Statistical Papers (2024) 65:5167–5201
https://doi.org/10.1007/s00362-024-01567-9

REGULAR ARTICLE

On the functional regression model and its finite-dimensional approximations

José R. Berrendero[1] · Alejandro Cholaquidis[2] · Antonio Cuevas[1]

Received: 9 May 2023 / Revised: 1 January 2024 / Published online: 10 July 2024

Abstract

The problem of linearly predicting a scalar response Y from a functional (random) explanatory variable $X = X(t),\ t \in I$ is considered. It is argued that the term "linearly" can be interpreted in several meaningful ways. Thus, one could interpret that (up to a random noise) Y could be expressed as a linear combination of a finite family of marginals $X(t_i)$ of the process X, or a limit of a sequence of such linear combinations. This simple point of view (which has some precedents in the literature) leads to a formulation of the linear model in terms of the RKHS space generated by the covariance function of the process $X(t)$. It turns out that such RKHS-based formulation includes the standard functional linear model, based on the inner product in the space $L^2[0, 1]$, as a particular case. It includes as well all models in which Y is assumed to be (up to an additive noise) a linear combination of a finite number of linear projections of X. Some consistency results are proved which, in particular, lead to an asymptotic approximation of the predictions derived from the general (functional) linear model in terms of finite-dimensional models based on a finite family of marginals $X(t_i)$, for an increasing grid of points t_j in I. We also include a discussion on the crucial notion of coefficient of determination (aimed at assessing the fit of the model) in this setting. A few experimental results are given.

Keywords Functional data analysis · Functional regression · RKHS methods · Comparison of linear models

✉ José R. Berrendero
joser.berrendero@uam.es

✉ Alejandro Cholaquidis
acholaquidis@hotmail.com

Antonio Cuevas
antonio.cuevas@uam.es

[1] Departmento de Matemática, Universidad Autónoma de Madrid and Instituto de Ciencias Matemáticas ICMAT (CSIC-UAM-UCM-UC3M), 28049 Madrid, Spain

[2] Facultad de Ciencias, Universidad de la República, Montevideo, Uruguay

1 Introduction

Linear regression is a topic of leading interest in statistics. The general paradigm is well-known: one aims to predict a response variable Y in the best possible way as a linear (or affine) function of some explanatory variable X. In the classical case of multivariate regression, where Y is a real random variable and X takes values in $\mathbb{R}^d$ there is little doubt about the meaning of "linear". However, this is not that obvious when X is a more complex object that can be modelled via different mathematical structures. The most important example arises perhaps in the field of Functional Data Analysis (FDA) where $X = X(t)$ is a real function; see e.g., Cuevas (2014) for a general survey on FDA and Horváth and Kokoszka (2012) for a more detailed account, including a short overview of functional linear models.

1.1 Some notation

More precisely, we will deal here with the scalar-on-function regression problem where the response Y is a real random variable and X is a random function (i.e., a trajectory of a stochastic process). In formal terms, let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and denote by $L^2(\Omega) = L^2(\Omega, \mathcal{F}, \mathbb{P})$ the space of square integrable random variables defined on $(\Omega, \mathcal{F}, \mathbb{P})$. Denote by $\langle X, Y\rangle_{L^2(\Omega)} = \mathbb{E}(XY)$ the inner product in this space and by $\|\cdot\|_{L^2(\Omega)}$ the corresponding norm. For $Y_1, Y_2 \in L^2(\Omega)$ the notation $Y_1 \perp Y_2$ will stand for $\mathbb{E}(Y_1 Y_2) = 0$.

Consider a response variable $Y \in L^2(\Omega)$ and a family of regressors $\{X(t) : t \in I\} \subset L^2(\Omega)$, where I is an arbitrary index set. For the sake of simplicity we will assume both the response and the explanatory variable are centred, so that $\mathbb{E}(Y) = \mathbb{E}(X(t)) = 0$, for all $t \in I$. The covariance function $K : I \times I \to \mathbb{R}$ of $\{X(t) : t \in I\}$ is the symmetric, positive semidefinite function given by $K(s,t) = \langle X(s), X(t)\rangle_{L^2(\Omega)} = \mathbb{E}(X(s)X(t))$. In what follows we will assume that $I = [0,1]$ and $K(s,t)$ is continuous on $I \times I$. We will also assume that all involved processes are separable. This can always be done without loss of generality, since Separability Theorem (Ash and Gardner 1975, p. 166) establishes that any process $\{X(t),\ t \in [0,1]\}$ has a separable version.

The trajectories $X(\cdot) = X(t)$ of the underlying process are assumed to live in the space $L^2[0,1]$ of square integrable real functions. Denote by $\langle\cdot,\cdot\rangle_2$ and $\|\cdot\|_2$, respectively, the usual inner product and norm in $L^2[0,1]$. The norm in the finite-dimensional Euclidean space $\mathbb{R}^p$ will be simply denoted by $\|\cdot\|$.

1.2 The aim of this work. Some motivation

Our purpose is to show that the term "linear" admits several interpretations in the functional case; all of them could be useful, depending on the considered context. We will provide a general formulation of the linear model and we will show that several useful formulations of functional linear models appear as particular cases. Some basic consistency results will be given regarding the estimation of the slope

(possibly functional) parameter. The theory of Reproducing Kernel Hilbert Spaces (RKHS) will be an important auxiliary tool in our approach; see Berlinet and Thomas-Agnan (2004).

In order to give some perspective and motivation, let us consider, for $I = [0, 1]$, the classical L^2-based functional regression model, as given by

$$Y = \int_0^1 \beta(t)X(t)dt + \varepsilon := \langle \beta, X \rangle_2 + \varepsilon, \tag{1}$$

where $X = X(t)$ is a process, ε is an error variable, with $\mathbb{E}(\varepsilon|X) = 0$ and $\mathrm{Var}(\varepsilon|X) = \sigma^2$, and $\beta \in L^2[0, 1]$ is the slope function. The usual aim in such a model is estimating β and σ^2 from an iid sample (X_i, Y_i), $i = 1, \ldots, n$. As we are assuming $\mathbb{E}(X(t)) = 0$ we omit as well the additional intercept additive term β_0 in the theoretical developments involving model (1). This term will be incorporated in the numerical examples of Sect. 6.

The vast majority of literature on functional linear regression is concerned with model (1); see, e.g., the pioneering book by Ramsay and Silverman (2005) (whose first edition dates back to 1997), as well as the paper by Cardot et al. (1999), the book by Horváth and Kokoszka (2012) and references therein. Though this model is, in several aspects, natural and useful, we argue here that this is not the only sensible approach to functional linear regression.

There are several reasons for this statement: first, unlike the finite dimensional model $Y = \beta_1 X_1 + \ldots + \beta_d X_d + \varepsilon$, there is no obvious, easy to calculate, estimator for β under model (1). The simple, elegant least squares theory is no longer available here. The optimality properties (Gauss–Markov Theorem) of the finite-dimensional least squares estimator do not directly apply to (1) either. Second, note that in the finite-dimensional situation, where $X = (X_1, \ldots, X_d)$, there is a strong case in favour of a model of type $Y = \beta_1 X_1 + \ldots + \beta_d X_d + \varepsilon$. The reason is that, as it is well-known, when the joint distribution of all the involved variables is Gaussian, the best approximation of Y in terms of $X = (X_1, \ldots, X_d)$ is necessarily of type $\beta_1 X_1 + \ldots + \beta_d X_d$. A similar motivating property does not hold for model (1). Third, some natural, linear-like functional approaches do not appear as particular instances of (1); this is the case, for example, with an approach of type *the response Y is (up to an additive noise) a linear combination of a finite subset of variables* $\{X(t),\ t \in I\}$.

Our goal here is to analyse a more general linear model which partially addresses these downsides and includes model (1) as a particular case. Perhaps more importantly, the finite dimensional models of type $Y = \beta_1 X(t_1) + \ldots + \beta_p X(t_p) + \varepsilon$, with $t_j \in I$, $\beta_j \in \mathbb{R}$ and $p \in \mathbb{N}$ are also included. This is particularly relevant in practice since, in some cases, the predictive power of such models may be larger than that of the L^2-based model (1); see the experiments in Sect. 6. The special points t_i used to define these finite-dimensional models are often called "impact points". Some interesting references on impact points-based functional regression (with no explicit use of an RKHS approach) are McKeague and Sen (2010), Kneip et al. (2016), Poß et al. (2020)

1.3 Some literature on RKHS methods in functional regression

The book by Hsing and Eubank (2015) is a good reference for the mathematical basis of functional data analysis, including the use of RKHS theory in this field.

If we focus on RKHS methods in functional regression models, we should refer to the papers by Hsing and Ren (2009) and Kneip and Liebl (2020).

The RKHS-based linear model (2) we will consider below has been previously analysed by other authors, from slightly different points of view. Thus (Shin and Hsing 2012, Eq. (2.3)), use that model with a particular emphasis in prediction. In fact, our Theorem 2 below provides a result similar to that Theorem 3.1 in Shin and Hsing (2012) under quite different conditions. Also, the RKHS formulation of the functional linear model explicitly appears in Hsing and Ren (2009) from a rather general perspective. Some closely related ideas appear as well in Kneip and Liebl (2020) focussing on the topic of reconstructing partially observed functional data.

The linear model is also considered from the RKHS perspective by Berrendero et al. (2019). Still, this work is only focussed in variable selection problems (i.e., on the estimation of the impact points) with no further theoretical development of the RKHS-based model.

From a completely different point of view, the RKHS methodology in functional regression has been previously addressed by Yuan and Cai (2010) and Shin and Lee (2016). Let us note, however, that these authors in fact deal with the classical L^2-model (1); the RKHS techniques are used in these papers to define the penalization term in a penalized approach to the estimation of the slope function β. See also Shang and Cheng (2015) in the context of generalized linear models.

1.4 The organization of this paper

In Sect. 2 our general linear RKHS-based model is defined. Section 3 is devoted to prove that some relevant examples of practical interest appear just as particular cases of such a model. Two results of consistent estimation of the slope function are given in Sect. 4. A discussion of the coefficient of determination on this setting is given in Sect. 5. Some experimental results are discussed in Sect. 6. Some conclusions are summarized in Sect. 7. The proof of Theorem 2 is given in the final Appendix.

2 A general formulation of the functional linear model

In the functional framework introduced in the previous section, a linear model might be defined as any suitable linear expression of the variables $X(t)$ aiming to explain (predict) the response variable Y. The L^2-model (1) is just one possible formulation of such idea.

In the present work, by "linear expression" we mean an element of the closed linear subspace L_X of $L^2(\Omega)$ spanned by the variables in $\{X(t) : t \in I\}$. In other words, L_X is the closure of the linear subspace of all finite linear combinations of variables

in the collection. Hence, L_X includes both finite linear combinations of the form $\sum_{j=1}^{p} \beta_j X(t_j)$ (where $p \in \mathbb{N}$, $\beta_1, \ldots, \beta_p \in \mathbb{R}$, and $t_1, \ldots, t_p \in I$) and rv's U such that there exists a sequence U_n of these linear combinations with $\|U_n - U\|_{L^2(\Omega)} \to 0$, as $n \to \infty$.

In more precise terms, our general linear model will be defined by assuming that Y and $\{X(t) : t \in I\}$ are related by

$$Y = U_X + \varepsilon, \tag{2}$$

where $U_X \in L_X$, and ε is a random variable with $\mathbb{E}(\varepsilon|X) = 0$ and $\mathrm{Var}(\varepsilon|X) = \sigma^2$ (a positive constant). Note that $\mathbb{E}(\varepsilon|X) = 0$ entails that ε belongs to $L_X^{\perp}$, the orthogonal complement of L_X, that is $\mathbb{E}(\varepsilon U) = 0$ for all $U \in L_X$. Let us also assume throughout, by simplicity, that $\mathbb{E}(X(t)) = 0$ for all $t \in I$; see Remark 1 below.

Since L_X is closed, we know that $L^2(\Omega) = L_X \oplus L_X^{\perp}$, and then the elements in the model are unambiguously given by the orthogonal projections $U_X = \mathrm{Proj}_{L_X}(Y)$ and $\varepsilon = \mathrm{Proj}_{L_X^{\perp}}(Y)$. Given a linear continuous operator $\mathcal{T}$, $\|\mathcal{T}\|_{op}$ stands for the usual operator norm $\|\mathcal{T}x\|_{op} = \sup_{\|x\|=1} \|\mathcal{T}x\|$.

2.1 An RKHS formulation of the proposed linear model

The aim of this subsection is to give a fairly natural parametrization of model (2), based on the RKHS theory. As a consequence of this alternative formulation we will show that several useful linear models appear as particular cases of (2).

Let us begin by briefly defining the notion of RKHS associated with a symmetric, positive semidefinite function $K : [0,1]^2 \to \mathbb{R}$; in our case, K will be the (continuous) covariance function of the process X whose trajectories provide the functional data. We first define the space H_K^0 of functions $g : [0,1] \to \mathbb{R}$ of the form $g(\cdot) = \sum_{j=1}^{n} \beta_j K(\cdot, t_j)$, for all possible choices of $n \in \mathbb{N}$, $t_1, \ldots, t_n \in [0,1]$ and $\beta_1, \ldots, \beta_n \in \mathbb{R}$. If $f \in H_K^0$ with $f(\cdot) = \sum_{i=1}^{m} \alpha_i K(\cdot, s_i)$, the inner product $\langle f, g \rangle_K = \sum_{i,j} \alpha_i \beta_j K(s_i, t_j)$ provides H_K^0 with a structure of pre-Hilbert space. Then H_K is defined as the completion of H_K^0, obtained by incorporating the pointwise limits of all Cauchy sequences of functions in H_K^0. The "completed" RKHS inner product fulfills the so-called *reproducing property* $\langle f, K(\cdot.t) \rangle_K = f(t)$, for all $t \in [0,1]$, $f \in H_K$. See Berlinet and Thomas-Agnan (2004), Cucker and Zhou (2007) and Janson (1997) for more details.

In the following paragraphs we recall a deep interpretation of the RKHS theory in statistical terms. Denote by $\mathbb{R}^I$ the set of all real functions defined on $I = [0,1]$. We are going to define a map $\Psi_X : L_X \to \mathbb{R}^I$ that will play a key role in the sequel: given $U \in L_X$, $\Psi_X(U)$ is just the function

$$\Psi_X(U)(t) = \mathbb{E}(UX(t)). \tag{3}$$

As we will next show, this transformation defines an isometry (often called Loève's isometry; see Lemma 1.1 in Lukić and Beder (2001)) between L_X and $\Psi_X(L_X)$. We

will also see that $\Psi_X(L_X)$ coincides in fact with the RKHS generated by K, that we have denoted H_K.

Let us recall here, for the sake of completeness, a simple lemma collecting two elementary properties of Ψ_X:

Lemma 1 *Let $\Psi_X(U)$ be as defined in (3). Then,*

(a) Ψ_X is injective.
(b) $\Psi_X(X(t))(\cdot) = K(\cdot, t)$, for all $t \in I$. Equivalently, $\Psi_X^{-1}[K(\cdot, t)] = X(t)$.

Proof To show (a), let $U, V \in L_X$ be such that $\Psi_X(U)(t) = \Psi_X(V)(t)$, for all $t \in I$. Then, $\mathbb{E}[(U - V)X(t)] = 0$, for all $t \in I$, what implies $U - V \in L_X^{\perp}$. Since we also have $U - V \in L_X$, we get $U = V$; recall that in L^2 spaces we identify those functions that coincide almost surely. Property (b) is obvious from the definition. □

As a consequence of this result $\Psi_X : L_X \to \Psi_X(L_X)$ is a bijection (Loève's isometry). Observe that by Lemma 1(b), all the finite linear combinations $\sum_{j=1}^{p} \beta_j K(\cdot, t_j)$ belong to $\Psi_X(L_X)$.

The inner product in L_X induces an inner product in $\Psi_X(L_X)$: given $f, g \in \Psi_X(L_X)$, define

$$\langle f, g \rangle := \langle \Psi_X^{-1}(f), \Psi_X^{-1}(g) \rangle_{L^2(\Omega)}.$$

It turns out that $\Psi_X(L_X)$, endowed with $\langle \cdot, \cdot \rangle$, is a Hilbert space. Once we endow $\Psi_X(L_X)$ with this structure, the mapping Ψ_X is a linear, bijective, and inner product preserving operator between L_X and $\Psi_X(L_X)$; this accounts for the word "isometry".

On the other hand, it is well-known (see, e.g., Appendix F in Janson (1997) for details) that, given a positive semidefinite function $K : I \times I \to \mathbb{R}$ (called "reproducing kernel"), there is a unique Hilbert space, generated by the linear combinations of the form $\sum_j \beta_j K(\cdot, t_j)$. This space is usually called the Reproducing Kernel Hilbert Space associated with K.

Let us recall also the following simple result, that shows that $\Psi_X(L_X)$ coincides in fact with the RKHS H_K generated by the covariance function K of the process $X = X(t)$.

Proposition 1 *The Hilbert space $\Psi_X(L_X)$ and the covariance function K satisfy the following two properties:*

(a) For all $t \in I$, $K(\cdot, t) \in \Psi_X(L_X)$.
(b) Reproducing property*: for all $f \in \Psi_X(L_X)$ and $t \in I$, $\langle f, K(\cdot, t) \rangle = f(t)$.*

Proof (a) follows directly from Lemma 1(b). To prove (b),

$$\langle f, K(\cdot, t) \rangle = \langle \Psi_X^{-1}(f), X(t) \rangle_{L^2(\Omega)} = \mathbb{E}[\Psi_X^{-1}(f) X(t)] = \Psi_X[\Psi_X^{-1}(f)](t) = f(t).$$

The first equality is also due to Lemma 1(b). □

Finally, from the uniqueness result mentioned above, we conclude that the space $(\Psi_X(L_X), \langle \cdot, \cdot \rangle)$ coincides with the RKHS $(H_K, \langle \cdot, \cdot \rangle_K)$ defined at the beginning of

this subsection. We are now in a position to recast the general model (2) into a sort of parametric formulation, where the "parameter" belongs to the RKHS generated by the covariance function K of the process $X = X(t)$. As we will see, this reformulation will be particularly useful to encompass several particular cases of practical relevance.

Theorem 1 *Model (2) can be equivalently established in the form*

$$Y = \Psi_X^{-1}(\alpha) + \varepsilon, \tag{4}$$

where $\alpha \in H_K$ *and* ε *is a random variable with* $\mathbb{E}(\varepsilon|X) = 0$ *and* $\mathrm{Var}(\varepsilon|X) = \sigma^2$.

In addition the "parameter" α *is the cross-covariance function* $\alpha(t) = \mathbb{E}(YX(t))$.

Proof Formulation (4) follows directly from the definition of Ψ_X and the fact that this transformation is a bijection between L_X and H_K; hence $U_X \in L_X$ if and only if there exists a (unique) $\alpha \in H_K$ such that $U_X = \Psi_X^{-1}(\alpha)$.

To prove the second assertion note that, by the reproducing property, $\alpha(t) = \langle \alpha, K(\cdot, t)\rangle_K$ for all $t \in I$, and hence

$$\alpha(t) = \langle \alpha, K(\cdot, t)\rangle_K = \mathbb{E}[\Psi_X^{-1}(\alpha)X(t)] = \mathbb{E}[(Y - \varepsilon)X(t)] = \mathbb{E}[YX(t)], \tag{5}$$

because $\varepsilon \in L_X^{\perp}$. □

As a consequence, the RKHS H_K is a fairly natural parametric space for our general linear regression model.

Remark 1 Let us note that model (4) was already considered, with a different notation, in the paper by Berrendero et al. (2019). Indeed, the inverse Loève transformation $\Psi_X^{-1}(\alpha)$ is sometimes denoted $\langle \alpha, X\rangle_K$ (or $\langle X, \alpha\rangle_K$). This is somewhat of a notational abuse, as typically the trajectories of the process do not belong to the RKHS H_K; see Berrendero et al. (2019) for details. Still, the notation is often convenient, so that we can also use the following expression to formulate the RKHS-model

$$Y = \langle X, \alpha\rangle_K + \varepsilon. \tag{6}$$

As mentioned above, we assume throughout, by simplicity, $\mathbb{E}(X(t)) = 0$. The general case can be treated by adding an intercept term β_0 on the right-hand side of (6) and using the estimator $\alpha \in H_K$ derived from (6) to estimate the additional parameter β_0 by a standard minimum squares procedure.

3 Some important particular cases

The above mentioned work by Berrendero et al. (2019) focusses in the model (4)–(6) from the point of view of its application to variable selection topics. In the present section, we go further in the study of such model by showing that other several commonly used models appear just as particular cases. In Sect. 4 we address the problem of estimating the "regression parameter" α and in Sect. 6 we carry out some numerical experiments.

3.1 Finite dimensional models: a setup for variable selection problems

When there are infinitely many regressors (which is the case in functional regression problems), several procedures of *variable selection* are available (see Berrendero et al. (2019) for details) for replacing the whole set of explanatory variables with a finite, carefully chosen, subset of these variables. The following proposition characterizes when it is possible to apply these procedures without any loss of information at all.

Proposition 2 *Under model (4)–(6), there exist* $X(t_1^*), \ldots, X(t_p^*) \in \{X(t) : t \in I\}$ *and* $\beta_1, \ldots, \beta_p \in \mathbb{R}$ *such that* $Y = \beta_1 X(t_1^*) + \cdots + \beta_p X(t_p^*) + \varepsilon$ *if and only if for all* $t \in I$, $\alpha(t) = \beta_1 K(t, t_1^*) + \cdots + \beta_p K(t, t_p^*)$.

Proof By Lemma 1(b), $Y = \beta_1 X(t_1^*) + \cdots + \beta_p X(t_p^*) + \varepsilon$ if and only if $\Psi_X^{-1}(\alpha) = \beta_1 \Psi_X^{-1}(K(\cdot, t_1^*)) + \cdots + \beta_p \Psi_X^{-1}(K(\cdot, t_p^*))$, what in turn happens if and only if $\alpha(t) = \beta_1 K(t, t_1^*) + \cdots + \beta_p K(t, t_p^*)$. □

3.2 The classical L^2-model

For $I = [0, 1]$ assume that $X = \{X(t) : t \in I\}$ is an L^2 random process and Y a response variable such that the RKHS linear model (2) or, equivalently (4) or (6), holds. To gain some insight, let us illustrate this with an example, beyond the finite-dimensional models considered in the previous subsection.

Example (Brownian regressors): When $X = \{X(t) : t \in [0, 1]\}$ is a standard Brownian Motion ($K(s, t) = \min\{s, t\}$) it can be shown

$$H_K = \{\alpha \in L^2[0, 1] : \alpha(0) = 0,\ \exists \alpha' \in L^2[0, 1] \text{ such that } \alpha(t) = \int_0^t \alpha'(s)ds\},$$

and $\langle \alpha_1, \alpha_2 \rangle_K = \langle \alpha_1', \alpha_2' \rangle_2$. It can also be proved that $\Psi_X^{-1}(\alpha)$ is given by Itô's stochastic integral, $\Psi_X^{-1}(\alpha) = \int_0^1 \alpha'(t) dX(t)$; for these results see Janson (1997), Example 8.19, p. 122. Thus, in this case, the linear model (2) or (4) reduces to $Y = \int_0^1 \alpha'(t) dX(t) + \varepsilon$.

Our goal in this subsection is to analyse under which conditions the RKHS model (2) or (4) entails the L^2-model (1). To do this, we need to recall some basic facts about the RKHS space associated with K. Let us denote by $\mathcal{K} : L^2[0, 1] \to L^2[0, 1]$ the integral operator defined by K, that is

$$\mathcal{K} f(t) = \int_0^1 K(t, s) f(s) ds.$$

Recall that we are assuming throughout that K is continuous. Under this condition, it is well-known that $\mathcal{K}$ is a compact and self-adjoint operator. The following proposition is a standard result in the RKHS theory. See, e.g., (Cucker and Zhou 2007, Corollary 4.13) for a proof and additional details.

Proposition 3 *Assume $I = [0, 1]$ and K is continuous. Let $\lambda_1 \geq \lambda_2 \geq \cdots$ be the non-null eigenvalues of $\mathcal{K}$ and let $e_1, e_2, \ldots$ be the corresponding unit eigenfunctions. Then, the RKHS corresponding to K is*

$$H_K = \left\{ f \in L^2[0,1] : \sum_{i=1}^{\infty} \frac{\langle f, e_i \rangle_2^2}{\lambda_i} < \infty \right\} = \mathcal{K}^{1/2}(L^2[0,1]), \tag{7}$$

endowed with the inner product $\langle f, g \rangle_K = \sum_{i=1}^{\infty} \langle f, e_i \rangle_2 \langle g, e_i \rangle_2 / \lambda_i$.

Thus, the membership to H_K can be understood as a "regularity property" established in terms of a very fast convergence to zero of the Fourier coefficients $\langle f, e_i \rangle_2$. This is just an alternative, equivalent formulation for the definition of H_K given at the beginning of Sect. 2.1. When the kernel K is continuous, both $\mathcal{K}$ and $\mathcal{K}^{1/2}$ can be considered as operators from $L^2[0, 1]$ to the space $\mathcal{C}[0, 1]$ of continuous functions. Expression (7) must be understood in this sense; see (Cucker and Zhou 2007, Th. 2.9 and Corollary 4.13).

Now, let us go back to the classical functional linear regression model (1). We will show that (1) appears as a particular case of our general model (4) if and only if the "slope function" α in (4) is regular enough to belong to the image subspace $\mathcal{K}(L^2[0, 1])$ which, by Proposition 3, is a subset of H_K. The formal statement is given in the following proposition.

Proposition 4 *If the L^2-based model (1) holds for some slope function $\beta \in L^2[0, 1]$, then it can be formulated as an RKHS-based model (4) whose slope function is $\alpha = \mathcal{K}\beta$. Conversely, if the RKHS model (4) holds for some $\alpha \in H_K$ and there exists $\beta \in L^2[0, 1]$ such that $\alpha = \mathcal{K}\beta$, then the model can be reformulated as an L^2-model such as (1) with slope function β.*

Proof The proof is essentially the same as that of Th. 1 in Berrendero et al. (2022), an analogous result in the framework of the functional logistic regression model.

□

Observe that the difference between (4) and (1) is not just a minor technical question. There are important values of the parameter α such that $\alpha \in H_K$ but $\alpha \neq \mathcal{K}\beta$ for all β. This is the case, for example, of finite linear combinations of the form $\beta_1 K(\cdot, t_1) + \cdots + \beta_p K(\cdot, t_p)$, which are important because they allow us to include finite dimensional regression models (also called impact point models in the literature on functional regression) as particular cases of the general model (see Proposition 2 above).

The procedures to fit model (1) very often involve to project $X = \{X(t) : t \in [0, 1]\}$ on a convenient subspace of $L^2[0, 1]$. More precisely, given $\{u_j : j = 1, 2, \ldots\}$, an arbitrary orthonormal basis of $L^2[0, 1]$, it is quite common to use as regressor variables the projections of $X = \{X(t) : t \in [0, 1]\}$ on the finite dimensional subspace spanned by the first p elements of the basis. This amounts to replace the whole trajectory with $\langle X, u_1 \rangle_2 u_1 + \cdots + \langle X, u_p \rangle_2 u_p$. This method will work fine whenever $\int_0^1 X(t)\beta(t)dt \approx \sum_{j=1}^{p} \beta_j \langle X, u_j \rangle_2$, where $\beta_j = \langle \beta, u_j \rangle_2$. More precisely, this projection-based model

would be as follows,

$$Y = \beta_1\langle X, u_1\rangle_2 + \cdots + \beta_p\langle X, u_p\rangle_2 + \varepsilon, \tag{8}$$

where $\beta_1, \ldots, \beta_p \in \mathbb{R}$ and $\varepsilon \in L_X^{\perp}$. A natural question to ask is when there is no loss in using the projection instead of the whole trajectory, and how is this situation characterized in terms of the parameter α in (4). The answer is given by Proposition 5 below. Its proof is completely similar to that of the analogous result Theorem 2 (b) in Berrendero et al. (2022).

Proposition 5 *If model (8) holds, then model (4) also holds. Conversely, if (4) holds and α belongs to the subspace spanned by $\{\mathcal{K}u_1, \ldots, \mathcal{K}u_p\}$ then (8) holds.*

An important particular case is functional principal component regression (FPCR). In FPCR, the orthonormal basis is given by $u_j = e_j$, the eigenfunctions of $\mathcal{K}$. Then, $\mathcal{K}e_j = \lambda_j e_j$, for $j = 1, \ldots, p$, and the condition in Proposition 5 simply states that α must belong to the subspace spanned by $\{e_1, \ldots, e_p\}$.

4 Estimation and prediction in the RKHS-model

We now focus on the main target of this work, that is, the RKHS-based functional model defined (with three alternative notations) in (2), (4) or (6). Our aim will be to estimate the functional parameter $\alpha \in H_K$ based on a sample of iid observations (X_i, Y_i), $i = 1, \ldots, n$ with $X_i = \{X_i(t) : t \in [0, 1]\}$. We will also address the prediction of the response variable based on the estimation of α.

4.1 Different approaches to the estimation of α

In short, our aim is to explore two "natural ways" of estimating α. We will first consider, in Sect. 4.2, an estimator based on regularization, denoted $\breve{\alpha}$. Though this method is conceptually meaningful and has some practical interest, it suffers from the serious limitation of assuming the knowledge of the covariance structure of the underlying process. Then, in Sect. 4.3 we will consider our main proposal, denoted $\hat{\alpha}_p$, which relies on the idea of approximating our model (4)–(6) by a sequence of finite-dimensional linear models. Its consistency is analysed in Theorem 2, the main result of this work.

4.2 An estimator based on regularization

The interpretation of α as a covariance given by Eq. (5) suggests a natural way to estimate it. We could just use the sample covariance function,

$$\tilde{\alpha}(t) = \frac{1}{n}\sum_{i=1}^{n} Y_i X_i(t).$$

Unfortunately, $\mathbb{P}(\tilde{\alpha} \in H_K) = 0$ (see Lukić and Beder (2001)) whereas we are assuming $\alpha \in H_K$. Also, the natural idea of projecting $\tilde{\alpha}$ on H_K (to obtain the element in H_K "closest" to $\tilde{\alpha}$) does not work since H_K is not in general closed and might even be dense in $L^2[0, 1]$; in fact, this is the case when all the eigenvalues of $\mathcal{K}$ are strictly positive, see (Cucker and Zhou 2007, Remark 4.9, p. 59). A common way of circumventing this problem is to get a sort of "quasi-projection", by penalizing the L^2-distance with a term accounting for the "roughness" of the quasi-projection, as measured by the norm in H_K. This idea is often referred to as Tikhonov regularization. It leads to the following estimator of α:

$$\check{\alpha} := \arg\min_{f \in H_K} \|\tilde{\alpha} - f\|_2^2 + \gamma_n \|f\|_K^2,$$

where $\gamma_n > 0$ is a sequence of regularization parameters depending on the sample size. It turns out that $\check{\alpha}$ has the following explicit expression:

$$\check{\alpha} = (\mathcal{K} + \gamma_n I)^{-1} \mathcal{K} \tilde{\alpha}, \tag{9}$$

where $\mathcal{K}$ is the integral operator defined by the kernel K; see (Cucker and Zhou 2007, p. 139).

Note that (9) relies on the previous knowledge of the true covariance operator $\mathcal{K}$. This could be the case in some particular models (e.g., Lindquist and McKeague (2009)) but, in general, such assumption is somewhat restrictive. In any case, the consistency in the RKHS norm of this estimator is established in the following result.

Proposition 6 *Let $\gamma_n \to 0$ such that $n\gamma_n^2 \to \infty$, then $\|\check{\alpha} - \alpha\|_K^2 \to 0$ in probability.*

Proof To prove that $\|\check{\alpha} - \alpha\|_K \to 0$ in probability, observe that since $\alpha \in H_K$, for all $\epsilon > 0$ there exists $N = N(\epsilon)$ such that

$$\sum_{j=N+1}^{\infty} \frac{1}{\lambda_j} \langle \alpha, e_j \rangle_2^2 < \epsilon. \tag{10}$$

For this value of N we have

$$\begin{aligned} \|\check{\alpha} - \alpha\|_K \leq & \Big\| \sum_{j=1}^{N} \Big(\frac{\lambda_j}{\gamma_n + \lambda_j} - 1 \Big) \langle \tilde{\alpha}, e_j \rangle_2 e_j \Big\|_K + \Big\| \sum_{j=1}^{N} \langle \tilde{\alpha} - \alpha, e_j \rangle_2 e_j \Big\|_K \\ & + \Big\| \sum_{j=N+1}^{\infty} \frac{\lambda_j}{\gamma_n + \lambda_j} \langle \tilde{\alpha} - \alpha, e_j \rangle_2 e_j \Big\|_K + \Big\| \sum_{j=N+1}^{\infty} \Big(\frac{\lambda_j}{\gamma_n + \lambda_j} - 1 \Big) \langle \alpha, e_j \rangle_2 e_j \Big\|_K. \end{aligned} \tag{11}$$

We will look at each term in the expression above. For the first one, we have:

$$\Big\| \sum_{j=1}^{N} \Big(\frac{\lambda_j}{\gamma_n + \lambda_j} - 1 \Big) \langle \tilde{\alpha}, e_j \rangle_2 e_j \Big\|_K \leq \Big\| \sum_{j=1}^{N} \Big(\frac{\lambda_j}{\gamma_n + \lambda_j} - 1 \Big) \langle \tilde{\alpha} - \alpha, e_j \rangle_2 e_j \Big\|_K +$$

$$\Big\| \sum_{j=1}^{N} \Big(\frac{\lambda_j}{\gamma_n + \lambda_j} - 1 \Big) \langle \alpha, e_j \rangle_2 e_j \Big\|_K .$$

Now, observe that from Mourier's SLLN (see e.g. Theorem 4.5.2 in Laha and Rohatgi (1979) p. 452) $\|\tilde{\alpha} - \alpha\|_2 \to 0$ almost surely (a.s.), and let us define

$$C_{N,n} := \max_{j=1,\ldots,N} \Big(\frac{\lambda_j}{\gamma_n + \lambda_j} - 1 \Big)^2 \|e_j\|_K^2 \leq \Big(\frac{\lambda_1}{\gamma_n + \lambda_1} - 1 \Big)^2 \frac{1}{\lambda_N} \to 0, \quad \text{as } n \to \infty.$$

Then, for large enough n, with probability one,

$$\Big\| \sum_{j=1}^{N} \Big(\frac{\lambda_j}{\gamma_n + \lambda_j} - 1 \Big) \langle \tilde{\alpha} - \alpha, e_j \rangle_2 e_j \Big\|_K^2 \leq N C_{N,n} \|\tilde{\alpha} - \alpha\|_2^2 < \epsilon.$$

We have used Cauchy–Schwarz inequality in the first inequality above. Similarly, we also have, for large enough n,

$$\Big\| \sum_{j=1}^{N} \Big(\frac{\lambda_j}{\gamma_n + \lambda_j} - 1 \Big) \langle \alpha, e_j \rangle_2 e_j \Big\|_K \leq N C_{N,n} \|\alpha\|_2^2 < \epsilon,$$

The second term in (11) satisfies $\| \sum_{j=1}^{N} \langle \tilde{\alpha} - \alpha, e_j \rangle_2 e_j \|_K^2 \leq N \lambda_N^{-1} \|\tilde{\alpha} - \alpha\|_2^2 < \epsilon$, for large enough n, with probability one. For the third term in (11), let $\sum_{j=1}^{\infty} \lambda_j := C < \infty$. Then,

$$\begin{aligned} \Big\| \sum_{j=N+1}^{\infty} \frac{\lambda_j}{\gamma_n + \lambda_j} \langle \tilde{\alpha} - \alpha, e_j \rangle_2 e_j \Big\|_K^2 &= \sum_{j=N+1}^{\infty} \frac{\lambda_j}{(\gamma_n + \lambda_j)^2} \langle \tilde{\alpha} - \alpha, e_j \rangle_2^2 \\ &\leq \frac{\|\tilde{\alpha} - \alpha\|_2^2}{\gamma_n^2} \sum_{j=N+1}^{\infty} \lambda_j \leq C \frac{n \|\tilde{\alpha} - \alpha\|_2^2}{n \gamma_n^2} \to 0, \end{aligned}$$

in probability, since we are assuming $n\gamma_n^2 \to \infty$ and $n\|\tilde{\alpha} - \alpha\|_2^2$ is bounded in probability by the Central Limit Theorem. Finally, the fourth term in (11) is also bounded by ϵ using (10):

$$\begin{aligned} \Big\| \sum_{j=N+1}^{\infty} \Big(\frac{\lambda_j}{\gamma_n + \lambda_j} - 1 \Big) \langle \alpha, e_j \rangle_2 e_j \Big\|_K^2 &= \sum_{j=N+1}^{\infty} \frac{1}{\lambda_j} \Big(\frac{\gamma_n}{\gamma_n + \lambda_j} \Big)^2 \langle \alpha, e_j \rangle_2^2 \leq \\ \sum_{j=N+1}^{\infty} \frac{1}{\lambda_j} \langle \alpha, e_j \rangle_2^2 &< \epsilon. \end{aligned}$$

□

4.3 RKHS-consistent estimation and prediction

In the previous subsection we have considered a penalized estimator $\check{\alpha}$ for the slope parameter α and we have established its RKHS-consistency. In this subsection we address the RKHS-based estimation of α using a different strategy: we will use a discrete approximation of the linear model itself, taking advantage of the RKHS structure. As it turns out, predictions based on such estimator can be made in the natural way with no need of knowing the covariance function, thanks to Loève's isometry. The idea is to approximate the RKHS model (4) by a sequence of finite dimensional models of type of those considered in Proposition 2, based on p_n-dimensional marginals $(X(t_{1,p}), \ldots, X(t_{p,p}))$, obtained by evaluating the process $X = \{X(t) : t \in [0,1]\}$ at the grid points $T_p = \{t_{j,p}\}$, where $p = p_n$. The corresponding sequence of least squares estimators of the slope function $\hat{\alpha}_p$ will hopefully provide a consistent sequence of estimators of the true slope function α in (4). This idea is next formalized.

We will use the following lemma (which follows from Theorem 6E in Parzen (1959)), that states that the function α can be approximated (in the RKHS norm) by a finite linear combination of the kernel function K, evaluated at points of a partition of $[0,1]$.

Lemma 2 *Let $\alpha \in H_K$. Let us consider $T_p = \{t_{j,p} : j = 1, \ldots, p\}$ where $0 \leq t_{1,p} \leq \cdots \leq t_{p,p} \leq 1$, is an increasing sequence of partitions of $[0,1]$, i.e, $T_p \subset T_{p+1}$, such that $\overline{\cup_p T_p} = [0,1]$. Then, there exist $\beta_{1,p}, \ldots, \beta_{p,p}$ such that, $\left\|\alpha(\cdot) - \sum_{j=1}^{p} \beta_{j,p} K(t_{j,p}, \cdot)\right\|_K^2 \to 0$, as $p \to \infty$.*

Now our estimator is defined by the ordinary least squares estimator of the coefficients $\beta_{1,p}, \ldots, \beta_{p,p}$. To be more precise, let us denote

$$\alpha_p(\cdot) = \sum_{j=1}^{p} \beta_{j,p} K(t_{j,p}, \cdot) \text{ and } \hat{\alpha}_p(\cdot) = \sum_{j=1}^{p} \hat{\beta}_{j,p} K(t_{j,p}, \cdot), \tag{12}$$

where $t_{1,p}, \ldots, t_{p,p}$ are chosen as indicated in Lemma 2 and, for $j = 1, \ldots, p$, $\hat{\beta}_{j,p}$ are the ordinary least squares estimators (based on a sample of size n) of the regression coefficients in the p-dimensional linear model

$$Y_i = \sum_{j=1}^{p} \beta_{j,p} X(t_{j,p}) + e_{i,p} = \langle \alpha_p, X \rangle_K + e_{i,p}, \quad i = 1, \ldots, n. \tag{13}$$

To prove the almost sure consistency of the estimator we will need to impose a condition of sub-Gaussianity. Let us recall that a random variable Y with $\mathbb{E}(Y) = 0$ is said to be sub-Gaussian with (positive) proxy constant σ^2 (we will denote $Y \in SG(\sigma^2)$) if the moment generating function of Y satisfies $\mathbb{E}(\exp(sY)) \leq \exp(\sigma^2 s^2/2)$, for all $s \in \mathbb{R}$. It can be seen that the tails of a random variable $Y \in SG(\sigma^2)$ are lighter than or equal to those of a Gaussian distribution with variance σ^2, i.e. $\mathbb{P}(|Y| > t) \leq 2\exp(-t^2/(2\sigma^2))$ for all $t > 0$. A p-dimensional random vector $\mathbf{X}$ is said to be sub-Gaussian with proxy constant σ^2 if $\mathbf{X}'v \in SG(\sigma^2)$ for all $v \in \mathbb{R}^d$ with $\|v\| = 1$.

Observe that if $\mathbf{X}$ is a p-dimensional random vector $\mathbf{X} = (X_1, \ldots, X_p)$ and the X_i are independent with $X_i \in SG(\sigma^2)$ and sub-Gaussian then $\mathbf{X}$ is sub-Gaussian with proxy constant σ^2 as well. See (Rigollet and Hütter 2017, Ch. 1) for details.

Theorem 2 *Assume the RKHS-based linear model $Y_i = \langle X, \alpha \rangle_K + \epsilon_i$ for $i = 1, \ldots, n$, as defined in (2), (4) or (6). Let us consider a sequence of approximating p-dimensional models (with $p = p_n$) as defined in (13). Assume that*

(i) The error variables $e_{i,p}$ in the p-dimensional models are iid and sub-Gaussian, $SG(\sigma_p^2)$ with $\sigma_p^2 \geq C_0$ for all p and some $C_0 > 0$.
(ii) The random variable $\sup_{t\in[0,1]} X(t)$ has sub-exponential tails, that is $\mathbb{P}(\sup_{t\in[0,1]} X(t) > s) \leq C_1 \exp(-C_2 s^2)$ for some constants $C_1, C_2 > 0$ and for all $s > 0$.
(iii) We have $p \to \infty$, as $n \to \infty$, in such a way that there exists $C_3 > 0$ such that $n(\gamma_{p,p})^2/(p^2 \log^3 n) \to C_3$, where $\gamma_{p,p}$ is the smallest eigenvalue of the covariance matrix, K_{T_p}, of $(X(t_{1,p}), \ldots, X(t_{p,p}))$.

Then, $v_n \|\hat{\alpha}_p - \alpha_p\|_K^2 \to 0$ a.s., for all $v_n \to \infty$ such that $n\gamma_{p,p}/(p^2 v_n \log n) \to C_4 > 0$. In addition, as a consequence of Lemma 2, $\|\hat{\alpha}_p - \alpha\|_K^2 = \max\{v_n^{-1}, \mathcal{O}(\|\alpha - \alpha_p\|_K^2)\}$ a.s.

The proof of this theorem is deferred to the appendix as it is a bit more technical than those of the previous results in the paper. Let us now discuss the real extent of this result by analysing how restrictive are the required assumptions.

4.4 Some remarks on the assumptions of Theorem 2

Clearly assumption (i) holds if the errors e_i are Gaussian, which is a common assumption in regression theory. But it is also satisfied by many other usual centred distributions such as those of compact support or finite mixtures of centred Gaussian distributions.

Regarding assumption (ii), it is fulfilled, for example, when the process $X = \{X(t) : t \in [0, 1]\}$ is Gaussian. To see this, define $Y = \sup_{s\in[0,1]} |X(s)|$ and note that, for any $t > 0$, $\mathbb{P}(\sup_{s\in[0,1]} X(s) > t) \leq \mathbb{P}(Y > t)$. Now, according to Theorem 5 in Landau and Shepp (1970), there is some $\epsilon > 0$ such that $\mathbb{E}(e^{\epsilon Y^2}) < \infty$. But this entails

$$\mathbb{P}\left(\sup_{s\in[0,1]} X(s) > t\right) \leq \mathbb{P}(Y > t) = \mathbb{P}(e^{\epsilon Y^2} > e^{\epsilon t^2}) \leq \frac{\mathbb{E}(e^{\epsilon Y^2})}{e^{\epsilon t^2}}.$$

Therefore, condition (ii) is fulfilled with $C_1 = \mathbb{E}(e^{\epsilon Y^2})$ and $C_2 = \epsilon$.

Finally, hypothesis (iii) in Theorem 2 is satisfied for the case of processes with stationary and independent increments and equispaced impact points $t_{j,p}$, as stated in the following proposition.

Proposition 7 *Let $\{W(t) : t \in [0, 1]\}$ be a stochastic process with stationary and independent increments, such that $\mathbb{E}(W^2(t)) < \infty$ and $\mathbb{E}(W(t)) = 0$ for all $t \in [0, 1]$.*

Then for all $\delta > 0$, $p^{1+\delta}\gamma_{p,p} \to \infty$, $\gamma_{p,p}$ *being the smallest eigenvalue of* K_{T_p}, *the covariance matrix of the random vector* $(W(1/p), \ldots, W(1))$.

Proof Let us denote, with some notational abuse, $W = (W(1/p), \ldots, W(1))$, $t_i = i/p$, and $v = (v_1, \ldots, v_p)$. Let us introduce the $p \times p$ matrix A, such that WA is the $1 \times p$ row vector $(W(1/p), W(2/p) - W(1/p), \ldots, W(1) - W(1 - 1/p))$, that is $A = (a_{ij})$ where $a_{ii} = 1$ for $i = 1, \ldots, p$, $a_{i-1,i} = -1$ for $i = 2, \ldots, p$ and $a_{ij} = 0$ otherwise. The coordinates of WA are independent random variables. Then, for all v,

$$\mathbb{E}|WAv|^2 = v'A'\mathbb{E}(W'W)Av = v'A'K_{T_p}Av = \|Av\|^2 \frac{v'A'}{\|v'A'\|} K_{T_p} \frac{Av}{\|Av\|}.$$

Since A is invertible there exists v with $\|v\| = 1$ such that $w := Av$ fulfils $K_{T_p}w = \gamma_{p,p}w$. Then

$$\min_{v:\|v\|=1} \|Av\|^2 \frac{v'A'}{\|v'A'\|} K_{T_p} \frac{Av}{\|Av\|} \le \|A\|_{op}^2 \min_{v:\|v\|=1} \frac{v'A'}{\|v'A'\|} K_{T_p} \frac{Av}{\|Av\|} \le$$
$$\|A\|_{op}^2 \frac{w'}{\|w\|} K_{T_p} \frac{w}{\|w\|} = \|A\|_{op}^2 \gamma_{p,p}.$$

From where it follows that $\gamma_{p,p} \ge \|A\|_{op}^{-2} \min_{v:\|v\|=1} \mathbb{E}|WAv|^2$. Since $\mathrm{Var}(W_{t+s} - W_t) = \sigma^2 s$ for all $0 \le t, s \le 1$, such that $s + t \le 1$, and for some $\sigma > 0$, then, if $\|v\| = 1$, it follows that

$$\mathbb{E}|WAv|^2 = \sum_{j=0}^{p-1} \mathbb{E}\big(W((j+1)/p) - W(j/p)\big)^2 v_{j+1}^2 = \frac{\sigma^2}{p}$$

Then $\gamma_{p,p} \ge \sigma^2/(p\|A\|_{op}^2)$. Lastly, $\|A\|_{op}^2 = 1/p + (4/p)(p-1)$, (because the maximum of $\|Av\|$ subject to $\|v\| = 1$ is attained at $v_i = (-1)^{i+1}/\sqrt{p}$). □

Remark 2 The class of processes with stationary independent increments includes many counting processes and the Brownian Motion. Putting together the condition $n(\gamma_{p,p})^2/(p^2 \log^3 n) \to C_3$ for some $C_3 > 0$, imposed in Theorem 2 and the conclusion obtained in Proposition 7, it turns out that a choice of type $p = (n/\log^3 n)^{1/5}$ for p would be sufficient to ensure the applicability of Theorem 2.

To conclude this section, we provide an interesting interpretation of Theorem 2 in terms of prediction.

Theorem 3 *Under the conditions of Theorem 2, we have that the general regression function of* Y *with respect to* X, $m(X) = \mathbb{E}(Y|X)$, *can be approximated in* $L^2(\Omega)$ *from the prediction functions derived from the finite-dimensional models, that is as* $p = p_n \to \infty$.

$$\|m(X) - \langle \hat{\alpha}_p, X\rangle_K\|_{L^2(\Omega)} = \|m(X) - \sum_{j=1}^{p} \hat{\beta}_{j,p} X(t_{j,p})\|_{L^2(\Omega)} \to 0, \ a.s. \quad (14)$$

Proof As a consequence of the assumed RKHS model, we have $m(X) = \mathbb{E}(Y|X) = \Psi_X^{-1}(\alpha) := \langle \alpha, X \rangle_K$. Now, the result follows from Loève's isometry, since

$$\|\hat{\alpha}_p - \alpha\|_K = \|\langle \hat{\alpha}_p, X \rangle_K - \langle \alpha, X \rangle_K\|_{L^2(\Omega)},$$

and $\|\hat{\alpha}_p - \alpha\|_K \to 0$, a.s., as a consequence of Theorem 2. □

In order to properly interpret this result, let us recall that the conditional expectation $g(X) = \mathbb{E}(Y|X)$ is known to be the projection of Y on the Hilbert subspace of $L^2(\Omega)$ of random variables Z of the form $Z = h(X)$ with $\mathbb{E}(Z^2) < \infty$; see e.g. (Laha and Rohatgi 1979, p. 382). In other words, $\mathbb{E}(Y|X)$ is the minimizer in Z of $\|Y - Z\|^2_{L^2(\Omega)} = \mathbb{E}(Y - Z)^2$ when Z is in the space of all rv's of type $Z = h(X)$ with $\mathbb{E}(Z^2) < \infty$. In this sense $m(X) = \mathbb{E}(Y|X)$ could be considered in a very precise way as the "best possible prediction (in the sense of quadratic mean error) of Y in terms of X". In Theorem 3 it is shown that we are able to asymptotically approach such $m(X)$. Note that these finite-dimensional predictions considered here do not require the knowledge of the covariance function K.

5 The coefficient of determination in the functional case

The coefficient of determination, often denoted R^2, is commonly used in regression analysis as a measure of the goodness of fit for the regression model under consideration. In this section we will define and motivate, in population terms, the notion of coefficient of determination for our RKHS-based regression model (2)–(4)–(6). We will show as well how this coefficient can be consistently approximated from a natural statistic, analogous to that used in the standard multivariate cases. Let us start by briefly recalling some essentials about the coefficient of determination in more classical situations.

5.1 The linear finite-dimensional case

In multivariate linear regression, the X_i are random vectors in $\mathbb{R}^d$. If $\hat{Y}_i$ stands for the prediction of Y_i obtained from the usual linear model $Y_i = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_p X_{ip}$, the coefficient of determination is given by

$$R^2 = \frac{\sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2}{\sum_{i=1}^n (Y_i - \bar{Y})^2}, \tag{15}$$

which can be interpreted as the portion of total variability explained by the model.

A thorough study of R^2 can be found in (Rencher and Schaalje 2008, Sect. 7.7, 10.4, 10.5). In summary, it can be seen that R^2 is the square (sample) linear correlation coefficient between the observations Y_i and the predictions $\hat{Y}_i$ (obtained with the standard least squares estimations of the parameters). It is as well the maximum

(sample) square linear correlation coefficient that can be obtained between the Y_i and all linear functions of the coordinates of the X_i.

This suggests a population version of R^2, not depending on the sample data, but on the underlying random variable (X, Y) with values in $\mathbb{R}^d \times \mathbb{R}$. It could be defined as the maximum linear correlation coefficient (denoted "Corr") between the response variable Y and linear functionals of X, of type $\beta' X = \beta_1 X_1 + \cdots + \beta_d X_d$. Thus

$$\rho^2_{\mathcal{L},Y|X} = \max_{\beta \in \mathbb{R}^d} \mathrm{Corr}^2(Y, \beta' X), \tag{16}$$

where the subscript $\mathcal{L}$ emphasizes that we are considering linear functions of X to predict Y.

5.2 The fully nonparametric case

By analogy with the linear multivariate case, one could consider the approximation of the scalar random variable Y in terms of a general measurable function of X. It is well-known that, if we assume $\mathbb{E}(Y^2) < \infty$, the function $m(x)$ which minimizes the square prediction error $\mathbb{E}[(Y - g(X))^2]$ within the class $\mathcal{G}$ of real measurable functions such that $\mathbb{E}(g^2(X)) < \infty$ is $m(x) = \mathbb{E}(Y|X = x)$. Thus, by analogy with (16), one might define

$$\rho^2_{\mathcal{G},Y|X} = \max_{g \in \mathcal{G}} \mathrm{Corr}^2(Y, g(X)) \tag{17}$$

In Doksum and Samarov (1995) the coefficient of determination is considered in this nonparametric setting under the name of Pearson's correlation ratio, and it is defined as the following quotient of variances

$$\eta^2 = \frac{\mathrm{Var}(m(X))}{\mathrm{Var}(Y)}. \tag{18}$$

In view of the ANOVA identity $\mathrm{Var}(Y) = \mathrm{Var}(m(X)) + \mathbb{E}(\mathrm{Var}(Y|X)))$, η^2 is nothing but the proportion of total variability explained by the regression model $m(X)$. But, actually, this interpretation is compatible with that behind definition (17), since it is easy to see that $\eta^2 = \rho^2_{\mathcal{G},y|X}$. Indeed, we have, for any $g \in \mathcal{G}$,

$$\begin{aligned}\mathrm{Corr}^2(Y, g(X)) &= \frac{\langle Y - \mathbb{E}(Y), g(X) - \mathbb{E}(g(X))\rangle_{L^2(\Omega)}}{\mathrm{Var}(Y)\mathrm{Var}(g(X))} \\ &\overset{(*)}{=} \frac{\langle m(X) - \mathbb{E}(m(X)), g(X) - \mathbb{E}(g(X))\rangle_{L^2(\Omega)}}{\mathrm{Var}(Y)\mathrm{Var}(g(X))} \overset{(**)}{\leq} \frac{\mathrm{Var}(m(X))}{\mathrm{Var}(Y)}.\end{aligned}$$

Note that (*) holds from the fact that $\mathbb{E}(Y) = \mathbb{E}(m(X))$ and $Y - m(X)$ is (by the projection properties of the conditional expectation) orthogonal to all functions in $\mathcal{G}$ and (**) holds from the Cauchy–Schwartz inequality. This shows $\eta^2 \geq \rho^2_{\mathcal{G},y|X}$. The converse inequality readily follows from the fact that $m(\cdot) \in \mathcal{G}$.

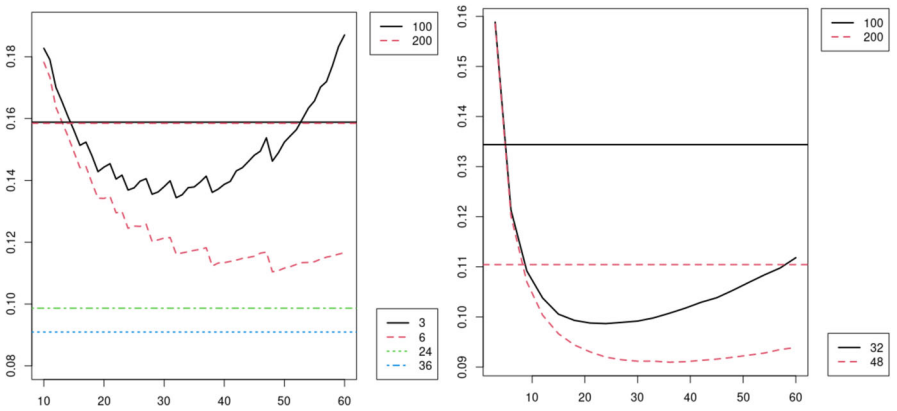

Fig. 1 Left panel: prediction errors under **Scenario 1** for the RKHS-method, plotted as a function of the number of explatanory variables. The horizontal lines correspond to the prediction errors for the principal components (PC) regression method for different choices of the number of components. Right panel: the dual graph for the PC method as a function of the number of components versus the RKHS-based method for a few choices of the number of components

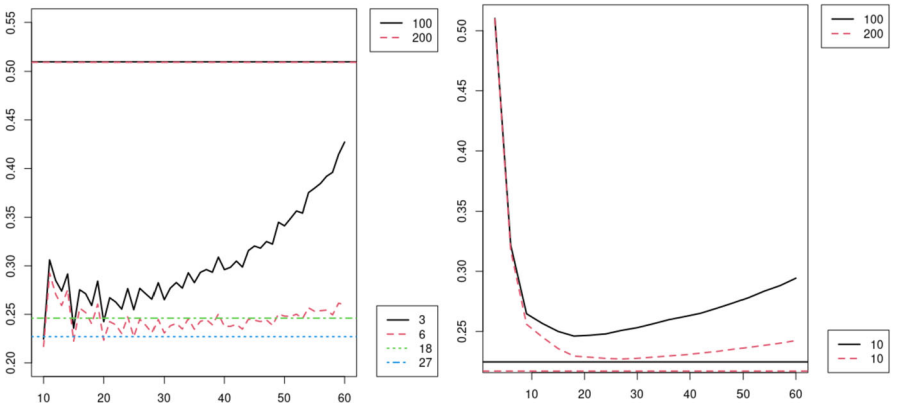

Fig. 2 Left panel: prediction errors under **Scenario 2a** for the RKHS-method, plotted as a function of the number of explatanory variables. The horizontal lines correspond to the prediction errors for the principal components (PC) regression method for different choices of the number of components. Right panel: the dual graph for the PC method as a function of the number of components versus the RKHS-based method for a few choices of the number of components

5.3 The coefficient of determination in the context of our RKHS-based linear functional model

Now, coming back to the linear functional model (2)–(4)–(6), as a consequence of the above discussion, we define the coefficient of determination by

$$\rho^2_{Y|X} = \max_{\alpha \in \mathcal{H}_K} \mathrm{Corr}^2(Y, \langle X, \alpha \rangle_K) = \max_{U \in L_X} \mathrm{Corr}^2(Y, U). \tag{19}$$

The following proposition, gives a simple expression for $\rho^2_{Y|X}$ in terms of the model (6),

Proposition 8 *Let us assume the validity of the RKHS-based regression model (2)–(4)–(6), i.e.* $Y = \Psi_X^{-1}(\alpha) + \varepsilon$, *with* $\mathbb{E}(\varepsilon|X) = 0$ *and* $Var(\varepsilon|X) = \sigma^2$. *Then, the coefficient of determination (19) can be expressed as*

$$\rho^2_{Y|X} = \frac{\|\alpha\|^2_K}{\sigma^2 + \|\alpha\|^2_K} = \frac{\mathrm{Var}[\mathbb{E}(Y|X)]}{\mathrm{Var}(Y)}. \tag{20}$$

Proof For any $U \in L_X$,

$$\mathrm{Corr}^2(Y, U) = \frac{\langle Y, U\rangle^2_{L^2(\Omega)}}{\|Y\|^2_{L^2(\Omega)}\|U\|^2_{L^2(\Omega)}} = \frac{\langle \Psi_X^{-1}(\alpha), U\rangle^2_{L^2(\Omega)}}{\|Y\|^2_{L^2(\Omega)}\|U\|^2_{L^2(\Omega)}}. \tag{21}$$

Now, observe that the RKHS-model entails $\mathbb{E}(Y|X) = \Psi_X^{-1}(\alpha)$ and

$$\|Y\|^2_{L^2(\Omega)} = \mathrm{Var}(Y) = \mathrm{Var}[\mathbb{E}(Y|X)] + \mathbb{E}[\mathrm{Var}(Y|X)] = \|\alpha\|^2_K + \sigma^2. \tag{22}$$

Indeed, $\mathrm{Var}[\mathbb{E}(Y|X)] = \mathrm{Var}[\Psi_X^{-1}(\alpha)] = \mathbb{E}(\Psi_X^{-1}(\alpha)^2)$ which, from Loève's isometry, equals $\|\alpha\|^2_K$ and $\mathbb{E}[\mathrm{Var}(Y|X)] = \mathrm{Var}(\varepsilon|X) = \sigma^2$.

Using again Loève's isometry in (21), we have

$$\mathrm{Corr}^2(Y, U) = \frac{\langle \alpha, \alpha_U\rangle^2_K}{(\|\alpha\|^2_K + \sigma^2)\|\alpha_U\|^2_K},$$

where α_U denotes the image of U in H_K by Loève's isometry, that is, $\alpha_U = \Psi_X(U)$. Finally, from Cauchy–Schwartz inequality

$$\mathrm{Corr}^2(Y, U) \leq \frac{\|\alpha\|^2_K}{\sigma^2 + \|\alpha\|^2_K},$$

and the bound in the right-hand side is attained when U is such that $\alpha_U = \alpha$. □

Note that, in view of expression (20), $\rho^2_{Y|X}$ can be interpreted again as the proportion of variance explained by the RKHS linear model (4).

We now address the problem of approximating the (population) functional coefficient of determination $\rho^2_{Y|X}$ with the corresponding quantities (that we will denote ρ^2_T for simplicity) in the approximating finite-dimensional models based on the observations $X(t_1), \ldots, X(t_p)$ on a grid $T_p := T = \{t_1, \ldots, t_p\} \subset [0, 1]$, with $t_1 < \ldots < t_p$. Denote $\alpha_T = (\alpha(t_1), \ldots, \alpha(t_p))'$ and $K_T \equiv K(t_i, t_j)$ the covariance matrix of $(X(t_1), \ldots, X(t_p))$. Let $L^2_T \equiv \mathrm{sp}\{X(t_1), \ldots, X(t_p)\}$, the space of all possible linear combinations of $X(t_1), \ldots, X(t_p)$.

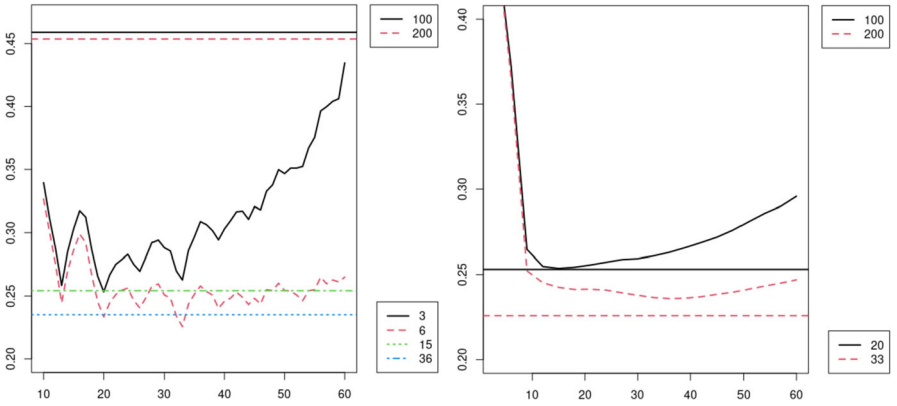

Fig. 3 Left panel: prediction errors under **Scenario 2b** for the RKHS-method, plotted as a function of the number of explanatory variables. The horizontal lines correspond to the prediction errors for the principal components (PC) regression method for different choices of the number of components. Right panel: the dual graph for the PC method as a function of the number of components versus the RKHS-based method for a few choices of the number of components

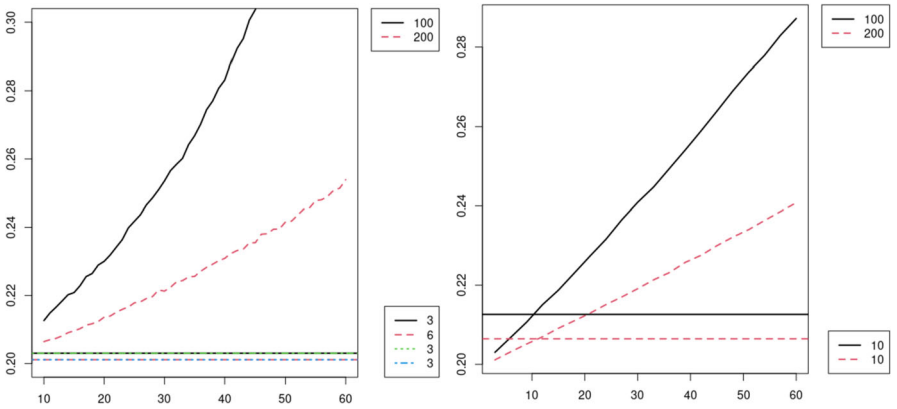

Fig. 4 Left panel: prediction errors under **Scenario 3** for the RKHS-method, plotted as a function of the number of explanatory variables. The horizontal lines correspond to the prediction errors for the principal components (PC) regression method for different choices of the number of components. Right panel: the dual graph for the PC method as a function of the number of components versus the RKHS-based method for a few choices of the number of components

In accordance to the previous discussion, the finite dimensional coefficient of determination, can be defined as

$$\rho_T^2 \equiv \max_{U \in L_T^2} \mathrm{Corr}^2(Y, U). \tag{23}$$

Then, the random variable U attaining this maximum is the best possible predictor of Y using a linear combination of the variables $X(t_1), \ldots, X(t_p)$. The following proposition establishes the convergence of ρ_T^2 to $\rho_{Y|X}^2$, as $p = p_n \to \infty$.

Proposition 9 *Under the indicated RKHS-linear model (2), assume that the covariance function K is continuous in $[0,1]^2$ and the matrix K_T defined above is invertible. Then,*

(a)

$$\rho_T^2 = \frac{\alpha_T' K_T^{-1} \alpha_T}{\sigma^2 + \|\alpha\|_K^2}, \tag{24}$$

where $\alpha_T = (\alpha(t_1), \ldots, \alpha(t_p))'$ and $\alpha(t) = \mathbb{E}(YX(t))$.

(b) Assume further that the sequence T_p is increasing ($T_p \subset T_{p+1}$) and the set $\bigcup_{p=1}^{\infty} T_p$ is dense in $[0,1]$. Then

$$\lim_{p\to\infty} \rho_T^2 = \rho_{Y|X}^2. \tag{25}$$

Proof (a) Recall that $\rho_T^2 \equiv \max_{U\in L_T^2} \mathrm{Corr}^2(Y,U)$. Since $U \in L_T^2$, there exists $\beta = (\beta_1, \ldots, \beta_p)$ such that $U = \sum_{j=1}^p \beta_j X(t_j)$. Then, from $\alpha(t) = \mathbb{E}(YX(t))$ and (22),

$$\mathrm{Corr}^2(Y,U) = \frac{\langle Y,U\rangle_{L^2(\Omega)}^2}{\|Y\|_{L^2(\Omega)}^2 \|U\|_{L^2(\Omega)}^2} = \frac{(\beta'\alpha_T)^2}{(\sigma^2 + \|\alpha\|_K^2)\beta' K_T \beta}.$$

We have to maximize on β the quotient $(\beta'\alpha_T)^2/(\beta' K_T \beta)$. Using (Mardia et al. 2021, Cor. A.9.2.2, p. 480) we get $\max_\beta (\beta'\alpha_T)^2/(\beta' K_T \beta) = \alpha_T' K_T^{-1} \alpha_T$, and the result follows.

(b) If $\alpha \in H_K$ is the maximizer of the expression (19) defining $\rho_{Y|X}^2$. Since $K_T := K(t_i,t_j)_{t_i,t_j\in T}$ is an invertible matrix, we have that, for each $p = p_n$ and t_j in the grid T, there exist constants $\beta_k^p := \beta_k$ such that

$$\alpha(t_j) = \sum_{k=1}^p \beta_k K(t_j, t_k),$$

so that $\alpha_T = K_T \beta_T$ and $\beta_T = K_T^{-1}\alpha_T$.

Now, the result is a consequence of (Parzen 1959, Th. 6E). Indeed, using expression (6.26) in that paper (for the particular case $f = g = \alpha$), we have

$$\lim_{p\to\infty} \sum_{k=1}^p \beta_k \alpha(t_k) = \langle \alpha, \alpha\rangle_K, \tag{26}$$

(note, that in Parzen's notation $\langle \alpha,\alpha\rangle_p := \sum_{k=1}^p \beta_k \alpha(t_k)$). Thus, from (26),

$$\sum_{k=1}^p \beta_k \alpha(t_k) = \beta_T' \alpha_T = \beta_T' K_T \beta_T = \alpha_T' K_T^{-1} K_T K_T^{-1} \alpha_T = \alpha_T' K_T^{-1} \alpha_T \to \|\alpha\|_K^2,$$

which proves (25).

□

We now consider the problem of estimating $\rho^2_{Y|X}$ from the sample data $(X_i(\cdot), Y_i)$, $i = 1, \ldots, n$. We will show that the sample versions of the coefficients of determination (23) corresponding to the "approximating models" provide a consistent estimation of $\rho^2_{Y|X}$.

First, recall that under the assumed model (4), $Y_i = \langle X_i, \alpha\rangle_K + \varepsilon_i$, we have $\mathrm{Var}(Y_i) = \|\alpha\|^2_K + \sigma^2$. Then, a natural estimator of $\rho^2_{Y|X}$ would be

$$R^2_p = \frac{\sum_{i=1}^n \hat{Y}_i^2}{\sum_{i=1}^n Y_i^2}, \tag{27}$$

where $\hat{Y}_i := \hat{Y}_{ip}$ is the standard prediction of Y_i obtained form the finite-dimensional approximating linear model based on $X(t_1), \ldots, X(t_p)$. We next show the almost sure consistency of this estimator.

Theorem 4 *Under the hypotheses of Theorem 2,*

$$R^2_p \to \frac{\|\alpha\|^2_K}{\sigma^2 + \|\alpha\|^2_K} \quad a.s.,$$

as $p = p_n \to \infty$.

Proof From (22), $\mathbb{E}(Y_i^2) = \sigma^2 + \|\alpha\|^2_K$. So, from the Strong Law of Large Numbers,

$$\frac{\sum_{i=1}^n Y_i^2}{n} \to \sigma^2 + \|\alpha\|^2_K, \text{ a.s.} \tag{28}$$

Let us now prove that $(1/n)\sum_{i=1}^n \hat{Y}_i^2 \to \|\alpha\|^2_K$ a.s. Indeed, following the notation in the proof of Theorem 2 (see the Appendix below),

$$\begin{aligned}(1/n)\sum_{i=1}^n \hat{Y}_i^2 &= \hat{\beta}_p'\Big(\frac{1}{n}\mathcal{X}_p'\mathcal{X}_p - K_{T_p}\Big)\hat{\beta}_p + \hat{\beta}_p' K_{T_p}\hat{\beta}_p \\ &= \hat{\beta}_p'\Big(\frac{1}{n}\mathcal{X}_p'\mathcal{X}_p - K_{T_p}\Big)\hat{\beta}_p + \|\hat{\alpha}_p\|^2_K\end{aligned}$$

From Theorem 3, $\|\hat{\alpha}_p\|^2_K \to \|\alpha\|^2_K$ a.s. Let $\epsilon > 0$, from Lemma 5 (see the proof of Theorem 2 in the Appendix),

$$\Big\|\frac{1}{n}\mathcal{X}_p'\mathcal{X}_p - K_{T_p}\Big\|_{op} \le \epsilon\gamma_{p,p}, \text{ eventually, with probability one.}$$

It follows that

$$\Big\|\hat{\beta}_p\Big(\frac{1}{n}\mathcal{X}_p'\mathcal{X}_p - K_{T_p}\Big)\hat{\beta}_p\Big\| \le \|\hat{\beta}_p\|^2\epsilon\gamma_{p,p}, \text{ eventually, with probability one.}$$

So it is enough to prove that $\|\hat{\beta}_p\|^2\gamma_{p,p}$ is bounded from above independently of p. But $\|\hat{\beta}_p\|^2\gamma_{p,p} \leq 2\|\hat{\beta}_p - \beta_p\|^2\gamma_{p,p} + 2\|\beta_p\|^2\gamma_{p,p}$. Then, since $\|\hat{\beta}_p - \beta_p\|^2 \to 0$ a.s., it is enough to bound $\|\beta_p\|^2\gamma_{p,p}$ from above. But $\|\alpha_p\|_K^2 = \beta_p' K_{T_p}\beta_p \geq \|\beta_p\|^2\gamma_{p,p}$. This, together with the conclusion of Theorem 2, $\|\hat{\alpha}_p\|_K^2 \to \|\alpha\|_K^2$, and (28), concludes the proof.

□

6 Some empirical results

We will consider here different examples of functional regression problems in which the goal is to predict a real random variable Y from a functional explanatory variable $X = \{X(t) : t \in I\}$. Hence, our sample information is given by n pairs (X_i, Y_i), $i = 1, \ldots, n$, where $X_i = \{X_i(t) : t \in I\}$ are sample trajectories of the process X, and Y_i are the corresponding response variables.

The overall aim of this section is to check the performance of different finite-dimensional models, based on a few one-dimensional marginals $X(t_1), \ldots, X(t_p)$, such as those whose asymptotic behaviour has been analysed in the previous section, versus that of a functional L^2-based counterpart. More precisely, we will compare the performance of a model of type

$$Y = \beta_0 + \beta_1 X(t_1) + \ldots + \beta_p X(t_p) + \varepsilon, \tag{29}$$

with that of

$$Y = \beta_0 + \int_0^1 \beta(t)X(t)dt + \varepsilon, \tag{30}$$

see the beginning of Sect. 6.1 for details. The word "performance" must be mostly understood in terms of "prediction capacity", as measured by appropriate estimations of the prediction error $\mathbb{E}[(\hat{Y} - Y)^2]$, $\hat{Y}$ being the predicted value for the response obtained from the fitted model; see Figs. 1, 2, 3, 4.

It is very important to note that the finite-dimensional models of type (29) are viewed here as functional models, in the sense that they are all considered as particular cases of the RKHS-model (4). In practice, this means that we do not assume any prior knowledge about the "impact points" t_i or the number p of variables. So, in principle, the whole trajectory is available in order to pick up the impact points t_i we will use. However, given the grid points t_i, the model (29) is handled as a problem of finite-dimensional multiple regression.

6.1 Simulation experiments

6.2 The models we use to generate the data

We analyse here three scenarios: the first scenario is more or less "neutral" in the comparison of a model based on finite-dimensional marginals versus a L^2-model. The

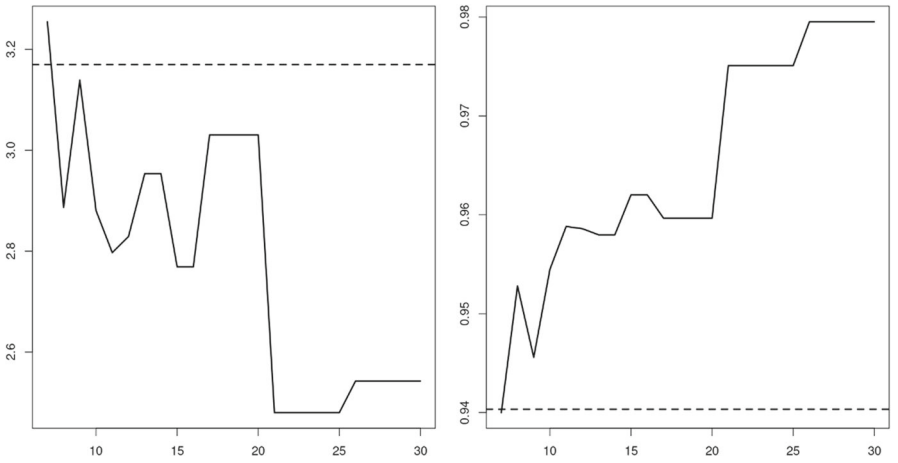

Fig. 5 Prediction errors (mean over 100 replications) and adjusted R_a^2, for different values of p for the Tecator data set using the second derivatives

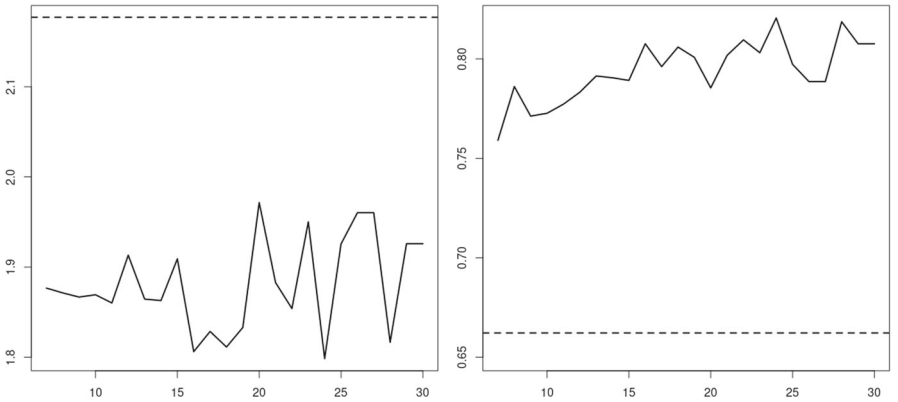

Fig. 6 Prediction errors and adjusted R_a^2, with different values of p for the sugar data set

second one is somewhat favourable to the finite-dimensional models, in the sense that one of these models is the "true one", though we have no advanced knowledge about the impact points t_i and the number of them. Finally, the third scenario clearly favours the L^2-choice since the data are generated according to a model of type (30). In all cases, the aim is to compare the prediction errors obtained with our RKHS-based approach (based on p discretization points), with those corresponding to the classical L^2 method or, more precisely, the popular version of this method based on q principal components; this is the so-called "principal components method" and will be denoted L_q^2 in what follows. We now define these scenarios in precise terms.

Scenario 1. We use the function `rproc2fdata` of the R-package `fda.usc` (Febrero-Bande and Oviedo de la Fuente 2012) to generate random trajectories according to a fractional Brownian Motion (fBM) $X = \{X(t) : t \in [0, 1]\}$ and the aim is to predict $Y \equiv X(1)$ from the

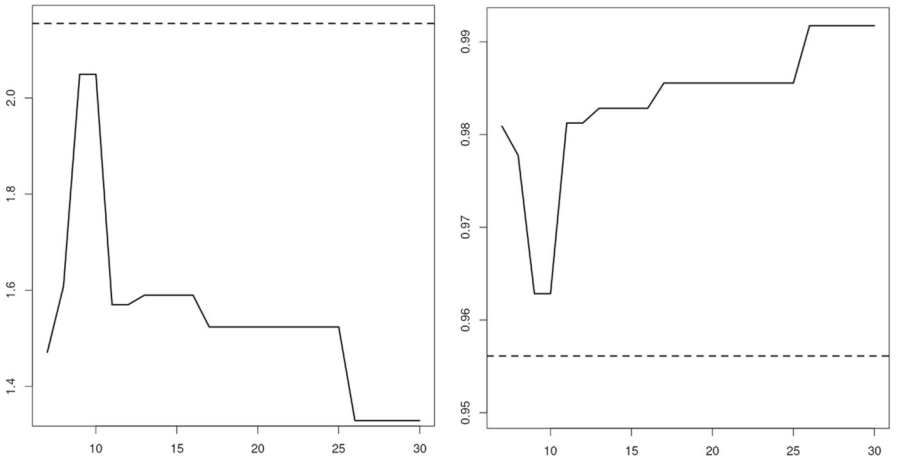

Fig. 7 Prediction errors and adjusted R_a^2 for the population-under-14 data set

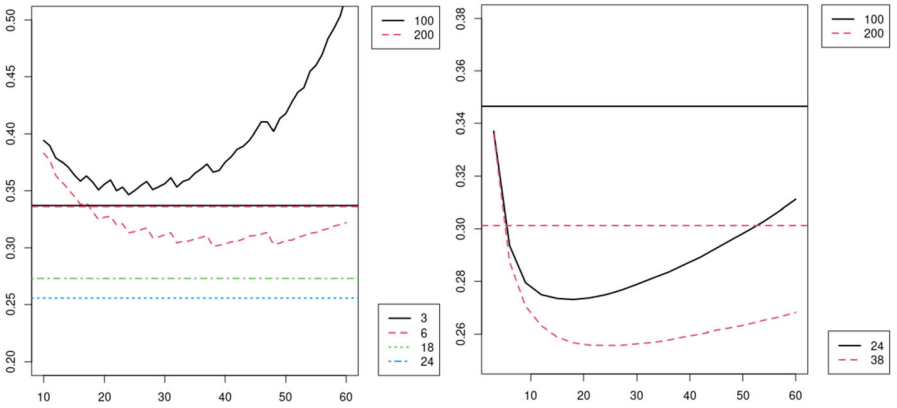

Fig. 8 Left panel: prediction errors under **Scenario 1** for the RKHS-method, plotted as a function of the number of explanatory variables. The horizontal lines correspond to the prediction errors for the principal components (PC) regression method for different choices of the number of components. Right panel: the dual graph for the PC method as a function of the number of components versus the RKHS-based method for a few choices of the number of components

observation of the sample trajectories $X(t)$ for $t \in [0, 0.95]$. Let us recall that the fBM is a Gaussian process whose covariance function is $K(s,t) = 0.5(|t|^{2H} + |s|^{2H} - |t-s|^{2H})$, H being the so-called "Hurst exponent". We have taken $H = 0.8$.

Scenario 2. We have generated the responses Y_i according to the following two finite-dimensional models (previously considered in Berrendero et al. (2019)),
Model 2a: $Y = 2X(0.2) - 5X(0.4) + X(0.9) + \varepsilon$.
Model 2b: $Y = 2.1X(0.16) - 0.2X(0.47) - 1.9X(0.67) + 5X(0.85) + 4.2X(0.91) + \varepsilon$, where in both cases the error variable ε has a distribution

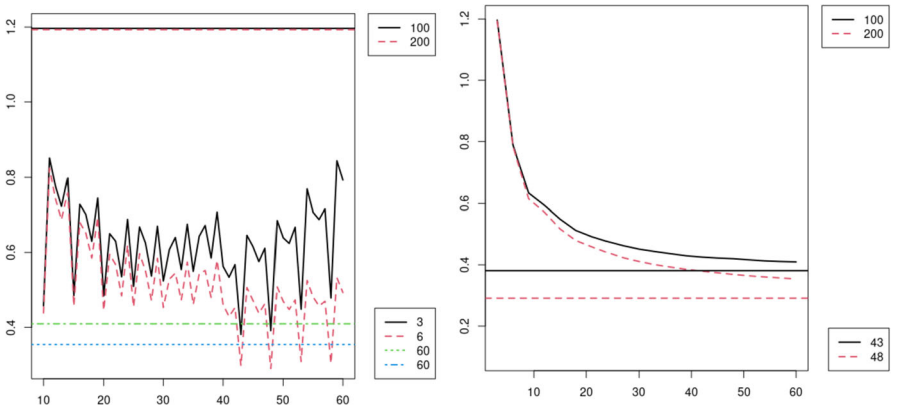

Fig. 9 Left panel: prediction errors under **Scenario 2a** for the RKHS-method, plotted as a function of the number of explanatory variables. The horizontal lines correspond to the prediction errors for the principal components (PC) regression method for different choices of the number of components. Right panel: the dual graph for the PC method as a function of the number of components versus the RKHS-based method for a few choices of the number of components

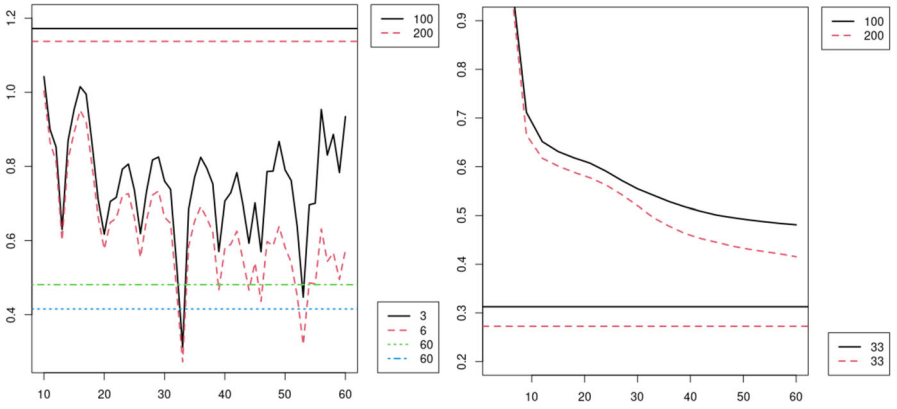

Fig. 10 Left panel: prediction errors under **Scenario 2b** for the RKHS-method, plotted as a function of the number of explanatory variables. The horizontal lines correspond to the prediction errors for the principal components (PC) regression method for different choices of the number of components. Right panel: the dual graph for the PC method as a function of the number of components versus the RKHS-based method for a few choices of the number of components

$N(0, \sigma)$ with standard deviation $\sigma = 0.2$. The process $\{X(t) : t \in [0, 1]\}$ follows a centred fBM with $H = 0.8$.

Scenario 3. The response variable Y is generated according to a L^2-based linear model with $Y = \int_0^1 \log(1+4s)X(s)ds + \varepsilon$, where, again, the trajectories $\{X(t) : t \in [0, 1]\}$ are drawn from the same fBM indicated above and ε has a $N(0, \sigma = 0.2)$ distribution.

Note that these models are only used to generate the data, so that none of the regression models we will compare in our simulations below will incorporate any prior knowledge on the true distribution of $(X(\cdot), Y)$ whatsoever.

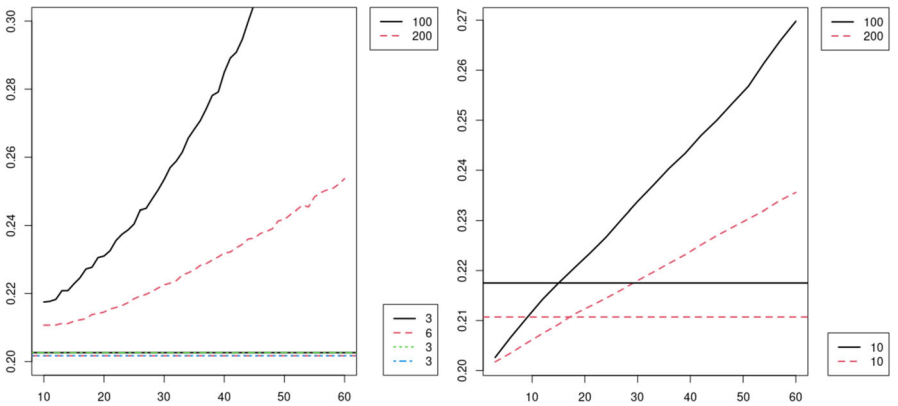

Fig. 11 Left panel: prediction errors under **Scenario 3** for the RKHS-method, plotted as a function of the number of explanatory variables. The horizontal lines correspond to the prediction errors for the principal components (PC) regression method for different choices of the number of components. Right panel: the dual graph for the PC method as a function of the number of components versus the RKHS-based method for a few choices of the number of components

6.3 The specific regression models and estimation methods we compare

Let us go back to our basic question: to what extent the finite-dimensional models (based on marginals $X(t_i)$) of type (29) are competitive against a standard, L^2-based regression model of type (30)? Since we do not assume any previous knowledge on the underlying models generating the data, we will take the "impact points" $t_1, \ldots, t_p$ equispaced in the observation interval [0, 1] (or, in the interval [0, 0.95] in Scenario 1 above). The coefficients β_i in this model are estimated by the ordinary least squares method for multiple regression, using the R-function `lm`. We will check several values of p, from 10 to 60.

As for the L^2-regression model (30), we will estimate the slope function β and the intercept β_0 by the so-called functional principal components (PC) method; this essentially amounts to approximate the model (30) with another finite-dimensional model obtained by projecting the functional data on a given number q of principal functions (i.e., eigenfunctions of the covariance operator). We use the function `fregre.pc` of the R-package `fda.usc`. The considered sample sizes are $n = 100$ (black solid lines in the figures), and 200 (red dotted lines). We report, under the different scenarios, the mean over 1000 replications of the "prediction errors" $\sqrt{(1/k)\sum(Y_i - \hat{Y}_i)^2}$, where $k = 0.2\,n$ is the size of the random "test sub-samples" we use to evaluate the predictions $\hat{Y}_i$. The "training subsamples", made of the remaining 80% of observations, are used to estimate the regression coefficients for $\hat{Y}_i$.

6.4 The simulation outputs

Our results are summarized in Figures 1-4 whose interpretation is as follows: in the left-hand panels, the wiggly curves show the prediction errors of our RKHS-based

models as a function of the number p of explanatory variables (which are just one-dimensional marginals of the underlying process). The horizontal lines correspond to the errors obtained with the standard regression model based on different numbers of principal components, including the "optimal one", assessed by simulation. The graphics in the right-hand panels, are, in some sense, dual: the curves show in this case the estimated prediction errors obtained with the principal components-based regression, as a function of the number of considered components. The horizontal lines correspond to the prediction errors obtained in the RKHS-based model with various "standard" choices of the number of explanatory one-dimensional marginals, including the "optimal" one in the considered equispaced grids. The small square legends in the lower right-hand side of the panels give the corresponding numbers of components (or variables) for the horizontal lines.

These graphics are, hopefully, self-explanatory. The RKHS-based models provide smaller prediction errors in those cases where the underlying model is of RKHS type. In any case, the differences are not very large. As a final, important, remark, let us point out that these comparisons are not completely fair for the RKHS-based models. Indeed, there is a considerable room for optimality in the grids $t_1, \ldots, t_p$, without any restriction of equispaced points; see Berrendero et al. (2019). However, this "variable selection approach" entails a heavier computation load and involves some theoretical challenges outside the scope of this work.

Overall, the results are to be expected: in Figs. 2 and 3, corresponding to Scenarios 2a and 2b (favourable to the finite-dimensional approximations) the predictions based on $\hat{\alpha}_p$ are better than those based on the L^2-based functional linear model, except for very large values of p where the collinearity effect hampers the estimation. In Scenario 3, see Fig. 4, the situation reverses but, still the finite-dimensional models appear to be competitive for small values of p and large sample sizes.

In Scenario 1 (Fig. 1) the U-shape of the curves of estimated prediction errors is more evident. Still, the finite-dimensional proposals are better than L^2_3 and L^2_6 for the central range of considered values of p. A similar example, included in Appendix B, considers the case of the standard Brownian Motion. Here the conclusions are not far from those of the fractional Brownian Motion (with Hurst index H=0.8) but the more irregular nature of the trajectories is reflected in a larger sensitivity with respect to the specific location of the "design points" t_i.

As a final remark, let us point out that our aim here is not to prove an overall superiority of the RKHS-based models in terms of prediction errors. This would require a much more exhaustive numerical study. Still, our experimental results suggest that the interpretability advantages of the RKHS-based models do not necessarily entail any serious loss in efficiency.

6.5 Real data examples

This is another natural playing field for a fair comparison on the prediction capacity of different regression models.

In all considered cases the sample is randomly divided in two parts: 80% of the observations is used for training (i.e. for parameter estimation) and the remaining

20% is used in order to check the accuracy of the predictions. This random splitting is repeated 100 times. Figures 5, 6 and 7 report the average prediction errors and the (average) adjusted coefficients of determination. In all cases we show in the left panel the average prediction errors of the RKHS model for different values of p across 100 replications. The horizontal line represents the error of the L^2 model, which was fitted by projecting onto the three principal components (this value has been chosen just as a reference). The right panel shows the adjusted R^2 values for both the RKHS and L^2 models, represented by horizontal lines.

6.6 The data sets under study

(a) *The Tecator data set*. This data set has been used and described many times in papers and textbooks; see, e.g., Ferraty and Vieu (2006). It is available in the R-package `fda.usc`, see Febrero-Bande and Oviedo de la Fuente (2012). After removing some duplicated data, we have 193 functions obtained from a spectrometric study performed on meat samples in which the near infrared absorbance spectrum is recorded. The response variable is the fat content of the meat pieces. The functions are observed at a grid of 100 points.

An important aspect of this data set is the fact that the derivatives of the sample functions seem to be more informative than the original data themselves. Thus, we have taken into account this feature, using the second derivatives to predict the response variable (obtained by preliminary smoothing of the data. Figure 5 displays the results. All the considered values of p are checked in every run.

(b) *The sugar data set*. This data set has been previously considered in functional data analysis by several authors; see e.g. Aneiros and Vieu (2014) for additional details. The functions $X(t)$ are fluorescence spectra obtained from sugar samples and the response Y is the ash content, in percentage of the sugar samples. The comparison results of finite-dimensional models versus the L^2-functional counterpart are shown in Fig. 6. Again the outputs correspond to the averages over 100 replications obtained by randomly selecting 214 (80%) data for training and 54 (20%) for testing, from the original data.

(c) *Population data*. For 237 countries and geographical areas, the percentage of population under 14 years for the period 1968-2018 (one datum per year) is recorded. In our experiment, we consider longitudinal data consisting of vectors $(X(1960), X(1961), \ldots, X(2010))$; the aim is to predict the value eight years ahead. Thus, the response variable is $Y = X(2018)$. Several theoretical assumptions (for example, independence), commonly used in the linear model, are violated here but, still, our comparisons make sense at an exploratory data level. The outputs can be found in Fig. 7 below. As in the previous examples, they correspond to 100 runs based on random partitions of the data set into 80% training data and 20% test data. Again p denotes the number of years (equispaced in the interval 1960-2010) used as explanatory variables in the finite-dimensional models. Thus for $p = 10$ we consider the years 1960, 1965,...,2010; for $p = 8$ we take 1960, 1967, 1974,...,2009.

7 Conclusions

We explore a mathematical framework, different from the classical L^2-approach (30), for the problem of linear regression with functional explanatory variable X and scalar response Y. This mathematical formulation includes, as particular cases, the finite-dimensional models (29) obtained by considering as explanatory variables a finite set of marginals $X(t_i)$, with $i = 1, \ldots, p$. This would allow us, for example, to compare such models for variable selection purposes (Berrendero et al. 2019) or considering, within a unified framework, the study of asymptotic behaviour of models as the number p of covariates grows to infinity; see e.g. Sur and Candès (2019) for a recent analysis in the logistic regression model. Note also that in the functional case the asymptotic analysis as $p \to \infty$ appears more naturally than in the case of general regression studies, since all co-variables $X(t_i)$ come from the unique, predefined reservoir of the one-dimensional marginals of the process $X = \{X(t) : t \in [0, 1]\}$.

While this model, based on the theory of RKHS spaces, has been considered (explicit or implicitly) in several other papers, as mentioned above, we contribute some insights and some new theory that, hopefully, will consolidate this RKHS option as a useful alternative.

From a practical point of view, the fact of encompassing all the finite-dimensional models under a unique super-model (4)–(6) is also relevant in view of the empirical results of Sect. 6: indeed, the outputs of the simulations and the real data examples there show that, somewhat surprisingly, there is often little gain in considering the L^2 functional model (30) instead of the simpler finite-dimensional alternatives (29).

Of course, we do not claim that the L^2-based regression model (30) should be abandoned in favour of the finite-dimensional alternatives of type (29), since the L^2 model is now well-understood and has proven useful in many examples. We are just suggesting that there are perhaps some reasons to consider the problem of linear functional regression under a broader perspective. In addition, note that the L^2 model appears as a particular case of the general formulation (4)–(6).

In any case, even if we are willing to incorporate the finite-dimensional models (29), according to our suggested approach, the functional character of the regression problem is not lost at all as the proposed global general formulation is unequivocally functional. In practice, this means that, according to our assumptions, the explanatory variables are still functions and we cannot get rid of this fact in the formulation of our problem.

A Proof of Theorem 2

In what follows, we denote $\mathcal{X}_p$ the $n \times p$ data matrix whose (i, j)-entry is $X_i(t_{j,p})$, $i = 1, \ldots, n$, $j = 1, \ldots, p$. Denote also by K_{T_p}, the covariance matrix of $(X(t_{1,p}), \ldots, X(t_{p,p}))$. Finally, we denote $\mathbf{Y} = (Y_1, \ldots, Y_n)'$ and $\mathbf{e} = (e_{1,p}, \ldots, e_{n,p})'$, where $'$ stands for the transpose.

The proof of Theorem 2 relies on the three lemmas stated below.

Lemma 3 *Let $\gamma_{1,p} \geq \gamma_{2,p} \geq \cdots \geq \gamma_{p,p}$ be the eigenvalues of K_{T_p} and $\hat{\gamma}_{1,p} \geq \hat{\gamma}_{2,p} \geq \cdots \geq \hat{\gamma}_{p,p}$ the eigenvalues of $(1/n)(\mathcal{X}_p'\mathcal{X}_p)$. Then, for $j = 1, \ldots, p$, $|\gamma_{j,p} - \hat{\gamma}_{j,p}| \leq \|(1/n)(\mathcal{X}_p'\mathcal{X}_p) - K_{T_p}\|_{op}$.*

Proof This result follows as a direct application of Lemma 3.1 in Bosq (1991), (this is also sometimes called Weyl's inequality in the literature). □

Lemma 4 *Let K be a continuous covariance function and let T_p be a set of grid points as in Lemma 2. Assume that all the eigenvalues of the covariance operator $\mathcal{K}$ are strictly positive. Then $\lim_{p\to\infty} \frac{1}{p}\|K_{T_p}\|_{op} = 0$.*

Proof Assume by contradiction that $\lim_{k\to\infty} \|K_{T_{p_k}}\|_{op} = \lim_{k\to\infty} \gamma_{1,p_k} = \infty$ for some sequence $p_k \to \infty$. Let us denote for simplicity $p_k = p$. Let the p-dimensional vector $f_p = (f(t_{1,p}), \ldots, f(t_{p,p}))$ be an eigenvector of $(1/p)K_{T_p}$ associated to $\gamma_{1,p}$ the largest eigenvalue of K_{T_p}, such that $\|f_p\|_{\max} = 1$ for all p. Let us define a polygonal function $g_p : [0,1] \to \mathbb{R}$ such that $g_p(t_{i,p}) = f(t_{i,p})$, observe that $\|g_p\|_\infty = \|f_p\|_{\max} = 1$. Let us prove that $\{g_p\}_p$ is an equicontinuous sequence. Since $K(s,t)$ is continuous, it is also uniformly continuous on $[0,1]^2$. Then, for all $\epsilon > 0$, there exists $\delta = \delta(\epsilon) > 0$ such that $|K(x,y) - K(x',y')| < \epsilon$ if $\|(x,y) - (x',y')\|_{\max} < \delta$. Let us denote $\|T_p\| = \max_{i=1,\ldots,p-1} |t_{i+1,p} - t_{i,p}|$. Let $\epsilon > 0$ and p large enough such that $\|T_p\| < \delta$. Then, we have

$$\begin{aligned}
&\gamma_{1,p}|f(t_{i,p}) - f(t_{i+k,p})| \\
&= \frac{1}{p}\Big|\sum_{j=1}^{p}\Big[K(t_{i,p},t_{j,p}) - K(t_{i+k,p},t_{j,p})\Big]f(t_{j,p})\Big| \\
&\leq \frac{1}{p}\sum_{j=1}^{p}\Big|K(t_{i,p},t_{j,p}) - K(t_{i+k,p},t_{j,p})\Big| \max_{j=1,\ldots,p}|f(t_{j,p})| \leq \epsilon,
\end{aligned}$$

where the last inequality follows from $|K(t_{i,p},t_{j,p}) - K(t_{i+k,p},t_{j,p})| < \epsilon$ and $\|f_p\|_{\max} = 1$ for all i,k such that $1 \leq i \leq p$, $1 \leq i+k \leq p$ and $|t_{i,p} - t_{i+k,p}| < \delta$.

Then for p large enough, for all i,k such that $1 \leq i \leq p$, $1 \leq i+k \leq p$ and $|t_{i,p} - t_{i+k,p}| < \delta$, $|f(t_{i,p}) - f(t_{i+k,p})| \leq \epsilon$. Hence, $\{g_p\}_p$ is equicontinuous.

Since $\{g_p\}_p$ is bounded, by Arzela–Ascoli Theorem there exists $p_k \to \infty$ and a continuous function g such that $\|g_{p_k} - g\|_\infty \to 0$. For ease of writing we will denote $g_{p_k} = g_p$, Let us fix $t_{i,p}$. Then, for all $\epsilon > 0$ and for p (which depends on $t_{i,p}$) large enough,

$$\begin{aligned}
&\Big|\gamma_{1,p}g_p(t_{i,p}) - \int_0^1 K(t_{i,p},t)g_p(t)dt\Big| \\
&= \Big|\frac{1}{p}\sum_{j=1}^{p}K(t_{i,p},t_{j,p})g_p(t_{j,p}) - \int_0^1 K(t_{i,p},t)g_p(t)dt\Big| < \epsilon. \qquad (31)
\end{aligned}$$

Since $g_p \to g$ uniformly, there exists $p_0 > 0$ such that Eq. (31) holds for all $p > p_0$. By continuity of K and g_p it can be seen that there exists $p_1 > p_0$ such that for all

$p > p_1$,

$$\max_{s\in[0,1]} \left|\gamma_{1,p} g_p(s) - \int_0^1 K(s,t) g_p(t) dt\right| < \epsilon.$$

Again, using that $g_p \to g$ uniformly, $\| \int_0^1 K(\cdot,t) g_p(t) dt - \int_0^1 K(\cdot,t) g(t) dt \|_\infty \to 0$. Then $\gamma_{1,p} g_p \to \lambda g$ for some $\lambda > 0$. Observe that $\|g\|_\infty = 1$. This proves that g is an eigenfunction of $\mathcal{K}$ with eigenvalue $\lambda > 0$, and contradicts that $\gamma_{1,p} \to \infty$. □

Lemma 5 *Under the hypotheses of Theorem 2. We have, for all $\epsilon > 0$,*

$$\left\| \frac{1}{n} \mathcal{X}_p' \mathcal{X}_p - K_{T_p} \right\|_{op} \leq \epsilon \gamma_{p,p}, \text{ eventually, with probability one.} \tag{32}$$

Proof Let us define $\mathcal{F}_n = \{\omega : \| \frac{1}{n} \mathcal{X}_p' \mathcal{X}_p - K_{T_p} \|_{op} > \epsilon \gamma_{p,p}\}$. To prove (32), by Borel-Cantelli lemma, it is enough to prove that, $\sum_n \mathbb{P}(\mathcal{F}_n) < \infty$. Let us denote

$$\mathcal{A}_i = \mathcal{A}_{i,n} = \{\omega : \max_{j=1,\ldots,p} |X_i(t_{j,p})| < \log n\},$$

then $\mathbb{P}(\mathcal{F}_n) \leq \mathbb{P}(\mathcal{F}_n \cap \bigcap_{i=1}^n \mathcal{A}_i) + n\mathbb{P}(\mathcal{A}_1^c) := I_{1,n} + I_{2,n}$.

To prove that $\sum_n I_{1,n} < \infty$, we will use Corollary 5.2 in Mackey et al. (2014). Let us define M_k the $p \times p$ random matrix whose entry i, j is $(X_k(t_{i,p}) X_k(t_{j,p}) - (K_{T_p})_{i,j}) \mathbb{I}_{\mathcal{A}_k}$. Let us denote $\mathbf{Z}_k = (X_k(t_{1,p}), \ldots, X_k(t_{p,p}))' \mathbb{I}_{\mathcal{A}_k}$, then $M_k = \mathbf{Z}_k \mathbf{Z}_k' - K_{T_p} \mathbb{I}_{\mathcal{A}_k}$, so

$$\|M_k\|_{op} \leq \|\mathbf{Z}_k \mathbf{Z}_k'\|_{op} + \|K_{T_p} \mathbb{I}_{\mathcal{A}_k}\|_{op} \leq \|\mathbf{Z}_k\|^2 + \|K_{T_p} \mathbb{I}_{\mathcal{A}_k}\|_{op}.$$

We have that $\|\mathbf{Z}_k\|^2 \leq p \log^2 n$ and

$$\begin{aligned} \|K_{T_p} \mathbb{I}_{\mathcal{A}_k}\|_{op} &= \max_{z\in S^{p-1}} \mathbb{E}[(z'\mathbf{Z}_k)(\mathbf{Z}_k' z)] = \max_{z\in S^{p-1}} \mathbb{E}[(z'\mathbf{Z}_k)^2] \\ &\leq \max_{z\in S^{p-1}} \|z\|^2 \mathbb{E}(\|\mathbf{Z}_k\|^2) \leq p \log^2 n. \end{aligned}$$

To bound $\eta^2 := \| \sum_k \mathbb{E}(M_k^2) \|_{op} \leq n \| \mathbb{E}(M_1^2) \|_{op}$, observe that, $\mathbb{E}[M_1^2] \leq \mathbb{E}[(\mathbf{Z}_1 \mathbf{Z}_1')^2] = \mathbb{E}[\|\mathbf{Z}_1\|^2 \mathbf{Z}_1 \mathbf{Z}_1'] \leq p(\log^2 n) \mathbb{E}[\mathbf{Z}_1 \mathbf{Z}_1'] \leq p(\log^2 n) K_{T_p}$. Then, $\eta^2 \leq np(\log^2 n)\gamma_{1,p}$. By Corollary 5.2 in Mackey et al. (2014),

$$\mathbb{P}\Big(\mathcal{F}_n \cap \bigcap_{i=1}^n \mathcal{A}_i\Big) \leq p \exp\left[- \frac{(n\epsilon\gamma_{p,p})^2}{3np\gamma_{1,p} \log^2 n + 4pn\epsilon\gamma_{p,p} \log^2 n} \right] := \exp(-a_n).$$

From Lemma 4, $\gamma_{1,p}/p \to 0$, then $\sum_n I_{1,n} < \infty$ follows from the assumption

$$n(\gamma_{p,p})^2/(p^2 \log(n)^3) \to C_3 > 0$$

since this implies $a_n/\log n \to \infty$. Finally, $\sum_n n\mathbb{P}(\mathcal{A}_1^c) < \infty$ follows from the assumption $\mathbb{P}(\sup_{t\in[0,1]} X(t) > s) \leq \exp(-Cs^2)$ for some constant $C > 0$ and for all $s > 0$. □

Now, to prove Theorem 2, let us take $p = p_n$ and T_p as in Lemma 4. Recall that

$$\alpha_p(\cdot) = \sum_{j=1}^{p} \beta_{j,p} K(t_{j,p}, \cdot) \text{ and } \hat{\alpha}_p(\cdot) = \sum_{j=1}^{p} \hat{\beta}_{j,p} K(t_{j,p}, \cdot),$$

and denote $\beta_p = (\beta_{1,p}, \ldots, \beta_{p,p})'$, and $\hat{\beta}_p = (\hat{\beta}_{1,p}, \ldots, \hat{\beta}_{p,p})'$. Observe that $\|\hat{\alpha}_p - \alpha_p\|_K^2 = (\hat{\beta}_p - \beta_p)' K_{T_p}(\hat{\beta}_p - \beta_p) = (\hat{\beta}_p - \beta_p)' K_{T_p}^{1/2} K_{T_p}^{1/2}(\hat{\beta}_p - \beta_p)$. Since $K_{T_p}^{1/2} = (K_{T_p}^{1/2})'$, $\|\hat{\alpha}_p - \alpha_p\|_K^2 = \|K_{T_p}^{1/2}(\hat{\beta}_p - \beta_p)\|^2 \leq \|K_{T_p}^{1/2}\|_{op}^2 \|\hat{\beta}_p - \beta_p\|^2$. By Lemma 4, for all $\lambda > 0$ we can take p large enough such that, $\|\hat{\alpha}_p - \alpha_p\|_K^2 \leq 2\lambda p \|\hat{\beta}_p - \beta_p\|^2$ for a finite value λ. So it is enough to prove that there exists $C < \infty$ such that

$$\nu_n p \|\hat{\beta}_p - \beta_p\|^2 < C \quad \text{a.s.} \tag{33}$$

Since $\hat{\beta}_p = \text{argmin}_v \|\mathbf{Y} - \mathcal{X}_p v\|$, we have $\|\mathbf{Y} - \mathcal{X}_p\hat{\beta}_p\| \leq \|\mathbf{Y} - \mathcal{X}_p\beta_p\| = \|\mathbf{e}\|$ and

$$\begin{aligned} \|\mathbf{Y} - \mathcal{X}_p\hat{\beta}_p\|^2 &= \|\mathcal{X}_p\beta_p + \mathbf{e} - \mathcal{X}_p\hat{\beta}_p\|^2 \\ &= \|\mathcal{X}_p(\hat{\beta}_p - \beta_p)\|^2 + \|\mathbf{e}\|^2 - 2\mathbf{e}'\mathcal{X}_p(\hat{\beta}_p - \beta_p), \end{aligned}$$

so $\|\mathcal{X}_p(\hat{\beta}_p - \beta_p)\|^2 \leq 2\mathbf{e}'\mathcal{X}_p(\hat{\beta}_p - \beta_p)$. Denote by Φ an $n \times p$ matrix whose columns form an orthonormal basis for the linear space spanned by the columns of $\mathcal{X}_p$. Then $\mathcal{X}_p(\hat{\beta}_p - \beta_p) = \Phi v$ for some unique $v \in \mathbb{R}^p$. Thus denoting by S^{p-1} the unit sphere of $\mathbb{R}^p$ and $\tilde{\mathbf{e}} = \Phi'\mathbf{e}$,

$$\|\mathcal{X}_p(\hat{\beta}_p - \beta_p)\| \leq 2\mathbf{e}' \frac{\mathcal{X}_p(\hat{\beta}_p - \beta_p)}{\|\mathcal{X}_p(\hat{\beta}_p - \beta_p)\|} = 2\tilde{\mathbf{e}}' \frac{v}{\|v\|} \leq 2 \sup_{u\in S^{p-1}} \tilde{\mathbf{e}}' u.$$

Let us denote $N_{1/2}$ a minimal covering of S^{p-1} with balls of radii $1/2$, centred at points in S^{p-1}. Its cardinality $|N_{1/2}|$ is bounded from above by 5^{p-1}. For all $u \in S^{p-1}$ there exists a point z in the set $C_{1/2}$ of centres of the balls in $N_{1/2}$ and $w \in \mathbb{R}^p$ such that $u = z + w$, with $\|w\| \leq 1/2$. Denote $W_{1/2}$ the set of such $w's$ so $\max_{u\in S^{p-1}} \tilde{\mathbf{e}}' u \leq \max_{z\in C_{1/2}} \tilde{\mathbf{e}}' z + \max_{w\in W_{1/2}} \tilde{\mathbf{e}}' w$, then

$$2 \sup_{u\in S^{p-1}} \tilde{\mathbf{e}}' u \leq 4 \max_{z\in C_{1/2}} \tilde{\mathbf{e}}' z. \tag{34}$$

Observe that $\|K_{T_p}^{1/2}(\hat{\beta}_p - \beta_p)\|^2 = (\hat{\beta}_p - \beta_p)' K_{T_p}(\hat{\beta}_p - \beta_p)$. Thus,

$$\|K_{T_p}^{1/2}(\hat{\beta}_p - \beta_p)\|^2 = (\hat{\beta}_p - \beta_p)'(K_{T_p} - (1/n)(\mathcal{X}_p'\mathcal{X}_p))(\hat{\beta}_p - \beta_p) +$$

$$(\hat{\beta}_p - \beta_p)'(1/n)(\mathcal{X}_p'\mathcal{X}_p)(\hat{\beta}_p - \beta_p).$$

Now, using Lemma 5 (for $\epsilon \in (0,1)$) together with the inequalities $|x'Ax| \leq \|A\|_{op}\|x\|^2$, for $x = (\hat{\beta}_p - \beta_p)$, $A = (1/n)(\mathcal{X}_p'\mathcal{X}_p) - K_{T_p}$ and $\|K_{T_p}\|_{op} \geq \gamma_{p,p}$ we get that, eventually a.s.,

$$\begin{aligned}\frac{1}{n}\|\mathcal{X}_p(\hat{\beta}_p - \beta_p)\|^2 \geq & \|K_{T_p}^{1/2}(\hat{\beta}_p - \beta_p)\|^2 - \epsilon\gamma_{p,p}\|\hat{\beta}_p - \beta_p\|^2 \\ \geq & \|\hat{\beta}_p - \beta_p\|^2\gamma_{p,p} - \epsilon\gamma_{p,p}\|\hat{\beta}_p - \beta_p\|^2.\end{aligned}$$

Let n large enough such that $n\gamma_{p,p}/(p^2\nu_n\log(n)) < 2C_4$, and C large enough such that $2CC_4(1-\epsilon)/16 > 1$. Then, from (34),

$$\mathbb{P}(\nu_n p\|\hat{\beta}_p - \beta_p\|^2 > C) \leq 5^{p-1}\max_{z\in S^{p-1}}\mathbb{P}\big(\tilde{\mathbf{e}}'z > \sqrt{nC\gamma_{p,p}(1-\epsilon)/(16p\nu_n)}\big).$$

Now, note that there is a sub-Gaussianity bound, not depending on z, for the tail probabilities $\mathbb{P}(\tilde{\mathbf{e}}'z > t)$ (see the remarks immediately before Theorem 2). Finally, (33) follows from Borel-Cantelli lemma.

B Simulations for the standard brownian motion

In Figures 8, 9, 10, 11 we present the results of the simulations for the same scenarios considered before and the same sample sizes, but when the process is the Brownian Motion.

Acknowledgements This research has been partially supported by Grants PID2019-109387GB-I00 from the Spanish Ministry of Science and Innovation, Grant CEX2019-000904-S funded by MCIN/AEI/10.13039/501100011033 and FCE_1_2019_1_156054 from ANII, Uruguay. The comments and criticisms from two reviewers and the Editors are gratefully acknowledged.

Funding Open Access funding provided thanks to the CRUE-CSIC agreement with Springer Nature.

References

Aneiros G, Vieu P (2014) Variable selection in infinite-dimensional problems. Stat Prob Lett 94:12–20

Ash R, Gardner M (1975) Topics in stochastic processes. Academic Press, Cambridge

Berlinet A, Thomas-Agnan C (2004) Reproducing kernel Hilbert spaces in probability and statistics. Kluwer Academic Publishers, New York

Berrendero J, Bueno-Larraz B, Cuevas A (2022) On functional logistic regression: some conceptual issues. Test 32:321–349
Berrendero J, Bueno-Larraz B, Cuevas A (2019) An RKHS model for variable selection in functional linear regression. J Multivar Anal 170:25–45
Bosq D (1991) Modelization, nonparametric estimation and prediction for continuous time processes. In: Roussas G (ed) Nonparametric functional estimation and related topics, NATO ASI Series. Mathematical and physical sciences series C. Springer, New York, pp 509–529
Cardot H, Ferraty F, Sarda P (1999) Functional linear model. Stat Prob Lett 45:11–22
Cucker F, Zhou DX (2007) Learning theory: an approximation theory viewpoint. Cambridge University Press, Cambridge
Cuevas A (2014) A partial overview of the theory of statistics with functional data. J Stat Plan Inference 147:1–23
Doksum K, Samarov A (1995) Nonparametric estimation of global functionals and a measure of the explanatory power of covariates in regression. Ann Stat 23:1443–1473
Febrero-Bande M, Oviedo de la Fuente M (2012) Statistical computing in functional data analysis: the R package fda.usc. J Stat Softw 51:1–28
Ferraty F, Vieu P (2006) Nonparametric functional data analysis: theory and practice. Springer Science and Business Media, New York
Horváth L, Kokoszka P (2012) Inference for functional data with applications. Springer Science and Business Media, New York
Hsing T, Eubank R (2015) Theoretical foundations of functional data analysis. Wiley, New York
Hsing T, Ren H (2009) An RKHS formulation of the inverse regression dimension-reduction problem. Ann Stat 37:726–755
Janson S (1997) Gaussian Hilbert spaces, vol 129. Cambridge University Press, Cambridge
Kneip A, Liebl D (2020) On the optimal reconstruction of partially observed functional data. Ann Stat 48:1692–1717
Kneip A, Poß D, Sarda P (2016) Functional linear regression with points of impact. Ann Stat 44:1–30
Landau HJ, Shepp LA (1970) On the supremum of a Gaussian process. Sankhyā 32:369–378
Laha RG, Rohatgi VK (1979) Probability theory. Wiley, New York
Lindquist MA, McKeague IW (2009) Logistic regression with Brownian-like predictors. J Am Stat Assoc 104:1575–1585
Lukić M, Beder J (2001) Stochastic processes with sample paths in reproducing kernel Hilbert spaces. Trans Am Math Soc 353:3945–3969
Mackey L, Jordan MI, Chen RY, Farrell B, Tropp JA (2014) Matrix concentration inequalities via the method of exchangeable pairs. Ann Prob 42:906–945
Mardia K, Kent J, Bibby J (2021) Multivariate analysis. Probability and mathematical statistics. Academic Press Inc, Cambridge
McKeague IW, Sen B (2010) Fractals with point impact in functional linear regression. An Stat 38:25–59
Parzen E (1959) Statistical inference on time series by Hilbert space methods. CA applied mathematics and statisticas labs. I. Stanford Univ, Stanford
Poß D, Liebl D, Kneip A, Eisenbarth H, Wager TD, Barrett LF (2020) Superconsistent estimation of points of impact in non-parametric regression with functional predictors. J R Stat Soc Ser B 82:1115–1140
Ramsay JO, Silverman BW (2005) Functional data analysis. Springer, New York
Rencher AC, Schaalje GB (2008) Linear models in statistics. Wiley, New York
Rigollet P, Hütter J (2017) High dimensional statistics. Lecture notes. Massachusetts Institute of Technology, Cambridge, Cambridge
Shang Z, Cheng G (2015) Nonparametric inference in generalized functional linear models. Ann Stat 43:1742–1773
Shin H, Hsing T (2012) Linear prediction in functional data analysis. Stoch Process Appl 122:3680–3700
Shin H, Lee S (2016) An RKHS approach to robust functional linear regression. Stat Sin 26:255–272
Sur P, Candès EJ (2019) A modern maximum-likelihood theory for high-dimensional logistic regression. Proc Nat Acad Sci 116:14516–14525
Yuan M, Cai TT (2010) A reproducing kernel Hilbert space approach to functional linear regression. Ann Stat 38:3412–3444