and

$$h_1(t) = \max\{6 - |t-11|, 0\}, \quad h_2(t) = h_1(t-4), \quad h_3(t) = h_1(t+4).$$

The sample curves were simulated at 101 equally spaced points on the interval $[1, 21]$. As binary response variable, we considered $Y = 0$ for the curves of the first class and $Y = 1$ for the ones of the second class. Finally, we considered 500 curves (250 of each class) as training sample and the rest as test sample. Some of the curves of each class can be seen in Fig. 4.

After simulating the data, we obtained a least squares approximation of the curves on the space spanned by the cubic B-spline functions defined on 70 equally spaced knots of the interval $[1, 21]$. The heuristic reason for choosing this high dimension is to obtain an accurate reconstruction of predictor sample paths with a suitable balance between smoothness and roughness. Then, we fitted the FLR model given by Eq. (3), by obtaining as estimated parameter function the one drawn in Fig. 5. The variance of this estimated parameter function was very large ($Var = 3.12E + 015$) which caused us to suspect the inaccuracy of this estimation. This inaccurate estimation can be caused by any of two problems, on

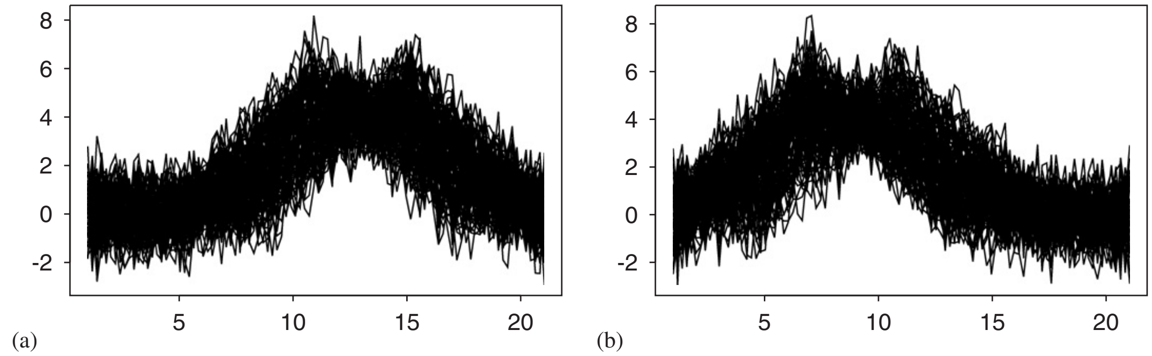

Fig. 4. Simulated sample curves of Example 3: (a) first class curves; (b) second class curves.

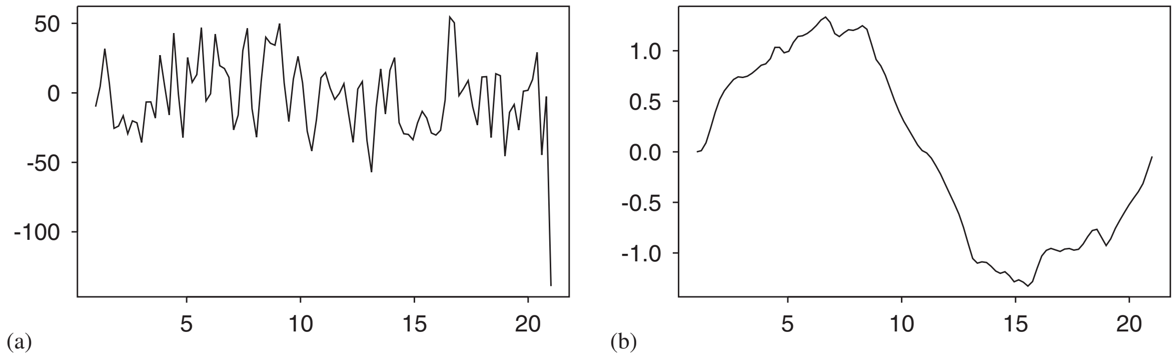

Fig. 5. Parameter function estimations of Example 3: (a) FLR; (b) FPLSLR.

the one hand the high multicollinearity that exists in the multiple model (3), and on the other hand the high number of variables (72) that the model has. The logistic regression model is very sensitive to the number of variables, in fact when the number of variables is high the algorithms of most statistical packages tend to fail giving estimated probabilities close to zero or one and making the likelihood function to be unstable or undefined.

In order to avoid these two problems we decided to use the FPLSLR model proposed in this paper, so we obtained the functional logit PLS components, finding that only one component was able to be extracted, and we fitted the logit model in terms of this component. The estimated parameter function provided by this model is in Fig. 5 and its estimated variance was $Var = 2.34$ far from the one obtained without using logit PLS. The goodness of fit measures associated to this model were $G^2 = 18.33$ (p-value $= 1.00$) and $CCR = 98.2$ what showed that the model fitted well.

Finally we used the test sample in order to evaluate how the model discriminates the curves. Therefore, we used the least squares approximation of the curves of the test sample and the estimated parameter function in order to predict the probabilities given by expression (1) for the test sample. Finally, we used 0.5 as cut point in order to predict the values of the response variable and calculated the correct classification rate ($CCR = 96.8$), providing that only 3.2% of curves were incorrectly classified.

In order to compare our methods with those existing for curve discrimination (Ferraty and Vieu, 2003) we repeated the simulation of the training and test samples 50 times, fitted the FPLSLR, FPCLR and MPLSLR models and obtained the misclassification rate and the ROC area in each repetition. Escabias et al. (2004) proved that for the FPCLR model the best possible estimation of the parameter function (the one with the lowest *IMSEB*) was achieved by the model previous to an outstanding increment of the variance of the parameter function, after including PCs in the model one by one according to the explained variance. Thus in each repetition we considered this criterion to select the most accurate parameter function estimation in FPCLR.

The mean of the ROC areas of the FPLSLR, FPCLR and MPLSLR models fitted to each one of the 50 simulations are 0.996, 0.999 and 0.939, with standard deviations 0.006, 0.001, and 0.071, respectively. The mean and variance of the test sample misclassifications provided by the adjusted FPLSLR, FPCLR and MPLSLR models over the 50 simulations can be seen in Table 4, next to the ones obtained by Ferraty and Vieu (2003) with alternative curve discrimination

Table 4
Comparison of different curves discrimination methods

Method	Error rates	Standard deviations
CART	0.262	1.8E − 2
FDA	0.123	1.4E − 2
NPCD	0.079	1.2E − 2
PDA	0.089	1.6E − 2
FPLSLR	0.030	8.7E − 3
FPCLR	0.032	8.3E − 3
MPLSLR	0.243	1.4E − 1

Mean and standard deviation of misclassifications after 50 simulations of Example 3.

methods on 50 different simulated data set of the same functional predictor. From this table, we can see that FPLSLR is the best model for curve discrimination because the error rate is much lower than the rest. Let us also observe how the functional PLS approach proposed in this paper overperforms the multiple PLS logit regression on the vectors of observed values of sample curves, and provides similar results (similar high CCR and ROC areas) to the functional PC approach.

5. Conclusions

In this paper, we have proposed a functional PLS logit regression model as a method to avoid the multicollinearity problem in functional logistic regression. We have shown that the method also avoids the problem of high dimensionality that precludes the logit model. The improvement of the proposed method with respect to the logit model is achieved by providing a more accurate parameter function estimation and by discriminating in an accurate way a set of curves of a functional variable.

From the three simulated examples developed in this paper, we can conclude that FPLSLR provides an accurate estimation of the parameter function estimation similar to the one given by the alternative FPCLR but with a bigger dimension reduction. On the other hand, the ability of FPLSLR for curve discrimination is also similar to FPCLR and better than other competitors.

Acknowledgments

This research has been supported by Project MTM2004-5992 from *Dirección General de Investigación*, *Ministerio de Ciencia y Tecnología*.

References

Aguilera, A.M., Gutiérrez, R., Ocaña, F.A., Valderrama, M.J., 1995. Computational approaches to estimation in the principal component analysis of a stochastic process. Appl. Stochastic Models Data Anal. 11 (4), 279–299.

Aguilera, A.M., Gutiérrez, R., Valderrama, M.J., 1996. Approximation of estimators in the PCA of a stochastic process using B-splines. Comm. Statist. Comput. Simulation 25 (3), 671–690.

Aguilera, A.M., Ocaña, F.A., Valderrama, M.J., 1999a. Forecasting with unequally spaced data by a functional principal component approach. Test 8 (1), 233–253.

Aguilera, A.M., Ocaña, F.A., Valderrama, M.J., 1999b. Stochastic modelling for evolution of stock-prizes by means of functional principal component analysis. Appl. Stochastic Models Bus. Industry 15 (4), 227–234.

Aguilera, A.M., Ocaña, F.A., Valderrama, M.J., 1999c. Forecasting Time series by functional PCA. Discussion of several weighted approaches. Comput. Statist. 14, 443–467.

Aguilera, A.M., Escabias, M., Valderrama, M.J., 2006. Using principal components for estimating logistic regression with high-dimensional multicollinear data. Comput. Statist. Data Anal. 50, 1905–1924.

Amato, U., Antoniadis, A., De Feis, I., 2006. Dimension reduction in functional regression with applications. Comput. Statist. Data Anal. 50 (9), 2422–2446.

Bastien, P., Esposito-Vinzi, V., Tenenhaus, M., 2005. PLS generalised linear regression. Comput. Statist. Data Anal. 48 (1), 17–46.

Bouzas, P.R., Valderrama, M.J., Aguilera, A.M., Ruiz-Fuentes, N., 2006. Modelling the mean of a doubly stochastic Poisson process by functional data analysis. Comput. Statist. Data Anal. 50, 2655–2667.

Cardot, H., Sarda, P., 2005. Estimation in generalized linear models for functional data via penalized likelihood. J. Multivariate Anal. 92, 24–41.

De Jong, S., 1993. SIMPLS: an alternative approach to partial least squares regression. Chemometrics Intell. Lab. Syst. 18, 251–263.

Escabias, M., Aguilera, A.M., Valderrama, M.J., 2004. Principal component estimation of functional logistic regression: discussion of two different approaches. J. Nonparametric Statist. 16 (3–4), 365–384.

Escabias, M., Aguilera, A.M., Valderrama, M.J., 2005. Modeling environmental data by functional principal component logistic regression. Environmetrics 16 (1), 95–107.

Ferraty, F., Vieu, P., 2003. Curves discrimination: a nonparametric functional approach. Comput. Statist. Data Anal. 44, 161–173.

Frank, I.E., Friedman, J.H., 1993. A statistical view of some chemometrics regression tools. Technometrics 32 (2), 109–135.

Garthwaite, P.H., 1994. An interpretation of partial least squares. J. Amer. Statist. Assoc. 89 (425), 122–127.

Geladi, P., Kowalski, B.R., 1986. Partial least squares regression: a tutorial. Anal. Chim. Acta 185, 1–17.

Hosmer, D.W., Lemeshow, S., 2000. Applied Logistic Regression. Wiley, New York.

Indahl, U., 2005. A twist to partial least squares regression. J. Chemometrics 19 (1), 32–44.

James, J.M., 2002. Generalized linear models with functional predictors. J. Roy. Statist. Soc. Ser. B 64 (3), 411–432.

Jollife, I.T., 2002. Principal Component Analysis. Springer, New York.

Marx, B.D., 1996. Iteratively reweighted partial least squares estimation for generalized linear regression. Technometrics 38, 374–381.

Müller, H.-G., Stadtmüller, U., 2005. Generalized functional linear models. Ann. Statist. 33 (2), 774–805.

Ocaña, F.A., Aguilera, A.M., Valderrama, M.J., 2006. A wavelet approach to functional principal component prediction model. Technical Report, University of Granada.

Preda, C., Saporta, G., 2005a. PLS regression on a stochastic process. Comput. Statist. Data Anal. 48 (1), 149–158.

Preda, C., Saporta, G., 2005b. PLS classification for functional data. In: Aluja, T., Casanovas, J., Esposito-Vinzi, V., Morineau, A., Tenenhaus, M. (Eds.), PLS and Related Methods. Proceedings of the PLS'05 International Symposium, pp. 164–174.

Ramsay, J.O., Silverman, B.W., 2005. Functional Data Analysis. Springer, New York.

Ratcliffe, S.J., Heller, G.Z., Leader, L.R., 2002. Functional data analysis with application to periodically stimulated foetal heart rate data. II: functional logistic regression. Statist. Med. 21, 1115–1127.

Rossi, N., Wang, X., Ramsay, J.O., 2002. Nonparametric item response function estimates with the EM algorithm. J. Behav. Educ. Sci. 27, 291–317.

Tenenhaus, M., 2002. La régression PLS. Théorie et pratique. Paris, Editions Technip.

Wold, H., 1975. Soft modelling by latent variables; the nonlinear iterative partial least squares approach. In: Gani, J. (Ed.), Perspectives in Probability and Statistics, Papers in Honour of M.S. Barlett, Academic Press, London.