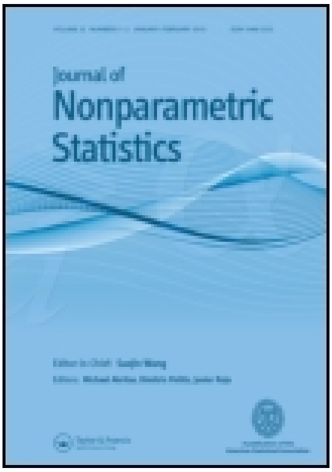


Journal of Nonparametric Statistics
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/gnst20

Principal component estimation of functional logistic regression: discussion of two different approaches

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

[a] Department of Statistics and Operation Research , University of Granada , Spain

Published online: 31 Jan 2007.

Nonparametric Statistics
Vol. 16(3–4), June–August 2004, pp. 365–384

PRINCIPAL COMPONENT ESTIMATION OF FUNCTIONAL LOGISTIC REGRESSION: DISCUSSION OF TWO DIFFERENT APPROACHES

M. ESCABIAS*, A. M. AGUILERA and M. J. VALDERRAMA

Department of Statistics and Operation Research, University of Granada, Spain

(Received 14 December 2002; Revised 21 July 2003; In final form 10 September 2003)

Over the last few years many methods have been developed for analyzing functional data with different objectives. The purpose of this paper is to predict a binary response variable in terms of a functional variable whose sample information is given by a set of curves measured without error. In order to solve this problem we formulate a functional logistic regression model and propose its estimation by approximating the sample paths in a finite dimensional space generated by a basis. Then, the problem is reduced to a multiple logistic regression model with highly correlated covariates. In order to reduce dimension and to avoid multicollinearity, two different approaches of functional principal component analysis of the sample paths are proposed. Finally, a simulation study for evaluating the estimating performance of the proposed principal component approaches is developed.

Keywords: Functional data analysis; Logistic regression; Principal components

1 INTRODUCTION

Data in many different fields come to us through a process naturally described as functions. This is the case of the evolution of a magnitude such as, for example, temperature in time. It usually happens that we only have discrete observations of functional data in spite of its continuous nature. In order to reconstruct the true functional form of data many approximation techniques have been developed, such as interpolation or smoothing in a finite dimensional space generated by a basis. A general overview of functional data analysis (FDA) can be seen in Ramsay and Silverman (1997, 2002) and Valderrama *et al.* (2000).

The great development on FDA in recent years has meant that many studies with longitudinal data historically raised from a multiple point of view are now analyzed on the basis of their functional nature. Some of these works have focused on modeling the relationship between functional predictor and response variables observed together at varying times. This is the case for example, in Liang and Zeger (1986) who proposed a set of estimating equations which take into account the correlation between discrete longitudinal observations in order to estimate a set of time-independent parameters in the generalized linear model context. With the objective of estimating a functional variable in a future period of time from its past evolution, Aguilera *et al.* (1997a) introduced the principal component prediction models (PCP)

* Corresponding author. E-mail: mescabias@ugr.es

ISSN 1048-5252 print; ISSN 1029-0311 online © 2004 Taylor & Francis Ltd
DOI: 10.1080/10485250310001624738

that have been adapted for forecasting a continuous-time series based on discrete observations at unequally spaced time points (Aguilera *et al.*, 1999a, b). More recently, Valderrama *et al.* (2002) have formulated mixed ARIMA-PCP models to solve a related problem.

On the other hand, in many practical situations it is necessary to model the relationship between a time-independent response and a functional predictor. A first approach to functional linear regression has been introduced by Ramsay and Silverman (1997) for the case of a continuous scalar response variable and the functional predictor measured at the same time points for each individual. When the response variable is the own functional predictor in a future time point, Aguilera *et al.* (1997b) proposed a principal component (PC) approach to functional linear regression.

The aim of this work is to predict a binary response variable, or equivalently the probability of occurrence of an event, from the evolution of a continuous variable in time, so that functional information is provided by a set of sample paths. This is the case in Ratcliffe *et al.* (2002), who used periodically stimulated foetal heart rate tracings to predict the probability of a high risk birth outcome. Taking into account that the analysis of such phenomenon has to be done by looking at its continuous nature we propose a functional regression model. On the other hand linear models are not the best ones when the response variable is binary as stated by Hosmer and Lemeshow (1989) in the multiple case, therefore we develop a functional logistic regression model.

As Ramsay and Silverman (1997) stated for the case of functional linear models, the estimation of such a model cannot be obtained by the usual methods of least squares or maximum likelihood. In the general context of generalized functional linear models, James (2002) assumes that each predictor can be modeled as a smooth curve from a given functional family as for example, natural cubic splines. Then the functional model can be equivalently seen as a generalized linear model whose design matrix is given by the unobserved spline basis coefficients for the predictor and the EM algorithm is used for estimating the model from longitudinal observations at different times for each individual. In order to assess the relationship between the functional predictor and the scalar response, either estimating the parameter function by considering an orthonormal spline basis or using as covariates all the PCs of the spline basis coefficients is also performed. In an unpublished paper, Müller and Stadmüller (2002) propose an estimation based on truncating an infinite basis expansion of the functional predictor whose sample curves are assumed to be in the space of square integrable functions. In that paper, asymptotic inference for the proposed class of generalized regression models is developed by using the orthonormal basis of eigenfunctions of the covariance function.

In this line of research the present paper provides some alternative methods for estimating the parameter function in the functional logistic regression model. A first estimation procedure is obtained by assuming that both the functional predictor and the parameter function belong to a finite space generated by a basis of functions that could be unorthonormal. Opposite to James (2002) that assumes unknown basis coefficients we fit a separate curve to each individual and approximate its basis coefficients from its observations at different time points. Then an approximated estimation is performed by estimating a standard multiple logistic regression model whose design matrix is the product of the matrix of resulting basis coefficients and the one of inner products between basis functions. This results in highly correlated covariates (multicollinearity) which causes the model to be unstable, as indicated by Ryan (1997), so that the parameter estimates provided by most statistical packages are not particularly accurate. To solve this problem in multiple logistic regression Aguilera and Escabias (2000) proposed to use a reduced set of PCs of the original predictor variables as covariates of the model and to obtain an estimation of the original parameters through the one provided by this PC logistic regression model. In order to reduce dimension and to avoid multicollinearity, in this paper we propose a second estimation procedure based on functional principal component analysis

(FPCA) of the original sample paths to improve the approximated estimation of the parameter function obtained when we consider that the sample paths and the parameter function belong to the same finite-dimension space.

Summarizing, the main advantage of this paper with respect to previous works in this area is to propose an estimation procedure of functional logistic regression, based on taking as covariates a reduced set of functional principal components of the predictor sample paths after their approximation on a finite space generated by a basis of no necessarily orthonormal functions. This is in contrast to the particular case of the orthonormal bases used in other papers, when the basis is unorthonormal standard FPCA is not equivalent to principal component analysis (PCA) of the resulting basis coefficients so that two different approaches to FPCA will be considered. Then their estimating performance will be evaluated on simulated data by proposing different criterions and some heuristic guidelines for choosing the optimal PCs to be included as covariates in the resulting functional PCs logistic regression model.

In Section 2, we briefly present the general ideas about how to reconstruct the true functional form of data by fitting a smooth curve to each individual from its observations at different times. Basic theory on FPCA is also summarized with special emphasis on the case of predictor curves in a finite space generated by a basis. The theoretical framework on functional logistic regression and a first estimation procedure based on approximating the basis coefficients of predictor curves on a finite space are developed in Section 3. In order to avoid multicollinearity, in Section 4 we propose a second estimation procedure for functional logistic regression that uses as covariates a reduced set of functional principal components of the predictor curves. Two different forms of FPCA and two different criterions for including the covariates in the model will be also considered. Simulation studies for testing the accuracy of the parameter function provided by these functional PC approaches are finally carried out in Section 5.

2 FUNCTIONAL DATA

The aim of this paper is to explain a binary response variable in terms of a functional predictor whose sample information is given by a set of functions $x_1(t), x_2(t), \ldots, x_n(t)$ that can be seen as observations of a stochastic process $\{X(t): t \in T\}$. From now on, we will suppose that this stochastic process is second order, continuous in quadratic mean and the sample functions belong to the Hilbert space $L^2(T)$ of squared integrable functions, with the usual inner product

$$\langle f, g\rangle_u = \int_T f(t)g(t)\,\mathrm{d}t, \quad \forall f, g \in L^2(T).$$

In practice it is impossible to observe a set of functions continuously in time: instead we will usually have observations of such functions at different times, $t_{i0}, t_{i1}, \ldots, t_{im_i} \in T$, $i = 1, \ldots, n$, and with a different number of observations for each individual. That is, the sample information is given by the vectors $x_i = (x_{i0}, \ldots, x_{im_i})'$, with x_{ik} being the observed value of the ith sample path at time $t_{ik} (k = 0, \ldots, m_i)$.

In order to reconstruct the functional form of sample paths from the discrete observed information we may use different methods depending on the way of obtaining such discrete-time data and the shape we expect the functions to have. In any case, it is usual to assume that sample paths belong to a finite-dimension space generated by a basis $\{\phi_1(t), \ldots, \phi_p(t)\}$, so that they can be expressed as

$$x_i(t) = \sum_{j=1}^{p} a_{ij}\phi_j(t), \quad i = 1, \ldots, n. \tag{1}$$

If the functional predictor is observed with error

$$x_{ik} = x_i(t_{ik}) + \varepsilon_k \quad k = 0, \ldots, m_i,$$

we can use some least square approximation approach after choosing a suitable basis as, for example, trigonometric functions (Aguilera *et al.*, 1995), B-splines or wavelets (see Ramsay and Silverman, 1997 for a detailed study). On the other hand if we consider that sample curves are observed without error

$$x_{ik} = x_i(t_{ik}) \quad k = 0, \ldots, m_i,$$

we could use some interpolation method, as for example, cubic spline interpolation (Aguilera *et al.*, 1996).

Both methods, smoothing and interpolation, allow us to obtain the functional form of the sample paths by approximating the basis coefficients $\{a_{ij}\}$ from the discrete-time observations of sample curves.

The different methods that we will propose later for estimating the functional logistic regression model are based on FPCA, so let us briefly look at the basic theory on functional PCs.

2.1 Functional Principal Component Analysis

We are going to define FPCA from a sample point of view. By analogy with multivariate case, the functional PCs of a stochastic process are defined as uncorrelated generalized linear span of its variables with maximum variance.

Let $x_1(t), \ldots, x_n(t) \in L^2(T)$ be a set of sample paths of a continuous stochastic process $\{X(t)\colon t \in T\}$. The sample mean and covariance functions are given by

$$\bar{x}(t) = \frac{1}{n}\sum_{i=1}^{n} x_i(t)$$

$$\hat{C}(s,t) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i(s) - \bar{x}(s))(x_i(t) - \bar{x}(t)).$$

We will consider without loss of generality that $\bar{x}(t) = 0\,\forall t \in T$.

The functional PCs of $x_1(t), \ldots, x_n(t)$ are defined as n-dimensional vectors ξ_j $(j = 1, \ldots, n-1)$ with components

$$\xi_{ij} = \int_T x_i(t) f_j(t)\,\mathrm{d}t, \quad i = 1, \ldots, n,$$

where the weight functions $f_j(t)$ that define the functional PCs are the $n-1$ solutions of the eigenproblem

$$\int_T \hat{C}(s,t) f(s)\,\mathrm{d}s = \lambda f(t), \quad \forall t \in T. \tag{2}$$

That is, $f_j(t)$ $(j = 1, \ldots, n-1)$ are the eigenfunctions associated to the corresponding positive eigenvalues λ_j $(j = 1, \ldots, n-1)$ that are the variances of the PCs and verify $\lambda_1 > \lambda_2 > \cdots > \lambda_{n-1} \geq 0$.

In terms of the eigenfunctions the covariance function admits the following expansion:

$$\hat{C}(s,t) = \sum_{j=1}^{n-1} \lambda_j f_j(s) f_j(t)$$

that provides the following orthogonal representation of sample paths in terms of PCs:

$$x_i(t) = \sum_{j=1}^{n-1} \xi_{ij} f_j(t), \quad i = 1, \ldots, n.$$

By truncating this expression we obtain a reconstruction of the sample paths in terms of a reduced number of PCs that accumulate a certain percentage of the total variance given by

$$\mathrm{TV} = \sum_{j=1}^{n-1} \lambda_j.$$

In order to obtain the functional PCs of a sample of curves $x_1(t), \ldots, x_n(t)$, we have to solve the eigenequation (2). The solution of such an equation is only possible for a few types of kernels (see Todorovic, 1992). As we can see in Aguilera *et al.* (2002), when the sample paths belong to a finite space of $L^2(T)$ generated by a basis $\{\phi_1(t), \ldots, \phi_p(t)\}$, the functional PCs of such sample paths are given by the standard PCs of the $A\Psi^{1/2}$ matrix, where $A = (a_{ij})_{n\times p}$ is the matrix that has as rows the basis coefficients of sample paths and $\Psi = (\psi_{jk})_{p\times p}$ the one of inner products between basis functions

$$\psi_{jk} = \int_T \phi_j(t)\phi_k(t)\,\mathrm{d}t, \quad j, k = 1, \ldots, p. \tag{3}$$

In that sense let us consider the sample paths as in Eq. (1), let $\Gamma = (\xi_{ij})_{n\times p}$ be the matrix whose columns are the PCs of the $A\Psi^{1/2}$ matrix, and G the one whose columns are its associated eigenvectors. Then $\Gamma = (A\Psi^{1/2})G$ and the weight functions that define the functional PCs are given by

$$f_j(t) = \sum_{t=1}^{p} f_{lj}\phi_l(t), \quad j = 1, \ldots, p \tag{4}$$

with $(f_{lj})_{p\times p} = F = \Psi^{-1/2}G$.

3 THE FUNCTIONAL LOGISTIC REGRESSION MODEL

Several authors have established different models to explain a single time-independent response variable based on a set of functions. A general overview of the functional linear model for a continuous response variable can be seen in Ramsay and Silverman (1997). As in the multiple case, linear models cannot be used when the response variable is binary; therefore we are going to formulate a functional version of logit regression.

Let us consider a sample of functional observations $x_1(t), \ldots, x_n(t)$ and let $y_1, \ldots, y_n$ be a random sample of a binary response variable Y associated to these sample paths, that is, $y_i \in \{0, 1\}$, $i = 1, \ldots, n$. Then the functional logistic regression model is given by

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

where π_i is the expectation of Y given $x_i(t)$ that will be modelized as

$$\begin{aligned}\pi_i &= P[Y = 1 | \{x_i(t) \colon t \in T\}] \\ &= \frac{\exp\{\alpha + \int_T x_i(t)\beta(t)\,\mathrm{d}t\}}{1 + \exp\{\alpha + \int_T x_i(t)\beta(t)\,\mathrm{d}t\}}, \quad i = 1, \ldots, n\end{aligned}$$

with α a real parameter, $\beta(t)$ a parameter function, and $\varepsilon_i, i = 1, \ldots, n$ independent errors with zero mean.

Equivalently the logit transformations can be expressed as

$$l_i = \ln\left[\frac{\pi_i}{1 - \pi_i}\right] = \alpha + \int_T x_i(t)\beta(t)\,\mathrm{d}t, \quad i = 1, \ldots, n, \tag{5}$$

where T is the support of the sample paths $x_i(t)$. So the functional logistic regression model can be seen as a functional generalized linear model with the logit transformation as link function (James, 2002).

As indicated by Ramsay and Silverman (1997) for the functional linear regression model, it is impossible to obtain an estimation of $\beta(t)$ by the usual methods of least squares or maximum likelihood. As we stated earlier, in order to solve this problem we will assume that both sample paths and parameter function belong to the same finite space.

3.1 An Approximated Estimation of the Parameter Function

Let us consider the sample paths as in Eq. (1), and the parameter function in terms of the same basis,

$$\beta(t) = \sum_{k=1}^{p} \beta_k \phi_k(t).$$

Once the basis coefficients of sample curves have been approximated from discrete-time data, their resulting approximated basis representations turn the functional model into a multiple one

$$l_i = \alpha + \int_T x_i(t)\beta(t)\,\mathrm{d}t = \alpha + \sum_{j=1}^{p}\sum_{k=1}^{p} a_{ij}\psi_{jk}\beta_k, \quad i = 1, \ldots, n$$

with ψ_{jk} the inner products defined in Eq. (3), a_{ij} the coordinates of sample paths, α a scalar parameter and β_k the parameters of the model that correspond to the parameter function basis coefficients which have to be estimated. In matrix form this multiple model can be expressed as

$$L = \alpha\mathbf{1} + A\Psi\beta, \tag{6}$$

where $A = (a_{ij})_{n\times p}$, $\Psi = (\psi_{jk})_{p\times p}$, $\beta = (\beta_1, \ldots, \beta_p)'$, $\mathbf{1} = (1, \ldots, 1)'$ and $L = (l_1, \ldots, l_n)'$.

Then, we will obtain an estimation of the parameter function by estimating this multiple logistic model. So let $\hat{\beta} = (\hat{\beta}_1, \ldots, \hat{\beta}_p)'$ the estimated parameters of this multiple model obtained by applying the Newton-Raphson method for solving the nonlinear likelihood equations

$$X'(Y - \Pi) = 0,$$

where $X = (\mathbf{1}|A\Psi)$, $Y = (y_1, \ldots, y_n)'$ and $\Pi = (\pi_1, \ldots, \pi_n)'$ (see for example, Hosmer and Lemeshow, 1989). Then the approximated parameter function estimate will be

$$\hat{\beta}(t) = \sum_{k=1}^{p} \hat{\beta}_k \phi_k(t). \tag{7}$$

In the more general context of functional generalized linear model, James (2002) has proposed a different estimation procedure where unobserved basis coefficients of sample paths are treated as missing data and the EM algorithm is used in order to optimize the observed likelihood. In that paper the relationship between predictor and response variable is investigated either considering an orthogonal spline basis ($\Psi = I_p$) and plotting the parameter function or using all the PCs of the coefficient matrix A as predictors and interpreting the resulting parameters.

As many authors have stated (Hosmer and Lemeshow, 1989 and Ryan, 1997 among others) the estimation of the logistic regression model obtained by most statistical packages is not very accurate when there is a great dependence between the variables in the model or a high correlation in the columns of the design matrix (multicollinearity).

The way of obtaining the design matrix of model (6) from correlated time-dependent data induces a great multicollinearity in this model. Then the approximated estimation proposed before will not be very good, as we will see at the end of this paper in the simulation study.

Aguilera and Escabias (2000) proposed to use a reduced number of PCs of the original variables as covariates of the multiple logistic model in order to obtain a better estimation of the parameters of this model in the presence of multicollinearity. In the next section we will use an analogous method of improving the estimation of the parameter function of the functional logistic model based on FPCA.

4 FUNCTIONAL PRINCIPAL COMPONENT LOGISTIC REGRESSION

In order to reduce dimension and to obtain better estimations of the parameter function, we propose to use two different approaches based on FPCA of sample paths, that is, we are going to reduce the functional logistic regression model to a multiple one with a reduced number of PCs as covariates.

On the one hand let us consider the multiple logistic model (6) obtained by considering that the sample paths and the parameter function of model (5) belong to the finite space generated by the basis $\{\phi_1(t), \ldots, \phi_p(t)\}$. As we have noted before, this model provides an estimation of the coordinates of the parameter function that may be very different to the real one due to the existence of multicollinearity. Following Aguilera and Escabias (2002) we propose to solve this problem by using as covariates of this model the PCs of its design matrix $A\Psi$.

Let $\Gamma^{(1)} = (\xi_{ij}^{(1)})_{n\times p}$ be the matrix of PCs of the $A\Psi$ design matrix and $V^{(1)}$ the one that has as columns the eigenvectors associated to them. Then the multiple logistic model (6) can be equivalently expressed in terms of all the PCs as

$$L = \alpha\mathbf{1} + A\Psi\beta = \alpha\mathbf{1} + \Gamma^{(1)}V^{(1)'}\beta = \alpha\mathbf{1} + \Gamma^{(1)}\gamma^{(1)},$$

and we can give an estimation of the parameters of model (6) (coordinates of $\beta(t)$) through the estimation of this one,

$$\hat{\beta} = V^{(1)}\hat{\gamma}^{(1)},$$

which is the same as the one obtained by using the original $A\Psi$ matrix.

This PCA is really a FPCA of a transformation of the original sample paths with respect to the usual inner product in $L^2(T)$. In fact, Aguilera *et al.* (2002) have proved that this transformation is given by $U(x_i)(t) = \Phi'\Psi^{1/2}a_i'$ with $a_i = (a_{i1}, \dots, a_{ip})'$ being the ith row of A and $\Phi = (\phi_1(t), \dots, \phi_p(t))'$.

On the other hand, let us denote by $\Gamma^{(2)} = (\xi_{ij}^{(2)})_{n\times p}$ the matrix of functional PCs of $x_1(t), \dots, x_n(t)$ obtained as in Section 2.1; then we can express

$$x_i(t) = \sum_{j=1}^{p} \xi_{ij}^{(2)} f_j(t), \quad i = 1, \dots, n$$

and model (5) can be equivalently expressed as

$$l_i = \alpha + \sum_{j=1}^{p} \xi_{ij}^{(2)} \gamma_j^{(2)}, \quad i = 1, \dots, n,$$

with

$$\gamma_j^{(2)} = \int_T f_j(t)\beta(t)\,\mathrm{d}t, \quad j = 1, \dots, p.$$

Equivalently, if we consider the multiple logistic model in terms of these PCs

$$\begin{aligned} l_i &= \alpha + \sum_{j=1}^{p} \xi_{ij}^{(2)} \gamma_j^{(2)} = \alpha + \sum_{j=1}^{p} \left(\int_T x_i(t) f_j(t)\,\mathrm{d}t \right) \gamma_j^{(2)} \\ &= \alpha + \int_T x_i(t) \left(\sum_{j=1}^{p} f_j(t)\gamma_j^{(2)} \right) \mathrm{d}t, \quad i = 1, \dots, n, \end{aligned}$$

we obtain the following expression for the parameter function

$$\beta(t) = \sum_{j=1}^{p} f_j(t)\gamma_j^{(2)}.$$

Under these considerations, we can obtain an estimation of the parameter function by estimating the multiple model in terms of all PCs and substituting $f_j(t)$, $j = 1, \dots, p$ with its basis expansion (4). Then its basis coefficients are given by

$$\hat{\beta}_j = \sum_{l=1}^{p} f_{lj}\hat{\gamma}_j^{(2)}, \quad j = 1, \dots, p$$

and $(f_{lj})_{p\times p} = F$ defined in Section 2.1 in terms of the eigenvectors of the covariance matrix of $A\Psi^{1/2}$. In matrix form $\hat{\beta} = F\hat{\gamma}^{(2)} = \Psi^{-1/2}G\hat{\gamma}^{(2)}$.

Summarizing, we have proved that the functional logistic regression model (5) can be equivalently expressed as

$$l_i = \alpha + \sum_{j=1}^{p} \xi_{ij}^{(k)} \gamma_j^{(k)}, \quad i = 1, \dots, n, \ \ k = 1, 2, \tag{8}$$

where the functional PCs $\xi_j^{(k)} (k = 1, 2)$ can be obtained by the following two different approaches:

1. Performing PCA of the design matrix $A\Psi$ of the multiple logistic model (6) with respect to the usual inner product in $\mathbb{R}^p$ that is equivalent to FPCA of a transformation of the sample paths with respect to the usual inner product in $L^2(T)$, that is, $\Gamma^{(1)} = A\Psi V^{(1)}$. We will call this principal component analysis PCA1.
2. Performing FPCA of the sample paths with respect to the usual inner product in $L^2(T)$ that is equivalent to PCA of the data matrix $A\Psi^{1/2}$, that is $\Gamma^{(2)} = A\Psi^{1/2}G$. We will call this principal component analysis PCA2.

In any case, an estimation of the parameter function can be given by

$$\hat{\beta}(t) = \sum_{j=1}^{p} \hat{\beta}_j \phi_j(t) = \hat{\beta}'\Phi, \tag{9}$$

where the vector of coordinates $\hat{\beta} = V^{(k)}\hat{\gamma}^{(k)}$ $(k = 1, 2)$, with $V^{(1)}$ the matrix whose columns are the eigenvectors of the covariance matrix of $A\Psi$ and $V^{(2)} = F\Psi^{-1/2}G$, with G being the matrix whose columns are the eigenvectors of the covariance matrix of $A\Psi^{1/2}$.

Let us observe that both approaches agree when the basis is orthonormal and then $\Psi = I_p$.

4.1 Model Formulation

The functional principal component logistic regression model is obtained by truncating model (8) in terms of a subset of PCs. Then, if we consider the matrices defined before partitioned as follows:

$$\Gamma^{(k)} = (\Gamma_{(s)}^{(k)} \mid \Gamma_{(r)}^{(k)}), \quad V^{(k)} = (V_{(s)}^{(k)} \mid V_{(r)}^{(k)}), \quad r + s = p, \quad k = 1, 2$$

the functional principal component logistic regression model is defined by taking as covariates the first s PCs

$$L_{(s)}^{(k)} = \alpha_{(s)}^{(k)}\mathbf{1} + \Gamma_{(s)}^{(k)}\gamma_{(s)}^{(k)}, \quad k = 1, 2,$$

where $\alpha_{(s)}^{(k)}$ is a real parameter and $L_{(s)}^{(k)} = (l_{1(s)}^{(k)}, \ldots, l_{n(s)}^{(k)})$ with $l_{i(s)}^{(k)} = \ln[\pi_{i(s)}^{(k)}/(1 - \pi_{i(s)}^{(k)})]$, and

$$\pi_{i(s)}^{(k)} = \frac{\exp\{\alpha_{(s)}^{(k)} + \sum_{j=1}^{s} \xi_{ij}^{(k)}\gamma_{j(s)}^{(k)}\}}{1 + \exp\{\alpha_{(s)}^{(k)} + \sum_{j=1}^{s} \xi_{ij}^{(k)}\gamma_{j(s)}^{(k)}\}}, \quad i = 1, \ldots, n, \quad k = 1, 2. \tag{10}$$

It will be seen in the simulation study that the estimation of the parameter function given by

$$\hat{\beta}_{(s)}^{(k)}(t) = \sum_{j=1}^{p} \hat{\beta}_{j(s)}^{(k)}\phi_j(t) = \hat{\beta}_{(s)}^{(k)'}\Phi, \quad k = 1, 2 \tag{11}$$

with the coefficient vector $\hat{\beta}_{(s)}^{(k)} = V_{(s)}^{(k)}\hat{\gamma}_{(s)}^{(k)}$ is more accurate than the one obtained with the original $A\Psi$ design matrix.

4.2 Model Selection

In spite of having included the PCs in the model according to explained variability, several authors, for example Hocking (1976), in PC regression, have established that PCs must be included based on a different criterion by taking into account their predictive ability. As Aguilera and Escabias (2000) established for the multiple case, it is better to include the PCs in the model in the order given by the stepwise method based on conditional likelihood ratio test. In the context of generalized linear models, Müller and Stadmüller (2002) select the number of predictors (base functions) in a sequence of nested models by using the standard Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC).

In the simulation study performed at the end of this paper we will consider two different methods of including PCs in the model. The first one (Method I) consists of including PCs in the natural order given by the explained varability. In the second one (Method II) PCs are included in the order given by the stepwise method based on conditional likelihood ratio test. This method consists of a stepwise procedure of including variables (PCs) into a regression model. Each step of the method tests whether each of the available variables ought to be in the present model by the likelihood ratio test for selecting one between two nested models: the ones with and without the corresponding variable. The procedure begins with the model with no variables.

In addition we will propose different criterions for selecting the optimal number of PCs based on different accuracy measures of the estimated parameters. First, we define the *integrated mean squared error of the beta parameter function* (IMSEB) as

$$\begin{aligned}\text{IMSEB}_{(s)}^{(k)} &= \frac{1}{T}\int_T (\beta(t)\text{—}\hat{\beta}_{(s)}^{(k)}(t))^2\,\mathrm{d}t\\ &= \frac{1}{T}(\beta-\hat{\beta}_{(s)}^{(k)})'\Psi(\beta-\hat{\beta}_{(s)}^{(k)}) \quad k=1,2.\end{aligned}$$

Second, we define the *mean squared error of beta parameters* (MSEB)

$$\text{MSEB}_{(s)}^{(k)} = \frac{1}{p+1}\left((\alpha-\hat{\alpha}_{(s)}^{(k)})^2+\sum_{j=1}^{p}(\beta_j-\hat{\beta}_{j(s)}^{(k)})^2\right), \quad k=1,2.$$

Let us observe that mean squared errors provide a good estimation when they are small enough. Therefore, in order to select the optimum number of PCs we will propose two optimum models for each method: the one with the smallest MSEB and the one with the smallest IMSEB.

We have to note that we cannot obtain IMSEB and MSEB with real data, in which case we need another measure that chooses the best estimation of the parameter function. In literature we can see several methods of choosing the best number of covariates to estimate the parameter of a regression model, some of which take into account the values of the variance of the estimated parameters (see for example Aucott *et al.* 2000). This *variance of the estimated parameters* is given in functional principal component logistic regression by

$$\text{var}_{(s)}^{(k)} = \text{var}[\hat{\beta}_{(s)}^{(k)}] = V_{(s)}^{(k)\prime}(\Gamma_{(s)}^{(k)\prime}W_{(s)}^{(k)}\Gamma_{(s)}^{(k)})^{-1}V_{(s)}^{(k)}, \quad k=1,2,$$

where $W_{(s)}^{(k)}=\text{diag}(\hat{\pi}_{i(s)}^{(k)}(1-\hat{\pi}_{i(s)}^{(k)}))$. We have observed in the simulated examples that the model previous to a significant increase in this variance provides a good estimation of the parameter function with mean squared errors similar to the optimum models. Then we will use this criterion for model selection in applications with real data. As heuristic rule for choosing the model with the variance criterion we propose to plot the variances and retain in the model the PCs previous to a kink in the curve. In addition we will take into account the parsimonious criterion so that between two significant increases in the variances we prefer the model with less PCs.

5 SIMULATION STUDY

In order to investigate the improvement in the parameter function estimation of the functional logistic regression model provided by the two proposed PC approaches, we have carried out a simulation study.

5.1 Simulation Process and Model Fitting

We have performed two simulations with the same drawing and a different procedure of sample curves simulation. Due to the fact that all presented methods consider that sample paths belong to a finite space generated by a basis, we have used the space of cubic splines generated by a basis of cubic B-splines defined by a set of nodes.

Before developing the simulated study let us briefly see the definition of the B-spline basis given by a set of nodes $t_0 < t_1 < \cdots < t_m$. Extending this partition as $t_{-3} < t_{-2} < t_{-1} < t_0 < \cdots < t_m < t_{m+1} < t_{m+2} < t_{m+3}$ the B-spline basis is recursively defined by

$$B_{j,1}(t) = \begin{cases} 1 & t_{j-2} < t < t_{j-1} \\ 0 & t \notin (t_{j-2}, t_{j-1}) \end{cases}, \quad j = -1, \ldots, m+3$$

$$B_{j,r}(t) = \frac{t - t_{j-2}}{t_{j+r-3} - t_{j-2}} B_{j,r-1}(t) + \frac{t_{j+r-2} - t}{t_{j+r-2} - t_{j-1}} B_{j+1,r-1}(t),$$
$$r = 2, 3, 4, \quad j = -1, \ldots, m - r + 5$$

(more details in De Boor 1978). From now we will omit the subscript corresponding to the degree of cubic B-spline functions ($r = 4$).

The first step of the simulation process in each case was to have a set of n sample curves that will explain the response. These curves are considered in the form (1) in terms of the basic cubic B-splines and their basis coefficients (A matrix). In the first example we have directly simulated these coefficients. In the second we have approximated them by least squares approximation from a set of observations of a known stochastic process simulated at a finite set of times.

After this we have chosen as parameter functions in the two examples the natural cubic spline interpolation of a known function on the set of nodes of definition of B-splines. Then the real basis coefficients $\beta = (\beta_1, \ldots, \beta_p)'$ of the parameter function are known.

Finally we have obtained the values of the response variable Y by simulating the functional logistic regression model. That is, first we have obtained the linear spans

$$c_i = \int_T x_i(t)\beta(t)\,\mathrm{d}t = a_i'\Psi\beta, \quad 1, \ldots, n,$$

where Ψ is the matrix of inner products between the basic B-splines. Then we have calculated the probabilities

$$\pi_i = \frac{\exp\{\alpha + c_i\}}{1 + \exp\{\alpha + c_i\}}, \quad i = 1, \ldots, n,$$

where α is fixed, and finally we have obtained n values of the response by simulating observations of a Bernouilli distribution with probabilities π_i.

Once we had the simulated data, the next step has been to obtain the approximated estimation of the parameter function in each example $\hat{\beta}(t) = \sum_{k=1}^{p} \hat{\beta}_k \phi_k(t)$ by estimating the parameters of the multiple model (6). In order to improve these estimations we have considered the two solutions seen in previous sections: PCA1 (PCA of $A\Psi$) and PCA2 (PCA of $A\Psi^{1/2}$). In both cases, we have fitted the logistic model taking as covariates the different set of PCs included by using the proposed Methods I and II. Then we have obtained for all these fittings the estimation of the parameter function given by Eq. (11) and the accuracy measures seen in Section 4.2.

In each fit we have calculated goodness of fit measures as the correct classification rate (CCR, rate of observations correctly classified using 0.5 as cutpoint) and the goodness of fit statistic G^2 with its associated p-value (more details in Hosmer and Lemeshow, 1989). All these measures show good fits, so we do not present their values.

Finally we have repeated all these steps a great number of times in each example. In order to select the optimum number of PCs we propose in each repetition two optimum models: the one with the smallest MSEB and the one with the smallest IMSEB. As resume we have calculated the mean of the different accuracy measures of optimum models and some measures of spread (variance and variation coefficient).

All calculations have been obtained by using the S-Plus2000 package, and the results are shown in the tables. In such tables the sub and upper scripts corresponding to the number of PCs and the type of PCA used have been omitted in the different measures. Such questions are indicated in captions.

5.2 Example 1

In the first example we have considered that the functional predictor has as sample observations cubic spline functions on the real interval $T = [0, 10]$ and can be expressed in terms of the cubic B-splines $\{B_{-1}(t), B_0(t), \ldots, B_{10}(t), B_{11}(t)\}$ defined by the partition of such interval $0 < 1 < \cdots < 9 < 10$. Then we have simulated $n = 100$ observations of this magnitude by simulating their basis coefficients (A matrix). Matrix A has been obtained by a linear transform of 100 vectors of 13 independent standard normal distributions, by using a 13×13 matrix of uniformly distributed values in the interval [0, 1].

The parameter function is the natural cubic spline interpolation of function $\sin(t - \pi/4)$ on the nodes $0 < 1 < \cdots < 10$ that define the basic elements. The basis coefficients $\beta = (\beta_1, \ldots, \beta_{13})'$ are in Table I. Finally the values of the response have been simulated as indicated in the previous section.

TABLE I Simulated Coordinates of the Parameter Function and Their Corresponding Estimations Without Using PCs for Example 1. Cumulative Variances of PCA1 and PCA2.

	Parameters			*Cumulative variance*	
	Simulated	*Estimated*	*PCs*	*PCA1*	*PCA2*
β_1	−1.63	1.92E+4	1	64.25	91.99
β_2	−0.71	−3.01E+3	2	76.10	95.56
β_3	0.22	1.10E+3	3	82.20	96.90
β_4	1.22	−5.04E+2	4	87.50	97.90
β_5	0.94	1.75E+2	5	92.53	98.80
β_6	−0.09	−2.03E+1	6	95.91	99.47
β_7	−1.04	−5.20E+1	7	97.82	99.75
β_8	−1.04	1.07E+2	8	98.99	99.90
β_9	−0.08	1.44E+2	9	99.43	99.95
β_{10}	0.95	2.95E+2	10	99.82	99.99
β_{11}	1.12	−6.39E+2	11	100.00	100.00
β_{12}	0.21	1.80E+3	12	100.00	100.00
β_{13}	−0.70	−1.19E+4	13	100.00	100.00

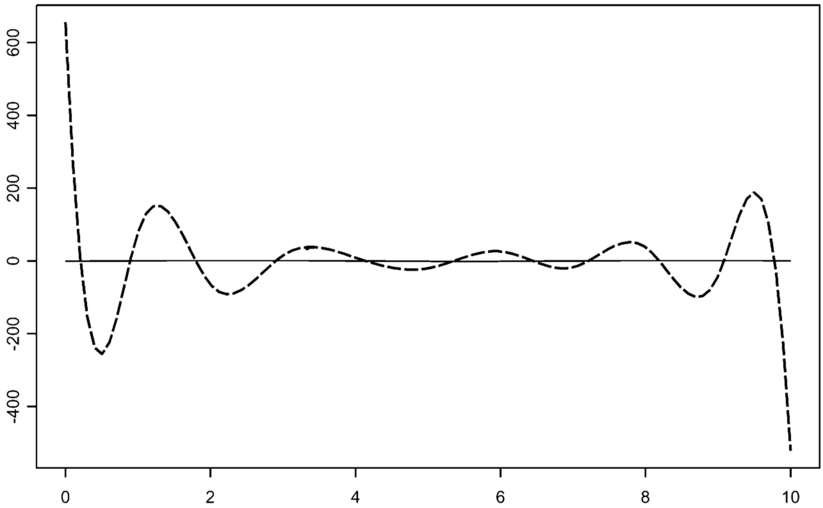


FIGURE 1 Graphic of the simulated parameter function (——) and its estimation (– – – –) without using PCs.

The estimation of the basis coefficients of the parameter function obtained without using PCs is given in Table I and as we can see they are very different to the simulated ones due to the great multicollinearity that exists in the approximated design matrix $(\mathbf{1}|A\Psi)$. Figure 1 shows the great difference between the simulated and estimated parameter functions. In spite of this bad estimation, the model fits well with CCR of 90%, $G^2 = 44.8$ and p-value $= 1.00$. The importance of obtaining a good estimation of the parameter function led us to improve it by using FPCA as proposed in previous section.

First, we have obtained the matrices $\Gamma^{(1)}$ and $\Gamma^{(2)}$ of PCs of the $A\Psi$ and $A\Psi^{1/2}$ matrices, respectively. These PCA1 and PCA2 approaches provide the explained variances given in Table I. Let us observe that the first five PCs explain more than 90% of the total variability and with the first eight we reach nearly 99% in PCA1. However in PCA2 more than 91% of total variability is explained with only one PC, and nearly 99% with the first five.

Now, we have fitted the multiple models with a different number of PCs by including them in the model according to the different methods seen in previous sections: Methods I and II. The results of the different measures of accuracy of estimations can be seen in Table II for Method I and Table III for Method II.

Let us observe from Table II the great increase of all measures obtained when we include the last PCs in the model, which shows the need of choosing a smaller number of PCs in order to obtain a good estimation of the parameter function. In addition all models fit well with CCR next to 90% in most cases and p-value of G^2 nearly 1.00.

We can see that in PCA1 the minimum values of IMSEB and MSEB are the ones corresponding to the model which has the first four PCs included by Method I (Tab. II) and the model with the first, third and fourth PCs included by Method II (Tab. III). Then these models provide the best possible estimations using this type of PCs with similar values of the accuracy measures. This is not the case in PCA2, which by using Method I (Tab. II) IMSEB indicates that the best estimation is the one obtained with the model that includes the first five PCs, whereas MSEB indicates that the best model is the one with only the first PC. Method II (Tab. III) shows again this difference between MSEB and IMSEB for PCA2, being the optimum models those with two PCs (first and fifth) and three (first, fifth and third) respectively. The estimated parameter functions given by these optimum models can be seen in Figure 2(a), (b) and (c). Let

TABLE II Accuracy Measures of Functional Logit Models with Different Number of PCs as Covariates for Example 1 by Method I.

	PCA1			*PCA2*		
s	*IMSEB*	*MSEB*	*Var*	*IMSEB*	*MSEB*	*Var*
1	4.555	0.710	0.108	4.565	**0.755**	0.110
2	4.299	0.689	0.166	4.395	0.755	0.173
3	1.485	0.393	0.618	2.626	0.976	0.931
4	**0.272**	**0.263**	1.242	2.555	1.060	1.577
5	1.297	0.373	1.986	**1.313**	1.298	3.774
6	1.263	0.368	3.228	1.793	1.772	6.216
7	4.737	1.189	**9.626**	3.956	1.987	**14.450**
8	4.895	1.233	21.481	3.908	2.025	31.450
9	8.526	5.062	89.599	7.964	10.774	2.4E+2
10	23.417	14.780	2.0E+2	24.357	48.959	4.6E+2
11	34.771	35.790	7.2E+2	34.628	46.064	1.4E+3
12	4.4E+3	4.3E+5	7.5E+2	4.4E+3	4.3E+5	2.4E+6
13	**3.7E+5**	**3.8E+7**	8.1E+2	3.7E+5	3.8E+7	1.7E+8

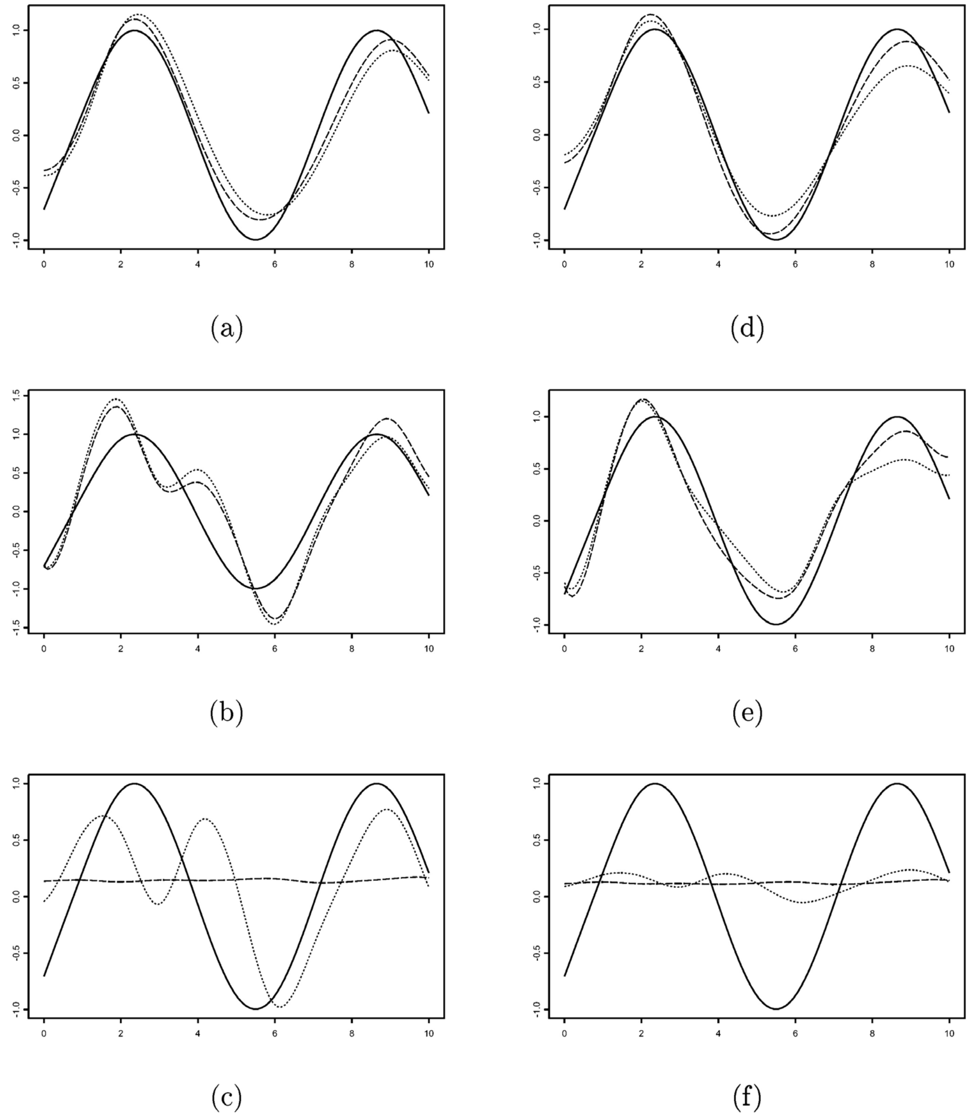
us observe from Figure 2(c) the bad estimation provided by the optimum models with PCA2 and MSEB criterion in both methods.

Looking at the estimated variance of the estimator of parameter function, we can see how it increases when we include the seventh PC in the model in both types of PCA and Method I of inclusion (Tab. II), in which case the best estimation that provides this criterion would be the one obtained with the model that contains the first 6 PCs. As we can see these models have values of MSEB and IMSEB similar to the optimum ones. These results can be clearly observed in Figure 3 where we can see a significant change (kink) in the linear trend of variances when we include the seventh PC. Let us observe how those extreme values that can confuse the interpretation have been removed in these plots. In Method II, (Tab. III) there is no increase in this variance, in which case the best model would be the one with all the PCs that enter in the model in each PCA. The values of MSEB and IMSEB are also similar to the optimum ones in this case. The estimated parameter functions provided by these models have been drawn in Figure 4.

From these results we can conclude that it is better to use PCA1 than PCA2 to improve the estimation of the parameter function in the functional logistic regression model. Moreover it is better to use Method II than Method I because similar estimations are obtained with less PCs. However if we decide to use PCA2 it is better to choose the IMSEB as criteria of selecting the best estimation. Finally, if our objective is to predict correctly, both PCAs are equally good.

Finally we have repeated the simulation and the fits 350 times. In each repetition, PCA approach and method of inclusion, we have considered two optimum models: the ones with the smallest IMSEB and MSEB. Then we have obtained, for each PCA, method of inclusion of PCs and measure (IMSEB and MSEB), the mean, variance and coefficient of variation (CV) of the 350 measures of the corresponding optimum models. The results can be seen in Tables IV and V.

Table IV, which corresponds to Method I, shows that the minimum MSEB and IMSEB provide the same mean results and variances in most measures with PCA1. For example, the mean number of optimum PCs is approximately 4 in both cases (MSEB and IMSEB) in PCA1, with similar variances and CV. The same happens in cumulative variance and the other accuracy measures. These results reveal that the conclusions provided in this situation by Table II are usual in this case.



(a) (d)

(b) (e)

(c) (f)

FIGURE 2 Example 1 – Graphics of the simulated parameter function (——) and its estimation with Method I (— — — —) and Method II (· · · · ·): (a) PCA1: 4 PCs in Method I and 3 in Method II (MSEB and IMSEB criteria), (b) PCA2: 5 PCs in Method I and 3 in Method II (IMSE criterion), (c) PCA2: 1 PC in Method I and 2 in Method II (MSEB criterion). Graphics of the simulated parameter function (——) and mean of its estimations with Method I (— — — —) and Method II (· · · · ·) after 350 repetitions: (d) PCA1 and MSEB criterion, (e) PCA2 and IMSEB criterion, (f) PCA2 and MSEB criterion.

TABLE III Accuracy Measures of Functional Logit Models with Different Number of PCs as Covariates for Example 1 by Method II.

	PCA1				*PCA2*			
s	*PCs*	*IMSEB*	*MSEB*	*Var*	*PCs*	*IMSEB*	*MSEB*	*Var*
1	1	4.555	0.710	0.108	1	4.565	0.755	0.110
2	3	1.848	0.422	0.521	5	2.968	**0.697**	0.980
3	4	**0.564**	**0.288**	1.137	3	**1.544**	1.236	2.858

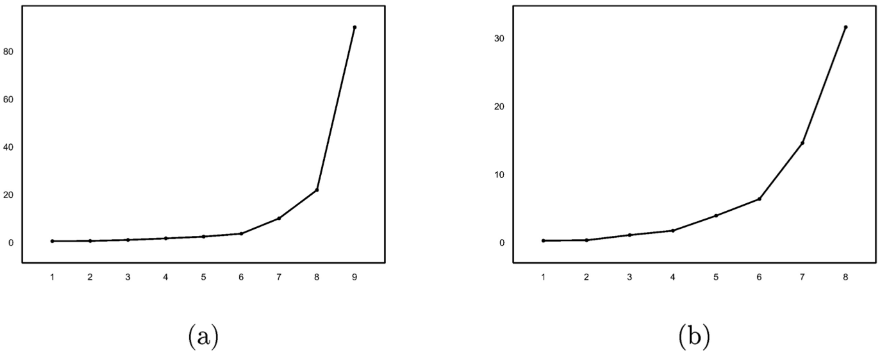


FIGURE 3 Example 1 – Graphics of the estimated variance of the parameters with Method I after removing extreme values: (a) PCA1, (b) PCA2.

In relation to PCA2 with Method I, the conclusions are the same as the ones we saw in the single simulation, now IMSEB and MSEB do not provide the same mean of optimum number of PCs (Tab. IV). Furthermore the mean number of PCs that provides IMSEB is similar to the one obtained with PCA1.

In Table V (Method II) the conclusions are the same as the ones presented for Table IV, in relation to the convenience of using the minimum IMSEB or MSEB for selecting the best estimation in PCA2. Moreover we can once again see a bigger reduction in the dimension of the problem than the one obtained in Table IV because the mean number of PCs needed is smaller than the one associated to Method I.

Finally we have calculated the mean of the optimum estimations of the parameter function for the optimum models provided by each type of PCA, method and criterion for selecting PCs (Fig. 2(d)–(f)). Let us observe again that PCA1 and Method II provide the best estimations.

5.3 Example 2

In the second example we have considered a different method of simulating the explicative sample paths. We have simulated observations of a set of 100 curves of a known stochastic process at the set of 21 equally spaced times of the [0, 10] interval (it is not necessary to take equally spaced times). More precisely we have considered the process used by Fraiman and Muñiz (2001) $X(t) = Z(t) + t/4 + 5E$ where $Z(t)$ is a zero mean Gaussian stochastic

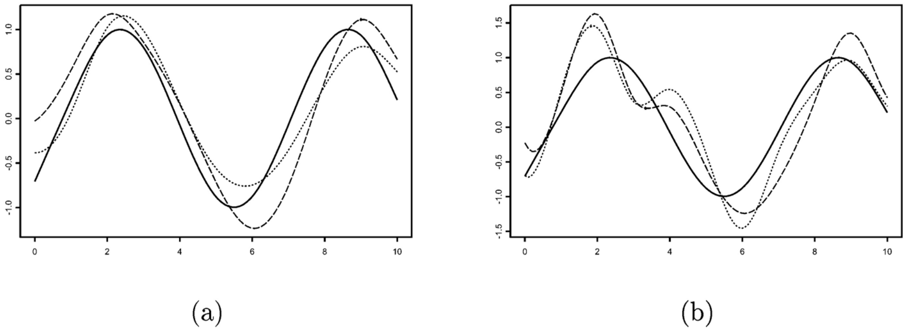


FIGURE 4 Example 1 – Graphics of the simulated parameter function (——) and its estimation given by Method I (– – – –) and Method II (· · · · ·) with variance criterion: (a) PCA1: 6 PCs in Method I and 3 in Method II, (b) PCA2: 6 PCs in Method I and 3 in Method II.

TABLE IV Means and Variabilities of Accuracy Measures for the Optimum Models of 350 Repetitions of Example 1 by Method I.

	Criterion for selecting the optimum model					
	MSEB			*IMSEB*		
Measures	*Mean*	*Variance*	*CV*	*Mean*	*Variance*	*CV*
PCA1						
No. PCs	**4.183**	**0.918**	**0.229**	**4.186**	**0.834**	**0.218**
Cum. Var.	**88.092**	19.477	0.050	**88.165**	18.395	0.049
IMSEB	0.964	0.589	0.796	0.950	0.580	0.801
MSEB	0.338	6.68E−3	0.242	0.339	6.76E−3	0.242
Var	0.005	7.65E−3	17.815	0.005	7.65E−3	17.641
PCA2						
No. PCs	**1.146**	0.325	0.498	**4.914**	1.620	0.259
Cum. Var.	**92.410**	1.592	0.014	98.550	0.900	0.010
IMSEB	4.552	0.064	0.055	1.531	0.737	0.561
MSEB	0.752	1.45E−4	0.016	1.676	0.370	0.363
Var	0.000	0.000	0.000	0.141	3.502	13.272

process with covariance function given by $C(t, s) = (1/2)^{80|t-s|}$ and E is a Bernouilly random variable with parameter $p = 0.1$.

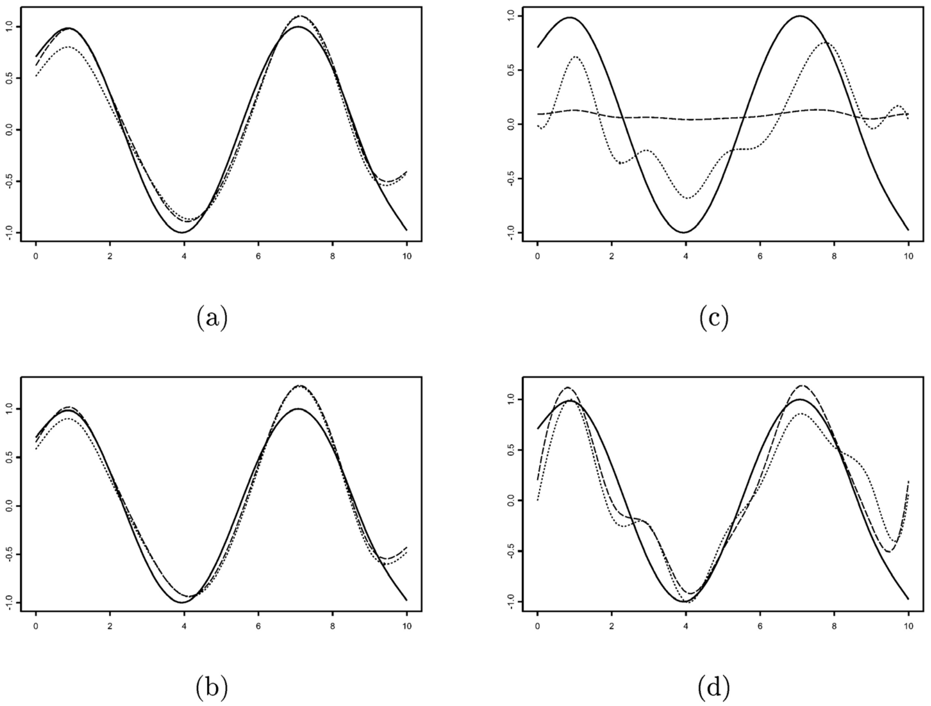
In this case we consider that the simulated observations are data measured with some error. In order to reconstruct the functional form of the curves we consider that they are expressed as in Eq. (1) in terms of the basic B-splines defined by 11 equally spaced nodes of the [0, 10] interval. The basis coefficients of the corresponding expansion are obtained by least squares approximation.

The parameter function considered in this example is the cubic spline interpolation of $\sin(t + \pi/4)$ on the nodes that define the B-splines.

The following steps for simulating the response and fitting the different models are the same as those seen in the first example. In this case we have also repeated the simulation and the fits 500 times and we have obtained the same resume measures for the optimum models with

TABLE V Means and Variabilities of Accuracy Measures for the Optimum Models of 350 Repetitions of Example 1 by Method II.

	Criterion for selecting the optimum model					
	MSEB			*IMSEB*		
Measures	*Mean*	*Variance*	*CV*	*Mean*	*Variance*	*CV*
PCA1						
No. PCs	**2.811**	**0.629**	**0.282**	**2.851**	**0.637**	**0.280**
IMSEB	1.624	0.995	0.614	1.618	0.992	0.615
MSEB	0.406	0.011	0.258	0.406	0.011	0.258
Var	0.514	0.197	0.863	0.519	0.200	0.861
PCA2						
No. PCs	**1.191**	0.190	0.365	**2.814**	0.633	0.283
IMSEB	4.295	0.392	0.146	2.025	0.944	0.480
MSEB	0.738	1.35E−3	0.050	1.450	0.214	0.319
Var	0.037	0.019	3.747	0.475	0.208	0.961



(a)

(c)

(b)

(d)

FIGURE 5 Graphics of the simulated parameter function (———) and the mean of its estimations with Method I (— — — —) and Method II (· · · · ·) after 500 repetitions of Example 2: (a) PCA1 and MSEB criterion, (b) PCA1 and IMSEB criterion, (c) PCA2 and MSEB criterion, (d) PCA2 and IMSEB criterion.

each PCA, method and criterion for choosing the PCs. The results are presented in Tables VI for Method I and Table VII for Method II. From such tables we can see again the great reduction of dimension that provides the best possible estimation of the parameter function. On the one hand, by using PCA1, we can see the great reduction of dimension obtained with a mean number of PCs in the optimum estimation rounding the three and four (Method II and

TABLE VI Means and Variabilities of Accuracy Measures for the Optimum Models of 500 Repetitions of Example 2 by Method I.

	Criterion for selecting the optimum model					
	MSEB			*IMSEB*		
Measures	*Mean*	*Variance*	*CV*	*Mean*	*Variance*	*CV*
PCA1						
No. PCs	**4.592**	2.622	0.353	**4.496**	**1.100**	**0.233**
Cum Var.	78.305	78.563	0.113	**78.550**	27.614	0.067
IMSEB	1.140	1.310	1.004	0.907	0.489	0.771
MSEB	0.411	0.058	0.585	0.446	0.113	0.754
Var	0.001	3.16E−7	0.978	0.001	1.09E−7	0.650
PCA2						
No. PCs	**1.340**	1.676	0.966	**7.472**	6.903	0.352
Cum Var.	79.443	11.110	0.042	**93.810**	22.026	0.050
IMSEB	4.778	0.533	0.153	1.743	0.329	0.329
MSEB	**0.854**	0.016	0.150	11.658	138.080	1.008
Var	0.002	3.92E−4	10.607	0.048	8.41E−3	1.897

TABLE VII Means and Variabilities of Accuracy Measures for the Optimum Models of 500 Repetitions of Example 2 by Method II.

	Criterion for selecting the optimum model					
	MSEB			*IMSEB*		
Measures	*Mean*	*Variance*	*CV*	*Mean*	*Variance*	*CV*
PCA1						
No. PCs	**3.298**	0.943	0.294	**3.550**	0.412	0.181
IMSEB	1.270	1.183	0.856	0.940	0.549	0.788
MSEB	0.380	0.029	0.449	0.482	0.127	0.738
Var	0.483	0.165	0.842	0.560	0.157	0.708
PCA2						
No. PCs	**1.230**	0.538	0.596	**3.192**	1.350	0.364
IMSEB	2.852	0.175	0.147	2.075	0.245	0.239
MSEB	**2.035**	3.496	0.919	12.273	0.013	0.913
Var	0.056	0.043	3.728	0.571	0.174	0.731

I respectively). On the other hand, by using PCA2 we obtain a great dimension reduction too, and it is more significant including PCs by Method II than by Method I. IMSEB is again the measure that informs better of the best possible estimation of the parameter function by using PCA2. This can be clearly seen in Figure 5(c) and (d) where the mean parameter function has been drawn for the corresponding optimum models. From Figure 5(a) and (b) we can see that using MSEB and IMSEB as criterions for choosing the best possible estimation with PCA1 provide similar results.

6 CONCLUSIONS

The objective of this study is to predict a binary response variable from a functional predictor through the functional logistic regression model. The estimation of such a model is not possible with the usual methods of least squares or maximum likelihood, therefore the most used method in literature consists of approximating the sample paths and the parameter function in a finite space generated by a basis and fitting the resultant multiple model. Due to the way of obtaining the design matrix of this multiple model there is a great multicollinearity which causes it to be unstable so that the estimation of the parameter function is not very accurate.

In order to reduce dimension and improve the estimation of functional logistic regression model in presence of multicollinearity we propose to use a reduced number of PCs as covariates of the model which provides a better estimation of the parameter function. Two different approaches to PCA have been considered: PCA of the design matrix obtained when we approximate the sample paths and the parameter function in a finite space (PCA1) and FPCA of the original sample paths (PCA2). In addition, different criterions for selecting the optimum PCs to be included in the model have been performed.

From the simulation studies developed in this paper, we can conclude that PCA1 provides a better estimation of the parameter function than PCA2. Moreover, including PCs in the order given by the stepwise method based on conditional likelihood ratio tests is better than including them by variability order as it provides an accurate estimation of the parameter function with a smaller number of PCs.

Acknowledgements

This research has been funded by BFM2000-1466 project from the Spanish Ministry of Science and Technology. The authors thanks an anonymous referee for helpful comments and suggestions that have contributed to improve the final version of this paper.

References

Aguilera, A. M. and Escabias, M. (2000). Principal component logistic regression. *Proceedings in Computational Statistics 2000*, Physica-Verlag, pp. 175–180.

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

Aguilera, A. M., Gutiérrez, R. and Valderrama, M. J. (1996). Approximation of estimators in the PCA of a stochastic process using B-splines. *Communications in Statistics: Computation and Simulation*, **25**(3), 671–690.

Aguilera, A. M., Ocaña, F. A. and Valderrama, M. J. (1997a). An approximated principal component prediction model for continuous-time stochastic processes. *Applied Stochastic Models and Data Analysis*, **13**(1), 61–72.

Aguilera, A. M., Ocaña, F. A. and Valderrama, M. J. (1997b). Regresión sobre componentes principales de un proceso estocástico con funciones muestrales escalonadas. *Estadística Española*, **39**(142), 5–21.

Aguilera, A. M., Ocaña, F. A. and Valderrama, M. J. (1999a). Forecasting time series by functional PCA. Discussion of several weighted approaches. *Computational Statistics*, **14**, 442–467.

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

Aguilera, A. M., Ocaña, F. A. and Valderrama, M. J. (2002). Estimating functional principal component analysis on finite-dimensional data spaces. *Computational Statistics and Data Analysis* (Submitted).

Aucott, L. S., Garthwaite, P. H. and Currall, J. (2000). Regression methods for high dimensional multicollinear data. *Communications in Statistics: Computation and Simulation*, **29**(4), 1021–1037.

De Boor, C. (1978). *A Practical Guide to Splines*, Springer-Verlag.

Fraiman, R. and Muñiz, G. (2001). Trimmed means for functional data. *Test*, **10**(2), 419–440.

Hocking, R. R. (1976). The analysis and selection of variables in linear regression. *Biometrics*, **32**, 1–49.

Hosmer, D. W. and Lemeshow, S. (1989). *Applied Logistic Regression*, Wiley.

Jackson, J. E. (1991). *An User's Guide to Principal Components*, Wiley.

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

Liang, K-Y. and Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, **73**(1), 13–22.

Müller, H-G. and Stadmüller, U. (2002). Generalized functional linear models. Unpublished paper.

McCullagh, P. and Nelder, J. A. (1983). *Generalized Linear Models*, Chapman & Hall.

Ramsay, J. O. and Silverman, B. W. (1997). *Functional Data Analysis*, Springer-Verlag.

Ramsay, J. O. and Silverman, B. W. (2002). *Applied Functional Data Analysis*, Springer-Verlag.

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

Ryan, T. P. (1997). *Modern Regression Methods*, Wiley.

Todorovic, P. (1992). *An Introduction to Stochastic Processes and their Applications*, Springer-Verlag.

Valderrama, M. J., Aguilera, A. M. and Ocaña, F. A. (2000). *Predicción Dinámica Mediante Análisis de Datos Funcionales*, Hespérides-La Muralla.

Valderrama, M. J., Ocaña, F. A. and Aguilera, A. M. (2002). Forecasting PC-ARIMA models for functional data. *Proceedings in Computational Statistics*, Physica-Verlag, pp. 25–36.