Journal of Multivariate Analysis 146 (2016) 191–208

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

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

Feature selection for functional data

Ricardo Fraiman [a], Yanina Gimenez [b,c], Marcela Svarc [b,c,*]

[a] Centro de Matemática, Facultad de Ciencias, Universidad de la República, Uruguay
[b] Departamento de Matemática, Universidad de San Andrés, Argentina
[c] CONICET, Argentina

ARTICLE INFO

Article history:
Received 30 January 2015
Available online 26 September 2015

AMS 2010 subject classifications:
62H30
62H25
62J05

Keywords:
Variable selection
Classification
Regression
Principal components

ABSTRACT

We herein introduce a general procedure to capture the relevant information from a functional data set in relation to a statistical method used to analyze the data, such as, classification, regression or principal components. The aim is to identify a small subset of functions that can "better explain" the model, highlighting its most important features. We obtain consistency results for our proposals. The computational aspects are analyzed, a heuristic stochastic algorithm is introduced and real data sets are studied.

1. Introduction

During the past years technological advances made possible the processing and storing of real time information. Consequently, functional data arose across several fields, such as, finance, meteorology, medicine, among others. Hence, to handle these data new statistical procedures began to be developed. Classical statistical tools had to be rethought under this framework from a theoretical and empirical approach. In addition, new problems concerning the nature of the data had to be tackled. There is a vast literature devoted to these topics, the most classical contributions are the books by Ramsay and Silverman [37,38] and Ferraty and Vieu [14]. More recently, the books by Ferraty and Romain [13] and Horváth and Kokoszka [24] establish the state of the art on functional data analysis, reviewing the advances towards principal components analysis, regression and classification, among other problems.

The problem of feature extraction for multivariate problems has extensively been studied. The methods that show better results usually assess the importance of each variable based on the role that it plays in the statistical procedure being followed (classification, regression, principal components, etc.). The list of strategies available to handle these problems is large, however it is worth to mention that there are at least two proposals that have been studied through several multivariate procedures, namely the Bayesian model averaging (BMA) and the "least absolute shrinkage and selection operator" (LASSO). The former approach is Bayesian proposals introduced by Fraley and Raftery [17,18]. Therein they analyze the problem of unsupervised classification. Hoeting et al. [23], extent those ideas to the linear model. The LASSO was proposed by Tibshirani [42], who stated that "It shrinks some coefficients and sets others to 0, and hence tries to retain the good features of both subset selection and ridge regression... The LASSO minimizes the residual sum of squares subject

* Correspondence to: Vito Dumas 284, Victoria, Buenos Aires, Argentina.
E-mail address: marcelasvarc@gmail.com (M. Svarc).

http://dx.doi.org/10.1016/j.jmva.2015.09.006
0047-259X/

to the sum of the absolute value of the coefficients being less than a constant. Because of the nature of this constraint it tends to produce some coefficients that are exactly 0 and hence gives interpretable models''. More recently Witten and Tibshirani [44,43] applied those principles to cluster and principal components. A detailed description of this method can be found in [22].

The scenario is completely different in the functional data framework where there is much less literature available. Several authors address the problem of variable selection in the regression model. James et al. [25] extend the ideas of variable selection for regression studying the regression model when the predictor is functional and the response is scalar. Their goal is to estimate the regression function smoothly and sparsely, controlling its derivative and extending the ideas of the LASSO estimate. They claim that regions where the regression function is not null correspond to places where there is a relationship between the predictor and response variables, alternatively when it vanishes there is no relationship. They divide the time period into a grid of points. Then they use variable selection methods to determine whether the dth derivative of the regression function is zero or not at each time point. They implicitly assume that the dth derivative will be sparse. Lee et al. [29] also tackle this problem. They consider a sparse functional linear regression model which is generated by a finite number of basis functions in an expansion of the coefficient function. They adopt the idea of variable selection in the linear regression setting adding a weighted L_1 penalty to the traditional least squares criterion. Zhou et al. [46] introduce a functional linear model with zero-value coefficient function at sub-regions extending the SCAD ideas to this framework. In the three references above mentioned, the procedures described are suitable under different notions of sparsity, an assumption that may be restrictive. Cuevas [7] highlights that functional data can be discretized where standard variable selection for high dimensional data could be applied. Aneiros and Vieu [3] study the problem of variable selection in the regression setting, when the covariates are functionals. They highlight that even though, due to the practical treatment of the data that involve discretization, these problems are usually treated as classical high dimensional problems, where $p \gg n$, hence ordinal variable selection procedures that are suitable for this setting are applied. However, they stress the importance of developing new variable selection techniques that take profit of the continuous nature of the variables. Along similar lines, Aneiros et al. [2] consider the problem of variable selection for high dimensional data in partial least squares, in a very general setup. In particular, since they work on an abstract setting they allow to incorporate functional data to the model. Gertheiss et al. [19] also consider the problem of variable selection in regression with scalar response and functional covariates, they propose a penalized likelihood approach that combines selection of the functional predictors and estimation of the smooth effects for the chosen subset of predictors. McKeague and Sen [35] and Kneip et al. [26], deal with the problems of functional regression models assuming the existence of an unknown number of points on the functional covariates that have special impact on the response. In [9] there are several recent results related to these topics. For instance, Aneiros and Vieu [4] introduce a variable selection procedure for the case of linear and partial linear regression when some covariates are functional, they take into account the natural specifications of the functional variables. Also, Abramowicz et al. [1] present a procedure for performing an ANOVA test for Functional Data that as a byproduct selects intervals where the groups significantly differ. The problem of variable selection for logistic regression has been studied by Matsui [34].

The problem of variable selection in classification has also been studied. Tian and James [41] introduce an interpretable dimension reduction procedure to classify functional data. On a first stage, they reduce the dimension of the data considering projections, then they proceed to the classification step. More precisely, let $\{X(t),\ t \in [a, b]\}$ be a stochastic process and Y a categorical response variable, without lost of generality $Y = \{0, 1\}$. Let $\{f_1, \ldots, f_d\}$ be a set of functions in $[a, b]$, and denote Z_j the projections of $X(t)$ on direction f_j, i.e. $Z_j = \int_a^b X(t)f_j(t)dt$ con $j = 1, \ldots, d$. They perform the classification procedure in the low dimensional space. Hence, their aim is to retain a subset of functions that minimizes the misclassification rate. The key step of the proposal relays on the choice of the set of functions $\{f_1, \ldots, f_d\}$, which are selected from the set of constant and linear functions defined on subintervals of $[a, b]$. They also tackle the computational problem describing a competitive stochastic algorithm for choosing f_j for $j = 1, \ldots, d$. Also, Martin-Barragan et al. [33] present a support vector machine based classifier for functional data that has good classification ability and it is easy to interpret. Fauvel et al. [10] present a nonlinear parsimonious feature selection algorithm for the classification of hyperspectral images and the selection of spectral variables.

Our approach is different, it is designed to be used following a ''satisfactory'' analysis of the data, by which we mean that the data have previously been subjected to classification, principal components or some other form of statistical analysis. Additionally, we have a set of functions, $\mathcal{A} = \{f_1, \ldots, f_p\}$, that provides relevant features of the data. Our aim is to choose a subset of $\mathcal{A}$ that retains almost all the relevant information of the data set related to the statistical procedure that is being studied. We extend the ideas introduced by Fraiman et al. [16] for several multivariate procedures. They assume that a statistical analysis of a multivariate data has been successfully carried out, hence their idea is to *blind* unnecessary or redundant variables (by means of the conditional expectation), keeping the subsets of variables that better reproduce the output of the statistical procedure in the high dimensional space. Our procedure can be applied to a broad family of statistical problems. In this paper we extend the *blinding* procedure to the functional framework for the problems of classification, principal components and regression.

The remainder of the paper is organized as follows. We introduce the main definitions in Section 2. In Section 3 we study the classification problem. Section 4 is devoted to principal components analysis and Section 5 to functional regression. Finally, in Section 6 we consider some numerical aspects and analyze several well known data sets. Our concluding remarks are made on Section 7. All proofs are given in the Appendix.

2. Main definitions and notation

We start by fixing some notation that will be used throughout the manuscript. Let $X : \Omega \to L^2[a, b]$ be a random element $\{X =: X(t) : t \in [a, b]\}$ defined on a rich enough probability space $(\Omega, \mathcal{A}, P)$ where $[a, b] \subset \mathbb{R}$ is a finite interval. As it is well known, in practice, the data will be available on a grid $a \leq s_1 < \cdots < s_N \leq b$ which will be also denoted by $[a, b]$.

A population statistical model is given by the function $\psi(P) := \psi(X, g)$, for $\psi : L^2[a, b] \times \mathbb{G} \to L^2[a, b]$, where $\mathbb{G}$ is a separable metric space. The output of the statistical procedure is a random function in $L^2[a, b]$ given by $\psi(X, g)$.

Let $\mathcal{A}$ be a set of known functions, $\mathcal{A} = \{f_1, \ldots, f_p, f_i : L^2([a, b], P) \to \mathbb{R}\}$, and we denote by

$$\mathbf{f}(X) =: \mathbf{f} =: (f_1, \ldots, f_p) := \{f_1(X), \ldots, f_p(X)\},$$

the random vector where each coordinate is the evaluation of the trajectory X on the function $f_j \in \mathcal{A}$.

Some examples of interest are the following:

1. *Pointwise evaluation:* $f_1(X) = X(t_1), \ldots, f_p(X) = X(t_p)$ for $a \leq t_1 < t_2 < \cdots < t_p \leq b$.
2. *Local Averages:* $f_1(X) = \frac{1}{|T_1|}\int_{T_1} X(t)dt, \ldots, f_p(X) = \frac{1}{|T_p|}\int_{T_p} X(t)dt$, where $\{T_1, \ldots, T_p\}$ are disjoint intervals of the interval $[a, b]$.
3. *Occupation Measure:* $f_1(X) = |\{t : X(t) \in T_1\}|, \ldots, f_p(X) = |\{t : X(t) \in T_p\}|$, where $T_1, \ldots, T_p$ are disjoint intervals (not necessarily bounded) in $\mathbb{R}$ and $|\cdot|$ stands for the Lebesgue measure. That is, the time the process spends at each interval, or above or below a barrier.
4. *Number of up crossings to a level:* Let $f_i(X) = N_i$, where N_i is the number of up crossings of X to a level $c_i \in \mathbb{R}$, $i = 1, \ldots, p$. Recall that the process $X(t)$ is said to have an up crossing of c at $t_0 > 0$ if for some $\epsilon > 0, X(t) \leq c$ if $t \in (t_0 - \epsilon, t_0)$ and $X(t) \geq c$ if $t \in (t_0, t_0 + \epsilon)$. If the trajectories are differentiable, then $N_i = card\{t \in [a, b] : X(t) = c_i, X'(t) > 0\}$.
5. *Moments of the norm of the process:* $f_1(X) = E(\|X\|), \ldots, f_p(X) = E(\|X\|^p)$.
6. *Moments of the process:* $f_1(X) = \int_a^b X(t, \omega)dP(\omega), \ldots, f_p(X) = \int_a^b X^p(t, \omega)dP(\omega)$.

Remark 1. The choice of the set of functions $\{f_1(X), \ldots, f_p(X)\}$, depends on the nature of the process $X(t)$ and on features of the statistical procedure under consideration that are relevant to achieve a better understanding of the practical problem. For instance, pointwise evaluation and local averages reveal information towards the domain of $X(t)$, to detect a subset of points or small intervals at the domain, which explains successfully the phenomena. A typical example is spectrometry studies. The occupation measure and the number of up crossings to a level, provide information concerning the image of $X(t)$, these characteristics are relevant when measuring the pulse oximetry in preterm infants. Finally, the last two proposals should be used if the aim is to retain global information of $X(t)$, these features are very important in trading financial problems.

These sets of functions provide relevant information of a process towards the statistical procedure that is being conducted. Our aim is to retain a subset of $\mathcal{A}$, of cardinality $d \ll p$ that contains "almost" all the relevant information provided by $\mathbf{f}(X)$ to a specific statistical analysis (classification, regression, etc.). To achieve this goal we extend the ideas introduced by Fraiman et al. [16]. Therein, they proposed to retrieve relevant information from a multivariate model *blinding* unnecessary variables. It is clear that in the functional data framework a componentwise approach does not make sense, hence the *blinding* procedure must be redefined.

Given a subset of indices $I = \{i_1 < \cdots < i_d\} \subset \{1, \ldots, p\}$ with $d \leq p$, consider the random vector $\mathbf{f}(I, X) := \mathbf{f}(I) := (f_{i_1}(X), \ldots, f_{i_d}(X)) \in \mathbb{R}^d$.

Definition 1. Let I be a subset of $\{1, \ldots, p\}$ denote the **stochastic blinded process** of X, based on $f(I)$, to the process $Z(I) : [a, b] \to \mathbb{R}$ given by

$$Z(I)(t) = E(X(t)|\mathbf{f}(I)) := \eta(t, \mathbf{f}(I, X)). \tag{1}$$

We denote $Q(I)$ to the distribution of $Z(I)$.

Remark 2. $Z(I)(t)$ is a stochastic process even though it is the conditional expectation given a random vector $\mathbf{f}(I) \in \mathbb{R}^d$.

Given a fixed integer d, $1 \leq d \ll p$, we let $\mathcal{I}_d$ be the family of all subsets of $\{1, \ldots, p\}$ with cardinality smaller than or equal to d.

We seek a small subset, I, such that $\psi(X, g)$ is as close as possible to $\psi(Z(I), g)$. The notion of closeness may vary from one problem to another, and is denoted by $h(I, P, Q(I), \psi) := h(I)$.

More precisely, $\mathcal{I}_0 \subset \mathcal{I}_d$ is defined as the family of subsets in which the minimum $h(I)$ is attained for $I \in \mathcal{I}_d$, i.e.,

$$\mathcal{I}_0 = \operatorname*{argmin}_{I \in \mathcal{I}_d} h(I). \tag{2}$$

In practice, we look for consistent estimates of the set I_0, $I_0 \subseteq \mathcal{I}_0$ based on a sample $X_1, \ldots, X_n$ of iid trajectories of the stochastic process X.

Given a subset $I \in \mathcal{I}_d$, the first step is to obtain the blinded version of the sample, $\hat{X}_1(I), \ldots, \hat{X}_n(I)$, based on $\mathbf{f}(I, X)$, using nonparametric estimates of the conditional expectation.

We denote by $Q_n(I)$ to the empirical distribution of $\{\hat{X}_j(I),\ 1 \le j \le n\}$.

For instance, we may consider the r-nearest neighbor (r-NN) estimates. We fix an integer value r, the number of nearest neighbors used. For each $j \in \{1, \ldots, n\}$, we find the set of indices C_j of the r nearest neighbors of $\mathbf{f}(I, X_j)$ among $\{\mathbf{f}(I, X_1), \ldots, \mathbf{f}(I, X_n)\}$. Next we define the predicted sample processes as

$$\hat{X}_j(I)(t) = \hat{E}_P(X_j(t)|\mathbf{f}(I)) = \frac{1}{r}\sum_{m \in C_j} X_m(t) \in L^2[a, b].$$

Note that the set C_j does not depend on t.

Then, given a subset of indices $I \in \mathcal{I}_d$, we define the empirical version of the objective function $h_n(I)$, and we seek the optimal subsets of variables $\mathcal{I}_0 \subset \mathcal{I}_d$, which are the family of subsets in which the minimum of $h_n(I)$ is attained, i.e.,

$$\mathcal{I}_n = \operatorname*{argmin}_{I \in \mathcal{I}_d} h_n(I). \tag{3}$$

This will become clear through the next sections where we consider different statistical problems and models under this approach.

In order to prove the consistency of (3) to (2) the following hypothesis must be satisfied.

H1. Let $\hat{X}(I)$ be a strong consistent estimate of $Z(I)$, i.e., $\left\|Z(I) - \hat{X}(I)\right\|_{L^2[a,b]} \to_{a.s.} 0$.

Liam et al. [30] establish general conditions to obtain consistent estimates of the conditional expectation on a space provided with a pseudometric. Hence, **H1** is a particular case of Liam et al.'s result, since we are dealing with variables in $\mathbb{R}^d$. To be more precise, the conditions imposed by Liam et al. can be relaxed, obtaining the following result:

Proposition 1. *Let $Z(t) = \eta(t, \mathbf{f}(I, X)) + e, Z(t) \in \mathcal{H}$, be a separable Hilbert space, $\mathbf{f}(I, X) \in \mathbb{R}^d$ and $e \in \mathcal{H}$ a random element with zero mean and independent of $\mathbf{f}(I, X)$. Let η_n be the non-parametric estimate of the conditional expectation given by the r nearest neighbors. Given that,*

1. *$\mathbf{f}(I, X)$ has density function g such that $0 < c_1 \le g(x) \le c_2$ for all x in the support of $\mathbf{f}(I, X)$,*
2. (a) $\|\eta(t, \mathbf{f}(I, X))\|_{\mathcal{H}} \le B$ *for all* $\mathbf{f}(I, X) \in \mathbb{R}^d$,
   (b) $\left\|\eta(t, \mathbf{f}(I, X)) - \eta(t, \mathbf{f}(I, X)')\right\|_{\mathcal{H}} \le M \left\|\mathbf{f}(I, X) - \mathbf{f}(I, X)'\right\|_2$ *(Lipschitz condition),*
3. *there exists $\delta > 0$ such that $E\left(\|e\|_{\mathcal{H}}^{2+\delta}\right) < \infty$,*
4. *$k/n \to 0$, $k/\log n \to \infty$, $\sum_{n=1}^{\infty} (\log n/k)^{\delta/2} < \infty$,*

then

$$\|\eta_n(t, \mathbf{f}(I, X)) - \eta(t, \mathbf{f}(I, X))\|_{\mathcal{H}} = O\left(\left(\frac{k}{n}\right)^{1/d} + \sqrt{\frac{\log n}{k}}\right) \quad a.s.$$

Remark 3. Proposition 1 is a direct consequence of Theorem 1 from Liam et al. [30] and also from the fact that, in $\mathbb{R}^d$, under condition 1 due to Lebesgue's Differentiation Theorem, we have that $P(B(x, h)) = O\left(h^d\right)$. In what follows $\mathcal{H} = L^2[a, b]$.

Remark 4. Kudaszow and Vieu [28] established general uniform consistency results and convergence rates for r-NN generalized regression estimators, where the observed variable takes values in an abstract space. We could have also applied these results to achieve consistency.

Once we have settled the general framework we proceed to define explicitly the theoretical and empirical objective functions for classification, principal components and regression.

3. Supervised and unsupervised classification

Let $X(t) \in L^2[a, b]$ be a random process and K the number of groups. For supervised classification, as well as unsupervised classification (when the number of clusters K is known) we have a function $g : L^2[a, b] \to \{1, \ldots, K\}$ that determines to which group (or cluster) each trajectory belongs to. We denote the space partition by $G_k = g^{-1}(k),\ k = 1, \ldots, K$.

For a fixed integer $d < p$, our aim is to find a set $I \subset \{1, \ldots, p\},\ \sharp I \le d$, where the population objective function, given by,

$$h(I) = 1 - \sum_{k=1}^{K} P(g(X) = k, g(Z(I)) = k), \tag{4}$$

attains its minimum. Function (4) measures the difference between the partition of the space considering the original and the blinded trajectory.

Remark 5. It is clear that instead of minimizing (4) one could maximize $h^*(I) = \sum_{k=1}^{K} P(g(X) = k, g(Z(I)) = k)$, which is the objective function defined by Fraiman et al. [16] for the multivariate case.

The empirical version of Eq. (4) is given by

$$h_n(I) = 1 - \frac{1}{n}\sum_{k=1}^{K}\sum_{j=1}^{n} I_{\{g_n(X_j)=k\}} I_{\{g_n(\hat{X}_j(I))=k\}}, \tag{5}$$

where the function $g_n : L^2[a, b] \to \{1, \ldots, K\}$ is the empirical space partitioning function, i.e., it determines to which group (partition subset) each empirical process belongs. In this case the partition of the space is given by $G_k^{(n)} = g_n^{-1}(k)$, $k = 1, \ldots, K$, meaning that (5) measures the proportion of observations that are classified in different groups by the original and the blinded processes.

To prove consistency results, in addition to **H1**, the following assumptions are requested.

HC1. (a) The partition of the space is strongly consistent, i.e., given $\epsilon > 0$, there exists a set $A(\epsilon) \subset L^2[a, b]$, with $P(X \in A(\epsilon)) > 1 - \epsilon$ such that, for all $r > 0$, $\sup_{x \in C(\epsilon, r)} \left| I_{\{g_n(x)=k\}} - I_{\{g(x)=k\}} \right| \to_{a.s.} 0$ as $n \to \infty$ for $k = 1, \ldots, K$, where $C(\epsilon, r) = A(\epsilon) \cap \mathcal{K}_r$ with $\mathcal{K}_r$ is an increasing sequence of compact sets, such that $P(X \in \mathcal{K}_r) \to 1$ when $r \to \infty$.

(b) $d(X, \partial G_k^n) - d(X, \partial G_k) \to_{a.s.} 0$ when $n \to \infty$, where ∂G_k (respec. ∂G_k^n) is the boundary of G_k (respec. G_k^n).

HC2. $P(d(Z(I), \partial G_k) < \delta) \to 0$ when $\delta \to 0$ for $k = 1, \ldots, K$.

HC3. The distribution is non-degenerated, i.e. for every $\delta > 0$, $P(X \in B(x, \delta)) > 0$ for almost every $x \in L^2[a, b]$.

Theorem 1. *Let $\{X_j(t) : t \in [a, b]\}$ be iid realizations of the stochastic process $X(t)$. Given d, $1 \le d \le p$, let $\mathcal{I}_d$ be the family of all the subsets $\{1, \ldots, p\}$ with cardinality smaller than or equal to d and let $\mathcal{I}_0 \subset \mathcal{I}_d$ be the family of all the subsets for which the minimum of Eq. (4) is attained. Under* **H1**, **HC1**, **HC2** *and* **HC3** *we have that for each $I_n \in \mathcal{I}_n$, there exists $n_0(\omega)$ such that, for every $n > n_0(\omega)$, with probability one $I_n \in \mathcal{I}_0$.*

The proof is given in the Appendix.

4. Principal components

Let $\{X(t) : t \in [a, b]\}$ be a stochastic process with finite second moment $\left(E\left(\|X^2\|_{L^2[a,b]}\right) < \infty\right)$, for almost every $t \in [a, b]$, with continuous trajectories. Without loss of generality we assume that it has zero mean ($E(X(t)) = 0$). We denote $v(t, s) = E(X(t)X(s))$ its covariance function. As in the finite-dimensional case, the covariance function has associated a linear operator, $\boldsymbol{\Gamma} : L^2[a, b] \to L^2[a, b]$ defined as

$$(\boldsymbol{\Gamma} u)(t) = \int_a^b v(t, s)u(s)ds \quad \text{for all } u \in L^2[a, b]. \tag{6}$$

We assume that $\|v\|^2_{L^2[a,b]\times L^2[a,b]} = \int_a^b \int_a^b v^2(t, s)dtds < \infty$. Cauchy–Schwarz's inequality implies that $|\boldsymbol{\Gamma} u|^2 \le \|v\|^2 |u|^2$, where $|u|$ stands for the standard norm in the space $L^2[a, b]$, while $\|v\|$ denotes the norm in the space $L^2([a, b] \times [a, b])$. Therefore, $\boldsymbol{\Gamma}$ is a self-adjoint continuous linear operator. Moreover, $\boldsymbol{\Gamma}$ is a Hilbert–Schmidt operator.

$\mathcal{F}$ stands for the Hilbert space of such operators with inner product given by

$$\langle \boldsymbol{\Gamma_1}, \boldsymbol{\Gamma_2} \rangle_{\mathcal{F}} = \text{trace}(\boldsymbol{\Gamma_1}\boldsymbol{\Gamma_2}) = \sum_{j=1}^{\infty} \langle \boldsymbol{\Gamma_1} u_j, \boldsymbol{\Gamma_2} u_j \rangle_{L^2[a,b]},$$

where $\{u_j : j \ge 1\}$ is any orthonormal basis of $L^2[a, b]$ and $\langle u, v \rangle$ denotes the ordinary inner product in $L^2[a, b]$.

If we consider the basis of the eigenfunctions of $\boldsymbol{\Gamma}$, $\{\phi_j : j \ge 1\}$, we have that $\|\boldsymbol{\Gamma}\|^2_{\mathcal{F}} = \langle \boldsymbol{\Gamma}, \boldsymbol{\Gamma} \rangle_{\mathcal{F}} = \sum_{j=1}^{\infty} \lambda_j^2 = \int_a^b \int_a^b v^2(t, s)dtds < \infty$, where $\{\lambda_j : j \ge 1\}$ are the corresponding eigenvalues of $\boldsymbol{\Gamma}$. In what follows we assume that all eigenvalues are different. For any random variable, U, defined as a linear combination of the process $\{X(s)\}$, i.e. $U = \int_a^b \alpha(t)X(t)dt = \langle \alpha, X \rangle$, $\alpha \in L^2[a, b]$, we have that $\text{var}(U) = E(U^2) = \int_a^b \int_a^b \alpha(t)v(t, s)\alpha(s)dsdt = \langle \alpha, \boldsymbol{\Gamma}\alpha \rangle$.

The first principal component is defined as the random variable $U_1 = \langle \alpha_1, X \rangle_{L^2[a,b]}$, such that,

$$\text{Var}(U_1) = \sup_{\|\alpha\|_{L^2[a,b]}=1} \text{Var}\left(\langle \alpha, X \rangle_{L^2[a,b]}\right) = \sup_{\|\alpha\|_{L^2[a,b]}=1} \langle \alpha, \boldsymbol{\Gamma}\alpha \rangle_{L^2[a,b]},$$

and the kth principal component as the variable

$$U_k = \langle \alpha_k, X \rangle_{L^2[a,b]}, \tag{7}$$

such that

$$\begin{aligned}\mathrm{Var}(U_k) &= \sup \mathrm{Var}\left(\langle \alpha, X\rangle_{L^2[a,b]}\right) = \sup \langle \alpha, \boldsymbol{\Gamma}\alpha\rangle_{L^2[a,b]} \\ &\quad \text{subject to } \|\alpha\|_{L^2[a,b]} = 1 \text{ and } \left\langle \alpha, \alpha_j\right\rangle_{L^2[a,b]} = 0 \quad \text{for } j = 1, \ldots, k-1.\end{aligned}$$

Therefore, if $\lambda_j > \lambda_{j+1}$, are the eigenvalues of Γ, Riesz Theorem [40] entails that the principal components are obtained from the corresponding eigenfunctions. Let $\{\alpha_k, k \geq 1\}$ be the basis of eigenfunctions of the covariance linear operator $\boldsymbol{\Gamma}$, then $U_k = \langle \alpha_k, X\rangle$ is the kth principal component and $\mathrm{Var}(U_k) = \lambda_k$.

Assuming that the first $l < p$ principal components yield a good representation of the original process, for each $I \in \mathcal{I}_d$, we define, $U_k(I) = \langle \alpha_k, Z(I)\rangle_{L^2[a,b]}$. Our goal is to find a subset I such that $U_k(I)$ is as closest as possible to U_k, for every $k = 1, \ldots, l$. The objective function is given by,

$$h(I) = \sum_{k=1}^{l} E\left((U_k - U_k(I))^2\right). \tag{8}$$

This function measures the mean value of the square distance between the projection with the original trajectory and with the blinded one. Given $d < p$, our aim is to find the set $I \in \mathcal{I}_d$ that minimizes the objective function (8).

In order to give the empirical version of (8), we shall give the empirical counterpart of the principal components. Then, let $v_n(t, s)$ be the empirical covariance function, i.e., $v_n(t, s) = \frac{1}{n}\sum_{j=1}^{n} X_j(t)X_j(s)$, and $\boldsymbol{\Gamma}_n$ its corresponding linear operator, given by

$$\boldsymbol{\Gamma}_n = \frac{1}{n}\sum_{j=1}^{n} \mathbf{V}_j, \tag{9}$$

where $\mathbf{V}_j$ is the linear operator defined as $\left(\mathbf{V}_j u\right)(t) = \int_a^b X_j(t)X_j(s)u(s)ds$. Hence, Fubini's Theorem entails that, $E\left(\mathbf{V}_j\right) = \boldsymbol{\Gamma}$ for all $1 \leq j \leq n$.

Then, the empirical version of function (8) is

$$h_n(I) = \sum_{k=1}^{l} \frac{1}{n}\sum_{j=1}^{n} \left(U_k^j - U_k^j(I)\right)^2, \tag{10}$$

where $U_k^j = \left\langle \alpha_k^n, X_j\right\rangle_{L^2[a,b]}$ and $U_k^j(I) = \left\langle \alpha_k^n, \hat{X}_j(I)\right\rangle_{L^2[a,b]}$. The weights of the kth empirical principal component for the random sample $\{X_1, \ldots, X_n\}$ are α_k^n. In addition, as a consequence of Riesz's Theorem, we have that, $\left\{\alpha_k^n, k \geq 1\right\}$ is the basis of eigenfunctions of $\boldsymbol{\Gamma}_n$ associated to $\left\{\lambda_k^n, k \geq 1\right\}$.

Our aim is to find the set $I \in \mathcal{I}_d$ that minimizes (10).

To obtain consistency results, in addition to **H1**, the following conditions are needed:

H2. $E\left(\|X - Z(I)\|^2_{L^2[a,b]}\right) < \infty$.

HP1. $E\left(\|X\|^2_{L^2[a,b]}\right) < \infty$ and $\|v\|_{L^2([a,b]\times[a,b])} = \int_a^b \int_a^b v^2(t, s)dtds < \infty$.

Theorem 2. *Let $\{X_j(t) : t \in [a, b]\}$ be a stochastic process satisfying (7). Given d, $1 \leq d \leq p$, let $\mathcal{I}_d$ be the family of all the subsets of $\{1, \ldots, p\}$ with cardinality smaller than or equal to d and let $\mathcal{I}_0 \subset \mathcal{I}_d$ be the family of subsets in which the minimum of Eq. (8) is obtained. Then, under* **H1**, **H2** *and* **HP1** *we have that for each $I_n \in \mathcal{I}_n$, $I_n \in \mathcal{I}_0$ ultimately, i.e. $I_n = I_0$ with $I_0 \in \mathcal{I}_0$ for all $n > n_0(\omega)$, with probability one.*

The proof is given in the Appendix.

5. The functional linear model

In the functional linear model the response can be either scalar or functional, in this section we analyze both cases.

5.1. Linear model with scalar response

Let $Y \in \mathbb{R}$ and $X \in L^2[a, b]$, the linear model with scalar response is given by,

$$Y = \int_a^b \beta(t)X(t)dt + \varepsilon, \tag{11}$$

where $\beta \in L^2[a, b]$ and ε is a random variable such that $E(\varepsilon) = 0$, $E(\varepsilon^2) < \infty$ and $E(X(t)\varepsilon) = 0$, for almost every $t \in [a, b]$.

Let $\beta_0 \in L^2[a,b]$ such that

$$\beta_0 = \arg\min_{\beta \in L^2[a,b]} E\left(\left(Y - \int_a^b \beta(t)X(t)dt\right)^2\right). \tag{12}$$

To ensure existence and uniqueness of β_0 under this setting additional conditions are required.

Assume that $E\left(\|X\|^2_{L^2[a,b]}\right) < \infty$. Without loss of generality, we may assume that the stochastic process is centered, i.e., $E(X(t)) = 0$ for almost every $t \in [a,b]$. As a consequence of Fubini's Theorem, is clear that $E(Y) = 0$, since $E(X(t)) = 0$ and $E(\varepsilon) = 0$.

Let $\boldsymbol{\Gamma}$ be the covariance operator of the random element X given in (6).

Denote $\overline{Im\,(\boldsymbol{\Gamma})}$ to the closure of $Im\,(\boldsymbol{\Gamma}) = \left\{\boldsymbol{\Gamma}u,\, u \in L^2[a,b]\right\}$. Under mild regular conditions the existence and uniqueness of (12) in $\overline{Im\,(\boldsymbol{\Gamma})}$ is ensured. For a detailed explanation see [13], Chapter 2.

We define the objective function as,

$$h(I) = E\left(\left(\int_a^b \beta_0(t)X(t)dt - \int_a^b \beta_0(t)Z(I)(t)dt\right)^2\right). \tag{13}$$

It measures the mean square distance between the predicted value considering the original variables and the blinded ones. Given $d < p$, our goal is to find a subset $I \in \mathcal{I}_d$ that minimizes Eq. (13).

In order to give the empirical counterpart of (13) we must provide an estimate of β. There are several ways to face this problem. One alternative is to estimate the covariance operator $\boldsymbol{\Gamma}$. Then project the data onto a finite dimensional space that grows, as the sample size grow, typically by means of the principal components. Sometimes this method is combined with a smoothing procedure. Another alternative is to represent β in a basis of functions of $L^2[a,b]$ satisfying (12) adding a penalty to obtain a regular solution. These basis may not be orthonormal, for instance Fourier or Spline basis.

As mentioned in the introduction this problem has been studied by several authors, among them we can mention Cardot et al. [6], Cai and Hall [5], Hall and Horowitz [21], Li and Hsing [31], James et al. [25]. We consider the estimator proposed by Cardot et al. [6], which is a strong consistent estimate of β.

The empirical version of the objective function (13) is given by

$$h_n(I) = \frac{1}{n}\sum_{j=1}^{n}\left(\int_a^b \beta_n(t)X_j(t)dt - \int_a^b \beta_n(t)\hat{X}_j(I)(t)dt\right)^2, \tag{14}$$

where β_n is the estimate of β_0. Our aim is to find the set $I \in \mathcal{I}_d$ that minimizes (14).

In order to obtain consistency results, in addition to **H1** and **H2**, the following hypotheses are needed:

HR1. $\|\beta_n - \beta_0\|_{L^2[a,b]} \to_{a.s.} 0$.

HR2. $E\left(\|X\|^2_{L^2[a,b]}\right) < \infty$.

It is important to note that the regression estimate, β_n, introduced in [6] satisfies **HR1**.

Theorem 3. *Let $\left\{\left(Y_j, X_j(t)\right) \in \mathbb{R} \times L^2[a,b]\right\}$ be iid stochastic processes satisfying (11). Given d, $1 \le d \le p$, let $\mathcal{I}_d$ be the family of all the subsets of $\{1,\ldots,p\}$ with cardinality smaller than or equal to d and let $\mathcal{I}_0 \subset \mathcal{I}_d$ be the family of all the subsets for which the objective function (13) attains its minimum. Under* **H1**, **H2**, **HR1** *and* **HR2** *we have that for each $I_n \in \mathcal{I}_n$, there exists $n_0(\omega)$ such that for each $n > n_0(\omega)$, with probability one, $I_n \in \mathcal{I}_0$.*

The proof is given in the Appendix.

5.2. Linear model with functional response

Let $Y \in L^2[c,d]$ and $X \in L^2[a,b]$, the regression model with functional response is given by,

$$Y(s) = \int_a^b \beta(t,s)X(t)dt + \varepsilon(s) \quad s \in [c,d], \tag{15}$$

where $\beta \in L^2([a,b]\times[c,d])$ and $\varepsilon(s)$ is a random variable for each s, such that $E(\varepsilon(s)) = 0$ and $E(X(t)\varepsilon(s)) = 0$, for almost every $t \in [a,b]$, $s \in [c,d]$.

Let $\beta_0 \in L^2([a,b]\times[c,d])$ such that

$$\beta_0 = \arg\min_{\beta \in L^2([a,b]\times[c,d])} E\left(\left\|Y - \int_a^b \beta(t,\cdot)X(t)dt\right\|^2_{L^2[c,d]}\right).$$

Under this setting existence and uniqueness of β_0 are not provided unless some additional hypothesis is established. Following the same ideas as in the case of scalar response these properties can be obtained.

Without loss of generality, we may assume that $E(X(t)) = 0$ for almost every $t \in [a, b]$, which entails that $E(Y(s)) = 0$, for almost every $s \in [c, d]$, and also that $E\left(\|X\|^2_{L^2[a,b]}\right) < \infty$ and $E\left(\|Y\|^2_{L^2[c,d]}\right) < \infty$. Let $\boldsymbol{\Gamma}_X$ be the covariance operator of X.

Then under mild conditions, $\beta_0 \in L^2([a, b] \times [c, d])$ and also the existence and uniqueness of the solution in the orthogonal space of the kernel of the covariance operator of X are obtained. See [13], Chapter 2.

Once more, following the same ideas as in the case of scalar response, the objective function is given by,

$$h(I) = E\left(\left\|\int_a^b \beta_0(t, s)X(t)dt - \int_a^b \beta_0(t, s)Z(I)(t)dt\right\|^2_{L^2[c,d]}\right). \tag{16}$$

This function measures the mean square distance between the predicted functions considering the original process and the blinded one. Then, given $d < p$, we search for a subset $I \in \mathcal{I}_d$ that minimizes the objective function (16).

To give the empirical version of (16), we need an estimator of β, this problem has been studied by several authors, for instance, Yao et al. [45] and Müller and Yao [36]. Yao et al., introduce the following estimate

$$\beta_n(t, s) = \sum_{j'=1}^{J'} \sum_{j=1}^{J} \frac{\sigma^n_{jj'}\alpha^n_{X,j}(t)\alpha^n_{Y,j'}(s)}{\lambda^n_{X,j}},$$

where $\sigma^n_{jj'}$ is an estimate of $E\left(\xi_{X,j}\xi_{Y,j'}\right)$. They show that β_n converges in probability to β_0. Hence, the empirical version of (16) is

$$h_n(I) = \frac{1}{n}\sum_{j=1}^{n} \left\|\int_a^b \beta_n(t, s)X_j(t)dt - \int_a^b \beta_n(t, s)\hat{X}_j(I)(t)dt\right\|^2_{L^2[c,d]}. \tag{17}$$

Our goal is to find a subset $I \in \mathcal{I}_d$, that minimizes (17). To obtain convergence results, in addition to **H1** and **H2**, we need the following conditions:

HRF1. $\|\beta_n - \beta_0\|_{L^2([a,b]\times[c,d])} \to_{p.} 0$.

HRF2. $E\left(\|X\|^2_{L^2[a,b]}\right) < \infty$ y $E\left(\|Y\|^2_{L^2[c,d]}\right) < \infty$.

It is important to remark that the estimate β_n, introduced by Yao et al. [45] satisfies **HRF1**.

Then we have the following consistency result.

Theorem 4. *Let $\left\{\left(Y_j(s), X_j(t)\right) \in L^2[c, d] \times L^2[a, b]\right\}$ be iid stochastic processes satisfying (15). Given d, $1 \le d \le p$, let $\mathcal{I}_d$ be the family of all the subsets of $\{1, \ldots, p\}$ with cardinality smaller than or equal to d and let $\mathcal{I}_0 \subset \mathcal{I}_d$ be the family of all the subsets where the minimum of the function (16) is attained. Under* **H1**, **H2**, **HRF1** *and* **HRF2** *given $I_n \in \mathcal{I}_n$ it can be shown that $P(I_n \in \mathcal{I}_0) \to 1$.*

The proof is given in the Appendix.

Remark 6. If β_n is a strong consistent estimate of β_0, then replacing hypothesis **HRF1** by

HRF1*. $\|\beta_n - \beta_0\|_{L^2([a,b]\times[c,d])} \to_{a.s.} 0$

strong consistency is attained in Theorem 4.

6. Numerical aspects

This section has two main parts. First we introduce a stochastic algorithm to find the optimal subset of functions. Then we illustrate the performance of our proposal analyzing some well known data sets.

6.1. The algorithm

As mentioned in Section 2 our aim is to select, I_d, a subset of $\mathcal{A} = \{f_1, \ldots, f_p\}$, that better captures the information of a stochastic process relative to a statistical procedure, i.e., $h_n(I_d)$ must be "small".

Before describing the algorithm we shall discuss when to consider that $h_n(I_d)$ is small enough. For the case of classification since $h_n(I_d)$ is the proportion of observations classified into different groups when we use the original or the blinded trajectories, i.e. the "matching error rate", it is clear how to establish that $h_n(I_d)$ is small enough. However for the cases of regression and PCA $h_n(I_d)$ depends on the units in which the functions are measured (regression and PCA). Therefore, our

proposal is to rescale the objective function, $\widetilde{h}_n$, to make it independent of the units of the data. Then, a subset of functions will be chosen if

$$\widetilde{h}_n(I_d) < \epsilon, \tag{18}$$

where ϵ is a positive constant that must be supplied by the user.

We exhibit the rescaled objective function for each of the problems analyzed in this work.

For the case of principal components we suggest to rescale the objective function as follows,

$$\widetilde{h}_n(I_d) = \frac{h_n(I_d)}{\sum_{k=1}^{l} \frac{1}{n} \sum_{j=1}^{n} \left(U_k^j\right)^2},$$

where $h_n(I_d)$ is given by Eq. (10). In an analogue way, the rescaled objective function for the linear functional model with scalar response is given by,

$$\widetilde{h}_n(I_d) = \frac{h_n(I_d)}{\frac{1}{n} \sum_{j=1}^{n} \left(\int_a^b \beta_n(t) X_j(t) dt\right)^2},$$

where $h_n(I_d)$ is given by Eq. (14). Finally, we consider the linear regression problem when the response is functional and the rescaled objective function is,

$$\widetilde{h}_n(I_d) = \frac{h_n(I_d)}{\frac{1}{n} \sum_{j=1}^{n} \left\| \int_a^b \beta_n(t,s) X_j(t) dt \right\|^2_{L^2[c,d]}},$$

where $h_n(I_d)$ is defined by Eq. (17). In the last three cases the rescaled objective functions measure the relative difference between the procedure carried out with the blinded processes and with the original processes.

Remark 7. For classification $\widetilde{h}_n(I_d) \equiv h_n(I_d)$, and ϵ is the maximum matching error rate allowed by the user.

It is clear that if p is moderate or large it is not feasible to analyze the $2^p - 1$ subsets of functions, hence we need to provide a numerical strategy to tackle this problem.

We introduce a forward algorithm that has two main steps. On the first one, an exhaustive search over all the subsets with cardinality d_1 (where $d_1 \ll d$) is performed. While the second step is a forward stochastic search.

Exhaustive Step

Consider all the subsets of functions with cardinality smaller than or equal to d_1. Keep the N_0 subsets with smaller $\widetilde{h}_n(\widetilde{I}_{\widetilde{d}})$, where $\widetilde{I}_{\widetilde{d}} = \widetilde{I}_1, \ldots, \widetilde{I}_{N_0}$.

If there is at least one of them satisfying condition (18), stop the algorithm.

Among all the subsets satisfying condition (18) retain the subset with smallest cardinality, in case of ties, keep the subset with smallest $\widetilde{h}_n$.

Otherwise, if condition (18) is not satisfied, proceed with the Stochastic Step, where the input is the subsets of functions given by $\widetilde{I}_1, \ldots, \widetilde{I}_{N_0}$.

Stochastic Step

Repeat until condition (18) is attained.

For each subset $I_{\widetilde{d}}$, where $d = 1, \ldots, N_0$.

Choose at random without replacement $N_1, i_1, \ldots, i_{N_1}$ functions from $I \backslash \widetilde{I}_{\widetilde{d}}$.

Consider the subsets $\widetilde{I}_{\widetilde{d},j} =: \widetilde{I}_{\widetilde{d}} \bigcup \{i_j\}$ for $j = 1, \ldots, N_1$.

Compute $\widetilde{h}_n(\widetilde{I}_{\widetilde{d},j})$ for $j = 1, \ldots, N_1$.

Retain the N_0, subsets with smaller $\widetilde{h}_n$, with a slight abuse of notation denote them $\widetilde{I}_1, \ldots, \widetilde{I}_{N_0}$.

If there is at least one of them satisfying condition (18), stop the algorithm. Among all the subset satisfying condition (18) retain the subset with smallest cardinality, in case of ties, keep the subset with smallest $\widetilde{h}_n$.

Otherwise, if condition (18) is not satisfied proceed with a revision step.

For each subset $I_{\widetilde{d}}$, where $d = 1, \ldots, N_0$.

Replace at random, one at a time, each element of $\widetilde{I}_{\widetilde{d}}$, if $\widetilde{h}_n$ decreases keep the best subset.

If condition (18) is attained stop the algorithm.

Otherwise, run the Stochastic Step from the top.

The algorithm depends on several parameters, namely, ϵ, d_1, N_0 and N_1. We have already discussed the role that ϵ plays. The other three parameters can be easily settled and are all in relation with the cardinality $\mathcal{A}$. d_1 denotes the maximum cardinality analyzed in the exhaustive step. This number should be small, specially if the cardinality of $\mathcal{A}$ is big. N_0 is the number of subsets retained after the exhaustive step and N_1 is the number of variables added to each potential subset chosen either in the exhaustive step or once the stochastic step has been run without achieving condition (18), this constants can be moderate, allowing several bifurcations we seek to avoid local minima.

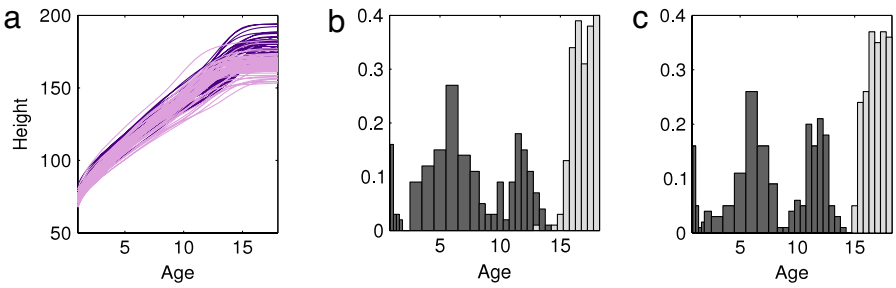

Fig. 1. (a) Height curves, the darker ones are from the boys and the lighter ones for the girls. (b) The dark grey histogram show the first age chosen by the blinding procedure, for $r = 3$, the light grey histogram exhibits the second age chosen. (c) Exhibits the same features that Fig. 1(b) for $r = 5$.

6.2. Real data examples

In this section we will analyze the behavior of our procedure on several real data examples. Routines are available upon request to the authors.

The growth data set

The growth data set has been analyzed in several opportunities, and corresponds to the Berkeley growth study. It is available in the functional data analysis packages for Matlab and R. The data set consists of the height measurements of 54 girls and 39 boys, that were measured on 31 opportunities from 1 to 18 years. The original data was smoothed by monotonic cubic regression spline, the curves are shown in Fig. 1(a).

Our aim is to classify the growth curves for a sample of boys and girls. Since there is no learning and test sample we shall proceed following the ideas from Lopez-Pintado and Romo [32]. Therein, they suggest to separate the data into four groups of the same size and consider one of them as test group. The remaining observations constitute the learning set. The data set is classified by fitting a functional generalized additive model (see [11]) and the number of misclassified observation is computed from the test group. All these steps are repeated changing the test group 4 times and the total error rate is defined as the number of misclassified observations over the total number of observations. It is clear that different partitions yield different misclassification rates, to avoid this effect we repeat this procedure 50 times. The misclassification rate is between 3.23% and 10.75%, the mean is 6.17% and the median is 6.45%, which are very good results in comparison to those obtained by Lopez-Pintado and Romo [32].

Our goal is to find which are the key time instants that determine that boys and girls have different growth patterns. Then, if the set $\mathcal{A}$ is conformed by the evaluations on the grid of time points, we apply the blinding procedure to these functions. The cardinality of $\mathcal{A}$ is 31, hence it is feasible to perform an exhaustive search. The nonparametric estimation is done by r-nearest neighbors, with $r = 3$ or 5. If $d_1 = 1$, the mean matching error is high 21.94% (respc. 22.09%) for $r = 3$ (respc. $r = 5$). Then we decided to set $d_1 = 2$, this means to look for the pair of functions with minimum matching error that belongs to $\mathcal{A}$. In these cases the mean matching error decreases to 4.75% (respc. 4.77%) for $r = 3$ (respc. $r = 5$). In Fig. 1(b) (respc. (c)) we show the histogram for the instants chosen by the blinding procedure, indicating with different tonalities of grey the first and second choice, in almost every case the final height is a relevant feature, boys tend to be taller than girls. For the other measure there are clearly three modes, the first one is the initial measurement (people tend to grow in the same percentile curve throughout their lives), the second one is close to 6 years (there is an acceleration in the growth speed) and the third one during the puberty (girls grow earlier than boys). Hence the time points chosen by the algorithm reflect important patterns on the growth charts.

It is important to remark that histograms corresponding to $r = 3$ or $r = 5$ are very similar.

The phoneme data set

In the following example, the data corresponds to the discretized log-periodograms, which are a widely used method for casting speech data in a form suitable for speech recognition. The learning data set, as well as the test data sets, contain 250 recordings of men pronouncing five phonemes. Each curve is recorded on a grid of 150 equally spaced time points. The phonemes are "sh" as in "she", "iy" as the vowel in "she", "dcl" as in "dark", "aa" as the vowel in "dark", and "ao" as the first vowel in "water". The label of each observation is known.

The data set is in the R package *fda.usc* (Functional Data Analysis and Utilities for Statistical Computing) and it is a part of the original one which can be found at http://www-stat.stanford.edu/ElemStatLearn.

The data set is classified fitting a functional generalized kernel additive models (see [11]) and only 6.4% observations of the test sample are misclassified.

The set of functions $\mathcal{A}$, is the evaluations on the 150 grid of time-points.

In order to apply the blinding procedure we must fix the values of the parameters involved at each stage in the algorithm. The maximum matching error rate is $\epsilon = 0.10$. For the exhaustive step we set $d_1 = 1$, since $\mathcal{A}$ has 150 functions. For the stochastic step we used $N_0 = 3$ or 6 and $N_1 = 5$ or 10. In addition, for the number of nearest neighbors for the nonparametric estimation we considered either $r = 5$ or 10. We performed 100 replicates for each parameter configuration.

Table 1
Mean number of functions chosen for each parameter configuration.

N_1	r	5		10	
	N_0	3	6	3	6
5		4.93	4.64	5.02	4.67
10		4.68	4.47	4.79	4.47

Table 2
Mean matching error rate.

N_1	r	5		10	
	N_0	3	6	3	6
5		8.48	8.31	8.44	8.34
10		8.32	8.20	8.34	8.36

Table 3
Functions chosen by an exhaustive search with their corresponding $\widetilde{h}$ values.

d_1	r			
	3		5	
	f_i	$\widetilde{h}$	f_i	$\widetilde{h}$
1	35	0.0013	37	0.0013
2	26, 35	0.0010	20, 37	0.00117
3	26, 36, 37	0.00094	19, 35, 37	0.00116
4	24, 35, 37, 38	0.00092	22, 24, 36, 37	0.00113

In Table 1 we can observe that the mean number of functions chosen is always between 4.5 and 5, in addition we may add that in every case between 3 and 6 functions were selected and that the median number of functions is always either four or five. Moreover, in Table 2 we report the mean matching error rate, which is in every case between 8.2% and 8.5%. The minimum matching error rate in every case is between 4 or 5%. The results obtained by the blinding procedure using different values for the parameters are practically the same.

The Canadian weather data

This data set has been introduced first by Ramsay and Silverman [37] and it is available in the R package *fda* (see [39]). The data contains daily temperature and rainfall observations over the course of a year measured on 35 monitoring stations from Canada. Several authors analyzed this data set in the context of scalar regression (see [20] and references therein), their aim is to predict the log annual precipitation from the temperature observations.

We estimate the functional linear model as proposed by Cardot et al. [6], i.e. considering an almost sure consistent estimate. This procedure has been numerically implemented by Goldsmith and Scheipl [20], they show that it has good empirical performance.

Our aim is to find the time periods where the temperature has greater influence in the prediction of the log annual precipitation amount. Hence, the set of functions $\mathcal{A}$, that we analyze is conformed by 40 local averages of the temperatures from non-overlapping intervals of 9 days, $f_1, \ldots, f_{40}$, and one local average with the remaining last five days, f_{41}. This means that f_1 is the mean temperature from day 1 to 9, f_2 is the mean temperature from day 10 to 18 and so on and so forth. We apply the blinding procedure to $\mathcal{A}$. The cardinality of $\mathcal{A}$ is 41, then it is feasible to perform an exhaustive search. The nonparametric estimation is done by r-nearest neighbors, with $r = 3$ or 5.

In Table 3 we exhibit the optimal functions chosen and also the value of the rescaled objective function, $\widetilde{h}$, for $d = 1, 2, 3$ or 4. We observe that $\widetilde{h}$ decreases on d, there is a bigger gain if we choose two functions instead of one, but practically no gain if more functions are chosen. The functions mainly retain information that corresponds with two larger periods of time, the first one during the summer and the other one during the autumn. In Fig. 2(a), the temperature functions are exhibited, the gray rectangles highlight the areas chosen by the exhaustive procedure. Even though, there are not much differences for $r = 3$ or 5 nearest neighbors the objective function shows better behavior for $r = 3$.

In addition, we also run the stochastic algorithm, first we fix the values of the parameters involved at each stage of the algorithm. For the exhaustive step, we set $d_1 = 1$. For the stochastic search, we fix, $\epsilon = 0.0012$, $N_0 = 3$ or 6 and $N_1 = 5$ or 10. The number of nearest neighbors remains the same as in the exhaustive search. We perform 100 replicates for each parameter configuration. For $r = 3$ in almost every case two functions are chosen (Fig. 2(b) to (e)), while for $r = 5$, even though the median number of functions is always 2, there is more dispersion (Fig. 2(f) to (i)).

In Fig. 3(a)–(d) we show the histograms for the functions chosen, for $r = 3$, by the algorithm for the different parameter configurations, when two functions are chosen, we indicate with different tonalities of grey each choice. In every case, one of the periods chosen is during the autumn and in most of the cases the other one is during the summer, this results are

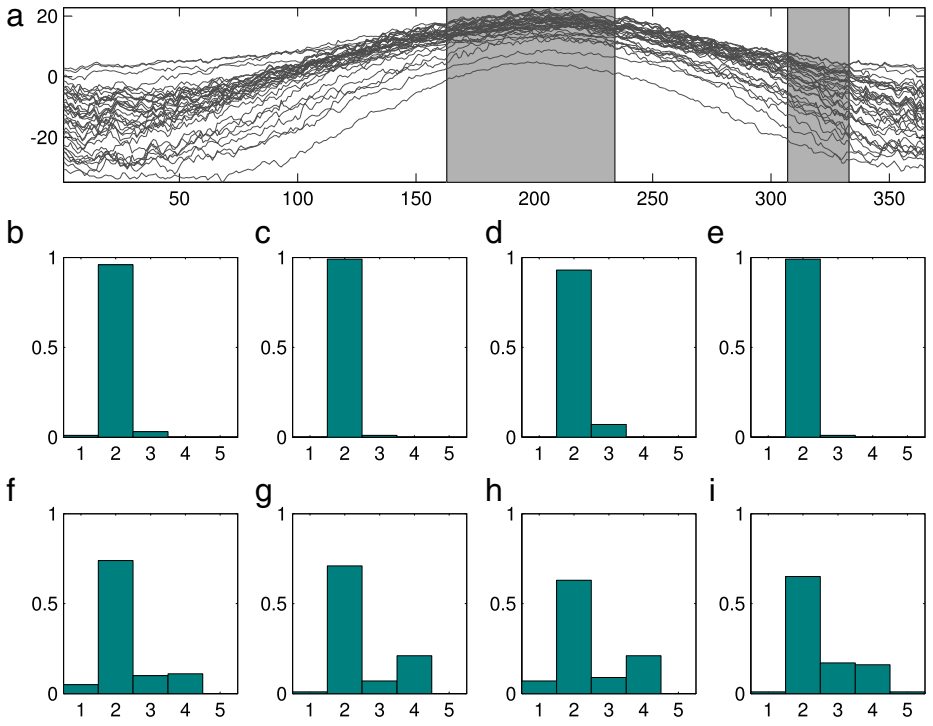

Fig. 2. (a) Temperature data set. (b)–(i) The histograms show the number of functions chosen by the algorithm for the different parameter choices. (b) $r = 3, N_0 = 3, N_1 = 5$. (c) $k = 3, N_0 = 3, N_1 = 10$. (d) $r = 3, N_0 = 6, N_1 = 5$. (e) $r = 3, N_0 = 6, N_1 = 10$. (f) $r = 5, N_0 = 3, N_1 = 5$. (g) $r = 5, N_0 = 3, N_1 = 10$. (h) $r = 5, N_0 = 6, N_1 = 5$. (i) $r = 5, N_0 = 6, N_1 = 10$.

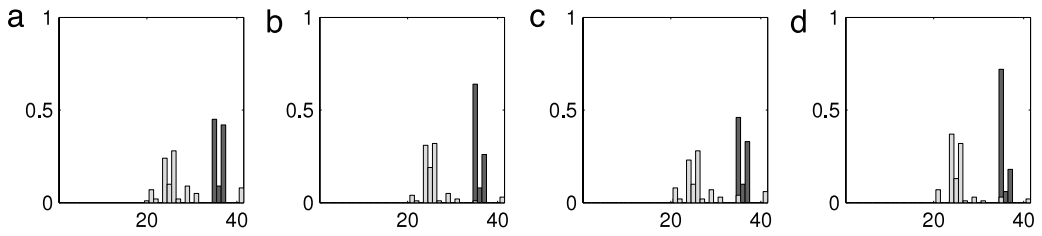

Fig. 3. The histograms show the distribution of functions chosen by the algorithm, for $r = 3$, when two functions are chosen for the different parameter configurations. (a) $N_0 = 3, N_1 = 5$. (b) $N_0 = 3, N_1 = 10$. (c) $N_0 = 6, N_1 = 5$. (d) $N_0 = 6, N_1 = 10$.

similar as those obtained by the exhaustive search. If more than two functions are chosen, in almost every case, two of them lay during the autumn and the remaining during the summer. The results for $r = 5$ are similar.

In addition, we change the set of functions considered in this example, we use the local averages during the twelve months, $\widetilde{\mathcal{A}}$. We run the exhaustive procedure, if $d_1 = 1$ the month chosen is November, which includes functions f_{35}, f_{36} and f_{37} from $\mathcal{A}$. If $d_1 = 2$, November is one of the months chosen while the other one is August for $r = 3$ and July for $r = 5$, these months are contained in the subsets of functions $f_{21}, \ldots, f_{27}$ from $\mathcal{A}$. In every case, the difference between the values of the $\widetilde{h}$ with $\mathcal{A}$ or with $\widetilde{\mathcal{A}}$ is smaller than 0.0002. These results are consistent with those obtained by Lee et al. [29] and by James et al. [25] that also analyzed this problem in the context of sparse regression.

Finally, we analyze this data set from a different perspective. Our goal is to determine temperature bands that have greater impact on the prediction of the logarithm of the annual precipitation. Hence, we consider the set of function $\mathcal{A} = \{f_1, \ldots, f_{12}\}$, where f_i denotes the occupation measure in a band of 5 Celsius degrees, i.e., $f_i = |\{t : (-35 + 5(i-1)) \leq X(t) < (-35 + 5i)\}|$, for $i = 1, \ldots, 12$. Since the cardinality of $\mathcal{A}$ is small we perform an exhaustive search, once more the nonparametric estimation has been performed for either $r = 3$ or $r = 5$ neighbors. In Table 4 we exhibit the optimal values of the rescaled objective function, $\widetilde{h}$. For $r = 5$ it is clear that the maximum gain is achieved if two functions are kept, for $r = 3$ it is not clear if two or three functions should be retained. In every case, one of the functions retained corresponds to an intermediate temperature, while the other one corresponds to a extreme value that can be either cold or hot, Fig. 4 is exhibited as example.

Table 4
Optimal values of $\widetilde{h}$ corresponding to the exhaustive search.

d_1	r	
	3	5
1	0.0036	0.0044
2	0.0030	0.0030
3	0.0027	0.0029
4	0.0025	0.0028

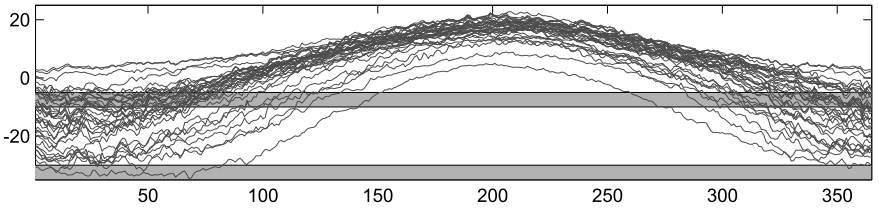

Fig. 4. Temperature bands chosen for $r = 3$.

7. Concluding remark

In this paper we address the problem of feature selection when the data is functional, we study several statistical procedures including classification, regression and principal components. One advantage of the blinding procedure is that it is very flexible since the features are defined by a set of functions, relevant to the problem being studied, proposed by the user. Our method is consistent under a set of quite general assumptions, and produces good results with the real data examples that we analyze. In any of the statistical procedure studies sparsity has been assumed.

From the numerical aspects, it is clear that if the cardinality of $\mathcal{A}$ is moderate or large an exhaustive search in unfeasible, hence a stochastic heuristic algorithm is introduced. The algorithm depends on several parameters, and an optimal way to choose them is beyond the scope of this paper.

A possible shortcut of our approach is that we are assuming that the cardinal p of the set of functions $\{f_1, \ldots, f_p\}$ is fixed, although it may be large. An extension to the case when $p \to \infty$ – besides a much more involved theory – would need in particular, some asymptotic results which ensure that assumption **H1** holds. Then, the regression function should fulfill Besicovitch condition, see for instance Forzani et al. [15]. An important issue is to extend our results on feature selection to this framework in the next future.

Several authors, among them Cuevas [7] and Aneiros et al. [2] suggest that the link between variable selection methods for high dimensional data and functional data should be deeply studied. In particular they focus on the idea considering a discrete representation of the functions and determine whether all the points are relevant for the statistical procedure carried out or if it is enough to observe some key points. Ferraty et al. [12] and Kneip and Sarda [27] consider different approaches to tackle this problem for the regression problem. We consider that the question the pose is very relevant (and even though in this work we deal with it), it should be deeply studied.

Appendix. Proofs

Proof of Theorem 1. Since $\mathcal{I}_d$ is finite it is enough to show that $\lim_{n\to\infty} h_n(I) = h(I)$ a.s., for all $I \in \mathcal{I}_d$. That is, for $k = 1, \ldots, K$,

$$\lim_{n\to\infty} \frac{1}{n} \sum_{j=1}^{n} I_{\{g_n(X_j)=k\}} I_{\{g_n(\hat{X}_j(I))=k\}} = P(g(X) = k, g(Z(I)) = k) \quad \text{a.s} \tag{19}$$

We define the unobservable trajectories $Z_1(I), \ldots, Z_n(I)$, where

$$Z_j(I)(t) = E_P(X_j(t)|\mathbf{f}(I, X_j)), \tag{20}$$

and denote by $Q_n^*(I)$ its empirical distribution.

Eq. (19) holds, if for every fixed k,

$$\lim_{n\to\infty} \frac{1}{n} \sum_{j=1}^{n} I_{\{g_n(X_j)=k\}} I_{\{g_n(Z_j(I))=k\}} = P(g(X) = k, g(Z(I)) = k) \quad \text{a.s.} \tag{21}$$

and

$$\lim_{n\to\infty}\frac{1}{n}\sum_{j=1}^{n} I_{\{g_n(X_j)=k\}}\left[I_{\{g_n(\hat{X}_j(I))=k\}}-I_{\{g_n(Z_j(I))=k\}}\right]=0 \quad \text{a.s.} \tag{22}$$

We derive (21) from the following statements,

$$\lim_{n\to\infty}\frac{1}{n}\sum_{j=1}^{n} I_{\{g_n(X_j)=k\}}I_{\{g_n(Z_j(I))=k\}}-I_{\{g(X_j)=k\}}I_{\{g(Z_j(I))=k\}}=0 \quad \text{a.s.} \tag{23}$$

and

$$\lim_{n\to\infty}\frac{1}{n}\sum_{j=1}^{n} I_{\{g(X_j)=k\}}I_{\{g(Z_j(I))=k\}}=P(g(X)=k,g(Z(I))=k) \quad \text{a.s.} \tag{24}$$

Eq. (24) follows from the Law of the Large Numbers. To proof (23) we observe that the left hand side of the equation is dominated by

$$\frac{1}{n}\sum_{\{X_j\in C(\epsilon,r)\}\cap\{Z(I)_j\in C(\epsilon,r)\}}|I_{\{g_n(X_j)=k\}}I_{\{g_n(Z_j(I))=k\}}-I_{\{g(X_j)=k\}}I_{\{g(Z_j(I))=k\}}|$$
$$+\frac{1}{n}\sum_{\{X_j\notin C(\epsilon,r)\}\cup\{Z(I)_j\notin C(\epsilon,r)\}}|I_{\{g_n(X_j)=k\}}I_{\{g_n(Z_j(I))=k\}}-I_{\{g(X_j)=k\}}I_{\{g(Z_j(I))=k\}}|,$$

where $C(\epsilon,r)$ is given in assumption **HC1(a)**. From **HC1(a)** it can be seen that the first term is clear that it converges to zero for every ϵ and r. On the other hand, the last term vanishes due to the Law of Large Numbers. Then (23) holds, and the proof of (21) is completed.

To prove Eq. (22), it is enough to show that

$$\sharp\{j: g_n(\hat{X}_j(I))=k, g_n(Z_j(I))\neq k, g_n(X_j)=k\}/n\to 0 \quad \text{a.s.}, \tag{25}$$

and

$$\sharp\{j: g_n(\hat{X}_j(I))\neq k, g_n(Z_j(I))=k, g_n(X_j)=k\}/n\to 0 \quad \text{a.s.} \tag{26}$$

We define the sets B, C and D as follows:

$$B=\{\omega\in\Omega: \|Z_j(I)-\hat{X}_j(I)\|^2_{L^2[0,1]}\to_{n\to\infty}0\ \forall j\},$$
$$C_j=\{\omega\in\Omega: d(X_j,\partial G_k^n)-d(X_j,\partial G_k)\to 0\} \quad \text{and} \quad C=\cap_{j=1}^{\infty}C_j,$$
$$D=\{\omega\in\Omega: d(\partial G_k^n,\partial G_k)\to_{n\to\infty}0\}.$$

From **HC1(b)** and **H1** we have that $P(B\cap C)=1$, and from **HC3** and **HC1(b)** we have that $P(D)=1$. Hence, $P(B\cap C\cap D)=1$. Then, given $\delta>0$ and $\omega\in B\cap C\cap D$, there exists $n_0=n_0(\omega,\delta)$ such that $max_{j=1,\dots,n}\|Z_j(I)-\hat{X}_j(I)\|^2_{L^2[0,1]}\leq\delta/2$.

Given $\omega\in B\cap C\cap D$, $\delta>0$ and $n\geq n_0(\omega,\delta)$ we have that:

$$\{j: g_n(\hat{X}_j(I))=k, g_n(Z_j(I))\neq k, g_n(X_j)=k\}\subseteq\{j: d(Z_j,\partial G_k)<2\delta\}.$$

This implies that the left hand side of (25) is dominated by $\frac{1}{n}\sharp\{j: d(Z_j,\partial G_k)<2\delta\}\leq\frac{1}{n}\sum_{j=1}^{n}I_{\{d(Z_j,\partial G_k)<2\delta\}}$, which converges a.s. to $P(d(Z_j,\partial G_k)<2\delta)$ when $n\to\infty$.

Finally, from **HC2** we get that $\lim_{\delta\to 0}P(d(Z,\partial G_k)<2\delta)=0$, which concludes the proof of (25). The proof of (26) is completely analog. □

Proof of Theorem 2. In order to prove our statement it is enough to show that (10) converges to (8) a.s. To simplify notation, without loss of generality, we consider only one principal component (i.e., l = 1), which will be denoted by $\alpha=\alpha_1$. Then we have,

$$h_n(I)=\frac{1}{n}\sum_{j=1}^{n}(U_1^j-U_1^j(I))^2=\frac{1}{n}\sum_{j=1}^{n}((\alpha^n,X_j)-(\alpha^n,\hat{X}_j(I)))^2=\frac{1}{n}\sum_{j=1}^{n}\left(\int_a^b\alpha^n(t)(X_j(t)-\hat{X}_j(I)(t))dt\right)^2,$$

and

$$h(I)=E_P\left((U_1-U_1(I))^2\right)=E_P\left(((\alpha,X)-(\alpha,Z(I)))^2\right)=E_P\left(\left(\int_a^b\alpha(t)(X(t)-Z(I)(t))dt\right)^2\right).$$

We define the unobservable functions $Z_1(I), \ldots, Z_n(I)$, where

$$Z_j(I)(t) = E\left(X_j(t)|(f_{i_1}(X_j)), \ldots, f_{i_d}(X_j)\right), \tag{27}$$

and denote by $Q_n^*(I)$ its empirical distribution.

$$h_n(I) = \frac{1}{n}\sum_{j=1}^{n}\left(\int_a^b \alpha^n(t)\left(X_j(t) - Z_j(I)(t)\right)dt\right)^2 \tag{28}$$

$$+\frac{1}{n}\sum_{j=1}^{n}\left(\int_a^b \alpha^n(t)\left(Z_j(I)(t) - \hat{X}_j(I)(t)\right)dt\right)^2 \tag{29}$$

$$+\frac{2}{n}\sum_{j=1}^{n}\left(\int_a^b \alpha^n(t)\left(X_j(t) - Z_j(I)(t)\right)dt\right)\left(\int_a^b \alpha^n(t)\left(Z_j(I)(t) - \hat{X}_j(I)(t)\right)dt\right). \tag{30}$$

First, we show that (28) converges to (8).

$$\frac{1}{n}\sum_{j=1}^{n}\left(\int_a^b \alpha^n(t)\left(X_j(t) - Z_j(I)(t)\right)dt\right)^2 = \frac{1}{n}\sum_{j=1}^{n}\left(\int_a^b \left(\alpha^n(t) - \alpha(t)\right)\left(X_j(t) - Z_j(I)(t)\right)dt\right)^2 \tag{31}$$

$$+\frac{1}{n}\sum_{j=1}^{n}\left(\int_a^b \alpha(t)\left(X_j(t) - Z_j(I)(t)\right)dt\right)^2 \tag{32}$$

$$+\frac{2}{n}\sum_{j=1}^{n}\left(\int_a^b \left(\alpha^n(t) - \alpha(t)\right)\left(X_j(t) - Z_j(I)(t)\right)dt\right)\left(\int_a^b \alpha(t)\left(X_j(t) - Z_j(I)(t)\right)dt\right). \tag{33}$$

By Cauchy–Schwarz's inequality we see that the right hand side of (31) converges a.s. to zero,

$$\frac{1}{n}\sum_{j=1}^{n}\left(\int_a^b \left(\alpha^n(t) - \alpha(t)\right)\left(X_j(t) - Z_j(I)(t)\right)dt\right)^2 \le \|\alpha^n - \alpha\|^2_{L^2_{[a,b]}}\frac{1}{n}\sum_{j=1}^{n}\|X_j - Z_j(I)\|^2_{L^2_{[a,b]}}.$$

Dauxois et al. [8], under mild regular conditions, show strong consistency of the eigenvalues and their associated eigenvectors (see Propositions 2 and 4 therein). They establish that it is enough to show the convergence of the covariance matrix in the operator space norm, assumption **HP1** is required. More specifically, they prove that if $\|\boldsymbol{\Gamma}_n - \boldsymbol{\Gamma}\|_F \to 0$ a.s., then $\|\alpha_k^n - \alpha\|_{L^2_{[a,b]}} \to 0$ a.s., for all $1 \le k \le p$, where α_k^n (respectively α_k) are the eigenvectors of $\boldsymbol{\Gamma}_n$, which is the empirical covariance matrix associated with P_n (respectively $\boldsymbol{\Gamma}$, which is the covariance matrix associated with P).

Then $\|\alpha_k^n - \alpha\|_{L^2_{[a,b]}}$ converges a.s. to zero. By the Law of Large Numbers $\frac{1}{n}\sum_{j=1}^{n}\|X_j - Z_j(I)\|^2_{L^2_{[a,b]}}$ is bounded, even more, since $\left\{\|X_j - Z_j(I)\|,\ \text{for } j = 1, \ldots, n\right\}$ are iid random variables with finite second moment (assumption **H2**), we have that,

$$\frac{1}{n}\sum_{j=1}^{n}\|X_j - Z_j(I)\|^2_{L^2_{[a,b]}} \to E_P\left(\|X - Z(I)\|^2\right) \quad \text{a.s.}$$

We will prove that (32) converges to (8). By the Law of Large Numbers, given that $\left\{\int_0^1 \alpha(t)(X_j(t) - Z_j(I)(t))dt,\ \text{for } j = 1, \ldots, n\right\}$ are iid random variables with finite second moment. Assumption **H2** and Cauchy–Schwarz's inequality ensure this statement since

$$\left(\int_0^1 \alpha(t)(X_j(t) - Z_j(I)(t))dt\right)^2 \le \|\alpha\|^2_{L^2_{[a,b]}}\|X_j - Z_j(I)\|^2_{L^2_{[a,b]}},$$

we have that,

$$\frac{1}{n}\sum_{j=1}^{n}\left(\int_a^b \alpha(t)(X_j(t) - Z_j(I)(t))dt\right)^2 \to E\left(\left(\int_a^b \alpha(t)(X(t) - Z(I)(t))dt\right)^2\right) \quad \text{a.s.}$$

By Cauchy–Schwarz's inequality it can be seen that (33) and (30) converges to zero. And also (29) vanishes by Cauchy–Schwarz's inequality and assumption **H1.** □

Proof of Theorem 3. We need to proof that

$$\frac{1}{n}\sum_{j=1}^{n}\left(\int_a^b \beta_n(t)X_j(t)dt - \int_a^b \beta_n(t)\hat{X}_j(I)(t)dt\right)^2 \to E\left(\left(\int_a^b \beta_0(t)X(t)dt - \int_a^b \beta_0(t)Z(I)(t)dt\right)^2\right) \quad \text{a.s.}$$

We consider the unobservable functions defined in (27) and observe that

$$\begin{aligned} h_n(I) &= \frac{1}{n}\sum_{j=1}^{n}\left(\int_a^b \beta_n(t)X_j(t)dt - \int_a^b \beta_n(t)\hat{X}_j(I)(t)dt\right)^2 \\ &= \frac{1}{n}\sum_{j=1}^{n}\left(\int_a^b \beta_n(t)\left(X_j(t) - Z_j(I)(t)\right)dt\right)^2 \qquad (34) \\ &\quad + \frac{1}{n}\sum_{j=1}^{n}\left(\int_a^b \beta_n(t)\left(Z_j(I)(t) - \hat{X}_j(I)(t)\right)dt\right)^2 \qquad (35) \\ &\quad + \frac{2}{n}\sum_{j=1}^{n}\left(\int_a^b \beta_n(t)\left(X_j(t) - Z_j(I)(t)\right)dt\right)\left(\int_a^b \beta_n(t)\left(Z_j(I)(t) - \hat{X}_j(I)(t)\right)dt\right). \qquad (36) \end{aligned}$$

First, we are going to prove that (34) converges to $h(I)$. It is clear that (34),

$$\begin{aligned} \frac{1}{n}\sum_{j=1}^{n}\left(\int_a^b \beta_n(t)\left(X_j(t) - Z_j(I)(t)\right)dt\right)^2 &= \frac{1}{n}\sum_{j=1}^{n}\left(\int_a^b (\beta_n(t) - \beta_0(t))\left(X_j(t) - Z_j(I)(t)\right)dt\right)^2 \qquad (37) \\ &\quad + \frac{1}{n}\sum_{j=1}^{n}\left(\int_a^b \beta_0(t)\left(X_j(t) - Z_j(I)(t)\right)dt\right)^2 \qquad (38) \\ &\quad + \frac{2}{n}\sum_{j=1}^{n}\left(\int_a^b (\beta_n(t) - \beta_0(t))\left(X_j(t) - Z_j(I)(t)\right)dt\right)\left(\int_a^b \beta_0(t)\left(X_j(t) - Z_j(I)(t)\right)dt\right). \qquad (39) \end{aligned}$$

We will show that the right hand side of (37) converges to zero. By Cauchy–Schwarz's inequality, we have that,

$$\frac{1}{n}\sum_{j=1}^{n}\left(\int_a^b (\beta_n(t) - \beta_0(t))\left(X_j(t) - Z_j(I)(t)\right)dt\right)^2 \le \|\beta_n - \beta_0\|^2_{L^2[a,b]}\,\frac{1}{n}\sum_{j=1}^{n}\left\|X_j - Z_j(I)\right\|^2_{L^2[a,b]}.$$

Since $\|\beta_n - \beta_0\|^2_{L^2[a,b]}$ converges a.s. to zero by assumption **HR1** and by the Law of Large Numbers we have that $\frac{1}{n}\sum_{j=1}^{n}\left\|X_j - Z_j(I)\right\|^2_{L^2[a,b]}$ is finite, because $\left\{\left\|X_j - Z_j(I)\right\|^2_{L^2[a,b]} \text{ for } j = 1,\ldots,n\right\}$ are iid random variables with finite second moment (assumption **H2**).

We will show that (38) converges to $h(I)$. By Cauchy–Schwarz's inequality we have,

$$\left(\int_a^b \beta_0(t)\left(X(t) - Z(I)(t)\right)dt\right)^2 \le \|\beta_0\|^2_{L^2[a,b]}\,\|X - Z(I)\|^2_{L^2[a,b]},$$

then,

$$E\left(\left(\int_a^b \beta_0(t)\left(X(t) - Z(I)(t)\right)dt\right)^2\right) \le \|\beta_0\|^2_{L^2[a,b]}\,E\left(\|X - Z(I)\|^2_{L^2[a,b]}\right).$$

Given that $\left\{\int_a^b \beta_0(t)\left(X_j(t) - Z_j(I)(t)\right)dt \text{ for } j = 1,\ldots,n\right\}$ are iid random variables with finite second moment (assumption **H2**), by the Law of Large Numbers we have that,

$$\frac{1}{n}\sum_{j=1}^{n}\left(\int_a^b \beta_0(t)\left(X_j(t) - Z_j(I)(t)\right)dt\right)^2 \to E_P\left(\left(\int_a^b \beta_0(t)X(t)dt - \int_a^b \beta_0(t)Z(I)(t)dt\right)^2\right) = h(I) \quad \text{a.s.}$$

By Cauchy–Schwarz's inequality it can be seen that (36), (35) and (39) converge a.s. to zero. □

Proof of Theorem 4. We need to prove that $h_n(I) \to h(I)$ a.s., that means that,

$$\frac{1}{n}\sum_{j=1}^{n}\left[\int_c^d\left[\int_a^b \beta_n(s,t)(X_j(s)ds - \hat{X}_j(I)(s))ds\right]^2 dt\right] \to E\left[\int_c^d\left[\int_a^b \beta(s,t)(X(s) - Z(I)(s))ds\right]^2 dt\right] \quad \text{a.s.}$$

Observe that

$$h_n(I) = \frac{1}{n}\sum_{j=1}^{n}\left[\int_c^d\left[\int_a^b \beta_n(s,t)(X_j(s)ds - \hat{X}_j(I)(s))ds\right]^2 dt\right]$$

$$= \frac{1}{n}\sum_{j=1}^{n}\left[\int_c^d\left[\int_a^b \beta_n(s,t)(X_j(s)ds - Z_j(I)(s))ds\right]^2 dt\right] \tag{40}$$

$$+ \frac{1}{n}\sum_{j=1}^{n}\left[\int_c^d\left[\int_a^b \beta_n(s,t)(Z_j(I)(s)ds - \hat{X}_j(I)(s))ds\right]^2 dt\right] \tag{41}$$

$$+ \frac{2}{n}\sum_{j=1}^{n}\left[\int_c^d\left[\int_a^b \beta_n(s,t)(X_j(s)ds - Z_j(I)(s))ds\right]\left[\int_a^b \beta_n(s,t)(Z_j(s)ds - \hat{X}_j(I)(s))ds\right]dt\right]. \tag{42}$$

First we are going to see that (40) converges to $h(I)$

$$\frac{1}{n}\sum_{j=1}^{n}\int_c^d\left[\int_a^b \beta_n(s,t)(X_j(s)ds - Z_j(I)(s))ds\right]^2 dt = \frac{1}{n}\sum_{j=1}^{n}\int_c^d\left[\int_a^b \beta(s,t)(X_j(s)ds - Z_j(I)(s))ds\right]^2 dt \tag{43}$$

$$+ \frac{1}{n}\sum_{j=1}^{n}\int_c^d\left[\int_a^b (\beta_n(s,t) - \beta(s,t))(X_j(s)ds - Z_j(I)(s))ds\right]^2 dt \tag{44}$$

$$+ \frac{2}{n}\sum_{j=1}^{n}\int_c^d\left[\int_a^b \beta(s,t)(X_j(s)ds - Z_j(I)(s))ds\right]\left[\int_a^b (\beta_n(s,t) - \beta(s,t))(X_j(s)ds - Z_j(I)(s))ds\right]dt. \tag{45}$$

We are going to prove that (43) converges to $h(I)$. Since $\int_c^d\left[\int_a^b \beta(s,t)(X_j(s)ds - Z_j(I)(s))ds\right]^2 dt$ has finite first moment, by Cauchy–Schwarz's inequality we have that

$$\int_c^d\left[\int_a^b \beta(s,t)(X_j(s)ds - Z_j(I)(s))ds\right]^2 dt \le \|X_j - Z_j(I)\|^2_{L^2[a,b]}\|\beta\|^2_{L^2([a,b]\times[c,d])}.$$

Then

$$E\left(\int_c^d\left[\int_a^b \beta(s,t)(X_j(s)ds - Z_j(I)(s))ds\right]^2 dt\right) \le \|\beta\|^2_{L^2([a,b]\times[c,d])}E(\|X_j - Z_j(I)\|^2_{L^2[a,b]}),$$

which is finite by assumption **H2**.

Since $\left\{\int_c^d\left[\int_a^b \beta(s,t)(X_j(s)ds - Z_j(I)(s))ds\right]^2 dt\right\}$ are iid variables with finite first moment, by Law of the Large Numbers, we have that (43) converges a.s. to (16).

Assumptions **H2** and **HRF1**, and Cauchy–Schwarz's inequality guarantee that (44) converges a.s. to zero.

As a consequence of Cauchy–Schwarz's inequality it can be seen that (45) converges a.s. to zero.

As a consequence of assumptions **H2** and **HRf1**, and by Cauchy–Schwarz's inequality it can be seen that (41) converges a.s. to zero. Finally it can be seen that assumptions **H1**, **H2** and **HRF1**, and Cauchy–Schwarz's inequality guarantee that (42) converges a.s. to zero. □

References

[1] K. Abramowicz, C. Häguer, K. Hébert-Losier, A. Pini, L. Schelin, J. Strandberg, S. Vantini, An inferential framework domain selection for functional ANOVA, in: E.G. Bongiorno, A. Goia, E. Sanelli, P. Vieu (Eds.), Contributions in infinite-dimensional statistics and related topics. Proceedings of IWFOS 2014, 2014.

[2] G. Aneiros, F. Ferraty, P. Vieu, Variable selection in partial linear regression with functional covariate, Statistics: J. Theor. Appl. Stat. (2015) http://dx.doi.org/10.1080/02331888.2014.998675.

[3] G. Aneiros, P. Vieu, Variable selection in infinite-dimensional problems, Statist. Probab. Lett. 24 (2014) 12–20. http://dx.doi.org/10.1016/j.spl.2014.06.025.

[4] G. Aneiros, P. Vieu, Selecting covariates coming from a continuos variable, in: E.G. Bongiorno, A. Goia, E. Sanelli, P. Vieu (Eds.), Contributions in infinite-dimensional statistics and related topics. Proceedings of IWFOS 2014, 2014.

[5] T.T. Cai, P. Hall, Prediction in functional linear regression, Ann. Statist. 34 (5) (2006) 2159–2179. http://dx.doi.org/10.1214/009053606000000830.

[6] H. Cardot, F. Ferraty, P. Sarda, Spline estimators for the functional linear model, Statist. Sinica 13 (2003) 571–591.

[7] A. Cuevas, A partial overview of the theory of statistics with functional data, J. Statist. Plann. Inference 147 (2014) 1–23. http://dx.doi.org/10.1016/j.jspi.2013.04.002.

[8] J. Dauxois, A. Pousse, Y. Romain, Asymptotic theory for the principal component analysis of a vector random function: Some applications to statistical inference, J. Multivariate Anal. 12 (1982) 136–154. http://dx.doi.org/10.1016/0047-259X(82)90088-4.
[9] Bongiorno E.G., Goia A., Sanelli E., Vieu P (Eds.), Contributions in Infinite-dimensional Statistics and Related Topics, Esculapio Pub., Bologna, Italy, 2014.
[10] M. Fauvel, A. Zullo, F. Ferraty, Nonlinear parsimonious feature selection for the classification of hyperspectral images, in: 6. Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing, Laussane, Switzerland, 2014.
[11] M. Febrero-Bande, M. Oviedo de la Fuente, Statistical computing in functional data analysis: The R package, J. Statist. Softw. 51 (4) (2012) 1–28.
[12] F. Ferraty, P. Hall, P. Vieu, Most-predictive design points for functional data predictors, Biometrika. 97 (2010) 807–824. http://dx.doi.org/10.1093/biomet/asq058.
[13] F. Ferraty, Y. Romain, The Oxford Handbook of Functional Data Analysis, Oxford University Press, 2011.
[14] F. Ferraty, P. Vieu, Nonparametric Functional Data Analysis, Springer, New York, 2006.
[15] L. Forzani, R. Fraiman, P. Llop, Consistent nonpaparametric regression for functional data under the Stone–Besicovitch conditions, IEEE Trans. Inform. Theory 58 (11) (2012) 6697–6708. http://dx.doi.org/10.1109/TIT.2012.2209628.
[16] R. Fraiman, A. Justel, M. Svarc, Selection of variables for cluster analysis and classification rules, J. Amer. Statist. Assoc. 103 (483) (2008) 1294–1303. http://dx.doi.org/10.1198/016214508000000544.
[17] C. Fraley, A.E. Raftery, Model-based clustering, discriminant analysis, and density estimation, J. Amer. Statist. Assoc. 97 (458) (2002) 611–631. http://dx.doi.org/10.1198/016214502760047131.
[18] C. Fraley, A.E. Raftery, MCLUST Version 3 for R: Normal Mixture Modeling and Model-based Clustering. Technical Report No. 504, Department of Statistics, University of Washington, 2009.
[19] J. Gertheiss, A. Maity, A.M. Staicu, Variable selection in generalized functional linear models, Statistics 2 (1) (2013) 86–101. http://dx.doi.org/10.1002/sta4.20.
[20] J. Goldsmith, F. Scheipl, Estimator selection and combination in scalar-on-function regression, Comput. Statist. Data Anal. 70 (2014) 362–372. http://dx.doi.org/10.1016/j.csda.2013.10.009.
[21] P. Hall, J.L. Horowitz, Methodology and convergence rates for functional linear regression, Ann. Statist. 35 (1) (2007) 70–91. http://dx.doi.org/10.1214/009053606000000957.
[22] T. Hastie, R. Tibshirani, J. Friedman, The elements of statistical learning. Data mining, inference and prediction, in: Springer Series in Statistics, Springer Verlag, New York, 2001.
[23] J. Hoeting, A.E. Raftery, D. Madigan, Bayesian variable and transformation selection in linear regression, J. Comput. Graph. Statist. 11 (2002) 485–507. http://dx.doi.org/10.1198/106186002501.
[24] L. Horváth, P. Kokoszka, Inference for Functional Data with Applications, Springer, New York, 2012, http://dx.doi.org/10.1007/978-1-4614-3655-3.
[25] G.M. James, J. Wang, J. Zhu, Functional linear regression that's interpretable, Ann. Statist. 37 (5A) (2009) 2083–2108. http://dx.doi.org/10.1214/08-AOS641.
[26] A. Kneip, P. Poss, P. Sarda, Functional linear regression with points of impact, Ann. Statist. (2015) in press. Currently available at http://www.e-publications.org/ims/submission/AOS/user/submissionFile/19648?confirm=10e645b3.
[27] A. Kneip, P. Sarda, Factor models and variable selection in high-dimensional regression analysis, Ann. Statist. 39 (5) (2011) 2410–2447. http://dx.doi.org/10.1214/11-AOS905.
[28] N. Kudaszow, P. Vieu, Uniform consistency of k-NN regressors for functional variables, Statist. Probab. Lett. 83 (2013) 1863–1870. http://dx.doi.org/10.1016/j.spl.2013.04.017.
[29] E.R. Lee, U. Byeong, B.U. Park, Sparse estimation in functional linear regression, J. Multivariate Anal. 105 (2012) 1–17. http://dx.doi.org/10.1016/j.jmva.2011.08.005.
[30] H. Lian, Convergence of functional k-nearest neighbor regression estimate with functional responses, Electron. J. Stat. 5 (2011) 31–40. http://dx.doi.org/10.1214/11-EJS595.
[31] Y. Li, T. Hsing, On rates of convergence in functional linear regression, J. Multivariate Anal. 98 (2007) 1782–1804. http://dx.doi.org/10.1016/j.jmva.2006.10.004.
[32] S. Lopez-Pintado, J. Romo, Depth-based classification for functional data, in: DIMACS Series in Discrete Mathematics and Theoretical Computer Science. Data Depth: Robust Multivariate Analysis, Computational Geometry and Appliations, vol. 72, American Mathematical Society, 2006, pp. 103–120.
[33] B. Martin-Barragan, R. Lillo, J. Romo, Interpretable support vector machines for functional data, European J. Oper. Res. 232 (1) (2014) 146–155. http://dx.doi.org/10.1016/j.ejor.2012.08.017.
[34] H. Matsui, Variable and boundary selection for functional data via multiclass logistic regression modeling, Comput. Statist. Data Anal. 78 (2014) 176–185. http://dx.doi.org/10.1016/j.csda.2014.04.015.
[35] I.W. McKeague, B. Sen, Fractals with impact points in functional linear regression, Ann. Statist. 38 (4) (2010) 2559–2586. http://dx.doi.org/10.1214/10-AOS791.
[36] H.G. Müller, F. Yao, Functional additive models, J. Amer. Statist. Assoc. 103 (2008) 1534–1544. http://dx.doi.org/10.1198/016214508000000751.
[37] J. Ramsay, B.W. Silverman, Applied Functional Data Analysis: Methods and Case Studies, Springer, New York, 2002, http://dx.doi.org/10.1007/b98886.
[38] J. Ramsay, B.W. Silverman, Functional Data Analysis, second ed., Springer, New York, 2005, http://dx.doi.org/10.1007/978-1-4757-7107-7.
[39] J.O. Ramsay, H. Wickham, S. Graves, G. Hooker, fda: Functional Data Analysis. R package version 2.3.2, 2012.
[40] F. Riesz, B. Nagy, Lecons d'analyse functionelle, Gauthiers-Villars, Paris, 1965.
[41] T.S. Tian, G.M. James, Interpretable dimension reduction for classifying functional data, Comput. Statist. Data Anal. 57 (1) (2013) 282–296. http://dx.doi.org/10.1016/j.csda.2012.06.017.
[42] R. Tibshirani, Regression shrinkage and selection via the lasso, J. R. Stat. Soc. Ser. B 58 (1) (1996) 267–288. http://dx.doi.org/10.1.1.35.7574.
[43] D.M. Witten, R. Tibshirani, Testing significance of features by lassoed principal components, Ann. Appl. Stat. 2 (3) (2008) 986–1012. http://dx.doi.org/10.1214/08-AOAS182.
[44] D.M. Witten, R. Tibshirani, A framework for feature selection in clustering, J. Amer. Statist. Assoc. 105 (490) (2010) 713–726. http://dx.doi.org/10.1198/jasa.2010.tm09415.
[45] F. Yao, H.G. Müller, J.L. Wang, Functional linear regression analysis for longitudinal data, Ann. Statist. 33 (6) (2005) 2873–2903. http://dx.doi.org/10.1214/009053605000000660.
[46] J. Zhou, N.Y. Wang, N. Naisyin Wang, Functional linear model with zero-value coefficient function at sub-regions, Statist. Sinica 23 (1) (2013) 25–50. http://dx.doi.org/10.5705/ss.2010.237.