Journal of Multivariate Analysis 170 (2019) 25–45

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

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

An RKHS model for variable selection in functional linear regression

José R. Berrendero, Beatriz Bueno-Larraz *, Antonio Cuevas

Departamento de Matemáticas, Facultad de Ciencias, Universidad Autónoma de Madrid, 28049, Madrid, Spain

ARTICLE INFO

Article history:
Received 18 September 2017
Available online 24 April 2018

AMS 2010 subject classifications:
62G05
62J99

Keywords:
Feature selection
Functional linear regression
Impact points
Variable selection

ABSTRACT

A mathematical model for variable selection in functional linear regression models with scalar response is proposed. By "variable selection" we mean a procedure to replace the whole trajectories of the functional explanatory variables with their values at a finite number of carefully selected instants (or "impact points"). The basic idea of our approach is to use the Reproducing Kernel Hilbert Space (RKHS) associated with the underlying process, instead of the more usual $L^2[0,\,1]$ space, in the definition of the linear model. This turns out to be especially suitable for variable selection purposes, since the finite-dimensional linear model based on the selected "impact points" can be seen as a particular case of the RKHS-based linear functional model. In this framework, we address the consistent estimation of the optimal design of impact points and we check, via simulations and real data examples, the performance of the proposed method.

1. Introduction: statement of the problem and motivation

The problem under study: variable selection in functional regression

The study of regression models is clearly among the leading topics in statistics. In particular, these models play a central role in the theory of statistics with functional data, often called Functional Data Analysis (FDA); see [7,15,16] for an overview on FDA.

Throughout this paper, we will consider "functional data" consisting of independent $X_1 = X_1(t), \ldots, X_n = X_n(t)$ observations (trajectories) drawn from a second-order (L^2) stochastic process $X = X(t)$, $t \in [0,\,1]$, with continuous trajectories and continuous mean and covariance functions, denoted by $m = m(t)$ and $K(s,\,t)$, respectively. All the involved random variables are supposed to be defined in a common probability space $(\Omega, \mathcal{A}, \Pr)$.

We are interested on functional regression models with scalar response, of type $Y_i = g(X_i) + \varepsilon_i$, where g is a real function defined on a suitable space $\mathcal{X}$ where the trajectories of our process are supposed to live. The random variables ε_i are independent errors (and also independent from the X_i) with mean zero and common variance σ^2.

More specifically, we are concerned with variable selection issues; see, [4, Sec. 1], [11] for additional information and references. Basically, a variable selection functional method is an automatic procedure that takes a function $\{x(t),\ t \in [0,\,1]\}$ to a finite-dimensional vector $(x(t_1), \ldots, x(t_p))$. The overall idea of variable selection is to choose the variables $x(t_i)$ (or, equivalently, the "impact points" $t_1, \ldots, t_p \in [0,\,1]$; see [22]), in an "optimal way" so that the original functional problem (regression, classification, clustering, ...) is replaced with the corresponding multivariate version, based on the selected variables. In the regression setting, this would amount to replace the functional model $Y_i = g(X_i) + \varepsilon_i$ with a finite

* Corresponding author.
E-mail addresses: joser.berrendero@uam.es (J.R. Berrendero), beatriz.bueno@uam.es (B. Bueno-Larraz), antonio.cuevas@uam.es (A. Cuevas).

https://doi.org/10.1016/j.jmva.2018.04.008
0047-259X/

dimensional version of type $Y_i = \phi\{X_i(t_1), \ldots, X_i(t_p)\} + e_i$. Nevertheless, note that still the problem is of a functional nature, since the methods to select the t_i are generally based upon the full data trajectories.

Some notation

A set of possible "impact" points $t_1, \ldots, t_p \in [0, 1]$ will be denoted T (sometimes S) or T_p when we want to stress the cardinality of T. Also, $X(T_p)$ will stand for $(X(t_1), \ldots, X(t_p))^\top$. The superindex * will be used to denote that the points t_i^* are the "true" ones, or the "optimal" ones according to some criterion.

Given a random variable Z (with finite variance) the notation Z_{T_p} will refer to the orthogonal projection of Z on the space spanned by the components of $X(T_p) - m(T_p)$.

If $p^* < p$, the notation $T_{p^*} \prec T_p$ will indicate that all the points in T_{p^*} belong also to T_p.

Finally, as usual in statistics, we use an upper hat sign to denote the estimated quantities (or the predicted variables). For instance, $\widehat{T}_p$ will denote a data-driven estimator of T_p and $\widehat{Y}_{\widehat{T}_p}$ will stand for the corresponding (fully data-driven) prediction of the response Y_{T_p}. The halfway notation $Y_{\widehat{T}_p}$ will represent the orthogonal projection of the response variable onto the space spanned by the marginal variables indexed by the estimated points $\widehat{T}_p$.

Some motivation. The drawbacks of the classical linear L^2-model for variable selection purposes

It is quite natural to assume that the explanatory functional variables $X_i = X_i(t)$ are members of the space $L^2[0, 1]$, endowed with the usual inner product $\langle x_1, x_2\rangle_2 = \int_0^1 x_1(t)x_2(t)dt$, for $x_1, x_2 \in L^2[0, 1]$. In this setting, the most popular choice for g is, by far, a linear (or affine) operator from $L^2[0, 1]$ to $\mathbb{R}$ which leads to a model of type

$$Y_i = \alpha_0 + \langle X_i, \beta\rangle_2 + \varepsilon_i,\ i = 1, \ldots, n, \tag{1}$$

where $X = X(t)$ is the explanatory functional variable, $\alpha_0 \in \mathbb{R}$ is the intercept constant and $\beta \in L^2[0, 1]$ denotes the slope function. As in the standard multivariate regression model, the aim here is to estimate α_0 and β in order to be able to make accurate predictions of the response variable Y.

The corresponding theory is outlined in several places; see, e.g., the article [6] or the books [13,16]. The Hilbert structure of the $L^2[0, 1]$ space allows us to keep ourselves as close as possible to the usual least squares framework in multivariate regression; for example, the projection $P(x)$ of an element x on a closed subspace H is characterized by the orthogonality condition $\langle x - P(x), a\rangle_2 = 0$, for all $a \in H$. However, other crucial differences with the finite-dimensional case (mostly associated with the non-invertibility of the covariance operator of the process $X(t)$) make the functional L^2 regression theory far from trivial. Most of these difficulties are intrinsic to the infinite-dimensional nature of the data, so that they cannot be overcome by just replacing $L^2[0, 1]$ with another function space. However, when it comes to variable selection applied to linear regression, it would be useful to have the finite dimensional linear model (based on the selected variables)

$$Y_i = \alpha_0 + \sum_{j=1}^{p} \beta_j X_i(t_j) + \varepsilon_i,\ i = 1, \ldots, n \tag{2}$$

as a particular case of our general model. Notice that (2) cannot be established in the L^2 framework, since a transformation of type $x \in L^2[0, 1] \mapsto \sum_{i=1}^{p} \beta_i x(t_j)$ is not a linear continuous functional in L^2. In heuristic terms, one would need to look for the slope function β in a suitable space, for which a "finite-dimensional" model such as (2) could make sense. More precisely, we will change the "habitat" space for the function β: instead of assuming $\beta \in L^2[0, 1]$ we will assume that β belongs to the Reproducing Kernel Hilbert Space (RKHS), $\mathcal{H}(K)$, associated with K.

As we will see below, the assumed membership of β to $\mathcal{H}(K)$ entails some additional restrictions of regularity on the slope function β (when compared to the simple assumption $\beta \in L^2[0, 1]$). In any case, such a situation is not unusual: some restrictions on β appear in different ways even when the classical L^2-model (1) is considered. The reason is that the space $L^2[0, 1]$ is in fact too large from several points of view. Hence, in spite of the advantages of the L^2-model commented above, one typically uses penalization or projection methods to exclude extremely rough solutions in the estimation of β.

Our proposal here, as presented in the next section, aims at reconciling two targets: first, we look for a functional linear model, wide enough to include finite-dimensional versions, such as (2), as particular cases. Second, we would like to achieve such a goal with a minimal change in the space where β lives.

Some related literature

A quite general RKHS-based approach to the problem of dimension reduction in functional regression has been proposed by [18]. These authors follow the inverse regression methodology to deal with a model of type $Y = \ell(\xi_1, \ldots \xi_d) + \varepsilon$ where ℓ is a link function and the ξ_j are linear functionals of the explanatory variable X, defined in RKHS terms. This pioneering reference shows very clearly the huge potential of the RKHS approach. However, as the authors point out, there are still many aspects not considered in that paper and worth of attention. Variable selection is one of them. In fact, the whole point of the present paper is to show that things become particularly simple when the RKHS machinery is applied to variable selection. A recent use of RKHS methods in the problem of functional binary classification is developed in [5]. See also [3,17] for a broader perspective of the applicability of RKHS methods in statistics.

Other variable selection methods, always aimed at selecting the "best points" $t_1, \ldots, t_d$ (or the "best variables" $X(t_1), \ldots, X(t_p)$) have been proposed as well, with no explicit reference of RKHS tools. Thus, the selection of the "best impact point" t_1 in a model of type (2) with $p = 1$ is addressed in [25]. Different variable selection methods have been suggested

by [1,9,12] for prediction and classification purposes. In addition, a non-parametric approach for non-linear functional regression models is presented in [2]. See the references therein for more information on non-linear models. Also, a criterion for "optimal design" in trajectory recovery is considered in [20].

A recent general proposal for dimension reduction (beyond variable selection and regression models) is [14].

Organization of the paper

In Section 2 we introduce and motivate (in population terms) our variable selection procedure. The asymptotic properties of the empirical version (when the parameters are estimated) are considered in Section 3. The problems associated with the choice of the number p of selected variables are analyzed in Sections 4 and 5. The empirical results (simulations and real data examples) are presented in Section 6. Section 7 includes some final comments and conclusions. The proofs of some results (stated but not proved in the previous sections) are included in Section 8.

2. An RKHS-based linear model suitable for variable selection

Our choice of the ambient space for the slope function β is, in some sense, "customized" for the problem at hand, since we will consider the Reproducing Kernel Hilbert Space (RKHS) associated with the process $\{X(t),\ t \in [0, 1]\}$.

The theory of RKHS goes back to the 1950s; see [19, Appendix F] for details and references. It has found a surprisingly large number of applications in different fields, including statistics, see [3].

2.1. RKHS spaces in a nutshell

Before establishing our RKHS-based regression model we need a minimal background on RKHS. Our starting point will be our underlying L^2-process $X = \{X(t),\ t \in [0, 1]\}$ with a continuous strictly positively definite covariance function $K(s, t)$ and a continuous mean function $m(t)$.

We first introduce an auxiliary space, associated with K, which we will denote by $\mathcal{H}_0(K)$. It is defined by the set of all finite linear combinations of type $\sum_i^n a_i K(s, t_i)$, that is,

$$\mathcal{H}_0(K) := \{f : f(s) = \sum_{i=1}^{n} a_i K(s, t_i),\ a_i \in \mathbb{R},\ t_i \in [0, 1],\ n \in \mathbb{N}\}.$$

In such space we define an inner product $\langle \cdot, \cdot \rangle_K$ by $\langle f, g \rangle_K = \sum_{i,j} \alpha_i \beta_j K(s_j, t_i)$, where $f(x) = \sum_i \alpha_i K(x, t_i)$ and $g(x) = \sum_j \beta_j K(x, s_j)$.

Now, the RKHS associated with K, denoted by $\mathcal{H}(K)$, is defined as the completion of $\mathcal{H}_0(K)$. More precisely, $\mathcal{H}(K)$ is the set of functions $f : [0, 1] \to \mathbb{R}$ obtained as t-pointwise limits of Cauchy sequences $\{f_n\}$ in $\mathcal{H}_0(K)$; see [3, p. 16]. Thus, in heuristic terms, one could say that $\mathcal{H}(K)$ is made of all linear combinations of type $f(s) = \sum_i^n a_i K(s, t_i)$ plus all the functions which can be obtained as limits of them. A natural question is when we can ensure that we have identifiability in this space. It is easy to see that the elements of $\mathcal{H}_0(K)$ have a unique representation in terms of K whenever K is strictly positive definite. For additional details we refer again to [19, Appendix F].

Among the many interesting properties of RKHS's, let us especially recall two which will be particularly useful in what follows.

Reproducing property. $f(t) = \langle f, K(\cdot, t) \rangle_K$, for all $f \in \mathcal{H}(K)$, $t \in [0, 1]$.

Natural congruence. Denote by $\mathcal{L}_X$, the linear (centered) span of X (i.e., the family of finite linear combinations of type $\sum_i a_i \{X(t_i) - m(t_i)\}$). Let $\overline{\mathcal{L}}_X$ be the L^2-completion of $\mathcal{L}_X$. It is clear that $\overline{\mathcal{L}}_X$ is a closed subspace of the usual Hilbert space $L^2(\Omega)$ of random variables with finite second moment and $\|Z\|_2^2 = \mathrm{E}(Z^2)$, for all $Z \in L^2(\Omega)$. This can be seen as the minimal Hilbert space including the (centered) variables $X(t)$. It can be proved, see [3, Th. 35], that $\Psi_X[\sum_i a_i \{X(t_i) - m(t_i)\}] = \sum_i a_i K(\cdot, t_i)$ defines (when extended by continuity) a congruence between $\overline{\mathcal{L}}_X$ and $\mathcal{H}(K)$. This means that the extension of Ψ_X (which we will denote also Ψ_X) is a linear bijective transformation which preserves the inner product. Such congruence is often called *Loève's isometry*. In explicit terms (see [24, Lemma 1.1]), Loève's isometry between $\overline{\mathcal{L}}_X$ and $\mathcal{H}(K)$ can be defined by

$$Y \mapsto \Psi_X(Y),\ \text{where } \Psi_X(Y) \text{ is the function } \Psi_X(Y)(t) = \mathrm{E}[Y\{X(t) - m(t)\}]. \tag{3}$$

Since $\mathcal{H}(K)$ is an isometric copy of $\overline{\mathcal{L}}_X$, both spaces can be identified. In particular the random variables $X(t_i) - m(t_i)$ are the inverse images of the functions $K(\cdot, t_i) \in \mathcal{H}(K)$ in such isometry.

There is, however, a not-so-nice feature in the RKHS associated with the process $X(t)$: under very general conditions, this space does not contain, with probability one, the trajectories of the process X; see, e.g., [24, Cor. 7.1], [28, Th. 11]. This will have some consequences in the formulation of our regression model, as pointed out below.

2.2. The RKHS functional regression model

We propose to replace the standard L^2 functional regression model (1) with the following RKHS counterpart

$$Y_i = \alpha_0 + \langle X_i, \beta\rangle_K + \varepsilon_i,\ i = 1, \ldots, n, \tag{4}$$

where $\beta \in \mathcal{H}(K)$ and $\langle\cdot,\cdot\rangle_K$ denotes the inner product in $\mathcal{H}(K)$.

Since the estimation of the intercept term α_0 is straightforward from those of β and m, we will assume, without loss of generality, that $\alpha_0 = 0$ in what follows.

As mentioned at the end of the previous subsection, it is important to keep in mind that the trajectories of the process X do not belong to $\mathcal{H}(K)$. Thus, the expression $\langle X_i, \beta\rangle_K$ has no direct meaning, unless it is appropriately interpreted: in what follows, $\langle X_i, \beta\rangle_K$ must be understood as $\Psi_{X_i}^{-1}(\beta)$, where Ψ_X is defined in (3). Then, with this definition, we might replace (for a given $\beta \in \mathcal{H}(K)$) the random process X with a specific trajectory x and in that case $\langle x, \beta\rangle_K$ would be well defined (as a constant) even if $x \notin \mathcal{H}(K)$.

Such an interpretation of $\langle X, \beta\rangle_K$ arises in the classical paper by Parzen [27, Th. 7A], aiming at different statistical purposes. In addition, note that in the context of the linear model (4), we assume $E\{X(t)\varepsilon\} = 0$ and $E(\varepsilon) = 0$, so that $\beta(t) = \mathrm{cov}\{Y, X(t)\}$; hence, $\langle X, \beta\rangle_K$ might be also defined as the solution $Z \in \overline{\mathcal{L}}_X$ of the functional equation $\Psi_X(Z)(t) = \mathrm{cov}\{Y, X(t)\}$.

The above commented problems to give a proper definition of $\langle X, \beta\rangle_K$ are reminiscent of those arising when defining Itô's stochastic integral. In fact, when $X(t)$ is a standard Brownian Motion in $[0, 1]$, model (4) with $\alpha_0 = 0$ can be expressed as

$$Y = \int_0^1 \beta'(s)dX(s) + \varepsilon,\ \text{with } \beta \in \mathcal{H}(K),\ \text{and } K(s,t) = \min(s,t),$$

where $\int_0^1 \beta'(s)dX(s)$ is Itô's integral and $\mathcal{H}(K)$ is the space of all real absolutely continuous functions β on $[0, 1]$ with $\beta' \in L^2[0, 1]$ and $\beta(0) = 0$ [19, Example 8.19, p. 122].

2.3. Variable selection in the RKHS functional regression model

Consider the RKHS functional regression model (4) introduced in the previous paragraph, where $\mathrm{E}(\varepsilon) = \mathrm{E}[\{X(t) - m(t)\}\varepsilon] = 0$ and $\mathrm{var}(\varepsilon) = \sigma^2$.

Our goal

Under this model, for fixed p, we aim at selecting p values $t_1, \ldots, t_p$ in order to use the p dimensional vector $(X(t_1), \ldots, X(t_p))$ instead of the whole trajectory $\{X(t) :\ t \in [0, 1]\}$ in our regression problem. Formally, we want to establish a transformation

$$\{X(t) :\ t \in [0, 1]\} \mapsto (X(t_1), \ldots, X(t_p)),$$

which should be "optimal" in the sense that the points $t_1, \ldots, t_p$ are chosen according to an optimality criterion, oriented to minimize the information loss in the passage from infinite to finite dimension.

In this section, we address this problem at the population level, that is, we assume that the parameters defining the model (the slope function β, the covariance function K of the process X, the mean function m and the variance of the error variable, σ^2) are known. Of course, the practical implementation will require using suitable estimators of the unknown parameters. This raises several questions concerning the sample behavior of the method which will be addressed in subsequent sections.

The optimality criterion Q_1

The first obvious question to address in such strategy is the choice of the optimality criterion. We will see that, in fact, different criteria can be used but, fortunately, they are all equivalent.

One of the basic goals of a functional regression model is to predict the value of the response variable Y for a given trajectory of the input process X. Then, a sensible approach for variable selection is to choose the p points $X(t_1), \ldots, X(t_p)$ that give the best linear prediction (in the sense of the L^2 norm) of Y. This implies to find the vector T_p that minimizes the function

$$Q_1(T_p) := \min_{(\beta_1,\ldots,\beta_p)\in\mathbb{R}^p} \|Y - \sum_{j=1}^{p} \beta_j\{X(t_j) - m(t_j)\}\|_2^2. \tag{5}$$

This natural criterion has been considered elsewhere, sometimes in slightly different contexts, see e.g., [20]. The contribution here is to interpret (5) in RKHS terms and, as a consequence, to show that the problem of finding the optimal value of p can be addressed in a meaningful way.

Where to look for the optimum

An important technical aspect is the choice of an appropriate subset $\Theta_p \subset [0, 1]^p$ to look for the optimum of the continuous function Q_1. This subset must be compact in order to guarantee the existence of the optimum. Moreover, if

we want to get a meaningful optimal value of $T_p = (t_1, \ldots, t_p)$ we should rule out those points including repeated values in the coordinates t_i. To this end, we will fix an arbitrarily small value $\delta > 0$, and will look for our optimum in the space

$$\Theta_p = \Theta_p(\delta) = \{T_p = (t_1, \ldots, t_p) \in [0, 1]^p : \ t_{i+1} - t_i \geq \delta, \ \text{for } i = 0, \ldots, p\}, \tag{6}$$

where $t_0 = 0$, $t_{p+1} = 1$. In practice, the restriction to the subset Θ_p is not relevant, since we observe the functions in a finite grid, and we can set $\delta > 0$ as small as required so that all the points in the grid belong to Θ_p.

The reason for the choice (6) of Θ_p is technical, very much in the same spirit of [20, Eq. (9)]. We need to work on a compact set and, at the same time, to avoid degeneracy problems in the choice of the points $(t_1, \ldots, t_p)$ that could lead to a singular covariance matrix in $(X(t_1), \ldots, X(t_p))$. Other choices are possible for Θ_p. For example, one could think of defining $t_0 = 0$, $t_{p+1} = 1$ and

$$\Theta_p = \{T_p = (t_1, \ldots, t_p) \in [0, 1]^p : \ t_i \leq t_{i+1}, \ \text{for } i = 0, \ldots, p\}. \tag{7}$$

This could lead to "degenerate" options with $t_i = t_{i+1}$ for some values of i. However, the theory we develop below using (6) could be carried out alternatively with (7) as long as we adopt the criterion of reducing the dimension of those vectors $(t_1, \ldots, t_p)$ with ties in the coordinate values by keeping just one coordinate for each different value. In this way, for example, $(0.2, 0.2, 0.5, 0.7)$ would be interpreted just as $(0.2, 0.5, 0.7)$.

Two additional, equivalent optimality criteria

A second optimality criterion, equivalent to that based on Q_1, arises if we take into account that, from the reproducing property, when the slope function is a finite linear combination of the form $\sum_{i=1}^p \beta_i K(t_i, \cdot)$, model (4) reduces to the usual finite dimensional multiple regression model:

$$Y = \sum_{i=1}^{p} \beta_j \{X(t_j) - m(t_j)\} + \varepsilon. \tag{8}$$

Then, another sensible approach for variable selection is to choose those points $t_1, \ldots, t_p$ giving the best approximation of the true slope function β in terms of a finite linear combination of the form $\sum_{i=1}^p \beta_i K(t_i, \cdot)$. It is quite natural to use the norm in $\mathcal{H}(K)$ to assess this approximation since both β and these finite linear combinations live in this RKHS. This approach amounts to find the vector $T_p \in \Theta_p$ that minimizes the function

$$Q_2(T_p) := \min_{(\alpha_1, \ldots, \alpha_p) \in \mathbb{R}^p} \|\beta - \sum_{j=1}^{p} \alpha_j K(t_j, \cdot)\|_K^2. \tag{9}$$

Proposition 1 shows that the variable selection procedures defined by (5) and (9), although apparently different, are indeed equivalent. Moreover, in the proof of Proposition 1 we will see that the minimum in the expressions of Q_1 and Q_2 is achieved at the value

$$(\alpha_1^*, \ldots, \alpha_p^*)^\top = \Sigma_{T_p}^{-1} c_{T_p},$$

where $c_{T_p} = (\text{cov}\{X(t_1), Y\}, \ldots, \text{cov}\{X(t_p), Y\})^\top$ and Σ_{T_p} is the covariance matrix of $X(T_p)$, for $T_p = (t_1, \ldots, t_p)$.

In addition, we show that the Q_1 and Q_2-based criteria are also both equivalent to a third criterion, defined in terms of a functional Q_0, which only depends on the covariances $K(t_i, t_j)$ and $\text{cov}\{X(t_i), Y\}$ for $i, j = 1, \ldots, p$. This Q_0 criterion turns out to be especially useful to implement the method in practice.

Proposition 1. *Assume that Y and X fulfill the RKHS functional regression model (4). Then,*

$$\operatorname*{argmin}_{T_p \in \Theta_p} Q_1(T_p) = \operatorname*{argmin}_{T_p \in \Theta_p} Q_2(T_p) = \operatorname*{argmax}_{T_p \in \Theta_p} Q_0(T_p), \tag{10}$$

where Q_1 and Q_2 are defined in (5) and (9) respectively, and

$$Q_0(T_p) := c_{T_p}^\top \Sigma_{T_p}^{-1} c_{T_p}, \tag{11}$$

with $c_{T_p} = (\text{cov}\{X(t_1), Y\}, \ldots, \text{cov}\{X(t_p), Y\})^\top$ and Σ_{T_p} the $p \times p$ matrix with entries $K(t_i, t_j)$.

Proof. Since $\mathrm{E}(\varepsilon) = 0$, $\mathrm{E}[\varepsilon\{X(t) - m(t)\}] = 0$ and $\langle X, \beta \rangle_K \in \overline{\mathcal{L}}_X$,

$$\|Y - \sum_{j=1}^{p} \alpha_j \{X(t_j) - m(t_j)\}\|_2^2 = \|\langle X, \beta \rangle_K - \sum_{j=1}^{p} \alpha_j \{X(t_j) - m(t_j)\}\|_2^2 + \sigma^2.$$

On the other hand, Loève's isometry implies

$$\|\langle X, \beta \rangle_K - \sum_{j=1}^{p} \alpha_j \{X(t_j) - m(t_j)\}\|_2^2 = \|\beta - \sum_{j=1}^{p} \alpha_j K(t_j, \cdot)\|_K^2.$$

From the last two equations, it follows that $Q_1(T_p) = Q_2(T_p) + \sigma^2$ and hence the first equality in (10).

By the reproducing property,

$$\|\beta - \sum_{j=1}^{p} \alpha_j K(t_j, \cdot)\|_K^2 = \|\beta\|_K^2 - 2\sum_{j=1}^{p} \alpha_j \beta(t_j) + \sum_{i=1}^{p}\sum_{j=1}^{p} \alpha_i \alpha_j K(t_i, t_j). \tag{12}$$

The function K is positive semidefinite so that the expression in (12) defines a convex function in $\alpha = (\alpha_1, \ldots, \alpha_p)$. By computing its gradient (with respect to α) it is very easy to see that the minimum is achieved at $\alpha^* = (\alpha_1^*, \ldots, \alpha_p^*)^\top = \Sigma_{T_p}^{-1}\beta(T_p)$, where $\beta(T_p) = (\beta(t_1), \ldots, \beta(t_p))^\top$. Then,

$$Q_2(T_p) = \|\beta - \sum_{j=1}^{p} \alpha_j^* K(t_j, \cdot)\|_K^2 = \|\beta\|_K^2 - \beta(T_p)^\top \Sigma_{T_p}^{-1}\beta(T_p). \tag{13}$$

Finally, using $\mathrm{E}[\{X(t) - m(t)\}\varepsilon] = 0$ and Eq. (3) we get

$$\mathrm{cov}\{Y, X(t)\} = \mathrm{E}\big[\langle X, \beta\rangle_K \{X(t) - m(t)\}\big] = \Psi_X(\langle X, \beta\rangle_K)(t) = \beta(t).$$

To obtain the last equality, recall that $\langle X, \beta\rangle_K = \Psi_X^{-1}(\beta)$. Therefore $\beta(T_p) = c_{T_p}$ and, by (13), $Q_2(T_p) = \|\beta\|_K^2 - Q_0(T_p)$. This implies the second equality in (10). □

The criterion provided by Q_0 (or Q_1, Q_2) for variable selection was already considered by [25], for $p = 1$, when $X(t)$ is a fractional Brownian Motion with Hurst exponent $H \in (0, 1)$, and by [20] for $p \geq 1$ in the usual L^2 functional regression model. The RKHS formalism we incorporate here provides a simple way to describe the scenario under which variable selection would lead to the optimal solution (with no loss of information). Variable selection is specially suitable when the true regression model is sparse, meaning that the response depends on the explanatory variables through their values at a finite small number of p^* points. As it was mentioned before, this is the case under (4) when

$$\beta(t) = \sum_{j=1}^{p^*} \beta_j K(t_j^*, t). \tag{14}$$

Let $T_{p^*}^* = (t_1^*, \ldots, t_{p^*}^*) \in \Theta_{p^*}$. Then, it is clear that, under (14),

$$Q_2(T_{p^*}^*) = 0 \leq Q_2(T_{p^*}), \quad \text{for all } T_{p^*} \in \Theta_{p^*}.$$

As a consequence, the true set of relevant variables $T_{p^*}^*$ is the one selected by the optimization of the functions in Proposition 1. In this reasoning we have considered the case when we know the actual number of points p^* to be selected. In practice, this is not usually the case. However, notice that if we make a conservative choice, taking a number of variables p larger than the true one ($p > p^*$), the true relevant variables $T_{p^*}^*$ will always be included among the selected ones. Indeed, if the true model is $Y = \sum_{i=1}^{p^*} \beta_i^* X(t_i^*) + \varepsilon$, this means that the orthogonal projection of Y on the space $\overline{\mathcal{L}}_X$ is $\sum_{i=1}^{p^*} \beta_i^* X(t_i^*)$. Then, assume that we try to fit (in the β_i's and the t_i's) an "overparameterized" model of type $Y = \sum_{i=1}^{p} \beta_i X(t_i) + \varepsilon$ with $p > p^*$. Since the projection is unique, the optimal fit under the second model must coincide with the optimum of the "true" model. So, it must necessarily include the variables $t_1^*, \ldots, t_p^*$ and the coefficients β_i for the remaining variables must be zero. Therefore the optimal set T_{p^*} of t_i^*'s in the first model must be included in the optimal set T_p for the second one.

Recall that we use the notation $T^* \prec T \in \Theta_p$ meaning that T^* is a sub-vector of T, that is, that the components of T^* are included within those of T. With this notation, what we have shown is that, under (14), $T^* = (t_1^*, \ldots, t_{p^*}^*) \prec \mathrm{argmax}\, Q_0(T_p)$, for $p \geq p^*$. In Section 4 we address the problem of estimating p^* when it is unknown.

2.4. A recursive expression

The function Q_0 defined in (11) can be rewritten in an alternative way, which is useful to analyze the gain when we add a new variable to a set of variables already selected. Moreover, this alternative expression paves the way for a sequential implementation of the variable selection method. Besides the notation c_T and Σ_T, introduced earlier, we will also use c_j to denote $\mathrm{cov}\{X(t_j), Y\}$, σ_j^2 to denote $\mathrm{var}\{X(t_j)\}$, and $c_{T_p,j}$ to denote the vector $(\mathrm{cov}\{X(t_1), X(t_j)\}, \ldots, \mathrm{cov}\{X(t_p), X(t_j)\})^\top$.

Proposition 2. *Given $T_{p+1} = (t_1, \ldots, t_{p+1}) \in \Theta_{p+1}$, $p \geq 1$, and $T_p \prec T_{p+1}$, for some $p \geq 1$ such that the covariance matrices $\Sigma_{T_{p+1}}$ of the process are invertible for all $T_{p+1} \in \Theta_{p+1}$,*

$$Q_0(T_{p+1}) = Q_0(T_p) + \frac{\left(c_{T_p}^\top \Sigma_{T_p}^{-1} c_{T_p,p+1} - c_{p+1}\right)^2}{\sigma_{p+1}^2 - c_{T_p,p+1}^\top \Sigma_{T_p}^{-1} c_{T_p,p+1}}. \tag{15}$$

Eq. (15) is useful to simplify other derivations in the paper and it is shown in Section 8.1. Actually, it can be also proved that the quotient in (15) tends to zero when t_{p+1} tends to one of the points in T_p, so that selecting a point too close to one of

those already selected is redundant and non-informative according to this criterion. The proof of this fact is quite technical so it is omitted.

Eq. (15) already appears in the well-known forward selection method for variable selection in multiple regression (see, e.g., [26, Section 3.2]). A modification of the resulting expression is also used in the variable selection method proposed by [30], still in the multivariate regression setting. In such alternative version, the usual covariance is replaced by the distance covariance, defined in [29].

The quotient in Eq. (15) can be written in a more insightful way, as shown in the following result.

Proposition 3. *In the above defined setup, denoting* $X(T_p) = \big(X(t_1), \ldots, X(t_p)\big)^\top$, *and* $Y_{T_p} = P_{\mathrm{span}\{X(t_i)-m(t_i), t_i\in T_p\}}Y$ *and* $X(t_{p+1})_{T_p} = P_{\mathrm{span}\{X(t_i)-m(t_i), t_i\in T_p\}}X(t_{p+1})$ *the projections on* $\mathrm{span}\{X(t_i) - m(t_i),\, t_i \in T_p\}$,

$$Q_0(T_{p+1}) = Q_0(T_p) + \frac{\mathrm{cov}^2\{Y - Y_{T_p}, X(t_{p+1})\}}{\mathrm{var}\{X(t_{p+1}) - X(t_{p+1})_{T_p}\}}. \tag{16}$$

The quotient of Eq. (16) is known as *part correlation coefficient* or *semi-partial correlation coefficient*, a quantity which appears in several techniques dealing with multivariate data.

The proofs of Proposition 3 can be found in Section 8.2.

3. Sample properties of the variable selection method

3.1. The proposed method

In order to carry out the variable selection in practice, we have to estimate the function Q_0 from a sample $(Y_1, X_1), \ldots, (Y_n, X_n)$ of independent observations drawn from the model (4). The most natural estimator is given by $\widehat{Q}_0(T_p) = \widehat{c}_{T_p}^\top \widehat{\Sigma}_{T_p}^{-1}\widehat{c}_{T_p}$, where $\widehat{c}_{T_p}$ and $\widehat{\Sigma}_{T_p}$ are the sample versions of c_{T_p} and Σ_{T_p}, respectively, based on the sample mean $\overline{X}(t) = n^{-1}\sum_{i=1}^n X_i(t)$ and the sample covariances

$$\widehat{\mathrm{cov}}\{X(s), X(t)\} = \frac{1}{n}\sum_{i=1}^n X_i(s)X_i(t) - \overline{X}(s)\overline{X}(t)$$

of the trajectories. Then, if we want to select p variables we propose to use $\widehat{T}_{p,n}$, where

$$\widehat{T}_{p,n} := \underset{T_p\in\Theta_p}{\mathrm{argmax}}\ \widehat{Q}_0(T_p) = \underset{T_p\in\Theta_p}{\mathrm{argmax}}\ \widehat{c}_{T_p}^\top \widehat{\Sigma}_{T_p}^{-1}\widehat{c}_{T_p}. \tag{17}$$

In practice, the number of combinations of variables is usually too large to carry out an exhaustive search to find the optimal p^* variables, even for small values of p^*. Then, we need to define a search strategy to perform the selection. That is, we must decide how to explore the space of all possible combinations of variables. We propose to use the sequential approach we describe below.

Observe that a proof analogous to that of Eqs. (15) and (16) also gives their corresponding sample versions:

$$\begin{aligned}
\widehat{Q}_0(T_{p+1}) &= \widehat{Q}_0(T_p) + \frac{\left(\widehat{c}_{T_p}^\top \widehat{\Sigma}_{T_p}^{-1}\widehat{c}_{T_p,p+1} - \widehat{c}_{p+1}\right)^2}{\widehat{\sigma}_{p+1}^2 - \widehat{c}_{T_p,p+1}^\top \widehat{\Sigma}_{T_p}^{-1}\widehat{c}_{T_p,p+1}},\\
\widehat{Q}_0(T_{p+1}) &= \widehat{Q}_0(T_p) + \frac{\widehat{\mathrm{cov}}^2\{Y - \widehat{Y}_{T_p}, X(t_{p+1})\}}{\widehat{\mathrm{var}}\{X(t_{p+1}) - \widehat{X}(t_{p+1})_{T_p}\}}.
\end{aligned} \tag{18}$$

These equations suggest a sequential way to carry out the variable selection. Initially it is selected the point $t_1 \in [\delta, 1]$ which maximizes the previous quotient for $p = 1$ (which equals $\widehat{\mathrm{cov}}^2\{Y, X(t_1)\}\widehat{\mathrm{var}}\{X(t_1)\}^{-1}$). Then, in each step, we find the variable $t_{p+1} \in [\delta, 1]$, $p > 1$, maximizing the equation above. In this way, we obtain nested subsets of variables, since $T_p \prec T_{p+1}$. This greedy method does not guarantee the convergence to the global maximum of $\widehat{Q}_0$, but it shows a good behavior in practice, as we will show later on.

3.2. Asymptotic results

In the following results, we will analyze the asymptotic behavior of the estimator proposed in (17). We start with three preliminary results that may be of some interest by themselves and whose proofs are included in Sections 8.3 to 8.5. First we prove that, under some moment conditions, the sample mean and covariance functions of X converge uniformly a.s. to their population counterparts:

Lemma 1. *Assume that the process X has continuous trajectories with continuous mean and covariance functions and that it fulfills that $\mathrm{E}\{\sup_{t\in[\delta,1]}X(t)^2\} < \infty$, for a certain $\delta \geq 0$. Then,*

$$\sup_{s,t\in[\delta,1]} |\widehat{\mathrm{cov}}\{X(s),X(t)\} - \mathrm{cov}\{X(s),X(t)\}| \xrightarrow{a.s.} 0. \tag{19}$$

Next, we prove that both $\widehat{Q}_0$ and Q_0 are continuous functions for any $p \geq 1$:

Lemma 2. *Assume that the process $X(t)$ has continuous mean and covariance functions. Let $p \geq 1$ and $\Theta_p = \Theta_p(\delta)$ be such that the assumptions of Lemma 1 hold. In addition, assume that the covariance matrix Σ_{T_p} is invertible for all $T_p \in \Theta_p$. Then, the functions $\widehat{Q}_0$ and Q_0 are continuous on Θ_p.*

The two previous lemmas allow us to prove the uniform convergence on Θ_p of the empirical criterion for variable selection to the theoretical one.

Lemma 3. *Under the assumptions of Lemma 2, it holds that*

$$\sup_{T_p\in\Theta_p} |\widehat{Q}_0(T_p) - Q_0(T_p)| \xrightarrow{a.s.} 0.$$

Now assume that the sparsity condition (14) holds. Then, $T^*_{p^*} = (t^*_1,\ldots,t^*_{p^*}) \in \Theta_{p^*}$ is "sufficient" in the sense that the response only depends on the regressor variable through the values $X(t^*_1),\ldots,X(t^*_{p^*})$. We have already seen that $T^*_{p^*}$ is a global maximum of Q_0 (see the remark below Eq. (14)). In fact, we are going to prove that under mild conditions it is the only global maximum of Q_0 on Θ_{p^*} and that the estimator $\widehat{T}_{p^*,n}$ (defined in (17) with $p = p^*$) converges a.s. to $T^*_{p^*}$. Then, our proposal is able to identify consistently the true relevant points.

Theorem 1. *Assume (14) holds and that the process $X(t)$ has continuous mean and covariance functions. Suppose also that the assumptions of Lemma 1 hold for $p = p^*$, the covariance matrix $\Sigma_{T_{p^*}}$ is invertible for all $T_{p^*} \in \Theta_{p^*}$ and the covariance matrix $\Sigma_{T_{p^*}\cup S_{p^*}}$ is invertible for all $T_{p^*}, S_{p^*} \in \Theta_{p^*}$, with $T_{p^*} \neq S_{p^*}$. Then,*

(a) *The point $T^*_{p^*} \in \Theta_{p^*}$, given by (14), is the only global maximum of Q_0 on Θ_{p^*}.*
(b) *If $\widehat{T}_{p^*,n} = \mathrm{argmax}_{T_{p^*}\in\Theta_{p^*}} \widehat{Q}_0(T_{p^*})$, then $\widehat{T}_{p^*,n} \to T^*_{p^*}$ a.s. as $n \to \infty$.*
(c) *$\widehat{T}_{p^*,n}$ converges to $T^*_{p^*}$ in quadratic mean, that is, $\mathrm{E}\|\widehat{T}_{p^*,n} - T^*_{p^*}\|^2 \to 0$, as $n \to \infty$, where $\|\cdot\|$ stands for the usual Euclidean norm in $\mathbb{R}^p$.*

Proof. (a) In view of (10), it is enough to prove that $T^* := T^*_{p^*}$ is the unique global minimum of

$$Q_1(T_{p^*}) = \|Y - Y_{T_{p^*}}\|_2^2 = \|Y_{T^*} - Y_{T_{p^*}}\|_2^2 + \mathrm{var}(\varepsilon).$$

The expression above readily shows that T^* minimizes Q_1. Suppose that there exists another minimum $S^* \in \Theta_{p^*}$ such that $S^* \neq T^*$. Then, we must have $\|Y_{T^*} - Y_{S^*}\|_2^2 = 0$ and hence $Y_{T^*} - Y_{S^*} = 0$ a.s. As a consequence, using the notation $\widetilde{X}(t) = X(t) - m(t)$, there exist coefficients β_j and α_j such that $\sum_{j=1}^{p^*}\beta_j\widetilde{X}(t^*_j) - \sum_{j=1}^{p^*}\alpha_j\widetilde{X}(s^*_j) = 0$ a.s., for $T^*, S^* \in \Theta_{p^*}$ with $S^* \neq T^*$. This fact contradicts the assumption that the covariance matrix $\Sigma_{T^*\cup S^*}$ must be invertible. Therefore, $T^* = S^*$.

(b) Since the functions $\widehat{Q}_0$ and Q_0 are continuous on Θ_{p^*} (by Lemma 2) and the sequence of functions $\widehat{Q}_0$ tends uniformly a.s. to Q_0 on Θ_{p^*} (by Lemma 3) the fact that Q_0 has a unique maximum on Θ_{p^*}(part (a)) implies that $\widehat{T}_{p^*,n}$ converges almost surely to $T^*_{p^*}$.

(c) From part (b), we have $\|\widehat{T}_{p^*,n} - T^*_{p^*}\| \to 0$ a.s. as $n \to \infty$. Moreover, since both $\widehat{T}_{p^*,n}$ and $T^*_{p^*}$ belong to Θ_{p^*},

$$\|\widehat{T}_{p^*,n} - T^*_{p^*}\| \leq \|\widehat{T}_{p^*,n}\| + \|T^*_{p^*}\| \leq 2p^*.$$

The result follows from dominated convergence theorem (using $2p^*$ as the integrable dominating function). □

Once we have selected p^* points, we can use them to predict the response variable. The optimal predictions (in a square mean sense) are given by:

$$\widehat{Y}_{\widehat{T}_{p^*}} = \widehat{\beta}_1\widetilde{X}(\widehat{t}_1) + \cdots + \widehat{\beta}_{p^*}\widetilde{X}(\widehat{t}_{p^*}),$$

where $\widetilde{X}(t) = X(t) - m(t)$, and $(\widehat{\beta}_1,\ldots,\widehat{\beta}_{p^*})^\top = \widehat{\Sigma}^{-1}_{\widehat{T}_{p^*}}\widehat{c}_{\widehat{T}_{p^*}}$. On the other hand, the prediction we would use under condition (14) if we knew the true relevant points and the true values of the parameters of the model would be

$$Y_{T^*} = \beta^*_1\widetilde{X}(t^*_1) + \cdots + \beta^*_{p^*}\widetilde{X}(t^*_{p^*}),$$

where now $(\beta^*_1,\ldots,\beta^*_{p^*})^\top = \Sigma^{-1}_{T^*_{p^*}}c_{T^*_{p^*}}$. The following result refers to the asymptotic behavior of the data-driven predictions $\widehat{Y}_{\widehat{T}_{p^*}}$. It is shown that they converge a.s. and in quadratic mean to the oracle values $Y_{T^*_{p^*}}$.

Theorem 2. *Under the assumptions of Theorem 1,* $\widehat{Y}_{\widehat{T}_{p^*}} \xrightarrow{a.s.} Y_{T^*_{p^*}}$. *If, in addition, there exists* $\eta > 0$ *such that* $\mathrm{E}\{\sup_{t\in[\delta,1]}|X(t)|^{2+\eta}\} < \infty$ *then* $\widehat{Y}_{\widehat{T}_{p^*}} \xrightarrow{L^2} Y_{T^*_{p^*}}$, *as* $n \to \infty$.

Proof. For simplicity, denote $\widehat{T} := \widehat{T}_{p^*}$ and $T^* := T^*_{p^*}$. Observe that

$$\begin{aligned}
|\widehat{Y}_{\widehat{T}} - Y_{T^*}| &= |\widehat{c}_{\widehat{T}}^{\top}\widehat{\Sigma}_{\widehat{T}}^{-1}\widetilde{X}(\widehat{T}) - c_{T^*}^{\top}\Sigma_{T^*}^{-1}\widetilde{X}(T^*)| \\
&\le |\widehat{c}_{\widehat{T}}^{\top}\widehat{\Sigma}_{\widehat{T}}^{-1}\widetilde{X}(\widehat{T}) - c_{\widehat{T}}^{\top}\Sigma_{\widehat{T}}^{-1}\widetilde{X}(\widehat{T})| + |c_{\widehat{T}}^{\top}\Sigma_{\widehat{T}}^{-1}\widetilde{X}(\widehat{T}) - c_{T^*}^{\top}\Sigma_{T^*}^{-1}\widetilde{X}(T^*)| \\
&\le \|\widehat{c}_{\widehat{T}}^{\top}\widehat{\Sigma}_{\widehat{T}}^{-1} - c_{\widehat{T}}^{\top}\Sigma_{\widehat{T}}^{-1}\|\,\|\widetilde{X}(\widehat{T})\| + |c_{\widehat{T}}^{\top}\Sigma_{\widehat{T}}^{-1}\widetilde{X}(\widehat{T}) - c_{T^*}^{\top}\Sigma_{T^*}^{-1}\widetilde{X}(T^*)| \\
&\le \sup_{T\in\Theta_{p^*}}\|\widehat{c}_T^{\top}\widehat{\Sigma}_T^{-1} - c_T^{\top}\Sigma_T^{-1}\| \sup_{T\in\Theta_{p^*}}\|\widetilde{X}(T)\| + |c_{\widehat{T}}^{\top}\Sigma_{\widehat{T}}^{-1}\widetilde{X}(\widehat{T}) - c_{T^*}^{\top}\Sigma_{T^*}^{-1}\widetilde{X}(T^*)|.
\end{aligned}$$

Then, to prove $\widehat{Y}_{\widehat{T}} \to Y_{T^*}$ a.s. it is enough to see that the two addends of the last expression go to 0 a.s. Observe that $\sup_{T\in\Theta_{p^*}}\|\widehat{c}_T^{\top}\widehat{\Sigma}_T^{-1} - c_T^{\top}\Sigma_T^{-1}\| \to 0$ a.s., as $n \to \infty$, by (19) and (27) (from proof of Lemma 3). Moreover, since $X(t)$ has continuous trajectories and continuous mean function, and Θ_{p^*} is compact, we have $\sup_{T\in\Theta_{p^*}}\|\widetilde{X}(T)\| < \infty$, a.s. Finally, the continuity of c_T, Σ_T and $\widetilde{X}(T)$, together with Theorem 1(b), imply that $|c_{\widehat{T}}^{\top}\Sigma_{\widehat{T}}^{-1}\widetilde{X}(\widehat{T}) - c_{T^*}^{\top}\Sigma_{T^*}^{-1}\widetilde{X}(T^*)| \to 0$ a.s., as $n \to \infty$.

In order to prove $\widehat{Y}_{\widehat{T}} \xrightarrow{L^2} Y_{T^*}$, as $n \to \infty$, we will check that there exists $\eta > 0$ such that $\sup_n \mathrm{E}|\widehat{Y}_{\widehat{T}}|^{2+\eta} < \infty$, which in turn implies that the sequence $\widehat{Y}^2_{\widehat{T}}$ is uniformly integrable. Then, we can apply that a uniformly integrable sequence of random variables which converges in probability, also converges in L^1 (see, e.g., Proposition 6.3.2 and the corollary of Proposition 6.3.3 in [23]).

By assumption, there exists $\eta > 0$ such that $\mathrm{E}\{\sup_{t\in[\delta,1]}|\widetilde{X}(t)|^{2+\eta}\} < \infty$. Observe that

$$|\widehat{Y}_{\widehat{T}}|^{2+\eta} = |\widehat{c}_{\widehat{T}}^{\top}\widehat{\Sigma}_{\widehat{T}}^{-1}\widetilde{X}(\widehat{T})|^{2+\eta} \le \|\widehat{c}_{\widehat{T}}^{\top}\widehat{\Sigma}_{\widehat{T}}^{-1}\|^{2+\eta}\,\|\widetilde{X}(\widehat{T})\|^{2+\eta}, \text{ a.s.}$$

We have seen that $\sup_{T\in\Theta_{p^*}}\|\widehat{c}_T^{\top}\widehat{\Sigma}_T^{-1} - c_T^{\top}\Sigma_T^{-1}\| \to 0$ a.s., as $n \to \infty$. Then, given $\epsilon > 0$, for large enough n,

$$\|\widehat{c}_{\widehat{T}}^{\top}\widehat{\Sigma}_{\widehat{T}}^{-1}\|^{2+\eta} \le \epsilon + \sup_{T\in\Theta_{p^*}}\|c_T^{\top}\Sigma_T^{-1}\|^{2+\eta} := C < \infty, \text{ a.s.}$$

From the last two displayed equations, for large enough n,

$$\mathrm{E}|\widehat{Y}_{\widehat{T}}|^{2+\eta} \le C\,\mathrm{E}\|\widetilde{X}(\widehat{T})\|^{2+\eta} \le C'\,\mathrm{E}\{\sup_{t\in[\delta,1]}|\widetilde{X}(t)|^{2+\eta}\} < \infty,$$

where $C' = C(p^*)^{(2+\eta)/2}$. Since the last upper bound does not depend on n, we get $\sup_n \mathrm{E}|\widehat{Y}_{\widehat{T}}|^{2+\eta} < \infty$. □

4. Estimating the number of variables

As discussed above, our method for variable selection works, whenever the slope function β belongs to the RKHS associated with the covariance function K. It is based on the idea of asymptotically minimizing (on $T_p = (t_1, \ldots, t_p)$) the residuals

$$Q_1(T_p) := \min_{(\beta_1,\ldots,\beta_p)\in\mathbb{R}^p}\|Y - \sum_{j=1}^{p}\beta_j\widetilde{X}(t_j)\|_2^2 = \|Y - \sum_{j=1}^{p}\beta_j^*\widetilde{X}(t_j)\|_2^2,$$

where $\widetilde{X}(t_j) = X(t_j) - m(t_j)$ and $(\beta_1^*, \ldots, \beta_p^*)^{\top} = \Sigma_T^{-1}c_T$, Σ_T being the covariance matrix of $(X(t_1), \ldots, X(t_p))^{\top}$ and c_T the vector whose jth component is $\mathrm{cov}\{X(t_j), Y\}$. As proved in Proposition 1, this amounts to asymptotically maximize the function Q_0 defined in (11), which in turn is equivalent to minimize the function Q_2, defined in (9). Also, the functions, Q_1 and Q_2 agree up to an additive constant and both agree with Q_0 up to a change of sign plus an additive constant.

Throughout this section we assume the validity of the sparsity assumption (8), that is, we assume that the slope function β has the form $\beta = \sum_{j=1}^{p^*}\beta_j K(t_j^*, \cdot)$, as stated in Eq. (14), for some constants $\beta_1, \ldots, \beta_{p^*} \in \mathbb{R}$ and for $T^*_{p^*} = (t_1^*, \ldots, t_{p^*}^*)$. In this case, we can properly speak of a specific target set of "true" variables $T^* = T^*_{p^*} = (t_1^*, \ldots, t_p^*)$ to be selected and, in particular, of a "true" number p^* of variables to select.

Keeping in mind these facts, the following comments provide some clues and motivation for the data-based selection of p^*. They will be formalized in the statement and proof of Lemma 4.

(a) On the one hand, any selection of type $T_p = (t_1, \ldots, t_p)$ with $p < p^*$ is clearly sub-optimal, since it would lack some relevant information, contributed by the variables in T^* not in T_p.
(b) Likewise, a choice T_p "by excess" with $T^* \prec T_p$ would not provide any benefit. To see this note that, under (8), the minimum of Q_2 is obviously attained at T^* and the value of Q_2 at such minimum is 0, which cannot be improved.

(c) As a consequence, the maximum value of Q_2 for points with p^*+1 coordinates is attained at some T_{p^*+1} such that $T^* \prec T_{p^*+1}$ (that is, T^* is a sub-vector of T_{p^*+1}) but, in any case, $Q_0(T_{p^*+1}) - Q_0(T^*) = 0$.

(d) Then, the optimal p^* is such that the maximum value of $Q_0(T_p)$ agrees with that of $Q_0(T_{p^*})$ for any T_p such that $T_{p^*} \prec T_p$. Thus p^* is in fact the "elbow" value in the plot of $p \mapsto Q_0(T_p^*)$ from which on the increase of the maximum values of Q_0 stops.

The following lemma will set the theoretical basis of our procedure of estimation of p^*. As a consequence of this result, a procedure to estimate p^* is proposed below.

Lemma 4. *Let us consider the model (4) under the assumption that β can be expressed as $\beta(t) = \sum_{j=1}^{p^*} \beta_j K(t_j^*, \cdot)$, where p^* is the minimal integer for which such representation holds. Define $\widehat{Q}_0^{\max}(p) = \max_{T_p \in \Theta_p} \widehat{Q}_0(T_p)$. Then, under the assumptions of Lemma 2 we have*

(a)

$$\widehat{Q}_0^{\max}(p^*+1) - \widehat{Q}_0^{\max}(p^*) \to 0, \ \text{a.s.}, \tag{20}$$

(b) for all $p < p^*$,

$$\lim_n \left\{\widehat{Q}_0^{\max}(p+1) - \widehat{Q}_0^{\max}(p)\right\} > 0, \ \text{a.s.}$$

The proof of this result can be found in Section 8.6. This suggests the following method to estimate p^*:

1. Define

$$\Delta = \min_{p<p^*} \left\{Q_0(T_{p+1}^*) - Q_0(T_p^*)\right\}. \tag{21}$$

Assume we are able to fix a value $\epsilon > 0$ such that $\epsilon < \Delta$.

2. Define

$$\widehat{p} = \min\left\{p : \widehat{Q}_0^{\max}(p+1) - \widehat{Q}_0^{\max}(p) < \epsilon\right\}, \tag{22}$$

In view of Eq. (18), this difference can be rewritten as the quotient involved in that equation.

Theorem 3. *Under the assumptions of Lemma 4 the estimator $\widehat{p}$ defined in (22) fulfills $\widehat{p} \to p^*$, almost surely.*

Proof. This result is a direct consequence of Lemma 4. □

In practice, the calculation of Δ defined in Eq. (21) is not feasible, since it is merely a theoretical bound. Thus, the restriction $\epsilon < \Delta$ should be understood as choosing a value ϵ small enough. In order to fix this value from the data, different approaches could be used. For instance in [9], where empirical methods to select both p and T_p in functional classification are given, the authors suggest to set ϵ equal to $\rho\widehat{Q}_0^{\max}(p-1)$ for a pre-determined small ρ. Nevertheless, we suggest to use techniques inspired in the change point detection methodology in time series, which avoid the need of an additional parameter. We could interpret the collection of values $L_n(p) = \ln\{\widehat{Q}_0^{\max}(p+1) - \widehat{Q}_0^{\max}(p)\}$ for $p = 1, \ldots$ as a time series and apply the usual k-means clustering algorithm to these values, with $k = 2$. Then, ϵ is fixed as $L_n(p)$, for p the largest value such that $L_n(p)$ belongs to the same cluster as $L_n(1)$. This is equivalent to estimate $\widehat{p}$ directly as the minimum value of p such that all the values $L_n(p)$ with $p \geq \widehat{p}$ belong to a different cluster than that of $L_n(1)$. This is the approach used in the experimental setting exposed in Section 6.

5. When p^* is not estimated: the conservative oracle property

Under the sparseness assumption (14), where p^* is unknown, another sensible approach for the choice of the number p of selected variables is to take a conservative, large enough value of p.

The basic idea of this section is easy to state: suppose that a "conservative oracle" gives us a value p such that $p > p^*$. Accordingly, we perform our variable selection procedure for such value p. This yields p variables $\widehat{t}_1, \ldots, \widehat{t}_p$. Then, we can be sure that the "true" variables $t_1^*, \ldots, t_{p^*}^*$ are very close to p^* variables in $\{\widehat{t}_1, \ldots, \widehat{t}_p\}$.

The next result formalizes this property.

Theorem 4. *Let us consider the model (4) under the assumption that β can be expressed as $\beta(t) = \sum_{j=1}^{p^*} \beta_j K(t_j^*, \cdot)$, where p^* is the minimal integer for which such representation holds. Let $\widehat{t}_1, \ldots, \widehat{t}_p$ be the variables selected by the method (17), where p is a given value larger than p^*. Then, for all $\epsilon > 0$,*

$$\Pr\left\{t_i^* \in \bigcup_{j=1}^{p} (\widehat{t}_j - \epsilon, \widehat{t}_j + \epsilon), \ i = 1, \ldots, p^*\right\} = 1, \ \text{eventually, as } n \to \infty. \tag{23}$$

Proof. Recall that the choice of the variables $T_p = (t_1, \ldots, t_p)$ is performed by asymptotically maximizing the function $Q_0(T_p)$, defined in (11). More precisely, as Q_0 depends on unknown population quantities, we in fact maximize the estimator $\widehat{Q}_0$ defined in Section 3.1.

Now, let us note that the maximum of Q_0 is not unique. Indeed, we assume that the "minimal" sparse representation of β has the form $\beta(t) = \sum_{j=1}^{p^*} \beta_j K(t_j^*, \cdot)$ but, of course, if $p > p^*$, we may formally put $\beta(t) = \sum_{i=1}^{p} \beta_i K(s_i, \cdot)$ as long as the "true" optimal points $t_1^*, \ldots, t_{p^*}^*$ are among the s_i's and all the coefficients β_i not matching with such t_i^*'s are null. On the other hand, from the uniqueness of the function β, all the maxima of $Q_0(T_p)$ must have an expression of this type.

Let us assume that conclusion (23) does not hold. Then, with positive probability, there exists a subsequence of maxima $\widehat{T}_{p,n}^*$ of $\widehat{Q}_0$ such that the point t_1^*, for instance, is not contained in the union of $(\widehat{t}_j - \epsilon, \widehat{t}_j + \epsilon), j = 1, \ldots, p$. Thus, with positive probability, there is a further subsequence (denoted again $\widehat{T}_{p,n}^*$) converging to some T_p^{**} whose coordinates are all at a distance of, at least, ϵ from t_1^*. According to Lemma 3, $\widehat{Q}_0(T_p)$ converges to $Q_0(T_p)$ uniformly a.s. in T_p. In particular this entails that, with positive probability, $\widehat{Q}_0(\widehat{T}_{p,n}^*)$ converges to $Q_0(T_p^{**})$, which contradicts the fact that T_p^{**} cannot be a maximum of Q_0. □

This result is reminiscent of the *Sure Screening Property* defined in [10], which is used to quantify the efficiency of multivariate variable selection methods. But, obviously, property (23) is adapted to cope with the functional nature of the data and the fact that the values t_i range on a continuous domain.

6. Experiments

The purpose of this section is to give some insights into the practical behavior of our proposal for variable selection, both in simulations and real data examples. We are aware that the design of these experiments is largely discretionary, as the range of possible models for simulation is potentially unlimited (especially in the case of functional data models) and there is also a considerable amount of real data examples currently available in the FDA literature. Still, our choices have not been completely arbitrary. We have tried to follow some objective criteria. First, the theoretical models chosen for the simulations must obviously include some situations in which our crucial "sparseness" assumption $\beta = \sum_j \beta_j K(t_j, \cdot)$ is fulfilled. As discussed above, such models are quite natural if we are willing to use variable selection techniques. Also, it looks reasonable to include at least one model in which this assumption is not valid. Regarding the real data, we have just chosen two examples used in the recent literature for the purpose of checking other variable selection methods in functional regression settings.

In any case, we would like to emphasize that we make here no attempt to draw any definitive conclusion on the performance of our method when compared with others. In our view, no unique empirical study can lead to safe, objective conclusions in this regard: the only reliable verdict should be given by the users community, after some time of practice with real data problems. Our purposes here are far more unassuming; we just want to provide some hints suggesting that our proposal

(a) has a satisfactory performance in the "sparse" models for which it has been designed,

(b) can be implemented in practice, with an affordable computational cost,

(c) could be hopefully competitive under other theoretical models, far from the ideal assumption $\beta = \sum_j \beta_j K(t_j, \cdot)$,

(d) has also a satisfactory practical performance in a couple of real data examples commonly used in the literature of variable selection.

The R code used in the experiments is accessible on-line and can be provided on request.

6.1. Simulation experiments

Keeping in mind the above general lines, we next define the simulation models under study. In our context a "model" is defined by three elements: a stochastic process (from which the functional data are generated), a regression equation, of type $Y = \langle X, \beta \rangle_K + \varepsilon$ (or, more generally, $Y = g(X) + \varepsilon$) and an error variable ε. In what follows, ε has been chosen in all cases as $\varepsilon \sim \mathcal{N}(0, \sigma)$ with $\sigma = 0.2$.

We have considered several processes, covering a broad range of different situations.

1. *Standard Brownian Motion* (Bm) $\{B(t),\ t \in [0, 1]\}$.
2. *Geometric Brownian Motion* (gBm). This non-Gaussian process is also known as exponential Brownian motion. It can be defined just by $X(t) = e^{B(t)}$.
3. *Integrated Brownian Motion* (iBm): it is obtained as $X(t) = \int_0^t B(s)\mathrm{d}s$. Note that the trajectories of this non-Markovian process are smooth.
4. *Ornstein–Uhlenbeck process* (OU). This is a Gaussian process $\{X(t)\}$ which satisfies the stochastic differential equation $\mathrm{d}X(t) = \theta\{\mu - X(t)\}\mathrm{d}t + \sigma \mathrm{d}B(t)$. In our simulations we have chosen $\theta = \mu = \sigma = 1$.
5. *Fractional Brownian Motion* (fBM). This process is a generalization of the Brownian motion $B(t)$ but, unlike $B(t)$, it does not have (in general) independent increments. The mean function of this Gaussian process is identically 0 and its covariance function is $K(t, s) = 0.5(|t|^{2H} + |s|^{2H} - |t - s|^{2H})$, where $H \in (0, 1)$ is the so-called Hurst exponent. Note that for $H = 0.5$, this process coincides with the standard Brownian Motion. Also, the trajectories of this process are still not differentiable at every point but the index H is closely related to the Hölder continuity properties of these trajectories. In particular, when $H > 0.5$, the trajectories look "more regular" than those of the Brownian motion, having a wilder appearance for $H < 0.5$. To cover both cases we have used $H = 0.2$ and $H = 0.8$ in our simulations.

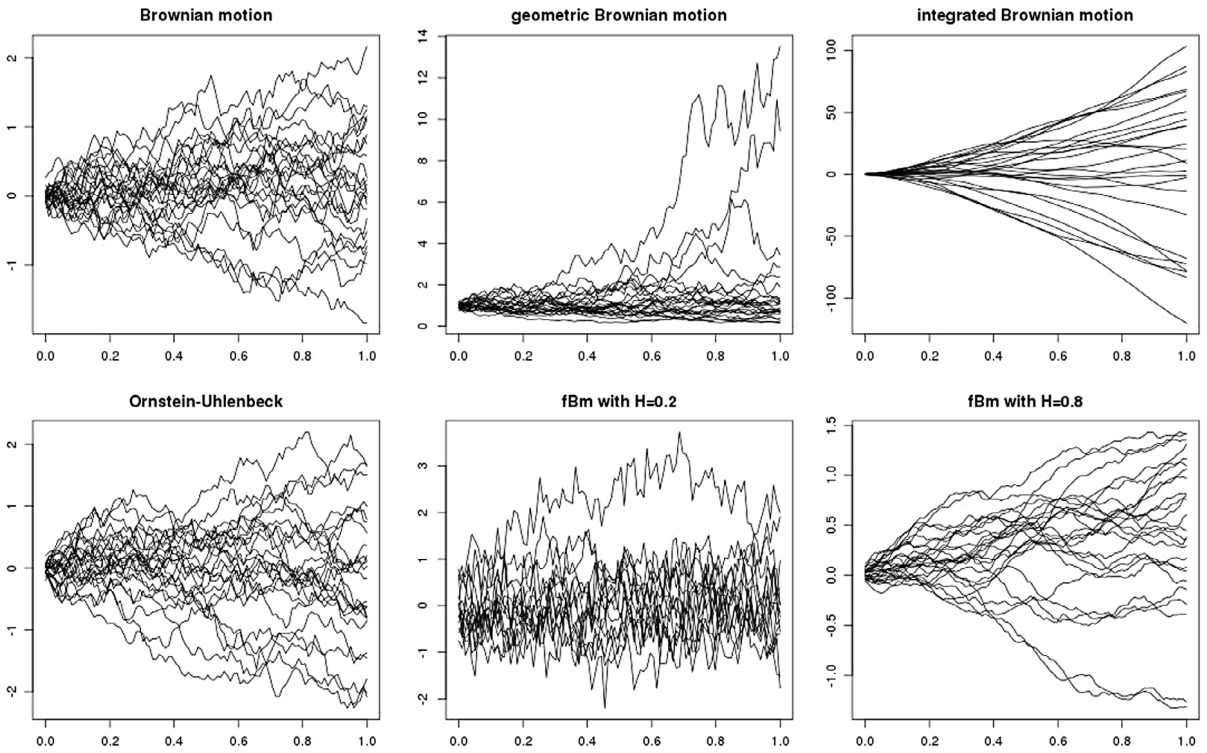

Fig. 1. 25 trajectories of each of the processes used in the simulations.

Fig. 1 shows some trajectories of each of these six processes, where the variables $X(t)$ for t in a neighborhood of 0 have been omitted to satisfy the non-degeneracy requirements of the method.

As for the regression function g we have considered the following three choices.

1. Two functions β in (4) of type $\beta(t) = \sum_j \beta_j K(t_j^*, \cdot)$ so that the regression model reduces to the "sparse version" (8). We have considered two different regression functions. For the first one, we have used the set of points $T^* = (0.2, 0.4, 0.9)$ with weights $(\beta_1, \beta_2, \beta_3) = (2, -5, 1)$; this is "Regression model 1" in the tables. For the second one, we have used $T^* = (0.16, 0.47, 0.6, 0.85, 0.91)$ and weights $(2.1, -0.2, -1.9, 5, 4.2)$ ("Regression model 2" in the tables). Therefore, the response variables in both cases are given, respectively, by
$$Y_1 = 2X(0.2) - 5X(0.4) + X(0.9) + \varepsilon,$$
$$Y_2 = 2.1X(0.16) - 0.2X(0.47) - 1.9X(0.67) + 5X(0.85) + 4.2X(0.91) + \varepsilon.$$
2. $\beta(t) = \ln(1 + t)$ and the regression model is (1) with $\alpha_0 = 0$. Thus, the sparse RKHS model (8) does not hold in this case. This is "Regression model 3" in the tables. It has been already used in [8]. Therefore, the corresponding response variable is generated by
$$Y_3 = \int_0^1 \ln(1 + t)X(t)\mathrm{d}t + \varepsilon.$$

6.2. Real data

We have also checked the different methods when applied to two real data sets. Since these data have been already considered in other recent papers of the FDA literature, we will give only brief descriptions of them.

1. *Tecator*. This data set has been widely used in the literature. The trajectories $X(t)$ are 100 channel spectrum of absorbances of 215 meat samples, and the response variable Y is the fat content. However, (in most versions of this data set) 15 of these observations are repeated, so we have removed them. The remaining 200 trajectories are discretized on a grid of 100 points. As usual when working with these measurements, we use the second derivative of the curves instead of the original ones. One version of this data set (including repeated curves) can be found as part of the `fda.usc` R-package. The version we have used in our experiment is available along with the R-code.

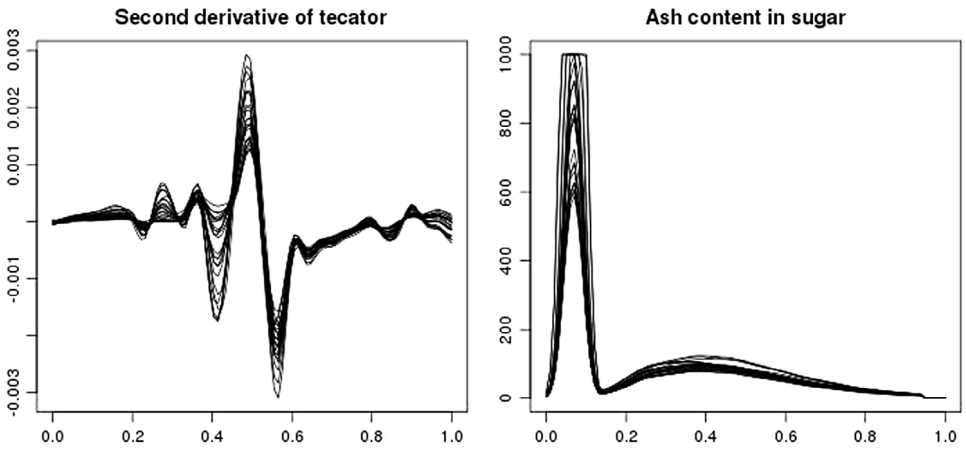

Fig. 2. 25 trajectories of each of the real data sets.

2. *Ash content in sugar samples*. This data set has been used, for example, in [1]. The version we use corresponds in fact to a subset of the whole data set, available in http://www.models.kvl.dk/Sugar_Process. The response variable Y is the percentage of ash content in 266 sugar samples. The trajectories $X(t)$ are the fluorescence spectra from 275 to 560 nm at excitation wavelength 290. These curves are discretized on a grid of 100 equispaced points.

Fig. 2 shows some trajectories of both original data sets, although we work with centered version of the curves.

6.3. Methods under study and methodology

We compare our proposal with other methods for variable selection recently considered in the literature. We now list the methods under study along with the notation used in Tables 1–6. For the first three methods we have used our own R implementations.

1. The method proposed in this paper (RKHS). It has been implemented using the iterative approximation described in Eq. (18). The number of relevant points is chosen as explained at the end of Section 4, by doing clustering on the values $L_n(p)$. Therefore, no validation technique is required.
2. The variable selection procedure proposed in [22] (KPS): in the original article, a mixed method for standard functional linear regression and variable selection technique is proposed. Since we are here concerned with variable selection, we have implemented just the corresponding part of the proposal. Essentially, the idea is to select the points (called "impact points" in [22]) maximizing the covariance between the response variable Y and a "decorrelated" version, $Z(t)$, of the original process. By construction, the decorrelated process $Z(t)$ is such that $Z(t)$ and $Z(s)$ are almost uncorrelated whenever $|t-s| \geq \delta$. The value δ and the number of selected variables are chosen using the BIC criterion, as proposed in the original paper.
3. *Partitioning Variable Selection* (PVS) with ML penalization, as proposed in [1]. The original sample must be split into two independent subsamples, which should be asymptotically of the same sizes. The basic idea is to apply some multivariate variable selection technique in this context, but taking advantage of the functional structure of the data. The procedure works in two steps. In the first step, one constructs an equispaced subgrid of variables among all the variables in the original grid (see below). Then a variable selection technique for multivariate data is applied on this subgrid, using the first subsample of the original data. For instance, we might use LASSO with ML penalization, as proposed in the original paper. Then, in the second step, this variable selection technique is applied again to an enlarged grid, constructed by taking all variables in an interval around those selected in step 1 (using the second subsample). We have used the R function "cv.glmnet" of the package `glmnet`. This function fits a generalized linear model via maximum likelihood with LASSO penalization and the amount of the penalization is fixed by 10-fold cross validation. Although the default implementation of the function standardizes the variates, we have decided not to do it in this case. This version of LASSO selects automatically the number of variables, thus the only smoothing parameter to be selected here is the grid step for the points used in first stage of the method. This parameter is also set by 10-fold cross-validation.
4. *Maxima Hunting* (MH), proposed in [4]: the original method was used in the setting of variable selection in supervised classification, but there is no conceptual restriction to apply the same procedure in a regression setting. The basic idea is to select the local maxima of the "distance covariance" (association measure for random variables introduced in [29]) between the response and the marginal variables of the process. In practice, the numerical identification of these maxima depends on a smoothing parameter h which is chosen by 10-fold cross-validation. The number of variables is also set by cross-validation. The code of this method was provided by the authors [4].

5. *Partial Least Squares* (PLS). This technique is well-known among the functional data practitioners. The goal of PLS is not to pick up a few variables but to select some appropriate linear functionals of the original data (very much in the spirit of principal components analysis). So PLS is not a variable selection procedure, but a dimension reduction method. This means that PLS is not directly comparable to the variable selection methods considered here, since its aims are not exactly the same. When we choose to use a variable selection procedure, it is understood that we want to perform some kind of dimension reduction still keeping the interpretability of the information directly given in terms of the original variables. By contrast, PLS might perhaps provide some gains in efficiency but at the expense of doing a dimensionality reduction with a more difficult interpretation. Anyway, we have included PLS in our study as a useful reference for comparisons, aimed at checking how much do we lose by restricting ourselves to variable selection methods. We have used the function "fregre.pls.cv" of the `fda.usc` R-package to compute the predictions. The number of components in the model is chosen using the "Akaike information criterion" (AIC). Moreover, for the real data sets, which are smoother than the simulated ones, we have found that it is better to penalize the norm of the second derivative of the slope function. The amount of penalization in these cases is also fixed using AIC model selection criterion.
6. *Base*. We denote by "base" the prediction methodology derived from the standard L_2 linear regression model (1). No variable selection or dimension reduction procedure is done. Hence, this method is incorporated again just for the sake of comparison, to assess the accuracy of the predictions based on some previous variable selection procedure with those provided by the standard functional regression model (1). For the real data sets and the integrated Brownian motion, which are quite smooth, we have used the function "fregre.basis.cv" of the `fda.usc` R-package (which relies on a spline basis representation of the trajectories) to compute the predictions. In this function the number of basis elements to retain is set by generalized cross-validation. As for PLS, we allow a penalization in the second derivative. However, for the remaining examples, we have found that it is better to use a data-derived basis, in order to capture the irregularities of the data. For these sets we have used the function "fregre.pc.cv" with no penalization, in which the number of components is chosen by the "Akaike information criterion".

For each model, all methods are checked for the sample size $n = 150$, which has been split on 100 observations used as training data and 50 employed as test data. As usual, the functional simulated data are discretized to $(X(t_1), \ldots, X(t_{100}))$, where t_i are equispaced points in $[0, 1]$, starting from $t_1 = 1/100$. The real data sets are already provided in a discretized fashion, so we use the corresponding grids. For all the experiments we obtain the Relative Mean Squared Error (RMSE) of each method, as defined by

$$\text{RMSE}\,(\widehat{Y}; Y) = \frac{\sum_{i=1}^{n} \left(\widehat{Y}_i - Y_i\right)^2}{\sum_{i=1}^{n} Y_i^2}. \tag{24}$$

Moreover, in those cases where β has a "sparse" form of type $\beta(t) = \sum_j \beta_j K(t_j, \cdot)$ we obtain two measures of the accuracy in the variable selection procedure. Namely, we calculate the Hausdorff distance between the set of estimated points and the set of the real ones T^* as well as the number of points selected $(\widehat{p})$ in order to compare it with the real number (p^*). We have imposed a maximum of 10 selected points to the methods, except to PVS, since our implementation of this method does not permit to decide the number of selected variables. The Hausdorff distance gives us an idea of the precision of the method since it takes into account the separation to the real points as well as the number of selected points. Each experiment has been replicated 100 times.

6.4. Numerical outputs

Tables 1–3 provide the performance of the methods measured in terms of the Relative Mean Squared Error (Eq. (24)) of the predictions, for each of the three regression models tested and for the two real data sets. Methods are presented in columns. Models appear in rows. Table 4 contains the Hausdorff distance between the selected and the "true" relevant points for the four variable selection methods and the two models with sparse β function. For these same experiments, Table 5 provides the number of selected points. In all the tables of the simulated data appear the mean and the standard deviation of each of the measured quantities.

In view of these tables, for the simulated data the proposed method seems to outperform the other variable selection procedures, according to all the considered criteria metrics (RMSE, Hausdorff distance and number of points) whenever a sparse model of type (2) holds. The PVS method also performs quite well. When the model is not satisfied, and the response variable depends on the whole trajectory, PLS and the base method are the best in general, as expected. The Maxima Hunting method outperforms these two methods in some cases. The order of magnitude of the error for the remaining methods is the same as that of PLS in most cases. That is, using a few number of variables instead of the whole trajectory does not significantly affect the prediction error, even if the response depends on the whole trajectory.

The exact points selected with each of the methods for the real data sets are plotted in Figs. 3 and 4. It seems that our method and Maxima Hunting are the ones that better manage the continuity of the data, in the sense that they do not choose close and highly correlated points. We can see also that there are some points in common among the ones selected by MH and our proposal. For the Tecator set, KPS obtains the best results, followed by our method. The results presented here for this data set might not coincide with those of previous works, since (as explained before) we have removed the repeated

Table 1
Mean and standard deviation of the RMSE for the response variable for simulated data sets with models 1 and 2 (scale of 10^{-2}).

	RKHS	KPS	PVS	MH	PLS	Base
Regression model 1						
Bm	1.09 (0.394)	2.38 (1.94)	1.2 (0.36)	5.2 (3.08)	4.62 (1.4)	5.9 (1.56)
gBm	0.42 (0.292)	5.5 (5.53)	0.55 (0.343)	6.55 (6.24)	3.71 (1.81)	5.02 (2.62)
iBM	0.0166 (0.0162)	0.106 (0.162)	0.13 (0.0553)	48.5 (10.9)	0.139 (0.0523)	0.0303 (0.00813)
OU	1.07 (0.433)	2.25 (1.51)	1.23 (0.462)	5.59 (3.57)	4.61 (1.63)	6 (2.07)
fBm 0.2	0.422 (0.134)	1.76 (2.52)	0.532 (0.185)	14.6 (5.27)	11.3 (3.78)	29.7 (7.83)
fBm 0.8	2.91 (0.929)	3.54 (1.41)	3.08 (0.99)	11.4 (7.04)	3.75 (1.03)	3.79 (1.04)
Regression model 2						
Bm	0.0976 (0.0481)	1.18 (0.595)	0.171 (0.0779)	2.48 (1.04)	0.596 (0.158)	0.693 (0.201)
gBm	0.0206 (0.0128)	1.03 (0.75)	0.295 (0.329)	2.17 (1.61)	0.429 (0.164)	0.462 (0.236)
iBm	0.000312 (0.000491)	0.00102 (0.000744)	0.102 (0.0368)	0.0525 (0.015)	0.00405 (0.00165)	0.000221 (0.0000625)
OU	0.0846 (0.0292)	1.19 (0.715)	0.163 (0.0783)	2.28 (1.25)	0.54 (0.166)	0.613 (0.167)
fBm 0.2	0.0831 (0.0215)	5.25 (4.01)	0.17 (0.11)	4.7 (4.26)	3.14 (0.999)	7.16 (2.162)
fBm 0.8	0.105 (0.0303)	0.168 (0.0592)	0.373 (0.183)	0.442 (0.132)	0.138 (0.0382)	0.122 (0.0294)

Table 2
Mean and standard deviation of the RMSE for the response variable for simulated data sets with model 3 (scale of 10^{-1}).

Regression model 3						
	RKHS	KPS	PVS	MH	PLS	Base
Bm	4.44 (1.29)	4.25 (1.08)	4.1 (1.1)	3.88 (1.05)	4.05 (1.13)	3.79 (0.969)
gBm	1.37 (1.25)	0.987 (0.23)	1.01 (0.343)	1.84 (0.603)	0.887 (0.323)	0.884 (0.324)
iBm	0.00379 (0.00118)	0.0033 (0.00109)	0.0127 (0.00477)	0.042 (0.0116)	0.00299 (0.000997)	0.00305 (0.00106)
OU	4.77 (1.2)	4.45 (1.13)	4.11 (0.982)	3.97 (0.928)	4.28 (1.24)	4.01 (0.998)
fBm 0.2	4.41 (1.01)	4.28 (0.992)	4.12 (1)	4.09 (0.95)	4.49 (1.18)	3.71 (0.835)
fBm 0.8	4.89 (1.15)	4.45 (0.988)	4.23 (1.1)	4.03 (0.867)	4.08 (0.897)	4.11 (0.921)

Table 3
RMSE for the response variable for real data sets.

	RKHS	KPS	PVS	MH	PLS	Base
2nd derivative of tecator	0.032	0.034	0.030	0.056	0.039	0.048
Ash content in sugar	0.321	0.185	0.222	0.246	0.401	0.465

observations. For the sugar data, our proposal is slightly outperformed, but it is better than the estimators that use the whole trajectories. In addition, RKHS method uses only 2 points in this case, in contrast with KPS and PVS (10 points each of them, which is the fixed maximum).

On the other hand, we also provide a couple of results regarding execution time. We have measured the execution time of the six methods for the third regression model when the functional data are drawn from the Ornstein–Uhlenbeck process and the fractional Brownian motion with $H = 0.8$. The results can be seen in Table 6. As we have already mentioned, the RKHS-based method does not require any validation step to determine the number of selected variables. Therefore, the

Table 4
Mean and standard deviation of the Hausdorff distance to the actual relevant points (Scale of 10^{-1}).

	RKHS	KPS	PVS	MH
Regression model 1				
Bm	**0.089** (0.141)	2.1 (0.374)	1.63 (0.648)	2.04 (0.431)
gBm	**0.135** (0.219)	2.17 (0.304)	1.71 (0.692)	2.13 (0.637)
iBm	**0.998** (0.0141)	2.14 (0.313)	1.05 (0.828)	6.01 (0.54)
OU	**0.139** (0.238)	2.03 (0.354)	1.64 (0.625)	2.03 (0.41)
fBm 0.2	**0** (0)	2.08 (0.362)	1.69 (0.592)	2.01 (0.621)
fBm 0.8	**0.264** (0.201)	2.09 (0.28)	1.6 (0.69)	2.9 (1.74)
Regression model 2				
Bm	1.29 (0.0731)	1.27 (0.181)	**0.994** (0.312)	1.51 (0.733)
gBm	1.25 (0.134)	1.43 (0.875)	**1.21** (0.869)	1.62 (0.929)
iBm	**0.775** (0.297)	1.26 (0.181)	7.25 (0.355)	6.53 (1.44)
OU	1.27 (0.127)	1.21 (0.238)	**1.03** (0.321)	1.44 (0.674)
fBm 0.2	1.3 (0)	1.27 (0.19)	**0.946** (0.419)	1.33 (0.272)
fBm 0.8	**1.18** (0.268)	1.2 (0.218)	4.29 (2.9)	2.77 (1.12)

Table 5
Mean and standard deviation of the number of selected points ($\widehat{p}$).

	RKHS	KPS	PVS	MH
Regression model 1 ($p^* = 3$)				
Bm	**3.14** (0.427)	8.3 (2.05)	11.1 (3.95)	6.42 (1.7)
gBm	**3.59** (0.753)	8.22 (2.01)	10.9 (3.61)	6.2 (1.75)
iBm	6.09 (0.473)	9.12 (1.35)	11.6 (4.32)	**1.17** (0.378)
OU	**3.23** (0.489)	8.34 (2.01)	10.9 (3.88)	5.91 (1.48)
fBm 0.2	**3** (0)	8.56 (1.78)	11.5 (3.79)	5.6 (2)
fBm 0.8	3.29 (0.478)	8.92 (1.73)	10.6 (4.21)	**2.92** (1.32)
Regression model 2 ($p^* = 5$)				
Bm	**4.88** (0.743)	9.61 (0.975)	8.75 (2.07)	6.15 (1.98)
gBm	**5.45** (1.09)	9.55 (0.869)	8.89 (2.74)	6.14 (2.2)
iBm	**5.68** (1.1)	9.28 (1.08)	3.03 (0.964)	1.3 (0.459)
OU	**5.05** (0.84)	9.69 (0.704)	8.11 (1.95)	5.77 (1.84)
fBm 0.2	**4** (0.086)	9.72 (0.727)	6.96 (1.88)	7.41 (1.92)
fBm 0.8	**5.07** (0.742)	9.29 (1.11)	6.96 (3.35)	3.17 (1.16)

execution time is significantly smaller than that of the other variable selection methods. We can also see that the execution time for the PVS method is much bigger than the others. Note however that this method has in general a good behavior in terms of prediction error.

Table 6
Mean and standard deviation of the execution time.

	RKHS		KPS		PVS		MH		PLS		Base	
OU	0.00566	(0.00136)	11.6	(1.11)	14.1	(1.95)	1.15	(0.0717)	0.173	(0.0447)	0.342	(0.0372)
fBm 0.8	0.00806	(0.00254)	8.39	(2.21)	35.9	(11.1)	1.65	(0.335)	0.24	(0.0831)	0.49	(0.123)

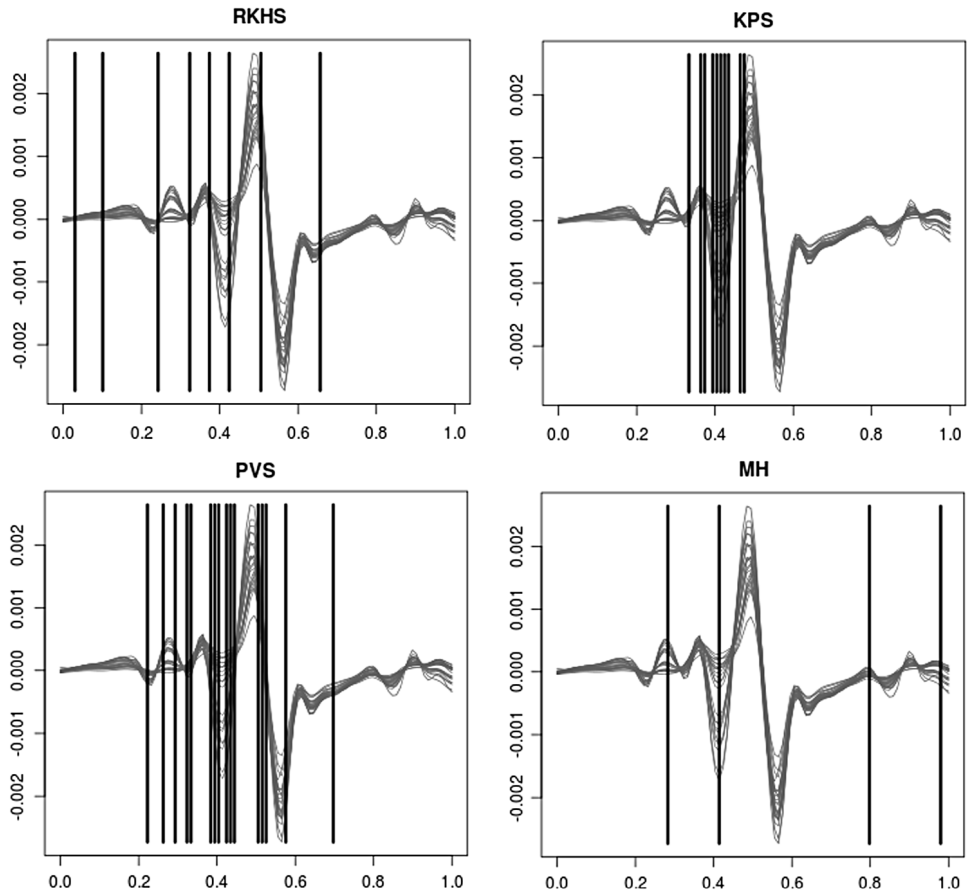

Fig. 3. Original curves of Tecator data set with the selected points for each of the methods.

7. Conclusions

The RKHS approach we have introduced in this paper provides a natural framework for a formal unified theory of variable selection for functional data. The "sparse" models (those where the variable selection techniques are fully justified) appear as particular cases in this setup. As a consequence, it is possible to derive asymptotic consistency results as those obtained in the paper. Likewise, it is also possible to consider the problem of estimating the "true" number of relevant variables in a consistent way, as we do in Section 4. This is in contrast with other standard proposals for which the number of variables is previously fixed as an input, or it is determined using cross validation and other computationally expensive methods. Then, our proposal is more firmly founded in theory and, at the same time, provides a much faster method in practice, which is important when dealing with large data sets.

The empirical results we have obtained are encouraging. In short, according to our experiments, the RKHS-based method works better than other variable selection methods is those sparse models that fulfill the ideal theoretical conditions we need. In the non sparse model considered in the simulations, the RKHS method is slightly outperformed by other proposals (but still behaves reasonably). Finally, in the "neutral" field of real data examples the performance looks also satisfactory and competitive.

Last but not least, from a general, methodological point of view, this paper represents an additional example of the surprising usefulness of reproducing kernels in statistics. Additional examples can be found in [3,5,17,21,31].

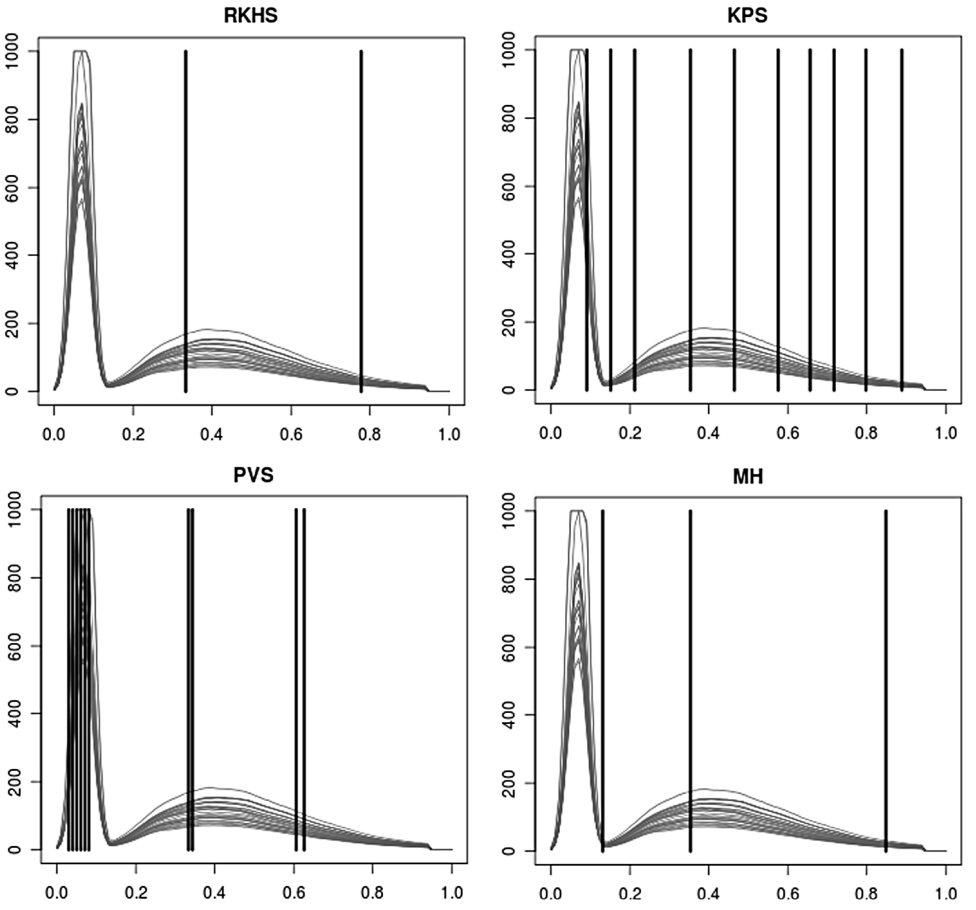

Fig. 4. Original curves of the ash content in sugar set with the selected points for each of the methods.

8. Some additional proofs

8.1. *Proposition 2*

We have to rewrite the expression $c_{T_{p+1}}^\top \Sigma_{T_{p+1}}^{-1} c_{T_{p+1}}$, where $p \geq 1$. We can write the matrix $\Sigma_{T_{p+1}}$ in block form as

$$\Sigma_{T_{p+1}} = \left(\begin{array}{ccc|c} & \Sigma_{T_p} & & \begin{array}{c}\mathrm{cov}\{X(t_1), X(t_{p+1})\} \\ \vdots \\ \mathrm{cov}\{X(t_p), X(t_{p+1})\}\end{array} \\ \hline \mathrm{cov}\{X(t_1), X(t_{p+1})\} & \ldots & \mathrm{cov}\{X(t_p), X(t_{p+1})\} & \mathrm{cov}\{X(t_{p+1}), X(t_{p+1})\} \end{array}\right)$$

$$\equiv \left(\begin{array}{c|c} \Sigma_{T_p} & c_{T_p,p+1} \\ \hline c_{T_p,p+1}^\top & \sigma_{p+1}^2 \end{array}\right).$$

Then its inverse matrix is

$$\Sigma_{T_{p+1}}^{-1} = \left(\begin{array}{c|c} \Sigma_{T_p}^{-1} + \frac{1}{a}\Sigma_{T_p}^{-1} c_{T_p,p+1} c_{T_p,p+1}^\top \Sigma_{T_p}^{-1} & -\frac{1}{a}\Sigma_{T_p}^{-1} c_{T_p,p+1} \\ \hline -\frac{1}{a} c_{T_p,p+1}^\top \Sigma_{T_p}^{-1} & \frac{1}{a} \end{array}\right),$$

where $a = \sigma_{p+1}^2 - c_{T_p,p+1}^\top \Sigma_{T_p}^{-1} c_{T_p,p+1}$. We can also write the vector of covariances as

$$c_{T_{p+1}}^\top = (\mathrm{cov}\{X(t_1), Y\}, \ldots, \mathrm{cov}\{X(t_{p+1}), Y\}) = (c_{T_p} \mid c_{p+1}).$$

Using this notation we can rewrite the original expression as follows,

$$\begin{aligned} c_{T_{p+1}}^\top \Sigma_{T_{p+1}}^{-1} c_{T_{p+1}} &= c_{T_p}^\top \Sigma_{T_p}^{-1} c_{T_p} + \frac{1}{a} c_{T_p}^\top \Sigma_{T_p}^{-1} c_{T_p,p+1} c_{T_p,p+1}^\top \Sigma_{T_p}^{-1} c_{T_p} - \frac{c_{p+1}}{a} c_{T_p,p+1}^\top \Sigma_{T_p}^{-1} c_{T_p} \\ &\quad - \frac{c_{p+1}}{a} c_{T_p}^\top \Sigma_{T_p}^{-1} c_{T_p,p+1} + \frac{c_{p+1}^2}{a} \\ &= c_{T_p}^\top \Sigma_{T_p}^{-1} c_{T_p} + \frac{1}{a} \left\{ \left(c_{T_p}^\top \Sigma_{T_p}^{-1} c_{T_p,p+1} \right)^2 - 2 c_{p+1} c_{T_p}^\top \Sigma_{T_p}^{-1} c_{T_p,p+1} + c_{p+1}^2 \right\} \\ &= c_{T_p}^\top \Sigma_{T_p}^{-1} c_{T_p} + \frac{\left(c_{T_p}^\top \Sigma_{T_p}^{-1} c_{T_p,p+1} - c_{p+1} \right)^2}{\sigma_{p+1}^2 - c_{T_p,p+1}^\top \Sigma_{T_p}^{-1} c_{T_p,p+1}}, \end{aligned}$$

since the product $c_{T_p}^\top \Sigma_{T_p}^{-1} c_{T_p,p+1}$ is actually a real number.

8.2. Proposition 3

Using the notation of the statement, if $\widetilde{X}(t)$ is the centered process, we have $Y_{T_p} = c_{T_p}^\top \Sigma_{T_p}^{-1} \widetilde{X}(T_p)$. Thus we can rewrite the numerator of the quotient of Eq. (15) as

$$\begin{aligned} \operatorname{cov}\left\{ Y - Y_{T_p}, X(t_{p+1}) \right\} &= c_{p+1} - \operatorname{cov}\left\{ Y_{T_p}, X(t_{p+1}) \right\} \\ &= c_{p+1} - c_{T_p}^\top \Sigma_{T_p}^{-1} \operatorname{cov}\left\{ \widetilde{X}(T_p), X(t_{p+1}) \right\} \\ &= c_{p+1} - c_{T_p}^\top \Sigma_{T_p}^{-1} c_{T_p,p+1}, \end{aligned}$$

since the covariances are not affected by the centering. We can also write $\operatorname{cov}\left\{ Y - Y_{T_p}, X(t_{p+1}) - X(t_{p+1})_{T_p} \right\}$ in the same way, since $(Y - Y_{T_p}) \perp X(t_{p+1})_{T_p}$. For the denominator, we have (taking into account that $X(t_{p+1})_{T_p} = c_{T_p,p+1}^\top \Sigma_{T_p}^{-1} \widetilde{X}(T_p)$),

$$\begin{aligned} &\operatorname{var}\left\{ X(t_{p+1}) - X(t_{p+1})_{T_p} \right\} \\ &\quad = \operatorname{var}\left\{ X(t_{p+1}) \right\} + \operatorname{var}\left\{ X(t_{p+1})_{T_p} \right\} - 2\operatorname{cov}\left\{ X(t_{p+1}), X(t_{p+1})_{T_p} \right\} \\ &\quad = \sigma_{p+1}^2 + c_{T_p,p+1}^\top \Sigma_{T_p}^{-1} c_{T_p,p+1} - 2 c_{T_p,p+1}^\top \Sigma_{T_p}^{-1} \operatorname{cov}\left\{ X(t_{p+1}), \widetilde{X}(T_p) \right\} \\ &\quad = \sigma_{p+1}^2 - c_{T_p,p+1}^\top \Sigma_{T_p}^{-1} c_{T_p,p+1}. \end{aligned}$$

From these two expressions the conclusion follows straightforwardly.

8.3. Lemma 1

Note that the assumption implies $\mathrm{E}\{\sup_{t\in[\delta,1]} |X(t)|\} < \infty$ and the stochastic process $\{X(t) : t \in [\delta, 1]\}$ has finite strong expectation with trajectories in $\mathcal{C}[\delta, 1]$, which is a separable Banach space. Then, we can apply Mourier's SLLN (see, e.g., Theorem 4.5.2 of [23, p. 452]) to conclude

$$\sup_{t\in[\delta,1]} |\overline{X}(t) - m(t)| \xrightarrow{a.s.} 0, \tag{25}$$

Similarly, the process $Z(s, t) := X(s)X(t)$, with trajectories in $\mathcal{C}([\delta, 1]^2)$, is such that its strong expectation exists. Indeed, since

$$0 \le \{|X(s)| - |X(t)|\}^2 = |X(s)|^2 + |X(t)|^2 - 2|Z(s, t)|,$$

it holds

$$\mathrm{E}\{ \sup_{s,t\in[\delta,1]} |Z(s, t)| \} \le \mathrm{E}\{ \sup_{t\in[\delta,1]} |X(t)|^2 \} < \infty.$$

Moreover, $\mathcal{C}([\delta, 1]^2)$ is separable since $[\delta, 1]^2$ is compact. Then, Mourier's SLLN and (25) imply (19).

8.4. Lemma 2

Fix $p \ge 1$. First, we prove that Q_0 is continuous. Since the process $X(t)$ has continuous mean and covariance functions we have that

$$c_{T_p} = (\operatorname{cov}\{X(t_1), Y\}, \ldots, \operatorname{cov}\{X(t_p), Y\})^\top$$

is continuous on Θ_p. On the other hand, since the entries of Σ_{T_p} are continuous on $[0,1]^2$, $\det(\Sigma_{T_p})$ is also continuous on Θ_p, where $\det(\Sigma)$ stands for the determinant of Σ. By assumption, $\det(\Sigma_{T_p}) > 0$ for all $T_p \in \Theta_p$. Since Θ_p is compact, $\inf_{T_p\in\Theta_p} \det(\Sigma_{T_p}) > 0$. Observe that

$$\Sigma_{T_p}^{-1} = \frac{\mathrm{adj}(\Sigma_{T_p})}{\det(\Sigma_{T_p})},$$

where $\mathrm{adj}(\Sigma)$ denotes the adjugate of Σ. As a consequence, the entries of $\Sigma_{T_p}^{-1}$ are continuous on Θ_p, and hence the function Q_0 is also continuous.

The proof for $\widehat{Q}_0$ is analogous with the only difference that in this case we must ensure that $\inf_{T_p\in\Theta_p} \det(\widehat{\Sigma}_{T_p}) > 0$ with probability 1. Notice that for all $n \geq 1$ and $T_p \in \Theta_p$, $\det(\widehat{\Sigma}_{T_p}) > 0$ a.s. On the other hand, from (19) it follows that

$$\sup_{T_p\in\Theta_p} |\det(\widehat{\Sigma}_{T_p}) - \det(\Sigma_{T_p})| \xrightarrow{a.s} 0. \tag{26}$$

We have seen before that $\inf_{T_p\in\Theta_p} \det(\Sigma_{T_p}) > 0$. Then, with probability 1, there exists n_0 such that if $n \geq n_0$, $\inf_{T_p\in\Theta_p} \det(\widehat{\Sigma}_{T_p}) > 0$.

8.5. Lemma 3

It is enough to establish the uniform convergence a.s. of the coordinates of $\widehat{c}_{T_p}$ and the entries of $\widehat{\Sigma}_{T_p}^{-1}$ to those of c_{T_p} and $\Sigma_{T_p}^{-1}$ respectively.

Eq. (25) and the same argument leading to (25) but applied to the process $Z(t) = X(t)Y$ yield

$$\sup_{t\in[\delta,1]} \left| \frac{1}{n}\sum_{i=1}^{n} \{X_i(t) - \overline{X}(t)\}(Y_i - \overline{Y}) - \mathrm{cov}\{X(t), Y\} \right| \xrightarrow{a.s} 0, \tag{27}$$

and hence the uniform convergence a.s. of the coordinates of $\widehat{c}_{T_p}$ to those of c_{T_p}.

Finally, observe that $\widehat{\Sigma}_{T_p}^{-1} = \det(\widehat{\Sigma}_{T_p})^{-1}\mathrm{adj}(\widehat{\Sigma}_{T_p})$. Then, from (19), (26) and $\inf_{T_p\in\Theta_p} \det(\Sigma_{T_p}) > 0$ we conclude the uniform convergence a.s. of the entries of $\widehat{\Sigma}_{T_p}^{-1}$ to those of $\Sigma_{T_p}^{-1}$.

8.6. Lemma 4

(a) Let us first prove

$$Q_0(T^*_{p^*+1}) - Q_0(T^*_{p^*}) = 0. \tag{28}$$

To see this, note that in the proof of Proposition 1 we have proved $Q_0(T_p) = \|\beta\|_K^2 - Q_2(T_p)$ with $Q_2(T_p) = \min_{(\beta_1,\ldots,\beta_p)\in\mathbb{R}^p} \|\beta - \sum_{j=1}^{p}\beta_j K(t_j,\cdot)\|_K^2$. Also, under (14), $Q_2(T^*_{p^*}) = 0$ so that $Q_0(T^*_{p^*}) = \|\beta\|_K^2$, which is the maximum possible value of Q_0. On the other hand, it is clear that $Q_2(T^*_{p^*+1}) \leq Q_2(T^*_{p^*})$ so that we must also have $Q_2(T^*_{p^*+1}) = 0$ and $Q_0(T^*_{p^*+1}) = \|\beta\|_K^2$. This proves (28). Now conclusion (20) follows directly from the uniform convergence of $\widehat{Q}_0$ to Q_0, as established in Lemma 3.

(b) Similarly to part (a) we only need to prove

$$Q_0(T^*_{p+1}) - Q_0(T^*_p) > 0 \text{ for all } p < p^*. \tag{29}$$

Indeed, assume we have $Q_0(T^*_{p+1}) - Q_0(T^*_p) = 0$ for some $p < p^*$. Then, since the prediction error $Q_1(T_p)$ defined in (5) satisfies $Q_1(T_p) = -Q_0(T_p) + \|\beta\|_K^2 + \sigma^2$ we would have that the prediction error $Q_1(T^*_p)$ obtained with p variables in the sparse model $Y = \sum_{j=1}^{q}\beta_j X(t_j) + \varepsilon$, with $\mathrm{var}(\varepsilon) = \sigma^2$ would be the same, for $q = p$ and $q = p+1$. Then, by recurrence, we get that the error would be in fact the same, irrespective of the number of q explanatory variables in the range $p, \ldots, p^*$. Thus, the linear model $Y = \langle\beta, X\rangle_K + \varepsilon$ holds for a slope function of type $\beta = \sum_{j=1}^{p}\beta_j K(t^*_j,\cdot)$ with $p < p^*$. This is a contradiction with the assumption that p^* is the minimal value for which such model holds.

Now the result follows from (29) and the a.s. uniform convergence of $\widehat{Q}_0$ to Q_0.

Acknowledgments

This work has been partially supported by Spanish Grant MTM2016-78751-P and the European Social Fund (Ayudas para contratos predoctorales para la formación de doctores 2014, Ministerio de Economía y Competitividad, Spain) (BES-2014-070460). Insightful comments from two referees are gratefully acknowledged.

References

[1] G. Aneiros, P. Vieu, Variable selection in infinite-dimensional problems, Statist. Probab. Lett. 94 (2014) 12–20.
[2] G. Aneiros, P. Vieu, Sparse nonparametric model for regression with functional covariate, J. Nonparametr. Stat. 28 (4) (2016) 839–859.
[3] A. Berlinet, C. Thomas-Agnan, Reproducing Kernel Hilbert Spaces in Probability and Statistics, Kluwer Academic, Boston, 2004.
[4] J.R. Berrendero, A. Cuevas, J.L. Torrecilla, Variable selection in functional data classification: a maxima-hunting proposal, Statist. Sinica 26 (2016) 619–638.
[5] J.R. Berrendero, A. Cuevas, J.L. Torrecilla, On the use of reproducing kernel Hilbert spaces in functional classification, J. Amer. Statist. Assoc. (2017). http://dx.doi.org/10.1080/01621459.2017.1320287 (in press).
[6] H. Cardot, P. Sarda, Functional Linear Regression, in: F. Ferraty, Y. Romain (Eds.), Handbook of Functional Data Analysis, Oxford University Press, Oxford, 2010, pp. 21–46.
[7] A. Cuevas, A partial overview of the theory of statistics with functional data, J. Statist. Plann. Inference 147 (2014) 1–23.
[8] A. Cuevas, M. Febrero, R. Fraiman, Linear functional regression: The case of fixed design and functional response, Can. J. Statist. / Rev. Can. Statist. 30 (2) (2002) 285–300.
[9] A. Delaigle, P. Hall, N. Bathia, Componentwise classification and clustering of functional data, Biometrika 99 (2) (2012) 299.
[10] J. Fan, J. Lv, Sure independence screening for ultrahigh dimensional feature space, J. R. Stat. Soc. Ser. B Stat. Methodol. 70 (5) (2008) 849–911.
[11] J. Fan, J. Lv, A selective overview of variable selection in high dimensional feature space, Statist. Sinica 20 (1) (2010) 101.
[12] F. Ferraty, P. Hall, P. Vieu, Most-predictive design points for functional data predictors, Biometrika 97 (4) (2010) 807–824.
[13] F. Ferraty, P. Vieu, Nonparametric Functional Data Analysis: Theory and Practice, Springer Science & Business Media, 2006.
[14] R. Fraiman, Y. Gimenez, M. Svarc, Feature selection for functional data, J. Multivariate Anal. 146 (2016) 191–208.
[15] A. Goia, P. Vieu, An introduction to recent advances in high/infinite dimensional statistics, J. Multivariate Anal. 146 (2016) 1–6 Special Issue on Statistical Models and Methods for High or Infinite Dimensional Spaces.
[16] L. Horváth, P. Kokoszka, Inference for Functional Data with Applications, Vol. 200, Springer Science & Business Media, 2012.
[17] T. Hsing, R. Eubank, Theoretical Foundations of Functional Data Analysis, With an Introduction to Linear Operators, John Wiley & Sons, 2015.
[18] T. Hsing, H. Ren, An RKHS formulation of the inverse regression dimension-reduction problem, Ann. Statist. 37 (2) (2009) 726–755.
[19] S. Janson, Gaussian Hilbert Spaces, Vol. 129, Cambridge university press, 1997.
[20] H. Ji, H.-G. Müller, Optimal designs for longitudinal and functional data, J. R. Stat. Soc. Ser. B Stat. Methodol. 79 (3) (2017) 859–876.
[21] H. Kadri, E. Duflos, P. Preux, S. Canu, A. Rakotomamonjy, J. Audiffren, Operator-valued kernels for learning from functional response data, J. Mach. Learn. Res. 16 (2015) 1–54.
[22] A. Kneip, D. Poss, P. Sarda, Funcional linear regression with points of impact, Ann. Statist. 44 (1) (2016) 1–30.
[23] R. Laha, V. Rohatgi, Probability Theory, in: Wiley Series in Probability and Mathematical Statistics, John Wiley & Sons, New York-Chichester-Brisbane, 1979.
[24] M.N. Lukić, J.H. Beder, Stochastic processes with sample paths in reproducing kernel Hilbert spaces, Trans. Amer. Math. Soc. 353 (10) (2001) 3945–3969.
[25] I.W. McKeague, B. Sen, Fractals with point impact in functional linear regression, Ann. Statist. 38 (4) (2010) 2559–2586.
[26] A.J. Miller, Subset Selection in Regression, in: Monographs on statistics and applied probability, Chapman & Hall/CRC, Boca Raton, 2002.
[27] E. Parzen, An approach to time series analysis, Ann. Math. Statist. (1961) 951–989.
[28] N.S. Pillai, Q. Wu, F. Liang, S. Mukherjee, R.L. Wolpert, Characterizing the function space for Bayesian kernel models, J. Mach. Learn. Res. 8 (Aug) (2007) 1769–1797.
[29] G.J. Székely, M.L. Rizzo, N.K. Bakirov, Measuring and testing dependence by correlation of distances, Ann. Statist. 35 (6) (2007) 2769–2794.
[30] C.D. Yenigün, M.L. Rizzo, Variable selection in regression using maximal correlation and distance correlation, J. Stat. Comput. Simul. 85 (8) (2015) 1692–1705.
[31] M. Yuan, T.T. Cai, A reproducing kernel Hilbert space approach to functional linear regression, Ann. Statist. 38 (6) (2010) 3412–3444.