Chemometrics and Intelligent Laboratory Systems 104 (2010) 289–305

Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

journal homepage: www.elsevier.com/locate/chemolab

Using basis expansions for estimating functional PLS regression Applications with chemometric data

Ana M. Aguilera [a,*], Manuel Escabias [a], Cristian Preda [b], Gilbert Saporta [c]

[a] Departamento de Estadística e I.O., Universidad de Granada, Granada, Spain
[b] Polytech'Lille, UMR 8524, Université des Sciences et Technologies de Lille, France
[c] Chaire de Statistique Appliquée, CNAM, Paris, France

ARTICLE INFO

Article history:
Received 31 December 2009
Received in revised form 7 September 2010
Accepted 20 September 2010
Available online 25 September 2010

Keywords:
Functional data
PLS regression
Basis expansion methods
B-splines

ABSTRACT

There are many chemometric applications, such as spectroscopy, where the objective is to explain a scalar response from a functional variable (the spectrum) whose observations are functions of wavelengths rather than vectors. In this paper, PLS regression is considered for estimating the linear model when the predictor is a functional random variable. Due to the infinite dimension of the space to which the predictor observations belong, they are usually approximated by curves/functions within a finite dimensional space spanned by a basis of functions. We show that PLS regression with a functional predictor is equivalent to finite multivariate PLS regression using expansion basis coefficients as the predictor, in the sense that, at each step of the PLS iteration, the same prediction is obtained. In addition, from the linear model estimated using the basis coefficients, we derive the expression of the PLS estimate of the regression coefficient function from the model with a functional predictor. The results provided by this functional PLS approach are compared with those given by functional PCR and discrete PLS and PCR using different sets of simulated and spectrometric data.

1. Introduction

Prediction models (regression and classification) are among the most widely used methodologies for scientific data analysis, and play an important role in chemometric applications, especially in chemistry, food industry and environmental studies. The main objective of these statistical techniques is to model and predict one or more response variables in terms of a set of related predictor variables. In many situations, the number of predictor variables or the dimension of the space spanned by them is much larger than the number of observations. This is generally the case of functional predictors for which the observations are curves or functions (functional data).

The main feature of a functional random variable is the infinite dimension of the space to which the observations belong. Accordingly, the estimation of functional regression models (in which some of the response and/or predictor variables are functional), and in particular the linear one, is, in general, an ill-posed problem. One standard method to estimate functional regression models is to use a roughness penalty approach that works well with noisy or unequally spaced observations of the curves [7,21]. Another common solution is to represent the functional data in terms of basis functions (B-splines, wavelets, trigonometrics) and to approximate the basis coefficients using a large number of discrete observations, which may be irregular and sparse [5,6,10]. Thus, the functional model is converted into a multiple one in terms of sample curve basis coefficients. The potential of functional data representation with B-splines has been discussed in the field of chemometrics using two practical data sets: near infra red (NIR) spectra from hog manure samples and NIR transmission spectra of Diesel data [32].

Reviews of common functional data analysis (FDA) methodologies and interesting real data applications in other fields (growth curves in medicine, financial series derived from stock-market movements, and rainfall and temperature curves in the environmental field, among others) are compiled in the well-known books by Ramsay and Silverman [29,30]. Nonparametric methods for analyzing functional data have also been discussed by Ferraty and View [15].

Popular estimation methods in chemometrics, such as Partial Least Squares (PLS) regression [16,35,39,40] and Principal Component Regression (PCR) [22,38], have been adopted in recent years to solve the problems of high dimensionality and multicollinearity encountered in functional regression models. With both these dimension reduction approaches, the problem is reduced to that of regression of the response variable on an optimum set of orthogonal (principal or partial least squares) components obtained as generalized linear combinations of the functional predictor that solve well known optimization criteria.

Functional Principal Component Analysis (FPCA) [25,34] of the predictor curves has been used previously to estimate the slope parameters of different types of functional regression models. First, principal component prediction (PCP) models were developed to forecast a stochastic process in the future (functional response) from

* Corresponding author. Departamento de Estadística e I.O., Universidad de Granada, Facultad de Ciencias, Campus de Fuentenueva, 18008-Granada, Spain. Tel.: +34 958 246 306.
E-mail address: aaguiler@ugr.es (A.M. Aguilera).

0169-7439/$ – see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.chemolab.2010.09.007

its evolution in the past (functional predictor) [1]. Subsequently, a weighted estimation of the PCP model was introduced to forecast a continuous time series [2]. These models have been extended to the case of a functional linear model in which both the response and the predictor are functional variables. An estimation approach for sparse data based on a nonparametric estimation of functional PCA of both functional variables has been proposed in [41]. A two-step functional regression approach has been applied to forecast curves of pollen concentration from curves of temperature [36]. The case of an scalar response variable and a functional predictor has been also studied in [3,18]. Functional generalized linear models (FGLM) were introduced in [17]. An estimation procedure of these models based on approximating the predictor variable by a truncated Karhunen–Loève expansion has been proposed in [24]. Furthermore, two different FPCA approaches have been developed for estimating the functional logistic regression model [10].

However, principal components do not take into account the relationship between response and predictor variables and thus, their choice for regression is not without drawbacks. To solve this problem, PLS regression has recently been generalized to the case of a functional predictor [27]. In order to obtain functional PLS components, Tucker's criterion is extended to functional data. Then, it is shown that the weight functions of PLS components are iteratively obtained by estimating the eigenvalues and eigenfunctions of a functional operator computed from the cross-covariance operators associated with the predictor and response variables. A combination of PLS and PCR with B-spline expansions and a roughness penalty have also been developed to estimate the functional linear regression model [19,31]. For the functional logit model with binary response, a PLS estimation approach was recently introduced and compared with functional principal component logistic regression [12].

In this paper, we consider the functional linear model with a functional predictor variable and a scalar response variable. We propose a new estimation procedure for functional PLS regression [27] based on using a basis expansion approximation of sample curves. Then, we prove that functional PLS regression is equivalent to PLS regression using as predictors the sample curve basis coefficients with a metric associated with the basis functions. This equivalence is expressed in terms of prediction via the PLS components. The expression of the coefficient regression function is also derived from the PLS estimate of the model with basis coefficients.

The paper is organized as follows. In Section 2 we introduce some basic theory on the linear model for functional data and the approximation within a finite dimensional function space. The functional PLS model and the particular case when the predictor is a functional variable with observations within a finite dimensional function space are presented in Section 3. Criteria are then proposed for the model selection procedure. In Section 4 simulation studies are performed to evaluate the capacity of standard methods for selecting PLS components (leave-one-out cross-validation) in order to accurately estimate the functional parameter. The results obtained are then compared with those provided by FPCR and classic PLS and PCR on the discrete observations of sample curves. To assess the performance of the proposed functional PLS regression with B-spline basis expansions, in Section 5 we discuss two applications using chemometric data. The aim of the first of these is to classify biscuits as good or bad on the basis of the resistance curves of dough during the kneading process (functional data classification). In the second application, the task is to accurately predict the amount of fat on meat pieces based on their spectrometry curves.

The proposed functional PLS approach with basis expansions is implemented using the fda R-package [23], which is available at http://cran.r-project.org.

2. Basic theory on the functional linear model

Let Y be a scalar random variable and $X=\{X(t)\}_{t\in[0,T]}$ be a second order stochastic process (functional predictor) whose sample paths belong to the space $L_2([0,T])$ of square integrable functions. Without loss of generality, we assume that $\mathbb{E}(Y)=0$ and $\mathbb{E}(X(t))=0$, $\forall t\in[0,T]$.

By analogy with the multiple linear model, the functional linear model is formulated as

$$Y=\int_0^T X(t)\beta(t)dt+\varepsilon, \tag{1}$$

where the slope parameter β is a square integrable function rather than a vector.

It is well known that the use of least squares criteria to estimate this model yields an ill posed problem because of the Wiener–Hopf equation

$$\mathbb{E}(YX(t))=\int_0^T \mathbb{E}(X(t)X(s))\beta(s)ds,$$

which, in general, does not possess a unique solution (see [33] for a detailed study).

In practice, besides the impossibility of the direct estimation of the functional parameter, a new problem appears. Normally, we only have discrete observations x_{ik} of each sample path $x_i(t)$ at a finite set of knots $\{t_{ik}: k=0,\ldots,m_i\}$. Because of this, the first step in FDA is often the reconstruction of the functional form of data from discrete observations. The most common solution to this problem is to consider that sample paths belong to a finite dimension space spanned by a basis of functions (see, for example, [30]). An alternative way of solving this problem is based on the nonparametric smoothing of functions [15].

Let us consider a basis $\{\phi_1(t),\ldots,\phi_K(t)\}$ and assume that the functional predictor admits the basis expansion

$$X(t)=\sum_{j=1}^{K}\alpha_j\phi_j(t). \tag{2}$$

Let us also assume that the functional parameter admits a basis representation like the sample paths $\beta(t)=\sum_{k=1}^{K}\beta_k\phi_k(t)$. Then, the functional model (1) becomes a multiple linear model for the response variable in terms of a transformation of the functional predictor basis coefficients. Thus,

$$Y=(\Phi\alpha)'\beta+\epsilon \tag{3}$$

where $\alpha=(\alpha_1,\ldots,\alpha_K)'$ and $\beta=(\beta_1,\ldots,\beta_K)'$ are the vectors of the basis coefficients of X and β, respectively, and Φ is the matrix of inner products between the basis functions, $\Phi_{K\times K}=(\phi_{jk})=\int_T\phi_j(t)\phi_k(t)dt$.

Then, the estimation procedure of the functional model based on the basis expansion of functional data has the following two steps:

1. Sample path basis coefficients are estimated from discrete-time observations by using an appropriate numerical method.
   If the sample curves are observed without error

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

   an interpolation procedure can be used. For example, Escabias et al. (2005) proposed quasi-natural cubic spline interpolation for reconstructing annual temperature curves from monthly values.
   On the other hand, if the functional predictor is observed with errors

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

then least squares smoothing is used. In this case, the basis coefficients of each sample path $x_i(t)$ are approximated by

$$\widehat{\alpha}_i = (\Theta_i'\Theta_i)^{-1}\Theta_i' x_i,$$

with $\Theta_i = (\theta_{kj}) = (\phi_j(t_{ik}))$ and $x_i = (x_{i0},\ldots,x_{im_i})'$.
An appropriate basis must be selected according to the main features of the sample curves. The most common examples are trigonometric functions for the case of periodic curves, B-splines for smooth curves and wavelets for curves with a strong local behavior (see [30] for a detailed study). Least squares smoothing on cubic B-splines is used in the application developed in this paper to approximate dough resistance curves during the kneading process and spectrometry curves of fine chopped meat pieces. In this case, the dimension of the basis depends on the number of definition knots. In many applications a number of equally spaced knots is selected by cross-validation. Alternatives procedures for solving the problem of knots selection have been addressed as for example [43].

2. Least squares estimation of the linear model (3) provides an estimation of the basis coefficients of the functional parameter. However, this estimation may be inaccurate, due to a strong correlation between the dependent variables (components of the random vector $\Phi\alpha$). In addition, the number of basis functions used in the approximation of the sample curves could be higher than the number of observations.

As in the multivariate case, these problems (multicollinearity and high dimension) that impede estimation of the functional linear model can be solved by using as predictors an uncorrelated set of variables (principal components or partial least squares components). An alternative solution to avoid excessive local fluctuation in the parameter function estimation is that of a roughness penalty approach based on maximizing a penalized likelihood function. Different spline estimators of the parameter function have been proposed in [5,6,8].

Solutions based on FPCA of X have been proposed by [9] and [4], extending principal component regression (PCR) to the functional case, and by [1] and [41] for the case where both the response and the predictors are functional variables. However, deciding upon the principal components (PCs) is not an easy task since these are computed without taking into account the relation between the response and the predictor variables. Therefore, one has to choose between robustness of the model (the most explanatory PCs) and its performance (the PCs most strongly correlated with the response).

As an alternative to functional PCR, [27] extended Partial Least Squares (PLS) regression to the case of a functional predictor. In practice, the main problem is the estimation of the functional PLS components from discrete-time observations of sample curves. In this paper, we propose a basis expansion approach that reduces functional PLS to ordinary PLS via a transformation of sample path basis coefficients.

3. PLS for functional linear regression

The PLS approach consists in penalizing the least squares criterion by maximizing the covariance (Tucker's criterion) instead of the correlation coefficient. The PLS approach is based on two simple ideas. The first one is to find in $L(X)$ – the linear space spanned by $\{X(t)\}_{t\in[0,T]}$ – latent variables similar to principal components but taking into account the response Y. The second idea, if Y is a multivariate or a functional response, is to use the correlation structure of Y, which is not the case for the least squares criterion. These ideas have been efficiently used in the finite dimensional case in the work of [39].

3.1. Functional PLS algorithm

The PLS components associated with the functional regression of a real random response Y in terms of a functional predictor $X = \{X(t)\}_{t\in[0,T]}$, are obtained as solutions of the Tucker's criterion extended to functional data as

$$\max_{w\in L_2([0,T]),\|w\|_{L_2([0,T])}=1} Cov^2\left(\int_0^T X(t)w(t)dt, Y\right).$$

Let us denote by W^X and W^Y, respectively, the Escoufier's operators associated with $\{X(t)\}_{t\in[0,T]}$, with respect to Y, defined by

$$W^X Z = \int_0^T \mathbb{E}(X(t)Z)X(t)dt, \quad W^Y Z = Y\mathbb{E}(YZ), \quad \forall Z\in L_2(\Omega).$$

The spectral analysis of this operator leads to the principal component analysis of the associated variable (see [13] for details).

Then, as shown in [27], the first PLS component of the regression of Y on X, t_1, is given by the eigenvector associated with the largest eigenvalue of the operator $W^X W^Y$

$$W^X W^Y t_1 = \lambda_{max} t_1.$$

Let $X_0(t) = X(t)$, $\forall t\in[0,T]$ and $Y_0 = Y$. Then, the first PLS-step is completed by ordinary linear regression of $X_0(t)$ and Y_0 on t_1. Let us denote by $X_1(t)$ $(t\in[0,T])$ and Y_1 the residuals of these linear regression models

$$X_1(t) = X_0(t) - p_1(t)t_1 \quad t\in[0,T],$$
$$Y_1 = Y_0 - c_1 t_1.$$

The weight function $w_1(t)$ associated with the first PLS component t_1 is given by

$$w_1(t) = \frac{\mathbb{E}(YX(t))}{\int_0^T \mathbb{E}(YX(t))dt}, t\in[0,T]$$

so that

$$t_1 = \int_0^T w_1(t)X(t)dt. \tag{4}$$

The PLS regression is an iterative method. At step h, $h\geq 1$, of the PLS regression of Y on $\{X(t)\}_{t\in[0,T]}$, we define the h^{th} PLS component, t_h, by the eigenvector associated with the largest eigenvalue of the operator $W^X_{h-1}W^Y_{h-1}$

$$W^X_{h-1}W^Y_{h-1}t_h = \lambda_{max}t_h,$$

where W^X_{h-1} and W^Y_{h-1} are the Escoufier's operators associated respectively with $\{X_{h-1}(t)\}_{t\in[0,T]}$ and Y_{h-1}. The PLS component t_h is also given by

$$t_h = \int_0^T w_h(t)X_{h-1}(t)dt, \tag{4}$$

where

$$w_h(t) = \frac{\mathbb{E}(Y_{h-1}X_{h-1}(t))}{\int_0^T \mathbb{E}(Y_{h-1}X_{h-1}(t))dt}, t\in[0,T].$$

Finally, the PLS step is completed by the ordinary linear regression of $X_{h-1}(t)$ and Y_{h-1} on t_h. We denote by $X_h(t)$ $(t\in[0,T])$ and Y_h the random variables which represent the error of these regressions

$$X_h(t) = X_{h-1}(t) - p_h(t)t_h, \quad t\in[0,T],$$
$$Y_h = Y_{h-1} - c_h t_h.$$

The properties of the PLS components are summarized by the following proposition:

Proposition 1. *For any* $h \geq 1$

a) $\{t_h\}_{h\geq 1}$ *forms an orthogonal system in* $L(X)$,
b) $Y = c_1 t_1 + c_2 t_2 + \ldots + c_h t_h + Y_h$,
c) $X(t) = p_1(t)t_1 + p_2(t)t_2 + \ldots + p_h(t)t_h + X_h(t), \quad t \in [0,T]$,
d) $\mathbb{E}(Y_h t_j) = 0, \quad \forall j = 1,\ldots,h$,
e) $\mathbb{E}(X_h(t)t_j) = 0, \quad \forall t \in [0,T], \quad \forall j = 1,\ldots,h$.

Thus, there is a simple way of computing PLS, involving only simple linear regression models.

The PLS linear approximation at iteration h is then given by

$$\hat{Y}^h = c_1 t_1 + c_2 t_2 + \ldots + c_h t_h.$$

Notice that the expression of the PLS component defined by Eq. (4) can be rewritten as an element of the linear space spanned by $\{X(t): t \in [0,T]\}$

$$t_h = \int_0^T v_h(t) X(t) dt,$$

with v_h being functions of $L_2[0,T]$. More precisely, simple calculus shows that $v_h \in span\{w_1,\ldots,w_h\}$. Thus, the PLS linear approximation at iteration h becomes

$$\hat{Y}^h = c_1 \int_0^T v_1(t)X(t)dt + \ldots + c_h \int_0^T v_h(t)X(t)dt = \int_0^T \beta^h(t)X(t)dt,$$

where β^h is the approximation provided after h iterations by the PLS approach for the slope parameter β in the functional linear model.

Finally, when $h \to \infty$, the convergence in quadratic mean of the PLS approximation to the least squares approximation, as well as the efficiency of the PLS approach with respect to PCR regression, is shown in [27].

Remark 1. Notice also that when Y is a binary response, because of the equivalence between linear discriminant analysis and linear regression, the functional PLS approach has been used by the authors in [28] for classification purposes. The discriminant function is the coefficient function of the linear regression of Y on $\{X(t): t \in T\}$ with Y recoded as

$$Y = \begin{cases} -\sqrt{p_0/p_1} & \text{if} \quad Y = 1 \\ \sqrt{p_1/p_0} & \text{if} \quad Y = 0 \end{cases}$$

with $p_0 = P[Y=0]$ and $p_1 = P[Y=1]$. The results of this functional PLS discrimination approach based on the approximation of kneading data by using cubic B-splines basis is presented in the application section.

3.2. PLS regression for basis expansion of functional data

Notice that, in practice, given a random sample $\{x_i(t): i = 1,\ldots,n\}$ from the functional predictor X, the Escoufier's operator W^X is estimated by the $n \times n$ matrix $\hat{W}^X$ with entries

$$\langle x_i, x_j \rangle = \int_0^T x_i(t)x_j(t)dt, \quad i,j \in \{1,\ldots n\}.$$

Because of this, the main problem related to the sample estimation of PLS components is that of approximating the inner products between sample curves from discrete-time observations. In this section, we propose to solve this problem by approximating the functional form of data in terms of basis expansions.

Let us consider that the functional predictor X is such that

$$X(t) = \sum_{j=1}^{K} \alpha_j \phi_j(t), \ \forall t \in [0,T],$$

where $\alpha = (\alpha_1,\ldots,\alpha_K)^t$ is a random column vector and $\{\phi_j\}_{j=1,\ldots,K}$ is a linear independent set of functions in $L_2([0,T])$.

The following result proves the relation between the PLS of Y on the functional predictor X and the PLS of Y on the vector α of basis coefficients.

Proposition 2. *Let* Φ *be the* $K \times K$ *matrix with entries given by the two-by-two inner products of the basis functions* $\{\phi_i\}_{i=1,\ldots,K}$. *Then, the functional PLS regression of* Y *on* $X = \{X(t)\}_{t\in[0,T]}$ *is equivalent to the PLS regression of* Y *on the finite random vector* $\Lambda = \Phi^{1/2}\alpha$,. *In this sense, at each step* h *of the PLS algorithm,* $1 \leq h \leq K$, *we have the same PLS component, and so, the same PLS approximation.*

Proof. We prove this result by induction.

Let $\Phi = (\phi_{i,j})_{1\leq i,j\leq K}$ be the symmetric matrix with entries

$$\phi_{i,j} = \langle \phi_i, \phi_j \rangle_{L_2([0,T])}$$

and let us denote by $\Phi^{1/2}$ the square root of matrix Φ such that $\Phi = \Phi^{1/2^t}\Phi^{1/2}$. Let us also denote by Λ the column random vector $\Lambda = \Phi^{1/2}\alpha$ and by ϕ the column vector of functions $\phi = (\phi_1,\ldots,\phi_K)^t$.

For $h = 1$ observe that for any random variable Z, we have

$$W^X Z = \int_0^T \sum_{i=1}^{K} \alpha_i \phi_i(t) \mathbb{E}\left(\sum_{j=1}^{K} Z\alpha_j \phi_j(t) \right) dt = \Lambda^t \mathbb{E}(\Lambda Z) = W^\Lambda Z$$

and thus the statement is true. Moreover, if after the first step, the residuals for the predictors (using t_1) are given by

$$\begin{cases} X_1 = X - t_1 p_1, & p_1 \in L_2([0,T]) \\ \Lambda_1 = \Lambda - t_1 \tilde{p}_1, & \tilde{p}_1 \in \mathbb{R}^K, \end{cases}$$

then

$$\Lambda_1 = \Phi^{1/2}\alpha_1,$$

where α_1 is the random vector of the basis coefficients of $X_1(t)$. Indeed,

$$X_1(t) = \sum_{i=1}^{K} \alpha_{1,i}\phi_i(t) = \alpha_1^t \phi(t), \ \forall t \in [0,T].$$

But from the PLS regression step we have

$$X_1 = \phi^t\alpha - t_1 \frac{\mathbb{E}\left(\phi^t \alpha t_1\right)}{\mathbb{E}(t_1^2)}$$

$$\Lambda_1 = \Phi^{1/2}\alpha - t_1 \frac{\mathbb{E}\left(\Phi^{1/2}\alpha t_1\right)}{\mathbb{E}(t_1^2)},$$

and hence $\alpha_1 = \alpha - t_1 \frac{\mathbb{E}(\alpha t_1)}{\mathbb{E}(t_1^2)}$.

Thus, at the second step of the PLS regression, $W^{X_1} = W^{\Lambda_1}$.

Assume that for each $s \leq h$ $W^{X_s} = W^{\Lambda_s}$ and that $W^{X_{h+1}} = W^{\Lambda_{h+1}}$. Then, at step $(h+1)$ we have

$$\begin{cases} X_{h+1} = X - t_1 p_1 - t_2 p_2 - \ldots - t_h p_h, & p_i \in L_2([0,T]) \\ \Lambda_{h+1} = \Lambda - t_1 \tilde{p}_1 - t_2 \tilde{p}_2 - \ldots - t_h \tilde{p}_h, & \tilde{p}_i \in \mathbb{R}^K, \end{cases}$$

As for $h = 1$ and using the orthogonality of t_i, $i = 1\ldots,h$, it is shown that $\Lambda_{h+1} = \Phi^{1/2}\alpha_{h+1}$, which concludes the proof.

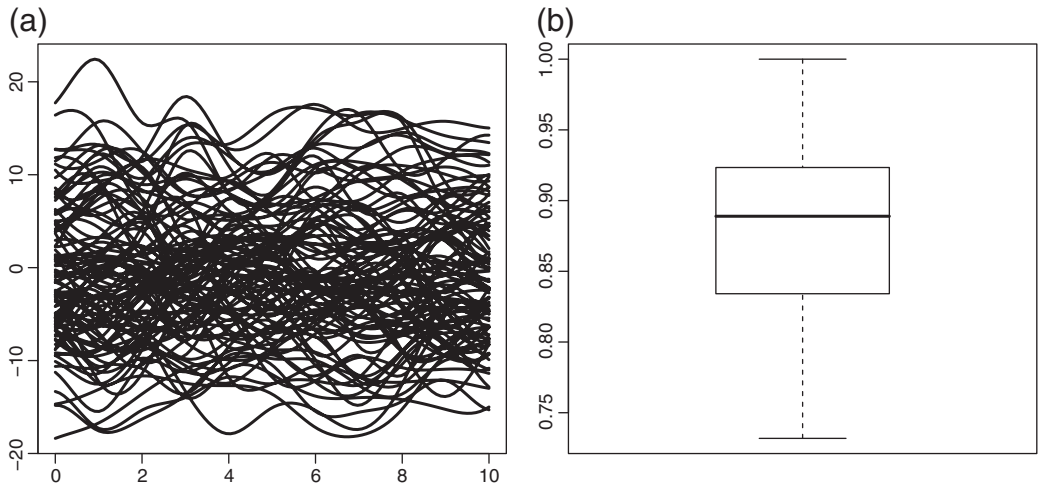

Fig. 1. Case I. (a) Sample of $n=100$ simulated curves. (b) Box and Whisker plot of correlation between columns of $A\Phi$ matrix.

Observe that the PLS approximation $\{\beta^h(t)(t\in[0,T])$ for the slope function parameter $\{\beta^h(t)(t\in[0,T])$ obtained after h iterations can be easily expressed in the basis ϕ. Let $\hat{\gamma}^h = (\hat{\gamma}_1^h,\ldots,\hat{\gamma}_K^h)^t$ be the regression coefficients provided by the PLS regression of Y on $\Lambda=\Phi^{1/2}\alpha$. Then,

$$\hat{Y}^h = \left(\Phi^{1/2}\alpha\right)^t\hat{\gamma}^h = \alpha^t\Phi^{1/2^t}(\Phi^{1/2}\Phi^{-1/2})\hat{\gamma}^h = \alpha\Phi\Phi^{-1/2}\hat{\gamma}^h = \int_0^T X(t)\beta^h(t)dt,$$

with

$$\beta^h(t) = \sum_{i=1}^{K}\left(\Phi^{-1/2}\hat{\gamma}^h\right)_{(i)}\phi_i(t), \quad t\in[0,T].$$

This result shows that using the particular metric Φ in the space of expansion coefficients α, the infinite dimension estimation problem is reduced to a simple finite PLS regression.

3.3. Sample functional PLS estimation

In applications with real data we have a random sample of pairs $\{(x_i(t),y_i): i=1,\ldots,n\}$ that can be seen as realizations of the functional predictor $X(t)$ and the response variable, Y, respectively.

The functional linear model is then expressed as

$$y_i = \beta_0 + <x_i,\beta>_{L^2[0,T]} = \beta_0 + \int_0^T x_i(t)\beta(t)dt + \varepsilon_i \tag{5}$$

where $\{\varepsilon_i: i=1,\ldots,n\}$ are independent and centered random errors.

The estimation procedure of the parameter function $\beta(t)$ using the basis expansion approach for functional PLS considered in this paper, has two main steps:

1. After choosing a suitable basis and assuming that each sample curve is represented as $x_i(t)=\sum_{j=1}^{K}\alpha_{ij}\phi_j(t)(i=1,\ldots,n)$, the basis coefficients are approximated by interpolation or smoothing as indicated in the previous section.
2. The functional PLS regression of Y on X is reduced to the PLS regression of $Y=(y_1,\ldots,y_n)'$ on matrix $A\Phi^{1/2}$, with A being the

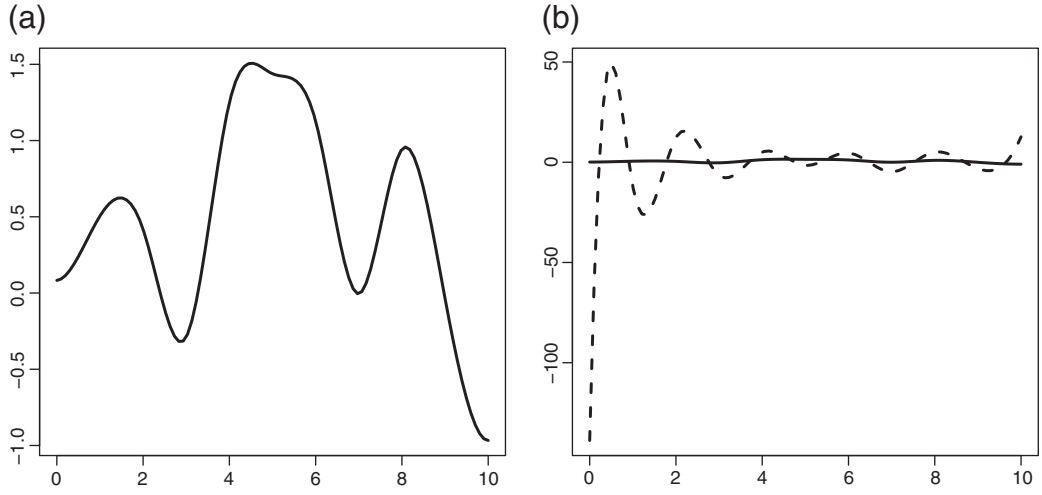

Fig. 2. Case I. (a) Simulated parameter function. (b) Simulated parameter function (solid line) and its estimation in terms of B-splines using least squares criterion.

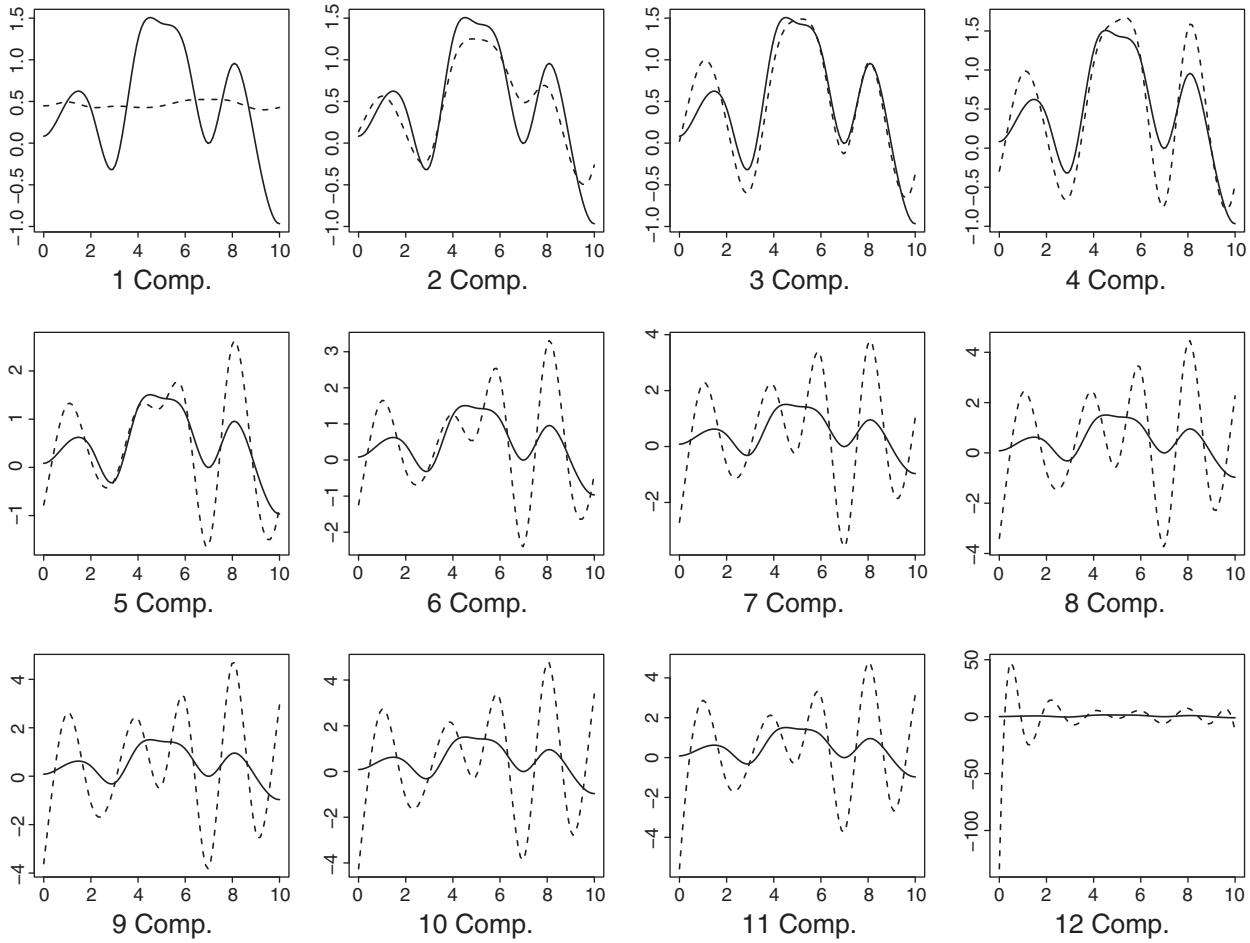

Fig. 3. Case I. Simulated functional parameter (solid line) and its estimations with different numbers of PLS components in the model (dashed).

matrix of sample curve basis coefficients $A=(\alpha_{ij})_{i=1,\ldots,n;j=1,\ldots,K}$. The PLS regression algorithm has the following steps:

- Computation of a set of h PLS components

$$T_{n\times h}=\left(A\Phi^{1/2}\right)_{n\times K}V_{K\times h}.$$

The columns of matrix T (PLS components) are computed by means of an iterative procedure for multivariate PLS similar to the one described in this paper for the functional case. The Escoufier's operators $W^{A\Phi^{1/2}}$ and W^{Y} are replaced by their sample estimations, given by the $n\times n$ matrices $\hat{W}^{A\Phi^{1/2}}=(A\Phi^{1/2})(A\Phi^{1/2})^t$ and $\hat{W}^{Y}=YY^t$, respectively.

- Linear regression of Y on the h PLS components

$$\hat{Y}^h=1\hat{\beta}_0^h+T\hat{\delta}^h.$$

- PLS regression in terms of the design matrix $A\Phi^{1/2}$

$$\hat{Y}^h=1\beta_0^h+A\Phi^{1/2}\hat{\gamma}^h,$$

with $\hat{\gamma}^h=V\hat{\delta}^h$.

- PLS regression in terms of the functional predictor

$$\hat{y}_i^h=\hat{\beta}_0^h+\langle(\Phi^{-1/2}\hat{\gamma}^h)^{\mathrm{t}}\phi,x_i\rangle_{L_2([0,T])},$$

so that the functional parameter is estimated from the PLS regression coefficients of Y in terms of the matrix A of sample path basis coefficients $\hat{\beta}(t)^h=(\Phi^{-1/2}V\hat{\delta}^h)\phi(t)$.

3.4. Model selection

In order to select the optimum number q of PLS components we considered three criteria.

— IMSE : minimizes the integrated mean squared error of the parameter function

$$IMSE(h)=\left(\frac{1}{T}\int_0^T\left(\beta(t)-\hat{\beta}^h(t)\right)^2dt\right)^{1/2},$$

which can be computed only for simulations where $\beta(t)$ is known.

Table 1
Case I. Goodness of fit measures for a simulated sample of size $n=100$.

NC	Exp.var.	IMSE(h)	CVMSE(h)	MSE(h)	$\frac{\text{CVMSE}(h)}{\text{MSE}(h-1)}$
1	87.29	0.6289	15.05	14.71307	
2	90.16	0.2504	14.41	13.55816	0.9796
3	93.17	0.2197	14.35	13.31779	1.0584
4	96.15	0.3377	14.32	13.16442	1.0753
5	97.38	0.6871	14.38	12.98355	1.0923
6	98.46	1.0142	14.36	12.87107	1.1060
7	98.96	1.5011	14.30	12.73760	1.1110
8	99.34	1.7216	14.27	12.69248	1.1203
9	99.75	1.8236	14.26	12.67645	1.1235
10	99.94	1.8825	14.23	12.66895	1.1225
11	100.00	1.9036	14.23	12.66734	1.1232
12	100.00	15.4125	14.31	12.56009	1.1297
13	100.00	15.9216	14.41	12.55843	1.1473

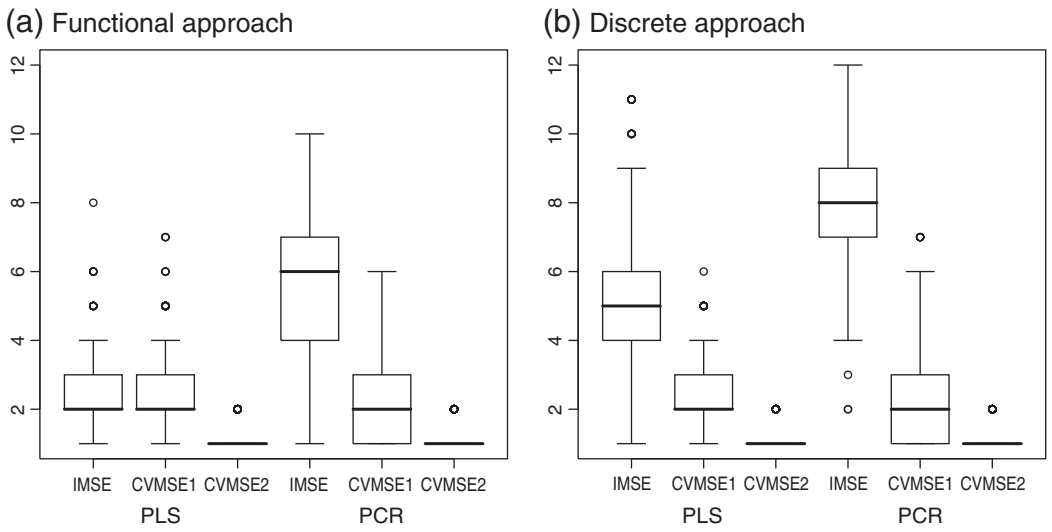

Fig. 4. Case I (NC). Box plots for the distribution of the number of components (NC) for the functional and discrete PLS and PCR models selected with the three criteria considered (IMSE, CVMSE1 and CVMSE2).

— CVMSE1 : minimizes the leave-one-out cross-validation mean square error

$$CVMSE(h) = \left(\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{y}_{(-i)}^{h}\right)^2\right)^{1/2},$$

with $\hat{y}_{(-i)}^{h}$ being the prediction for the i-th observation provided by the model with h PLS components, estimated from a sample of size *n-1* obtained by eliminating the i-th observation.
As the components are obtained iteratively, a classical rule to stop the process is the first time a minimum in CVMSE is found.

— CVMSE2 : leave-one-out cross-validation with threshold α (see [35]). The h^{th} PLS component is retained in the model if

$$CVMSE(h) \leq \alpha MSE(h-1), \quad 0 \leq \alpha \leq 1,$$

with $MSE(h) = \left(\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{y}_i^{h}\right)^2\right)^{1/2}$.

The classical way to select the number of PLS components is leave-one-out cross validation. See [42] for its natural extension, known as multifold cross validation, which allows the deletion of more than one observation. In this paper, we have added the IMSE criterion because our aim is not only to forecast the response but also to obtain a good estimation of the parameter function. Therefore, we examine whether the CVMSE1 and CVMSE2 criteria also provide good estimations of the parameter function.

3.5. Functional parameter interpretation

In practice, it is very important to obtain an accurate estimation of the parameters of a functional linear regression model, because then the relationship between the response and the predictor variable can be interpreted in terms of the estimated parameter function. In fact, the value of the functional parameter at each time gives the weight of the functional predictor $X(t)$ in the value of the response Y. This means that high absolute values of the parameter function represent

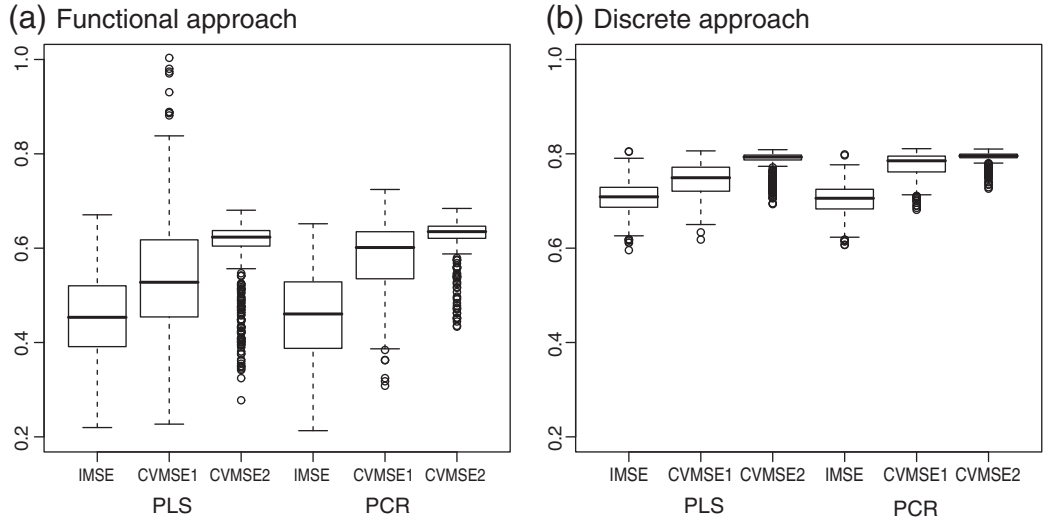

Fig. 5. Case I (IMSE). Box plots for the distribution of the integrated mean squared error (IMSE) for the functional and discrete PLS and PCR models selected with the three criteria considered (IMSE, CVMSE1 and CVMSE2).

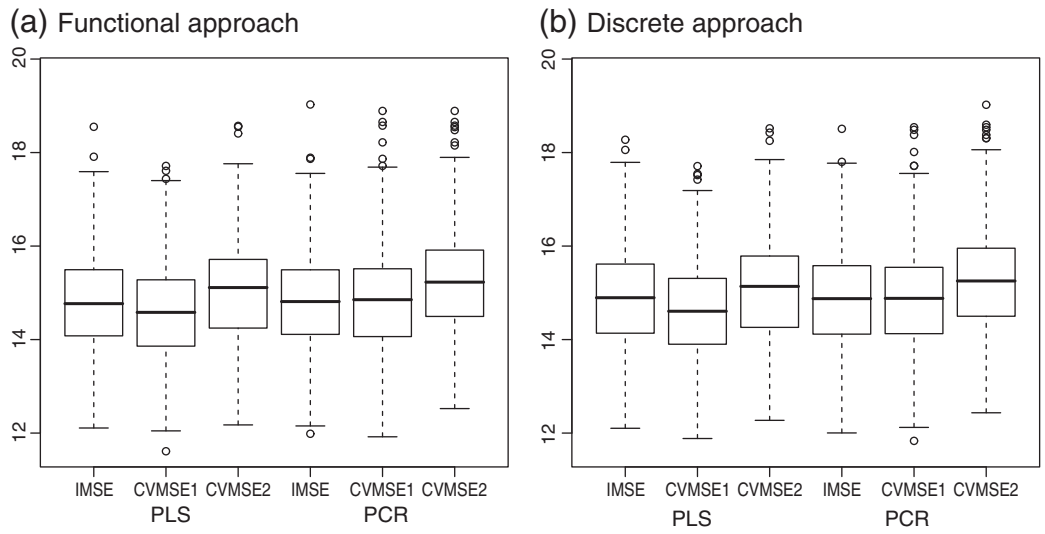

Fig. 6. Case I (CVMSE). Box plots for the distribution of the cross-validation mean squared error (CVMSE) for the functional and discrete PLS and PCR models selected with the three criteria considered (IMSE, CVMSE1 and CVMSE2).

periods with a large influence on the response whereas small values correspond to periods with little influence. Therefore, high levels of the predictor variable in periods where the functional parameter is positive are associated with high levels of the response. See [17] for a detailed study of different real-data applications and functional regression models.

Following the ideas introduced in [11] for interpreting the functional logit model in the case of an environmental data application, zones can be found in the functional observation domain in which a constant increase in a functional observation causes a specific change in the response. More specifically, the integral of the parameter function under a specific interval, multiplied by a constant K, can be interpreted as the additive change in the response when the functional observation associated with it in that interval is constantly increased by K units. If this integral is positive there will be an increase in the response while if it is negative there will be a decrease.

In the particular case of the functional PLS model proposed in this paper, a K-units increase in the jth PLS component ($\triangle t_j = K$) will cause the functional predictor $X(t)$ to increase by K times the associated functional loading $p_j(t)$ above the population average. Then, the average increase in the response variable Y is given by K times the estimated regression coefficient $\hat{\delta}_j$ associated with the jth PLS component in the linear regression of Y on the first h PLS components.

4. Experiments and discussion

The ability of the proposed functional PLS approach to estimate the parameter function and to predict the response is tested on simulated data, and the results obtained are compared with multivariate PLS and PCR on the discrete observations of sample curves, and also with the functional PCR, which is equivalent to the PCR of the response Y on matrix $A\Phi^{1/2}$. The proof of the theoretical relation between functional and multivariate PCA in the case of basis expansion of sample curves can be seen in [26].

4.1. Case I: simulation with B-spline functions

The considered functional variable has as curves the linear spans of the cubic B-spline functions $\{\phi_j : j = 1,\ldots,13\}$ defined by the knots

$$\{0,1,2,3,4,5,6,7,8,9,10\}.$$

In order to obtain a sample of curves, we simulated $n = 100$ vectors of a 13-dimensional centered multivariate normal distribution with covariance matrix $\sum = I_{13}$. Then, the matrix of the basis coefficients was obtained as a linear transform of these vectors, given by a 13×13 matrix of uniformly distributed values in the interval $[0,4]$. From these basis coefficients, we obtained the curves shown in Fig. 1(a).

As parameter function, we considered $\beta(t) = \sum_{k=1}^{13} \beta_k \phi_k(t)$, by simulating the basis coefficients $\beta = (\beta_1,\ldots,\beta_{13})$ with a uniform distribution in the interval [0,1]. This functional parameter can be seen in Fig. 2(a).

Finally, the non-functional response variable is simulated by using the functional model (5) with $T = 10$, which is converted into the multiple linear model

$$Y = 1\beta_0 + A\Phi\beta + \varepsilon, \tag{6}$$

Table 2
Case I. Mean and standard deviation of the number of components (NC), IMSE and CVMSE, for the functional and discrete PLS and PCR models selected with the three criteria considered (IMSE, CVMSE1 and CVMSE2).

		Functional PLS		Functional PCR	
Measure	Criterion	Mean	StDev	Mean	StDev
NC	IMSE	2.594	0.850	5.528	1.762
	CVMSE1	2.358	1.026	2.116	1.153
	CVMSE2	1.194	0.396	1.098	0.298
IMSE	IMSE	0.456	0.090	0.458	0.091
	CVMSE1	0.549	0.171	0.580	0.072
	CVMSE2	0.598	0.074	0.626	0.040
CVMSE	IMSE	14.784	1.082	14.822	1.082
	CVMSE1	14.595	1.067	14.841	1.136
	CVMSE2	15.018	1.130	15.221	1.144
		Discrete PLS		Discrete PCR	
Measure	Criterion	Mean	StDev	Mean	StDev
NC	IMSE	4.980	1.979	7.924	1.791
	CVMSE1	2.418	0.989	2.146	1.221
	CVMSE	1.198	0.399	1.094	0.292
IMSE	IMSE	0.661	0.032	0.659	0.029
	CVMSE1	0.697	0.035	0.729	0.026
	CVMSE2	0.736	0.026	0.746	0.014
CVMSE	IMSE	14.872	1.099	14.852	1.113
	CVMSE1	15.051	1.132	15.269	1.156
	CVMSE2	13.465	0.992	13.518	0.993

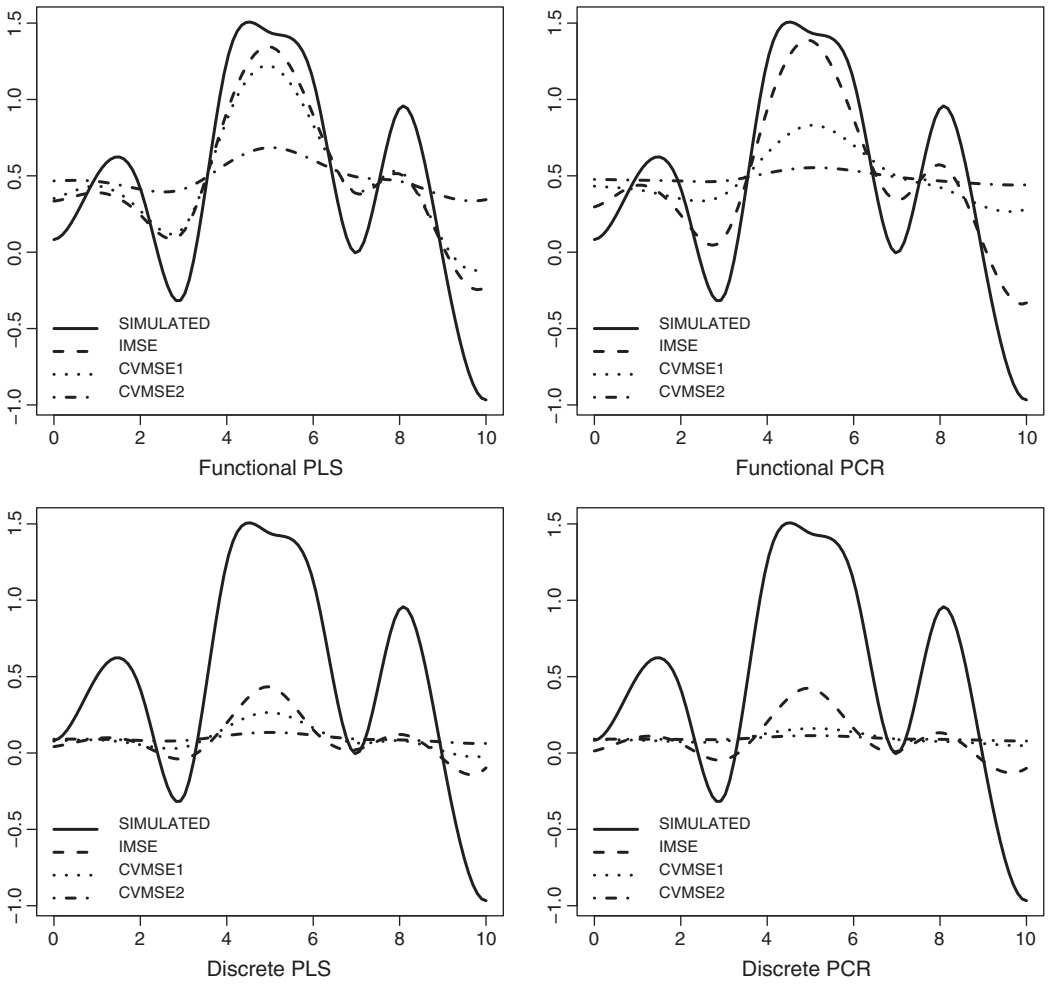

Fig. 7. Case I. Simulated parameter function and the mean of its optimum estimations with the three criteria considered (IMSE, CVMSE1 and CVMSE2) on $N=500$ repetitions.

where A is the matrix of basis coefficients of sample paths and Φ is the matrix of inner products between the B-spline basis functions. After computing the Φ matrix by an exact quadrature formula, the response values are generated from the previous equation with $\beta_0=1.4$ and ε_i simulated random values of a normal distribution with zero mean and variance $\sigma^2=200$. The signal to noise ratio $\mathrm{Var}(E[Y/X])/\mathrm{Var}(\epsilon)$ for this value of σ^2 is 5.55 and so the correlation coefficient R^2 associated with this linear model is $R^2=0.85$. The simulation study was also repeated with higher values of R^2, and in all cases similar results were obtained.

The multiple linear model (6) is affected by multicollinearity (high correlations between the columns of matrix $A\Phi$). The distribution of these correlations is shown in Fig. 1(b), where it can be seen that almost all of them are greater than 0.9. This fact means that the least squares estimation of the parameter function of this model is highly inaccurate (Fig. 2(b)). In order to solve the multicollinearity problem, we performed PLS regression of Y on $A\Phi^{1/2}$ and fitted the linear model with different numbers of PLS components as predictors. Then, the vector of the basis coefficients β was reconstructed, using the appropriate matrix of loadings. The functional parameter estimated by the models with different numbers of components can be seen in Fig. 3.

The goodness of fit measures used for selecting the optimum number of PLS components are shown in Table 1. Note that the best models are the one with 3 PLS components for the first model selection criterion (minimizing IMSE), the one with four components for the second (first minimum of CVMSE) and the one with one component for the third (CV with threshold 0.95). All these models fitted well with high R^2 values.

In order to validate the simulation and compare the results with those obtained by using functional PCR, discrete PCR and discrete PLS, we repeated the previously described process $N=500$ times setting the knots and the functional parameter, and simulating each time the basis coefficients of the sample curves and the response. The discrete PCR and PLS models were estimated by regressing the response variable Y on the components (principal or PLS) associated with the values of the sample curves on a set of 50 unequally spaced knots uniformly distributed in the interval [0,10]. The functional parameter was then reconstructed by least squares approximation on the parameters estimated by the discrete models on the 50 unequally spaced knots established in the simulation.

For each repetition and for each of the four different approaches considered, we obtained the corresponding PLS or principal components. The linear model was then fitted in terms of the different numbers of components of each type. Finally, in each repetition we selected the optimum PLSR and PCR models according to the three criteria previously considered (IMSE, CVMSE1, CVMSE2). Figs. 4–6 show box plots for the distributions of number of components, and properly errors IMSE and CVMSE after 500 repetitions. Table 2 also shows the mean and standard deviations of these measures. The means of the optimum estimations of the functional parameters for PLS and PCR regression can be seen in Fig. 7.

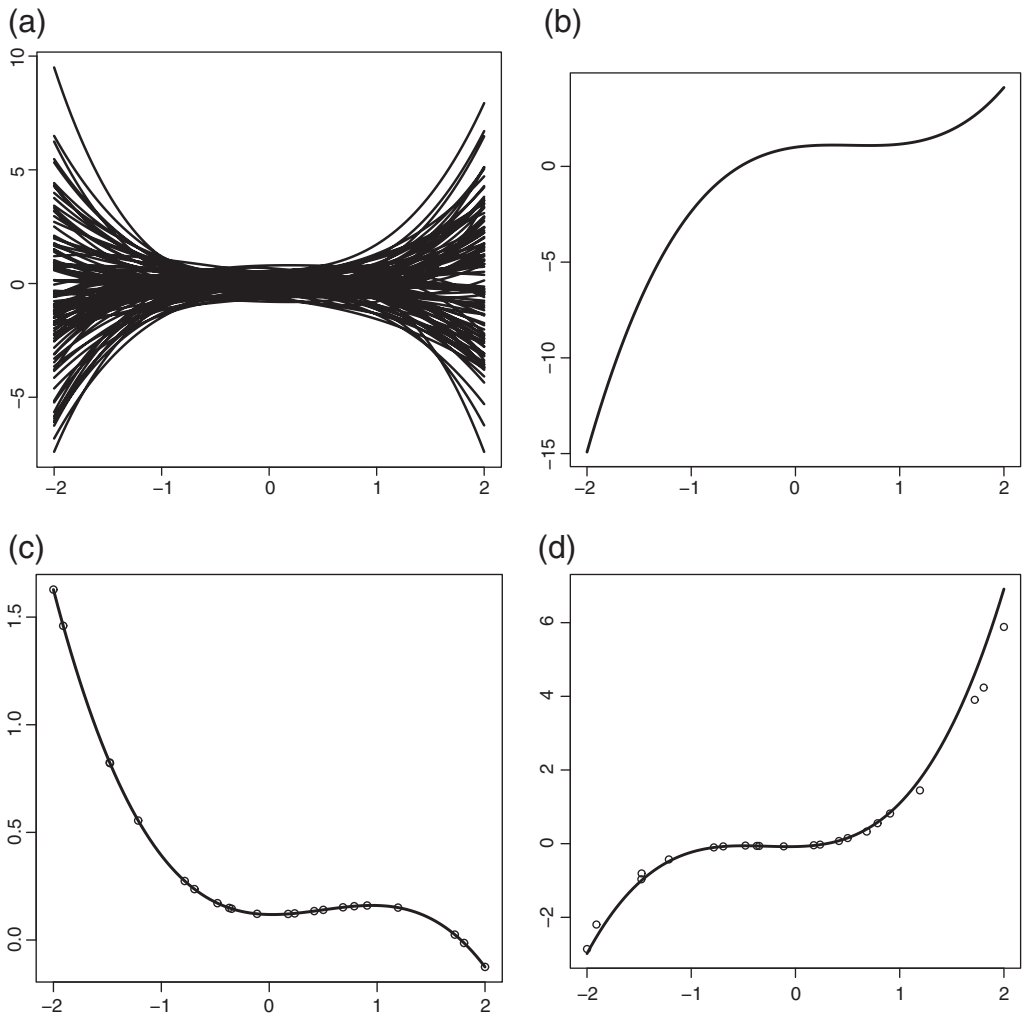

Fig. 8. Case II. (a) Sample of $n=100$ simulated curves. (b) Simulated functional parameter. (c) (d) Cubic spline approximation of two specific sampled curves.

4.2. Case II: simulation with polynomial sample curves

The functional variable considered in this simulated example has as curves polynomials of the form

$$x(t) = a_0 + a_1 t + a_2 t^2 + a_3 t^3, \quad t \in [-2,2].$$

The functional parameter is also a polynomial defined as $\beta(t)=1+0.76t-1.6t^2+t^3, t\in[-2,2]$ (see Fig. 8(b)). Following the same process that in previous examples, 500 simulations were performed. In each repetition a sample of size $n=100$ has been considered with coefficients a_{ij} independently simulated by using a centered normal distribution with variance equal to 0.1. Fig. 8(a) shows the curves of one of the 500 simulations. In each case the response variable was simulated by using the functional model (6) with error variance $\sigma^2=66.94$. The multiple correlation coefficient associated to this linear model is approximately $R^2=0.8$.

For each repetition discrete evaluation of the curves on 22 unequally spaced points on the interval $[-2,2]$ was performed. After discrete data simulation, the functional form of curves was reconstructed by least squares approximation on the basis of the cubic B-splines defined by the knots $\{-2.00, -1.72, -0.90, 0.51, 1.04, 1.84, 2.00\}$ (see Fig. 8(c) and (d)). Different linear regression models are fitted in terms of different number of functional and multivariate PLS and PCR components. Finally, in each one of the 500 repetitions and each one of the four considered approaches (FPLS, FPCR, discrete PLS and discrete PCR) one optimum model is selected with the three different criteria (IMSE, CVMSE1 and CVMSE2). Figs. 9–11 show box plots for the distributions of NC, IMSE and CVMSE, respectively, on the 500 repetitions. The means of the optimum estimations of the functional parameter for the functional and discrete PLS and PCR regression models can be seen in Fig. 12.

4.3. Discussion of simulation results

From these results, we conclude that with both the PLS and the PCR approaches, we obtain an evident dimension reduction, as is necessary for an accurate estimation of the functional parameter. In practice the reduction in the dimension and the prediction errors are similar with the four functional and discrete PLS and PCR approaches. Unlike other approaches that have been considered, in the case of the FPLS regression model, the number of components that minimizes the IMSE is similar to the number of components selected by leave-one-out cross validation. Therefore, FPLS regression and the CVMSE1 criterion provide the most accurate estimation of the parameter function. This issue is important when the aim is an accurate estimation of the functional parameter, because it provides a measure to identify the optimum number of components when there are non-simulated

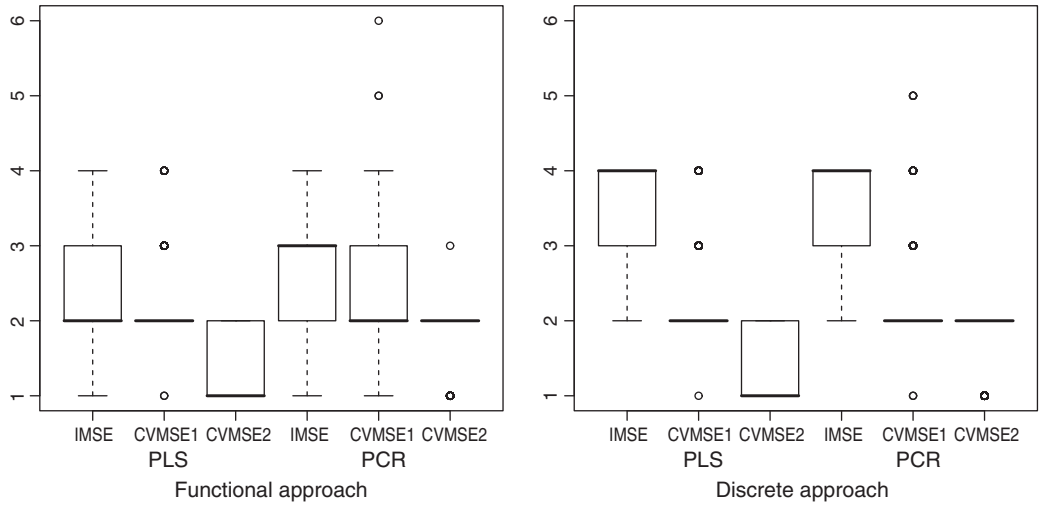

Fig. 9. Case II (NC). Box plots for the distribution of the number of components (NC) for the functional and discrete PLS and PCR models selected with the three considered criteria (IMSE, CVMSE1 and CVMSE2).

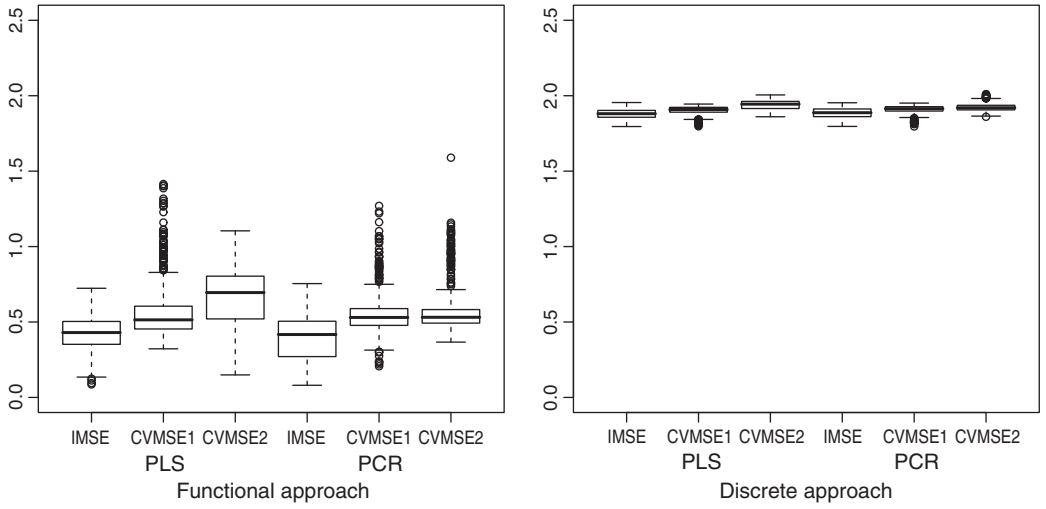

Fig. 10. Case II (IMSE). Box plots for the distribution of IMSE for the functional and discrete PLS and PCR models selected with the three considered criteria (IMSE, CVMSE1 and CVMSE2).

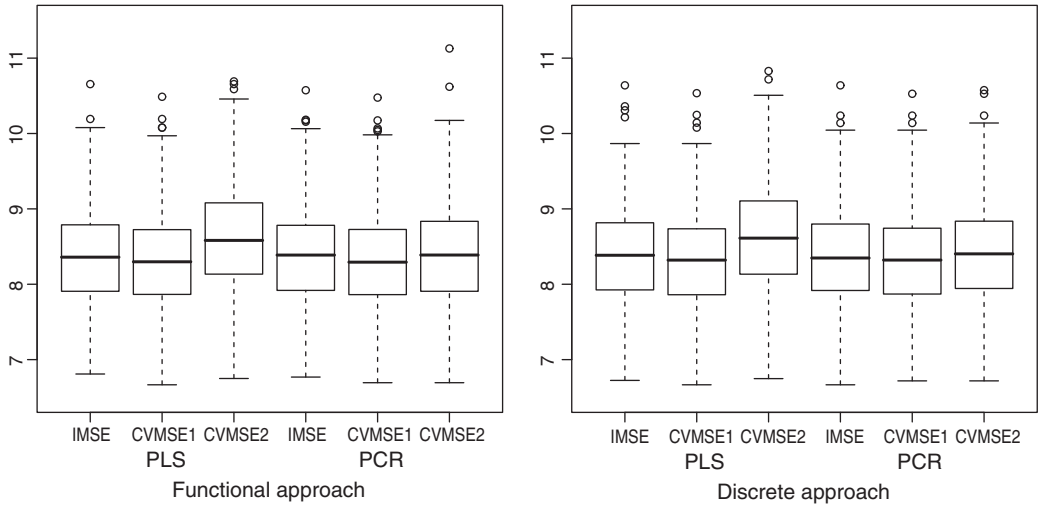

Fig. 11. Case II (CVMSE). Box plots for the distribution of CVMSE for the functional and discrete PLS and PCR models selected with the three considered criteria (IMSE, CVMSE1 and CVMSE2).

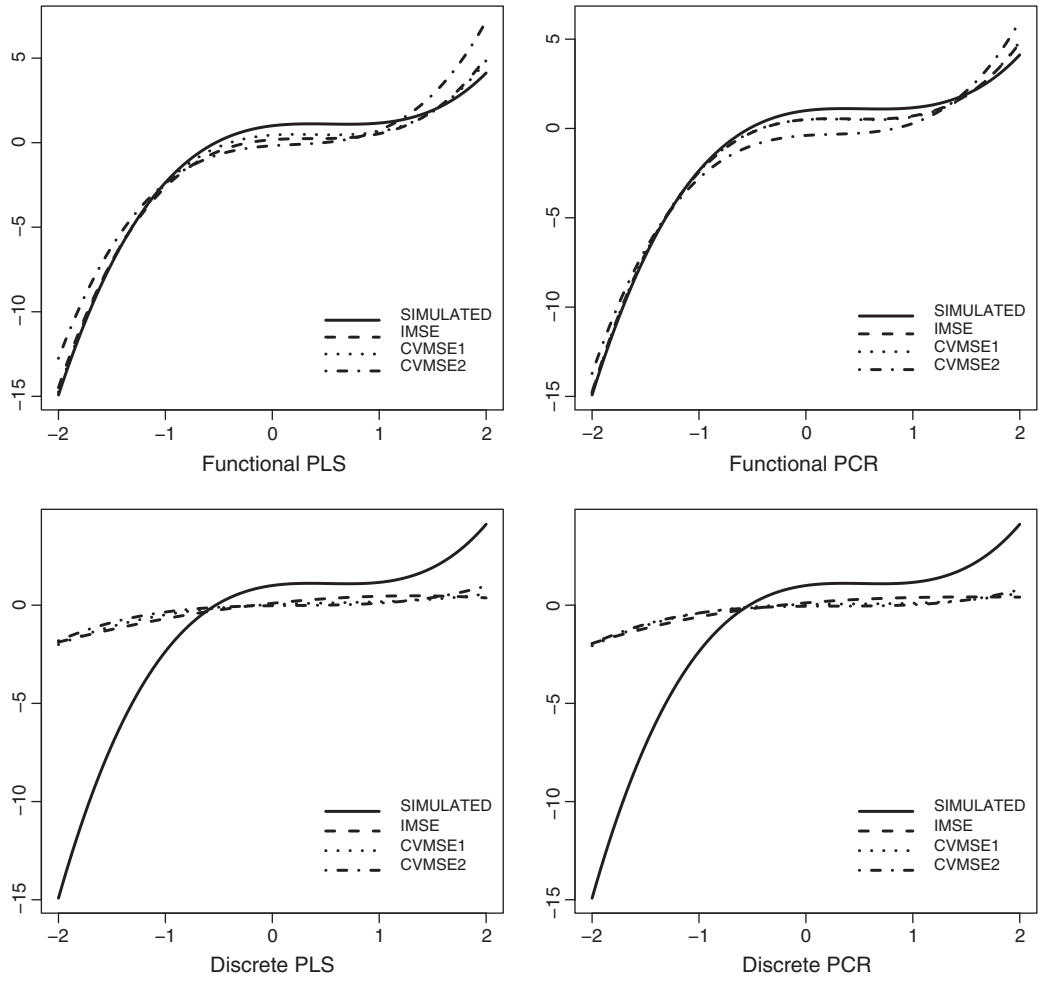

Fig. 12. Case II. Simulated parameter function and the mean of its optimum estimations with the three considered criteria (IMSE, CVMSE1 and CVMSE2)on $N=500$ repetitions.

examples, in which the real functional parameter is unknown. If the goal is only to predict the response, criterion CVMSE2 is preferable to CVMSE1 because it produces similar prediction errors with fewer components.

With respect to the discrete PLS and PCR approaches, the prediction errors are similar to those of the functional versions but the errors in the estimation of the functional parameter are much larger. Observe from Figs. 7 and 12 that the difference between the parameter functions estimated by functional and discrete approaches is in amplitude but not in shape. This difference in amplitude may be due to the fact that the scaling in the predictor variables in functional and discrete models may not be equivalent. This scaling does not affect the prediction errors, but provides a less accurate estimate of the functional parameter.

In order to determine whether this difference is really so, we considered a new selection model criterion based on the corrected IMSE defined as

$$CIMSE(h,k)=\left(\frac{1}{T}\int_0^T\left(\beta(t)-k\hat{\beta}^h(t)\right)^2dt\right)^{1/2}.$$

For each number of principal components h the value of k that minimizes $CIMSE(h,k)$ is given by

$$\hat{k}(h)=\frac{\int_T\beta(t)\hat{\beta}^h(t)dt}{\int_T\hat{\beta}^h(t)\hat{\beta}^h(t)dt}.$$

Then, we select as optimum the model that minimizes $CIMSE(h,\hat{k}(h))$.

In order to study the accuracy of the estimations of the parameter function given by the CIMSE criterion, it was used to select the optimum models in each of the simulations developed in this section. The means of the optimum estimations of the functional parameter for the functional and discrete PLS regression models are shown in Fig. 13. Observe that in the functional PLS models the estimation provided by the CIMSE criterion is almost the same as that provided by the IMSE criterion. On the other hand, an appropriate scaling of the estimation of the functional parameter given by the discrete PLS model provides an estimation of the functional parameter similar to that given by the functional PLS model. In practice, this is a problem because the functional parameter is unknown and the CIMSE criterion cannot be used to find the proper scaling when discrete PLS is used. Therefore, we conclude that in real-data applications, the functional model provides a more accurate estimation of the functional parameter. This is very important when the aim is to assess the relationship between the predictor and the response, because the parameter function at each moment of time gives the weight of the predictor in the value of the response.

5. Real data applications

Finally, we test the performance of the proposed functional PLS approach, using two chemometric data sets. In the first application, functional PLS are used to predict the quality of biscuits (binary classification) from the resistance curves of dough during the

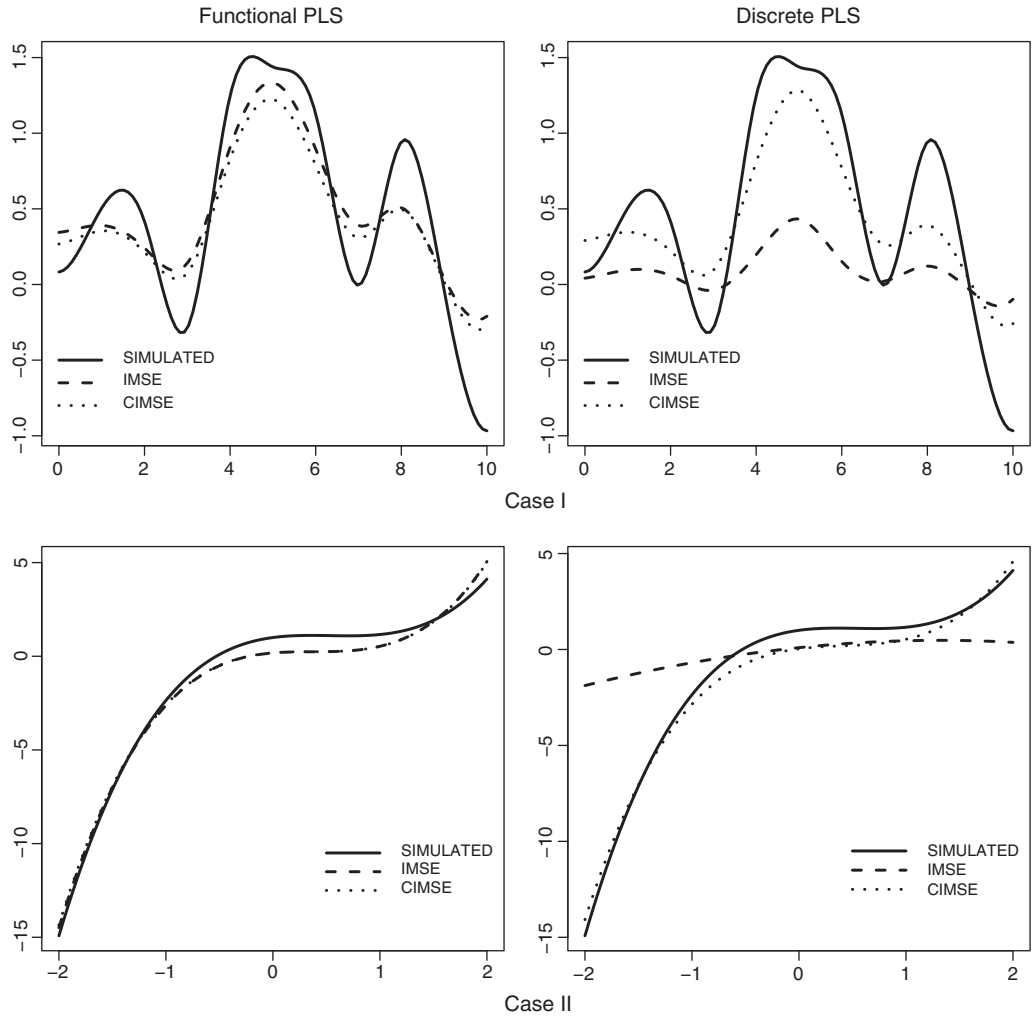

Fig. 13. Simulated parameter function and the mean of its optimum estimations with the two criteria based on the integrated mean squared error (IMSE and CIMSE) on $N=500$ repetitions.

kneading process. In the second application, functional PLS is applied to predict the quantity of fat contained in meat pieces from their spectrometric curves.

5.1. Biscuit quality and kneading functional data

The quality of a biscuit depends essentially on the quality of the flour used to make it. There are several kinds of flours, which are distinguished by their composition. The manufacturer, Danone, aims to use only flours that guarantee good product quality. Of course, one could test each and every flour, but this would be a lengthy process. Our idea in this analysis is to make use of a parameter observed during the kneading process in order to predict the final quality of the biscuit produced. For this purpose, we use the resistance (density) of the dough, observed in a certain interval of time. In our application, for a given flour, the resistance of dough is recorded during the first 480 s of the kneading process. Thus, we obtain a set of curves observed at 240 equally spaced time points in the interval [0,480]. Thus, a kneading curve is represented by the set of 241 points $\{(t_i,X(t_i)),i=0,\ldots,240\}$. After kneading, the dough is processed to obtain biscuits. For each flour, we have the quality (Y) of the biscuits, a quality that may be *Good* or *Bad*. We now use the proposed functional PLS regression approach to classify the biscuits as good or bad on the basis of the dough resistance curves defined above.

For 90 different flours, we have 90 curves, which can be considered as sample paths of an L_2-continuous stochastic process, $\mathbf{X}=\{X(t):t\in[0,480]\}$. This sample contains 50 observations for $Y=Good$, and 40 for $Y=Bad$ (left hand side in Fig. 14). Taking into account that the resistance of dough is a smooth curve measured with error, least squares approximation on a basis of cubic B-spline functions is used to reconstruct the true functional form of each sample curve. In order to compare the results with those presented in [20] the following 16 knots $\{10,42,84,88,108,134,148,\ 200,216,284,286,328,334,380,388,478\}$ has been considered to define the cubic B-spline basis. Thus, each curve $x_i=\{x_i(t):t\in[0,480]\}$ is represented by a set of 18 coefficients $\alpha_i=\{\alpha_{i,1},\ldots,\alpha_{i,18}\}$ which best approximates the real curve under the least squares criterion (right hand side in Fig. 14).

Taking into account the relation between linear discriminant analysis and linear regression, functional PLS regression can be used to estimate the discriminant function. To do so, the response $Y\in\{Good,Bad\}$ is recoded by $Y\in\left\{-\sqrt{\frac{40}{50}},\sqrt{\frac{50}{40}}\right\}$ and functional PLS regression of Y on a cubic B-spline approximation of the kneading curves is performed (see [28] for a detailed explanation). The sample of 90 flours is randomly divided into a learning sample of size 60 and a test sample of size 30. In the test sample, the two classes have the same number of observations. The PLS approach on cubic B-spline expansions of sample curves was then used to regress Y on the

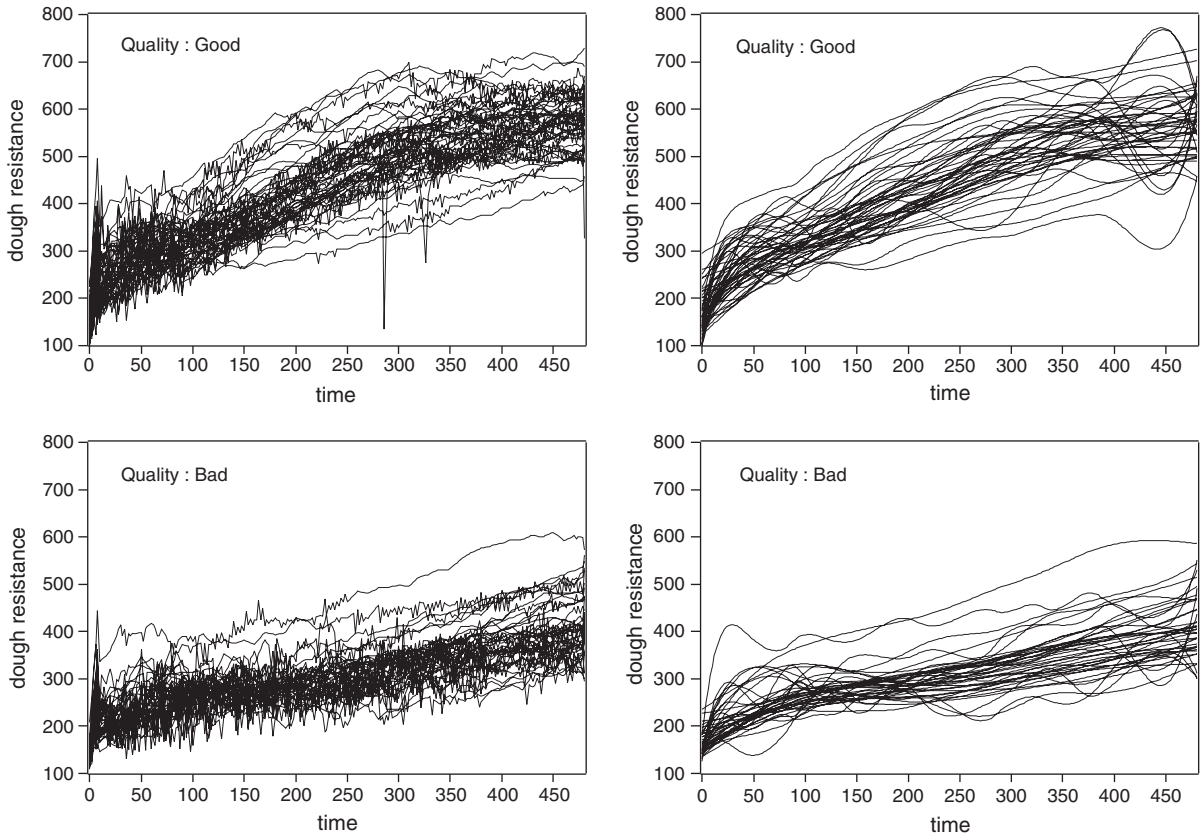

Fig. 14. Kneading data: 90 flours observed for 480 s. Left: observed data. Right: smoothed data using cubic B-splines.

kneading curves. For comparison purposes, the functional PCR on the cubic B-spline expansions was also considered.

The results were validated by randomly simulating 100 training and test samples and by fitting the functional linear model using the FPLS and FPCR approaches each time. For both approaches, the number of components retained for regression was obtained by leave-one-out cross validation with and without a threshold (CVMSE1 and CVMSE2 with $\alpha = 0.95$). Table 3 shows the results for the training samples. The average and standard deviations over 100 test samples of the mean squared errors given by the optimum models, with both criteria, are shown in Table 4. The average of the optimum functional parameters over the 100 training samples can be seen in Fig. 15. It can again be seen that the FPLS and FPCR approaches have a similar prediction ability. Although there are few differences, in the case of the FPLS criterion, CVMSE1 is preferable to CVMSE2 because the dimension reduction is similar and the prediction errors are lower.

Table 3
Kneading data. Mean and standard deviation of the number of components (NC), *CVMSE* and *MSE* of 100 training samples for the *FPLS* and *FPCR* models selected with the two considered criteria (*CVMSE1* and *CVMSE2*).

		FPLS		FPCR	
Measure	Criterion	Mean	StDev	Mean	StDev
NC	CVMSE1	2.32	0.469	3.03	0.502
	CVMSE2	1.7	0.461	1.39	0.490
CVMSE	CVMSE1	0.547	0.033	0.547	0.034
	CVMSE2	0.560	0.043	0.585	0.044
MSE	CVMSE1	0.512	0.033	0.519	0.033
	CVMSE2	0.534	0.046	0.567	0.045

Finally, in order to compare the proposed functional PLS approach with the functional linear discriminant PLS approach (FLD-PLS) developed in [28] to predict biscuit quality, we estimated the error rates by using both selection model criteria. The results in Table 5 show that the estimation of FPLS and FPCR in terms of the B-spline basis provides a considerable reduction in the classification error.

5.2. Spectrometric data

The spectrometric data analyzed in this paper consist of curves of spectrometry (absorbance measured in terms of wavelength) of fine chopped meat pieces. These data were recently used by [14], who proposed a nonparametric functional data analysis approach to predict the percentage of fat contained in these pieces of meat. The data can be downloaded from http://lib.stat.cmu.edu/datasets/tecator and consist of 215 pieces of finely chopped meat in which the spectrometric curve (absorbance measured at 100 wavelengths) was measured $\{x_i = (x_i(\lambda_1),\ldots,x_i(\lambda_{100})) : i = 1,\ldots,215\}$. The fat content y_i was obtained by analytical chemical processing. Our aim in this

Table 4
Kneading data. Mean and standard deviation of the *MSE* of 100 test samples for the *PLS* and *PCR* models with the optimum number of components selected with the two considered criteria (*CVMSE1* and *CVMSE2*).

	FPLS		FPCR	
Criterion	Mean	StDev	Mean	StDev
CVMSE1	0.516	0.082	0.510	0.086
CVMSE2	0.536	0.081	0.559	0.076

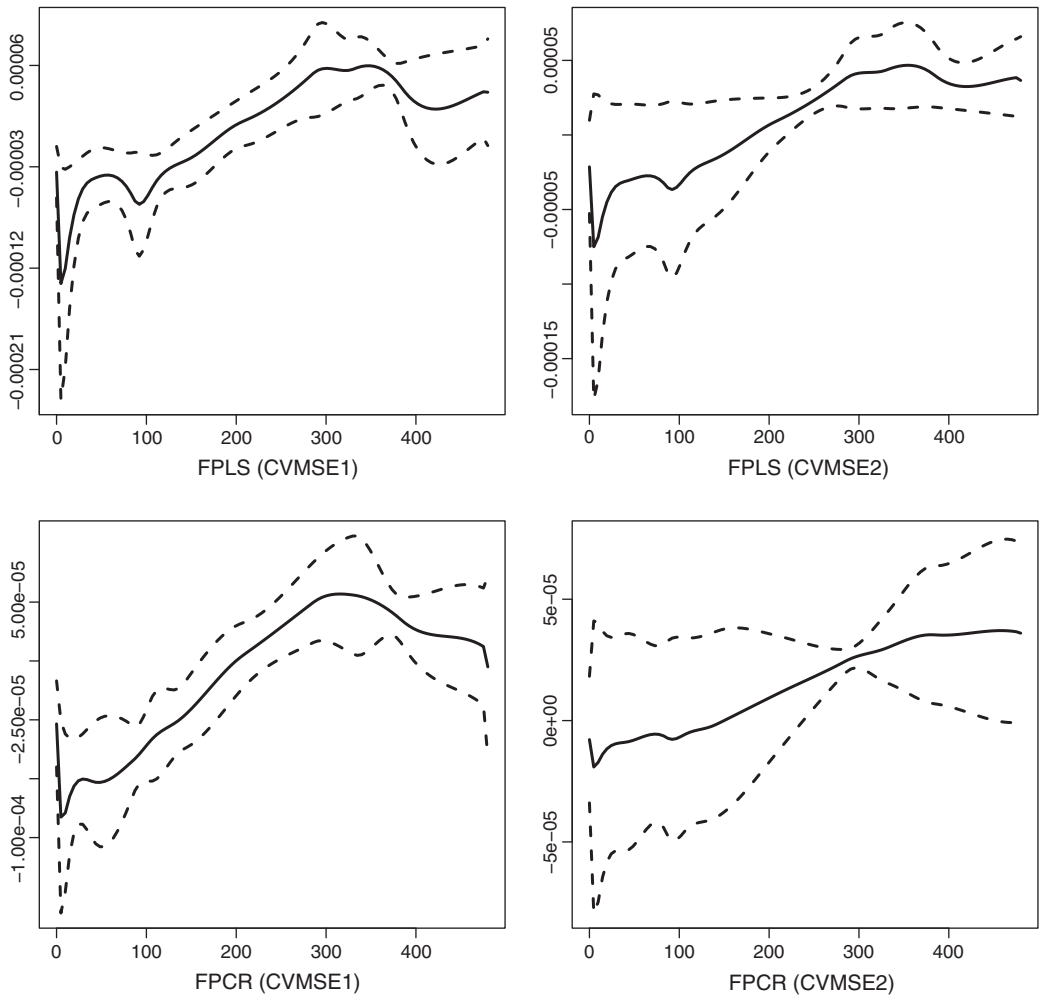

Fig. 15. Kneading data. Average of the optimum functional parameters (solid line) together with ± 2 times the standard deviation (broken line) estimated by using the CVMSE1 and CVMSE2 criteria with the FPLS and FPCR approaches.

example is to predict the fat content from the spectrometry curves by using PLS functional linear regression onto a basis expansion of the spectrometric curves.

The spectrometric curves have been smoothed by using basis expansion methods on cubic B-spline functions (see Fig. 16 (a)). To do so, the wavelengths {850.00, 854.04, 858.08, 862.12, 864.14, 882.32, 888.38, 890.40, 902.53, 908.59, 910.61, 918.69, 924.75, 926.77, 942.93, 944.95, 948.99, 991.41, 1001.52, 1017.68, 1035.86, 1050.00 } are considered to define the cubic B-splines. In the first attempt at fitting the functional linear model by means of the multiple model (6), we found high multicollinearity (see correlations between columns of $A\Phi$ matrix in Fig. 16(b)). This problem was resolved by using functional PLS regression onto a basis expansion and comparing the results with those obtained by functional PCR on the same basis expansion.

As in the kneading data example, we considered training and test samples of sizes 160 and 55 respectively, and this time repeated the random selection of these samples 150 times. As in the previous example, the functional PLSR and PCR approaches to cubic B-spline expansions of sample curves were used to regress Y on the training curves. On the one hand, it can be appreciated that the dimension reduction with CVMSE2 is much greater than with CVMSE1. On the other hand, the number of components used in PLSR is much larger than in PCR, but the leave-one-out mean squared errors and the classical mean squared errors of prediction are significantly higher when PCR rather than PLSR is used. Therefore, it is better to achieve lower mean squared errors in the prediction even though more components are used in the model (PLSR and CVMSE1). Table 6 shows these results. The average and standard deviation over 150 test samples of the mean squared errors given by the optimum models with both criteria are shown in Table 7.

Finally, in order to compare the proposed methods with those of [14], we applied the model to classify spectrometric curves in two groups. One was composed of the observations with less than 20% fat and the other of observations with at least 20% fat. The percentage of misclassifications on the test samples was computed for the FPLS regression models selected with both criteria, CVMSE1 and CVMSE2, and is shown in Table 8. The distribution of the error rate averaged over 150 test samples is similar to that produced by the nonparametric approach proposed in [14]. However, the error rate arising from the FPCR approach is much higher with both criteria, and especially with the second (CVMSE2) (Fig. 17).

Table 5
Kneading data. Misclassification rate averaged over 100 test samples.

	FPLS	FPCR
CVMSE1	0.071	0.078
CVMSE2	0.071	0.065
FLD-PLS (Preda et al., 2007)	0.112	0.142

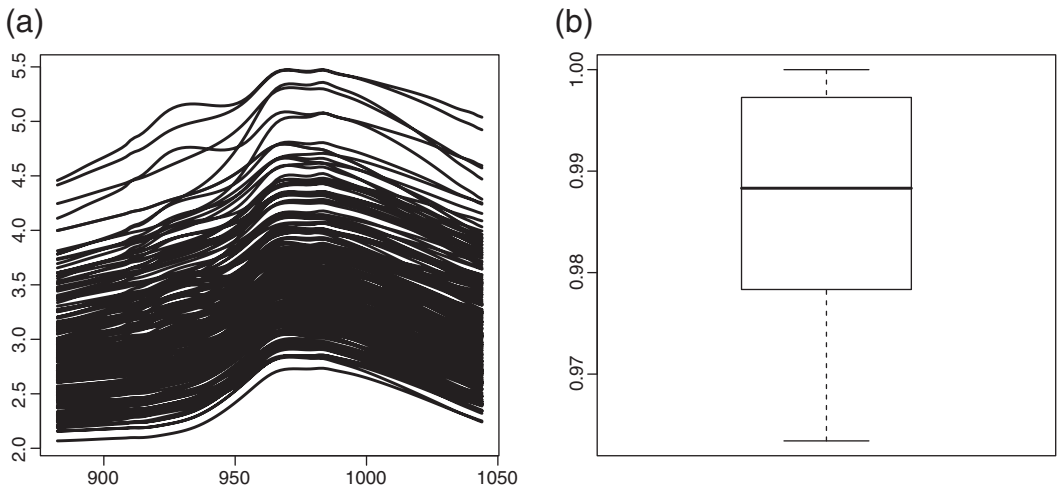

Fig. 16. Spectrometric data. (a) Sample of $N=215$ pieces of fine chopped meat. (b) Box and Whisker plot of correlation between columns of $A\Phi$ matrix in spectrometric data.

Table 6
Spectrometric data. Mean and standard deviation of the number of components (NC), *CVMSE* and *MSE* of 150 training samples for the *FPLS* and *FPCR* models selected with the two considered criteria (*CVMSE1* and *CVMSE2*).

		FPLS		FPCR	
Measure	Criterion	Mean	StDev	Mean	StDev
NC	CVMSE1	12.207	3.536	5.507	2.830
	CVMSE2	5	0	1	0
CVMSE	CVMSE1	2.560	0.295	5.243	3.744
	CVMSE2	3.198	0.098	11.511	0.287
MSE	CVMSE1	2.230	0.340	5.060	3.779
	CVMSE2	3.011	0.093	11.386	0.284

Table 7
Spectrometric data. Mean and standard deviation of the *MSE* of the 150 test samples for the *PLS* and *PCR* models with the optimum number of components selected with the two considered criteria (*CVMSE1* and *CVMSE2*).

	FPLS		FPCR	
Criterion	Mean	StDev	Mean	StDev
CVMSE1	2.673	0.467	5.248	3.6779
CVMSE2	3.167	0.380	11.412	0.989

6. Conclusions

There are many chemical applications such as spectroscopy where the objective is to account for a scalar response from a functional variable (the spectrum) whose observations are functions of wavelengths. In the majority of applications, this problem is solved by considering the spectrum as a vector whose values are the observations of the curve at a set of points. Multivariate data analysis techniques such as PCR and PLS are then used to solve the regression problem.

Taking into account the functional form of data, in this paper we have proposed a new estimation procedure for functional PLS regression based on a basis expansion of the sample curves. This approach reduces the functional PLS to the multivariate PLS of the response on a transformation of the matrix of sample path basis coefficients. Various leave-one-out cross-validation procedures have been considered in selecting the number of PLS components. The capability of these procedures to provide an accurate estimation of the parameter function and to forecast the response were tested in different simulation studies. Two applications using real chemometric data sets were also performed, demonstrating the good performance of the proposed methodology. With both simulated and real data, the results were compared with functional PCR on a basis expansion of sample curves and classical PLS and PCR on the discrete values of sample paths.

From these case studies, we can conclude that FPLS provides better estimations of the parameter function than do FPCR and similar predictions. As regards the comparison with the PLS and PCR discrete models, it has been shown that the predictive ability of discrete and functional models is almost the same. However, the ability of discrete approaches to provide an accurate estimation of the functional parameter is much lower in practice than that of functional approaches. In addition, the best model selection method is cross-validation without threshold because this provides an accurate estimation of the parameter function, very similar to that obtained by minimizing the IMSE. However, if the aim is only to predict the response, cross-validation with threshold is also a good choice because it provides a greater degree of dimension reduction, and the increase in errors is not excessive compared to the others.

These results corroborate other comparisons between PLS and PCR that have been carried out using chemometric data ([37]). Most such studies have concluded that PLS almost always requires fewer latent variables than PCR, and that there are no significant differences in the prediction errors reported by PLS and PCR. However, the present paper shows that in the functional case the parameter function estimated with PLS is much more accurate than with PCR. It also highlights the necessity for functional data analysis to accurately estimate the parameter function.

Table 8
Spectrometric data. Error rate averaged over 150 test samples.

	FPLS		FPCR	
	Mean	Sdev	Mean	Sdev
CVMSE1	0.0268	0.02111	0.1000	0.1378
CVMSE2	0.0225	0.0189	0.3108	0.0577

Acknowledgements

This research has been funded by project P06-FQM-01470 from *Consejería de Innovación, Ciencia y Empresa. Junta de Andalucía*, *Spain* and project MTM2007-63793 from *Dirección General de Investigación*, *Ministerio de Educación y Ciencia, Spain.* The authors would like to thank Caroline Lévéder for providing us with the data from Danone. The authors wish to thank an anonymous reviewer for helping to improve the work.

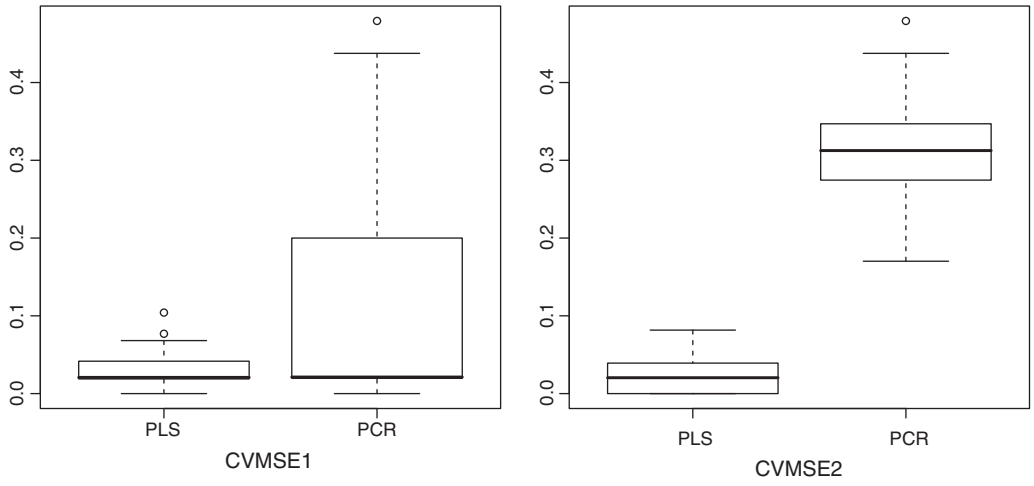

Fig. 17. Spectrometric data. Box plots for the distribution of the misclassification rate with FPLS and FPCR approaches and both model selection criteria on 150 test samples.

References

[1] A.M. Aguilera, F.A. Ocaña, M.J. Valderrama, Forecasting with unequally spaced data by a functional principal component approach, Test 8 (1) (1999) 233–254.
[2] A.M. Aguilera, F.A. Ocaña, M.J. Valderrama, Forecasting time series by functional PCA. Discussion of several weighted approaches, Comp. Stat. 14 (3) (1999) 443–467.
[3] T.T. Cai, P. Hall, Prediction in functional linear regression, Ann. Stat. 34 (2006) 21592179.
[4] H. Cardot, F. Ferraty, P. Sarda, Functional linear model, Stat. Probabil. Lett. 45 (1999) 11–22.
[5] H. Cardot, F. Ferraty, P. Sarda, Spline estimators for the functional linear model, Stat. Sinica 13 (2003) 571–591.
[6] H. Cardot, C. Crambes, A. Kneip, P. Sarda, Smoothing splines estimators in functional linear regression with errors-in-variables, Comput. Stat. Data Anal. 51 (2007) 4832–4848.
[7] H. Cardot, P. Sarda, Estimation in generalized linear models for functional data via penalized likelihood, J. Multivariate Anal. 92 (2005) 24–41.
[8] C. Crambes, A. Kneip, P. Sarda, Smoothing splines estimators for functional linear regression, Ann. Stat. 37 (1) (2009) 35–72.
[9] J.C. Deville, Analyse et prévision des séries chronologiques multiples non stationnaires, Statistique et Analyse des Données 3 (1978) 19–29.
[10] M. Escabias, A.M. Aguilera, M.J. Valderrama, Principal component estimation of functional logistic regression: discussion of two different approaches, J. Nonparametr. Stat. 16 (3–4) (2004) 365–384.
[11] M. Escabias, A.M. Aguilera, M.J. Valderrama, Modeling environmental data by functional principal component logistic regression, Environmetrics 16 (2005) 95–107.
[12] M. Escabias, A.M. Aguilera, M.J. Valderrama, Functional PLS logit regression model, Comput. Stat. Data Anal. 51 (10) (2007) 4891–4902.
[13] Y. Escoufier, Echantillonnage dans une population de variables aléatoires réelles, Publications de l'Institut de Statistique de l'Université de Paris 19(4) (1970) 1-47.
[14] F. Ferraty, P. View, The functional nonparametric model and application to spectrometric data, Comp. Stat. 17 (2002) 545–564.
[15] F. Ferraty, P. View, Nonparametric functional data analysis. Theory and practice, Springer, 2006.
[16] A. Höskuldsson, PLS regression methods, J. Chemom. 2 (1988) 211–228.
[17] G.M. James, Generalized linear models with functional predictors, J. R. Stat. Soc. B 64 (3) (2002) 411–432.
[18] G.M. James, J. Wang, J. Zhu, Functional linear regression that's interpretable, Ann. Stat. (2009) (To appear).
[19] N. Krämer, A.-L. Boulesteix, G. Tutz, Penalized Partial Least Squares with applications to B-spline transformations and functional data, Chemom. Intell. Lab. 94 (2008) 60–69.
[20] C. Lévéder, C. Abraham, P.A. Cornillon, E. Matzner-Lober, N. Molinari, Discrimination de courbes de pétrissage, Chimiométrie (2004) 37–43.
[21] B.D. Marx, P.H.C. Eilers, Generalized linear regression on sampled signals and curves. A p-spline approach, Technometrics 41 (1999) 1–13.
[22] W.F. Massy, Principal components regression in exploratory statistical research, J. Am. Stat. Assoc. 60 (309) (1965) 234–256.
[23] B.-H. Mevik, R. Wehrens, The PLS package: principal component and partial least squares regression in R, J. Stat. Softw. 18 (2) (2007) 1–24.
[24] H.-G. Müller, U. Stadtmüller, Generalized functional linear models, Ann. Stat. 33 (2) (2005) 774–805.
[25] F.A. Ocaña, A.M. Aguilera, M.J. Valderrama, Functional Principal Components Analysis by choice of norm, J. Multivariate Anal. 71 (2) (1999) 262–276.
[26] F.A. Ocaña, A.M. Aguilera, M. Escabias, Computational considerations in functional principal component analysis, Comp. Stat. 22 (3) (2007) 449–465.
[27] C. Preda, G. Saporta, PLS regression on a stochastic process, Comput. Stat. Data Anal. 48 (2005) 149–158.
[28] C. Preda, G. Saporta, C. Lévéder, PLS classification for functional data, Comp. Stat. 22 (2007) 223–235.
[29] J.O. Ramsay, B.W. Silverman, Applied Functional Data Analysis, Springer, 2002.
[30] J.O. Ramsay, B.W. Silverman, Functional Data Analysis, Second edition, Springer, 2005.
[31] P.T. Reiss, R.T. Ogden, Functional Principal Component Analysis and Functional Partial Least Squares, J. Am. Stat. Assoc. 102 (479) (2007) 984–996.
[32] W. Saeys, B. De Ketelaere, P. Dairus, Potential applications of functional data analysis in chemometrics, J. Chemometr. 22 (2008) 335–344.
[33] G. Saporta, Méthodes exploratoires d'analyse de données temporelles, Cahiers du B.U.R.O., Université Pierre et Marie Curie, Paris, 1981, pp. 37–38.
[34] B.W. Silverman, Smoothed functional principal component analysis by choice of norm, Ann. Stat. 24 (1) (1996) 1–24.
[35] M. Tenenhaus, La régression PLS. Théorie et pratique, Editions Technip, 2002.
[36] M.J. Valderrama, F.A. Ocaña, A.M. Aguilera, F.M. Ocaña-Peinado, Forecasting pollen concentration by a two-step functional model, Biometrics 66 (2) (2010) 578–585.
[37] P.D. Wentzell, L. Vega Montoto, Comparison of principal components regression and partial least squares through generic simulations of complex mixtures, Chemom. Intell. Lab. 65 (2003) 257–279.
[38] S. Wold, K. Esbensen, P. Geladi, Principal component analysis, Chemom. Intell. Lab. 2 (1987) 37–52.
[39] S. Wold, A. Ruhe, H. Wold, W.J. Dunn III, The collinearity problem in linear regression. The partial least squares (PLS) approach to generalized inverses, SIAM J. Sci, Stat. Comput. 5 (3) (1984) 735–743.
[40] S. Wold, M. Sjöström, L. Eriksson, PLS-regresion: a basic tool of chemometrics, Chemom. Intell. Lab. 58 (2001) 109–130.
[41] F. Yao, H.-G. Müller, J.-L. Wang, Functional linear regression analysis for longitudinal data, Ann. Stat. 33 (6) (2005) 2873–2903.
[42] P. Zhang, Model selection via multifold cross validation, Ann. Stat. 21 (1) (1993) 299–313.
[43] S. Zhou, X. Shen, Spatially adaptative regression splines and accurate knot selection, J. Am. Stat. Assoc. 96 (2001) 247–259.