

PRESMOOTHING IN FUNCTIONAL LINEAR REGRESSION

Frédéric Ferraty¹, Wenceslao González-Manteiga²,
Adela Martínez-Calvo² and Philippe Vieu¹

¹*Université Paul Sabatier* and ²*Universidade de Santiago de Compostela*

Abstract: In this paper, we consider the functional linear model with scalar response, and explanatory variable valued in a function space. In recent literature, functional principal components analysis (FPCA) has been used to estimate the model parameter. We propose to modify this approach by using presmoothing techniques. For this new estimate, consistency is stated and efficiency by comparison with the standard FPCA estimator is studied. We have also analysed the behaviour of our presmoothed estimator by means of a simulation study and data applications.

Key words and phrases: Functional linear regression (FLR), functional principal components analysis (FPCA), nonparametric kernel estimator, presmoothing.

1. Introduction

Nowadays computational tools allow us to create and store very large databases. In many cases, the dataset is made up of observations from a finite-dimensional distribution, measured over a period of time or recorded at different spatial locations. When the temporal or spatial grid is fine enough, the sample can be considered as an observation of a random variable on a certain function space (see examples in Ramsay and Silverman (2002, 2005) or Ferraty and Vieu (2006)). Analysing this kind of data with standard multivariate methods and ignoring its functional feature may fail dramatically (curse of dimensionality, strong collinearities, etc). In these cases, specific statistical techniques are required in order to draw relevant underlying information.

With regard to regression models, classical multivariate methods have been adapted to the functional context where the response Y and/or the explanatory variable X are valued in a function space (see Ramsay and Silverman (2005) for a *parametric* state of art, and Ferraty and Vieu (2006) for a *nonparametric* one). Particularly, the functional linear model with scalar response has been the subject of numerous studies in the recent literature, as it is here.

From now on, let $(H, \langle \cdot, \cdot \rangle)$ be a real separable Hilbert space and let $\|\cdot\|$ denote the induced norm ($\|x\| = \langle x, x \rangle^{1/2}$, $x \in H$). We consider the functional linear model

$$Y = m(X) + \epsilon = \langle \theta, X \rangle + \epsilon, \quad (1.1)$$

where Y is a real random variable, $m(\cdot) = \langle \theta, \cdot \rangle$ is a linear regression operator with $\theta \in H$ and $\|\theta\|^2 < \infty$, X is a random zero-mean variable valued in H satisfying that $E(\|X\|^2) < \infty$, and ϵ is a real random variable such that $E(\epsilon) = 0$, $Var(\epsilon) = \sigma^2$, and $E(\epsilon X) = 0$. When the variables X and Y are not centered, we can take $\tilde{Y} = Y - E(Y)$ and $\tilde{X} = X - E(X)$ and consider the regression model $\tilde{Y} = \langle \theta, \tilde{X} \rangle + \epsilon$, which is equivalent to

$$Y = \theta_0 + \langle \theta, X \rangle + \epsilon, \quad (1.2)$$

with $\theta_0 = E(Y) - \langle \theta, E(X) \rangle$. Thus we focus on (1.1) with zero-mean variables, without loss of generality.

Although we have fixed the broad framework of Hilbert spaces, functional data are often valued in well-known common spaces. For example, the space $H = L^2([0, 1])$ is usually considered when the functional data are curves, that is, when $X = \{X(t), t \in [0, 1]\}$. In this situation, the general regression model (1.1) becomes

$$Y = \int_0^1 \theta(t)X(t)dt + \epsilon, \quad (1.3)$$

when one takes the inner product $\langle x, y \rangle = \int_0^1 x(t)y(t)dt$ for all $x, y \in L^2[0, 1]$.

Many authors have proposed procedures for estimating θ , both in the general model (1.1) and in the particular case given by (1.3). Common methods use basis systems such as Fourier series, wavelets, or splines (see a general review in Ramsay and Silverman (2005), and see Cardot, Ferraty, and Sarda (2003) or Crambes, Kneip, and Sarda (2009) for the popular splines basis). Second methodology is based on functional principal components analysis (FPCA), as developed and analyzed by Cardot, Ferraty, and Sarda (1999) Cardot, Ferraty, and Sarda (2003), Cai and Hall (2006), Hall and Hosseini-Nasab (2006), Hall and Horowitz (2007), and Cardot, Mas, and Sarda (2007). The FPCA estimator is revisited here in order to improve its behaviour in terms of conditional mean square errors by introducing presmoothing techniques.

We decided on presmoothing methods in light of Faraldo-Roca and González-Manteiga (1987) and Cristóbal-Cristóbal, Faraldo-Roca, and González-Manteiga (1987). These authors proposed the application of the least-squares principle on the pairs $(X_i, \hat{m}_n(X_i))$ instead of (X_i, Y_i) , where \hat{m}_n is a nonparametric kernel-type estimate of the regression function m . This alteration to the minimization problem produced efficient estimates that reduced the mean squared error of

the classical least-squares estimates for the linear regression model, except in the compact support case. Later, Janssen, Swanepoel, and Veraverbeke (2001) showed that the inefficiency problem in the compact support case could be rectified using *boundary kernels*. A similar procedure was developed by Akritas (1996) to fit polynomial regression models to data with incomplete observations. Since then, the presmoothing methodologies have been successfully applied in such areas as model selection procedures (see Aerts, Hens, and Simonoff (2010), who smoothed the response data prior to model selection by Akaike's Information Criterion), and censored/truncated survival data analysis (see Cao-Abad et al. (2005), Jácome and Iglesias-Pérez (2008), and Jácome, Gijbels, and Cao (2008), who replaced censoring indicator variables by values of a nonparametric regression estimator).

In functional data analysis, presmoothing processes are usually included as preliminary steps in such a way that the observations are replaced by their smoothed approximations. For instance, Hitchcock, Casella, and Booth (2006) examined the effect of this substitution on estimating the dissimilarities among elements in the dataset. Another way to use presmoothing methods has been considered by Zhang and Chen (2007) in dealing with the regression model $Y_i(t) = X_i^t \theta(t) + V_i(t) + \epsilon_i(t)$, $i = 1, \dots, n$, where the covariate X_i is multi-dimensional and time-independent, the process V_i represents the i th individual variation, and Y_i is the i th response process. These authors proposed an estimate of θ based on a local polynomial kernel reconstruction of $f_i(t) = X_i^t \theta(t) + V_i(t)$.

We are not interested in presmoothing as a preprocessing tool, but rather as a way to build a new efficient FPCA-type estimator, that reduces the conditional mean square errors of the standard FPCA one, along the lines of Faraldo-Roca and González-Manteiga (1987) in the real case. The idea here is similar to one that motivates multivariate ridge regression: circumvent the problem of an ill-conditioned covariance operator by means of an artificial perturbation of its eigenvalues. To see the usefulness of our approach, here is an example that shows the instability of the FPCA estimator when the eigenvalues are close to 0; it will be analyzed in detail in the simulation study (see Section 4). Consider the regression model (1.3) when the explanatory curves are

$$X(t) = a_1 \sqrt{2} \sin(\pi t) + a_2 \sqrt{2} \cos(\pi t) + a_3 \sqrt{2} \sin(2\pi t) + a_4 \sqrt{2} \cos(2\pi t)$$

with $a_l \sim \mathcal{U}(-1/3^{l-1}, 1/3^{l-1})$ for all $l \in \{1, \dots, 4\}$, the model parameter is $\theta(t) = 2\sqrt{2} \cos(2\pi t)$, and $\epsilon \sim \mathcal{N}(0, \sigma^2)$ with $\sigma = 0.2\sqrt{E(\langle X, \theta \rangle^2)}$. The calculation of the standard FPCA estimator involves the eigenelements of the second moment operator of X (see (2.1) in Section 2) and, in this example, it can be shown that only the first four eigenvalues are different from zero. To analyse the effect of null eigenvalues, we simulated 200 samples of 100 observations and

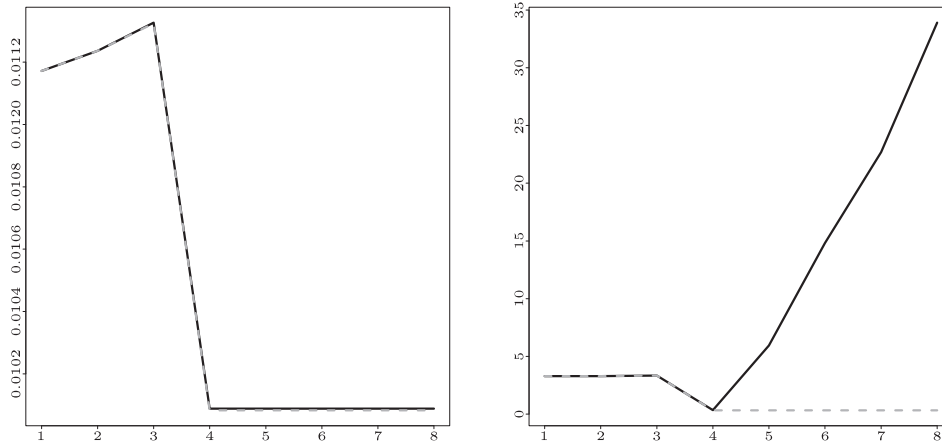


Figure 1. Mean of squared errors of prediction (left panel) and estimated mean square error of estimation (right panel) for the standard FPCA estimator (solid black line) and for the FPCA-type estimator with perturbed eigenvalues (dashed gray line).

computed the mean of squared errors of prediction and the mean square error of estimation (see (4.1) in Section 4) when $k_n \in \{1, \dots, 8\}$ eigenvalues of the second order operator are involved in the FPCA estimator. Figure 1 presents the results (black solid line). We also calculated the errors when the eigenvalues were slightly perturbed, adding $\alpha = 10^{-5}$ to them (see grey dashed line in Figure 1). It can be seen that the presence of null eigenvalues ($k_n > 4$), which hardly affects prediction error, considerably increases the estimation error of the FPCA estimator. Moreover, it seems that perturbation of eigenvalues allows us to keep small estimation errors, even when null eigenvalues are involved. We are thus led to developing a FPCA-type estimator based on a presmoothing method that avoids the inconvenience of ill-conditioning.

Some definitions and notations related to FPCA are given in Section 2, and we have listed results already present in literature for the conditional mean square errors of FPCA estimator, to facilitate comparison with our estimator. In Section 3, we propose a new estimate based on FPCA and presmoothing ideas. The section also contains theoretical results (consistency and mean square error expressions) that allow us to compare it with the standard FPCA estimator. Section 4 includes a simulation study to analyze the behaviour of our proposal from a practical point of view, and Section 5 presents data applications. Some conclusions are in Section 6, and an appendix compiles technical lemmas and the proofs of our results. In particular, Lemma A.1 in Appendix gives us expressions for the conditional errors of a wide class of estimators, including those discussed here. The generality of this lemma has two advantages: it has Theorem 1 and

Theorem 3 as almost direct corollaries, and it may be useful in further research on functional linear regression.

2. Functional Linear Model and FPCA

From now on, we are going to consider the functional linear model given by (1.1). Before presenting the presmoothing ideas that we have developed, we recall briefly some recent results on FPCA-type estimates. In particular, Theorem 1 is useful in comparing the theoretical performances of the new estimate with the standard one.

To analyse the standard FPCA estimator proposed by Cardot, Ferraty, and Sarda (1999), we need to define the second moment operator Γ , and the cross second moment operator Δ . The former is a nuclear, self-adjoint, and positive operator

$$\begin{aligned} \Gamma : H &\rightarrow H \\ x &\rightarrow \Gamma x = E(\langle X, x \rangle X), \end{aligned}$$

with $\{(\lambda_j, v_j)\}_{j=1}^{\infty}$ as its eigenvalues and eigenfunctions, respectively (and with $\lambda_1 \geq \lambda_2 \geq \dots$). The cross second moment operator is

$$\begin{aligned} \Delta : H &\rightarrow \mathbb{R} \\ x &\rightarrow \Delta x = E(\langle X, x \rangle Y). \end{aligned}$$

Given $\{(X_i, Y_i)\}_{i=1}^n$, a sample from (X, Y) , Γ and Δ are estimated by their empirical counterparts, $\Gamma_n x = n^{-1} \sum_{i=1}^n \langle X_i, x \rangle X_i$ and $\Delta_n x = n^{-1} \sum_{i=1}^n \langle X_i, x \rangle Y_i$. We denote by $\{(\hat{\lambda}_j, \hat{v}_j)\}_{j=1}^{\infty}$ the eigenvalues and eigenfunctions of Γ_n (with $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \dots \geq \hat{\lambda}_n \geq \hat{\lambda}_{n+1} = 0 = \dots$).

To estimate the model parameter, Cardot, Ferraty, and Sarda (2003) studied the optimization problem

$$\min_{\beta \in H} E[(Y - \langle \beta, X \rangle)^2].$$

When $\text{Ker}(\Gamma) = \{x \in H / \Gamma x = 0\} = \{0\}$ and $\sum_{j=1}^{\infty} (\Delta v_j / \lambda_j)^2 < +\infty$, the model parameter θ is the unique solution to this minimization problem, and it satisfies $\Delta x = \langle \theta, \Gamma x \rangle$ for all $x \in H$. This fact allows us to express the model parameter as

$$\theta = \sum_{j=1}^{\infty} \frac{\Delta v_j}{\lambda_j} v_j.$$

There is no bounded inverse of Γ , so Cardot, Ferraty, and Sarda (1999) projected the data on the subspace spanned by the first k_n eigenfunctions of Γ_n , and proposed the estimator

$$\hat{\theta}_{k_n} = \sum_{j=1}^{k_n} \frac{\Delta_n \hat{v}_j}{\hat{\lambda}_j} \hat{v}_j, \tag{2.1}$$

where $\{k_n\}_{n=1}^\infty$ is a sequence of positive integers such that $k_n \rightarrow +\infty$, $k_n \leq n$, and $\hat{\lambda}_{k_n} > 0$. Note that (2.1) is the truncated $\beta \in H$ that satisfies

$$\Delta_n x = \langle \beta, \Gamma_n x \rangle, \quad \forall x \in \text{Im}(\Gamma_n), \quad (2.2)$$

and that the estimator (2.1) converges almost surely (see Cardot, Ferraty, and Sarda (1999)).

Remark 1. The estimator $\hat{\theta}_{k_n}$ can be obtained by projecting the observations onto a finite subspace of H , and the same arguments involved in Faraldo-Roca and González-Manteiga (1987) allow one to reduce the conditional mean square error of the standard least squares estimator. Consider $\{e_j\}_{j=1}^\infty$ an orthonormal basis of H , and fix $k_n < n$. For all $x \in H$, the corresponding boldfaced letter \mathbf{x} denotes the k_n -vector given by $\mathbf{x} = (\langle x, e_1 \rangle, \dots, \langle x, e_{k_n} \rangle)^t$. Following the steps given by Cristóbal-Cristóbal, Faraldo-Roca, and González-Manteiga (1987), one considers the generalized optimization problem

$$\min_{\mathbf{b}} E_{\mu_n} [(\hat{m}_{k_n}(\mathbf{X}) - \mathbf{X}^t \mathbf{b})^2], \quad \text{with } \mathbf{b} = (\langle \beta, e_1 \rangle, \dots, \langle \beta, e_{k_n} \rangle)^t, \forall \beta \in H, \quad (2.3)$$

where $\hat{m}_{k_n}(\mathbf{x}) = \sum_{i=1}^n Y_i \delta(\mathbf{x}, \mathbf{X}_i) / \sum_{i=1}^n \delta(\mathbf{x}, \mathbf{X}_i)$ is a nonparametric estimator of the regression function $m_{k_n}(\mathbf{x}) = E(Y | \mathbf{X} = \mathbf{x})$ (here $\delta(\cdot, \cdot)$ is a measurable function from $\mathbb{R}^{k_n} \times \mathbb{R}^{k_n}$ into \mathbb{R}), and $\mu_n(\mathbf{x}) = \int_{-\infty}^{\mathbf{x}} \hat{f}_n(\mathbf{t}) d\mathbf{t}$ is a weighting function, with $\hat{f}_n(\mathbf{x}) = n^{-1} \sum_{i=1}^n \delta(\mathbf{x}, \mathbf{X}_i)$ a nonparametric estimator of the density f of \mathbf{X} . Solving (2.3) for the special case $\delta(\mathbf{u}, \mathbf{w}) = 1$ (resp. 0) if $\mathbf{u} = \mathbf{w}$ (resp. $\mathbf{u} \neq \mathbf{w}$) leads to the k_n -dimensional normal equation $\Delta_n = \Gamma_n \mathbf{b}$ where $\Gamma_n = n^{-1} \sum_{i=1}^n \mathbf{X}_i \mathbf{X}_i^t$ and $\Delta_n = n^{-1} \sum_{i=1}^n \mathbf{X}_i Y_i$. This amounts to k_n equations $\Delta_n e_l = \sum_{j=1}^{k_n} \langle \Gamma_n e_l, e_j \rangle \langle \beta, e_j \rangle$, for all $l \in \{1, \dots, k_n\}$. Taking $e_l = \hat{v}_l$, one gets $\langle \beta, \hat{v}_l \rangle = (\Delta_n \hat{v}_l) / \hat{\lambda}_l$ for all $l \in \{1, \dots, k_n\}$, and the estimator (2.1) appears as the projection of β onto the subspace of H spanned by $\hat{v}_1, \dots, \hat{v}_{k_n}$ (i.e. $\hat{\theta}_{k_n} = \sum_{j=1}^{k_n} \langle \beta, \hat{v}_j \rangle \hat{v}_j$); $\hat{\theta}_{k_n}$ is the solution of (2.3) for a basic nonparametric estimator (i.e. $\hat{m}_{k_n}(\mathbf{X}_i) = Y_i$). A natural extension of $\hat{\theta}_{k_n}$ consists of investigating solution of (2.3) for a general nonparametric estimator, equivalent to presmoothing the responses before estimating the functional parameter θ .

The next theorem gives the conditional mean square error of prediction for a new response $Y_{n+1} = \langle X_{n+1}, \theta \rangle + \epsilon_{n+1}$, and the conditional mean square error of estimation. For any $n \in \mathbb{N}^*$, set $\mathcal{X}^n = \{X_1, \dots, X_n\}$ and let $E_{\mathcal{X}^n}(\cdot)$ be the expectation conditionally on \mathcal{X}^n . Furthermore, let

$$\hat{R}_{k_n} = \sum_{j>k_n} \langle \theta, \hat{v}_j \rangle \hat{v}_j. \quad (2.4)$$

Theorem 1.

$$E_{\mathcal{X}^{n+1}}(Y_{n+1} - \langle X_{n+1}, \hat{\theta}_{k_n} \rangle)^2 = \sigma^2 + \frac{\sigma^2}{n} \sum_{j=1}^{k_n} \frac{\langle X_{n+1}, \hat{v}_j \rangle^2}{\hat{\lambda}_j} + \langle X_{n+1}, \hat{R}_{k_n} \rangle^2,$$

$$E_{\mathcal{X}^n}(\|\theta - \hat{\theta}_{k_n}\|^2) = \frac{\sigma^2}{n} \sum_{j=1}^{k_n} \frac{1}{\hat{\lambda}_j} + \|\hat{R}_{k_n}\|^2.$$

Remark 2. The theorem comes from a direct application of Lemma A.1 (see the Appendix), which allows us to obtain the conditional errors for a general type of estimate simply, and to compare their behaviours from a theoretical point of view. The conditional estimation error of $\hat{\theta}_{k_n}$ given in Theorem 1 has already been studied in literature. In fact, Theorem 5 in Hall and Hosseini-Nasab (2006) gives conditions in order to find that

$$E(\|\theta - \hat{\theta}_{k_n}\|^2) \sim \frac{\sigma^2}{n} \sum_{j=1}^{k_n} \frac{1}{\lambda_j} + \|R_{k_n}\|^2,$$

where $R_{k_n} = \sum_{j>k_n} \langle \theta, v_j \rangle v_j$, and $W_n \sim Z_n$ means that the ratio of W_n and Z_n converges to 1 when $n \rightarrow +\infty$.

3. Presmoothing via Covariance Structure

In this section, we are going to perturb the normal equation (2.2), and look for a function β in H such that

$$\Delta_n x = \langle \beta, (\Gamma_n + \alpha_n I)x \rangle, \quad \forall x \in H, \tag{3.1}$$

where α_n is a positive real sequence satisfying $\alpha_n \rightarrow 0$ when $n \rightarrow \infty$, and I is the identity operator in H . From (3.1), we derive the next estimator for θ ,

$$\hat{\theta}_{k_n}^{\alpha_n} = \sum_{j=1}^{k_n} \frac{\Delta_n \hat{v}_j}{\hat{\lambda}_j + \alpha_n} \hat{v}_j. \tag{3.2}$$

Note that both (3.1) and (3.2) can be seen as functional versions of the normal equation, and the derived estimator for the ordinary multivariate ridge regression with penalization term is α_n times the usual norm of the model parameter.

Remark 3. In order to build $\hat{\theta}_{k_n}^{\alpha_n}$, we can solve the optimization problem (2.3) using $\delta(\mathbf{u}, \mathbf{w}) = h_n^{-k_n} K^*(h_n^{-1}(\mathbf{u} - \mathbf{w}))$, where $K^*(\mathbf{x}) = \prod_{j=1}^{k_n} K(x_j)$ for all $\mathbf{x} = (x_1, \dots, x_{k_n})^t \in \mathbb{R}^{k_n}$, with $K : \mathbb{R} \rightarrow \mathbb{R}$ a symmetric positive kernel such that $\int K(z)dz = 1$, $\int zK(z)dz = 0$ and $\int z^2K(z)dz = c(K) < \infty$, and h_n a strictly

positive sequence of bandwidths. In this situation, the associated normal equation is

$$\Delta_n = (\Gamma_n + h_n^2 c(K) I_{k_n}) \mathbf{b}, \quad \text{with } I_{k_n} \text{ the } k_n \times k_n\text{-identity matrix.}$$

This equation can be expressed as $\Delta_n e_l = \sum_{j=1}^{k_n} \langle \Gamma_n e_l, e_j \rangle \langle \beta, e_j \rangle + h_n^2 c(K) \langle \beta, e_l \rangle$, for all $l \in \{1, \dots, k_n\}$. Selecting $e_l = \hat{v}_l$, we have $\Delta_n \hat{v}_l = (\hat{\lambda}_l + h_n^2 c(K)) \langle \beta, \hat{v}_l \rangle$, and (3.2) can be derived with $\alpha_n = h_n^2 c(K)$.

3.1. Consistency

In order to state convergence results, we need some notation. Let H' be the dual space of H and $\|\cdot\|_{H'}$ be the associated norm $\|T\|_{H'} = (\sum_{k=1}^{\infty} (Te_k)^2)^{1/2}$ for all $T \in H'$, where $\{e_k\}_{k=1}^{\infty}$ is an orthonormal basis in H . Let $\|T\|_{\infty} = \sup_{\|x\|=1} \|Tx\|$ for all $T \in \mathcal{H}$, where \mathcal{H} is the space of Hilbert-Schmidt operators defined on H . The following assumptions are required.

$$(H.1) \quad \lambda_1 > \lambda_2 > \dots > 0, \text{ and } \hat{\lambda}_1 > \hat{\lambda}_2 > \dots > \hat{\lambda}_{k_n} > 0 \text{ a.s..}$$

$$(H.2) \quad \|X\| \leq c_1 \text{ a.s..}$$

$$(H.3) \quad \exists c_2 > 0, \forall l \geq 1, E(|\epsilon|^l) < l! c_2 < +\infty.$$

$$(H.4) \quad n\lambda_{k_n}^4 / \log n \rightarrow +\infty \text{ and } n\lambda_{k_n}^2 / ((\sum_{j=1}^{k_n} a_j)^2 \log n) \rightarrow +\infty, \text{ where}$$

$$a_j = \begin{cases} \frac{2\sqrt{2}}{(\lambda_1 - \lambda_2)} & \text{if } j = 1, \\ \frac{2\sqrt{2}}{\min(\lambda_{j-1} - \lambda_j, \lambda_j - \lambda_{j+1})} & \text{if } j \neq 1. \end{cases}$$

$$(H.5) \quad \lambda_{k_n}^2 / \alpha_n \rightarrow \infty.$$

Theorem 2. *Under (H.1)–(H.5), $\|\hat{m}_{k_n}^{\alpha_n} - m\|_{H'} \rightarrow 0$ a.s., where $\hat{m}_{k_n}^{\alpha_n} = \langle \hat{\theta}_{k_n}^{\alpha_n}, \cdot \rangle$ and $m = \langle \theta, \cdot \rangle$.*

Remark 4. Cardot, Ferraty, and Sarda (1999) obtained consistency for the standard FPCA estimator $\hat{\theta}_{k_n}$ using (H.1), (H.2), (H.3') $|\epsilon| \leq c_2$ a.s., and (H.4). We have replaced the assumption (H.3') by (H.3) in order to consider non-bounded errors as gaussian. This modification can also be done in Cardot, Ferraty, and Sarda (1999), extending their results to a larger class of errors.

Remark 5. If $k_n = o(\log n)$, (H.4) is satisfied when $\lambda_j = c_1 c_2^j$ or $\lambda_j = c_1 j^{-c_3}$, with $c_1 > 0$, $0 < c_2 < 1$, and $c_3 > 1$.

3.2. Conditional errors

To obtain the conditional errors of prediction and estimation, we need the assumptions

$$(H.6) \quad \hat{\lambda}_{k_n}/\alpha_n \rightarrow \infty \text{ a.s.},$$

$$(H.7) \quad n\alpha_n \rightarrow 0.$$

Theorem 3. *Under (H.6), we have*

$$\begin{aligned} & E_{\mathcal{X}^{n+1}}(Y_{n+1} - \langle X_{n+1}, \hat{\theta}_{k_n}^{\alpha_n} \rangle)^2 - E_{\mathcal{X}^{n+1}}(Y_{n+1} - \langle X_{n+1}, \hat{\theta}_{k_n} \rangle)^2 \\ &= (-2\alpha_n \frac{\sigma^2}{n} \sum_{j=1}^{k_n} \frac{\langle X_{n+1}, \hat{v}_j \rangle^2}{\hat{\lambda}_j^2} + 2\alpha_n \langle X_{n+1}, \hat{R}_{k_n} \rangle \langle X_{n+1}, \hat{T}_{k_n} \rangle + \alpha_n^2 \langle X_{n+1}, \hat{T}_{k_n} \rangle^2) \\ &\quad \times (1 + o_{a.s.}(1)), \end{aligned} \quad (3.3)$$

$$\begin{aligned} & E_{\mathcal{X}^n}(\|\theta - \hat{\theta}_{k_n}^{\alpha_n}\|^2) - E_{\mathcal{X}^n}(\|\theta - \hat{\theta}_{k_n}\|^2) \\ &= (-2\alpha_n \frac{\sigma^2}{n} \sum_{j=1}^{k_n} \frac{1}{\hat{\lambda}_j^2} + \alpha_n^2 \|\hat{T}_{k_n}\|^2)(1 + o_{a.s.}(1)), \end{aligned} \quad (3.4)$$

where $\hat{T}_{k_n} = \sum_{j=1}^{k_n} \hat{\lambda}_j^{-1} \langle \theta, \hat{v}_j \rangle \hat{v}_j$, and \hat{R}_{k_n} is defined at (2.4).

Corollary 1. *Under (H.6) and (H.7),*

$$\begin{aligned} & E_{\mathcal{X}^{n+1}}(Y_{n+1} - \langle X_{n+1}, \hat{\theta}_{k_n}^{\alpha_n} \rangle)^2 - E_{\mathcal{X}^{n+1}}(Y_{n+1} - \langle X_{n+1}, \hat{\theta}_{k_n} \rangle)^2 \\ &= (-2\alpha_n \frac{\sigma^2}{n} \sum_{j=1}^{k_n} \frac{\langle X_{n+1}, \hat{v}_j \rangle^2}{\hat{\lambda}_j^2} + 2\alpha_n \langle X_{n+1}, \hat{R}_{k_n} \rangle \langle X_{n+1}, \hat{T}_{k_n} \rangle)(1 + o_{a.s.}(1)), \\ & E_{\mathcal{X}^n}(\|\theta - \hat{\theta}_{k_n}^{\alpha_n}\|^2) - E_{\mathcal{X}^n}(\|\theta - \hat{\theta}_{k_n}\|^2) = (-2\alpha_n \frac{\sigma^2}{n} \sum_{j=1}^{k_n} \frac{1}{\hat{\lambda}_j^2})(1 + o_{a.s.}(1)). \end{aligned}$$

Remark 6. Corollary 1 shows that the bias term \hat{R}_{k_n} plays a fundamental role in the conditional mean square error for prediction. In fact, our presmoothed estimator gives better or worse results than the standard FPCA estimator $\hat{\theta}_{k_n}$ depending on the order of $\langle X_{n+1}, \hat{R}_{k_n} \rangle$. Thus if $\langle X_{n+1}, \hat{R}_{k_n} \rangle (\alpha_n \langle X_{n+1}, \hat{T}_{k_n} \rangle)^{-1} = o_{a.s.}(1)$, then (3.3) becomes

$$\begin{aligned} & E_{\mathcal{X}^{n+1}}(Y_{n+1} - \langle X_{n+1}, \hat{\theta}_{k_n}^{\alpha_n} \rangle)^2 - E_{\mathcal{X}^{n+1}}(Y_{n+1} - \langle X_{n+1}, \hat{\theta}_{k_n} \rangle)^2 \\ &= (-2\alpha_n \frac{\sigma^2}{n} \sum_{j=1}^{k_n} \frac{\langle X_{n+1}, \hat{v}_j \rangle^2}{\hat{\lambda}_j^2} + \alpha_n^2 \langle X_{n+1}, \hat{T}_{k_n} \rangle^2)(1 + o_{a.s.}(1)). \end{aligned}$$

With $f(\alpha_n) = -2\alpha_n \frac{\sigma^2}{n} \sum_{j=1}^{k_n} \hat{\lambda}_j^{-2} \langle X_{n+1}, \hat{v}_j \rangle^2 + \alpha_n^2 \langle X_{n+1}, \hat{T}_{k_n} \rangle^2$, take $\alpha_n^{opt,1} = \arg \min_{\alpha} f(\alpha)$. It can be shown that $\alpha_n^{opt,1} = (\sigma^2/n) \sum_{j=1}^{k_n} \hat{\lambda}_j^{-2} \langle X_{n+1}, \hat{v}_j \rangle^2 \langle X_{n+1},$

$\hat{T}_{k_n}\rangle^{-2}$ and, in this case,

$$\begin{aligned} & E_{\mathcal{X}^{n+1}}((Y_{n+1} - \langle X_{n+1}, \hat{\theta}_{k_n}^{\alpha_n^{opt,1}} \rangle)^2) - E_{\mathcal{X}^{n+1}}((Y_{n+1} - \langle X_{n+1}, \hat{\theta}_{k_n} \rangle)^2) \\ &= \left(-\frac{\sigma^4}{n^2} \left(\sum_{j=1}^{k_n} \frac{\langle X_{n+1}, \hat{v}_j \rangle^2}{\hat{\lambda}_j^2}\right)^2 \langle X_{n+1}, \hat{T}_{k_n} \rangle^{-2}\right) (1 + o_{a.s.}(1)). \end{aligned}$$

Hence, there is a decrease in the prediction error using $\hat{\theta}_{k_n}^{\alpha_n^{opt,1}}$, and this reduction is more important when σ^2 is large and/or sample size n is small.

Remark 7. Under the assumptions of Corollary 1, we have that $\hat{\theta}_{k_n}^{\alpha_n}$ improves against $\hat{\theta}_{k_n}$ in terms of the conditional mean square error for estimation, when above all, σ^2 is large and/or n is small. Moreover, we can look for the value of α_n that minimizes $g(\alpha_n) = -2\alpha_n \frac{\sigma^2}{n} \sum_{j=1}^{k_n} \hat{\lambda}_j^{-2} + \alpha_n^2 \|\hat{T}_{k_n}\|^2$ in (3.4). This is $\alpha_n^{opt,2} = \sigma^2 n^{-1} (\sum_{j=1}^{k_n} \hat{\lambda}_j^{-2}) \|\hat{T}_{k_n}\|^{-2}$, for which we get

$$E_{\mathcal{X}^n}(\|\theta - \hat{\theta}_{k_n}^{\alpha_n^{opt,2}}\|^2) - E_{\mathcal{X}^n}(\|\theta - \hat{\theta}_{k_n}\|^2) = \left(-\frac{\sigma^4}{n^2} \left(\sum_{j=1}^{k_n} \hat{\lambda}_j^{-2}\right)^2 \|\hat{T}_{k_n}\|^{-2}\right) (1 + o_{a.s.}(1)).$$

Previous remarks ensure second order efficiency, that is, $\hat{\theta}_{k_n}^{\alpha_n}$ performs better than the standard FPCA estimator in small samples, and the same as $\hat{\theta}_{k_n}$ in large ones. Second order efficiency has already been achieved for linear regression estimators based on presmoothing in the real case (see Faraldo-Roca and González-Manteiga (1987) and Janssen, Swanepoel, and Veraverbeke (2001)). Presmoothing techniques have also allowed this kind of gain in efficiency in other contexts (see the presmoothed Nelson-Aalen estimator versus the classical Nelson-Aalen estimator in Cao-Abad et al. (2005)).

4. Simulation Study

This section is devoted to the presentation of two simulation studies for the regression model (1.3). The first case is characterized by the existence of null eigenvalues: $\lambda_j > 0$ if $j \in \{1, \dots, 4\}$, and $\lambda_j = 0$ otherwise (it corresponds to the motivation example introduced in Section 1). In the second case, the eigenvalues of Γ decrease quickly, but all of them are different from zero.

In both studies, we simulated 1,000 samples, each containing $2n$ observations drawn from the model (1.3) with $\epsilon \sim \mathcal{N}(0, \sigma^2)$ and signal-to-noise ratio $r = \sigma / \sqrt{E(\langle X, \theta \rangle^2)}$. We fixed different values for the signal-to-noise ratio r ($r = 0.02, 0.2, 0.5, 1, 2$), for the sample size n ($n = 25, 50, 100, 200$), and for different percentages of outliers ($out = 0\%, 10\%, 20\%$), constructed by changing the original model error by $\epsilon^* \sim \mathcal{N}(7, \sigma^2)$.

For each case, we calculated the standard estimator $\hat{\theta}_{k_n}$ (see (2.1)) and the presmoothed estimator $\hat{\theta}_{k_n}^{\alpha_n}$ (see (3.2)). We also computed the penalized B-spline estimator

$$\hat{\theta}_{PS} = \sum_{j=1}^{q+k} \hat{\beta}_j B_{k,j},$$

where $\{B_{k,j}, j = 1, \dots, k+q\}$ is the normalized B-splines basis of the space S_{qk} of splines with degree q and $k-1$ equispaced interior knots, and $\hat{\beta}$ is the solution of $\min_{\beta \in \mathbb{R}^{q+k}} (1/n) \sum_{i=1}^n (Y_i - \sum_{j=1}^{q+k} \langle \beta_j B_{k,j}, X_i \rangle)^2 + \rho \|B_k^{(m)'} \beta\|^2$ (see Cardot, Ferraty, and Sarda (2003)). We fixed $k = 20$, $q = 4$ and $m = 2$.

To compare the behaviour of these estimators, we used the risk functions

$$R(Y) = \frac{1}{n} \sum_{i=n+1}^{2n} (Y_i - \hat{Y}_i)^2, \quad \text{and} \quad R(\theta) = \int (\theta(t) - \hat{\theta}(t))^2 dt, \quad (4.1)$$

the mean of squared errors of prediction and the mean square error of estimation of θ , respectively. To calculate the first, for each simulated sample, we built each estimator $\hat{\theta}$ using $\{(X_i, Y_i)\}_{i=1}^n$ as a learning sample. The test sample $\{(X_i, Y_i)\}_{i=n+1}^{2n}$ then produced the prediction $\hat{Y}_i = \langle \hat{\theta}, X_i \rangle$ for each $i \in \{n+1, \dots, 2n\}$, and the corresponding value of $R(Y)$.

We used *GCV* for the choice of k_n and α_n , and calculated the ‘‘optimal’’ estimators, for the k_n and α_n chosen, to minimize $R(Y)$ and $R(\theta)$. For the penalized B-spline estimator, *GCV* was used to select ρ .

To present the results in a clear way, we show just the median of $R(Y)$ and $R(\theta)$ when

- r varies with $n = 100$ and $out = 10\%$,
- n varies with $r = 0.2$ and $out = 10\%$, and
- out varies with $n = 100$ and $r = 0.2$.

Remark 8. Since curves are not recorded continuously, X and θ were discretized on $p = 500$ equispaced design points (t_1, \dots, t_{500}) so that integrals involved had to be approximated. We used quadrature weights of p^{-1} , though more complex integral approximations are available. As a result, the eigenelements of Γ_n were approximated by those of the $p \times p$ -matrix $\mathbf{\Gamma}_n = (\gamma_{l_1, l_2})_{l_1=1, \dots, p; l_2=1, \dots, p}$ with $\gamma_{l_1, l_2} = (np)^{-1} \sum_{i=1}^n X_i(t_{l_1}) X_i(t_{l_2})$, and Δ_n was approximated by the p -vector $\mathbf{\Delta}_n = (\delta_l)_{l=1, \dots, p}$ where $\delta_l = (np)^{-1} \sum_{i=1}^n X_i(t_l) Y_i$. The risk functions (4.1) also depend on integral calculations and we took $R(Y) \approx n^{-1} \sum_{i=1}^n (Y_i - p^{-1} \sum_{l=1}^p \theta(t_l) X_i(t_l))^2$ and $R(\theta) \approx p^{-1} \sum_{l=1}^p (\theta(t_l) - \hat{\theta}(t_l))^2$.

4.1. Case A

For the curves

$$X(t) = a_1\sqrt{2}\sin(\pi t) + a_2\sqrt{2}\cos(\pi t) + a_3\sqrt{2}\sin(2\pi t) + a_4\sqrt{2}\cos(2\pi t)$$

with $a_l \sim \mathcal{U}(-1/3^{l-1}, 1/3^{l-1})$ for all $l \in \{1, \dots, 4\}$, and $\theta(t) = 2\sqrt{2}\cos(2\pi t)$, the eigenvalues of Γ are $\lambda_j = \text{Var}(a_j) = 1/3^{2j-1}$ for $j \in \{1, \dots, 4\}$, and $\lambda_j = 0$ for $j > 4$. Moreover, the model parameter is twice the fourth eigenfunction of the second moment operator. This is then a favourable case for the FPCA-type estimators if we select an adequate k_n .

The medians of the risk functions for the different cases are compiled in Table 1, Table 2, and Table 3. From them, one can see that both the FPCA-type estimators give better estimations for θ than the penalized B-spline estimator, as was expected given the connection between the model parameter and the fourth functional principal component. As for $R(Y)$, the three estimators have values of the same order.

If the *GCV* method is compared with the optimal selections, it can be seen that errors when the *GCV* choice is used are similar to the $\text{opt}_{R(Y)}$ ones in terms of prediction error, and generally higher than $\text{opt}_{R(\theta)}$ errors when the estimation error is concerned. This is reasonable since *GCV* techniques minimize criteria strongly linked with prediction error.

Centering on the two FPCA-type estimators, the results show that $\hat{\theta}_{k_n}^{\alpha_n}$ obtains smaller errors than $\hat{\theta}_{k_n}$ in most cases. Nevertheless, we must remark that the improvement is almost negligible for $R(Y)$, whereas it can be very significant for $R(\theta)$ when the signal-to-noise ratio is high (see Table 1), when the sample size is small (see Table 2), or when there is an important presence of outliers (see Table 3). These results coincide with the theoretical conclusions in Remark 7.

4.2. Case B

We simulated the model (1.3) with X a Brownian motion, and

$$\theta(t) = \log(15t^2 + 10) + \cos(4\pi t), \quad \forall t \in [0, 1].$$

It is well-known that the eigenelements of the second moment operator of a Brownian motion are

$$\lambda_j = \frac{1}{(j - 0.5)^2\pi^2}, \quad v_j(t) = \sqrt{2}\sin((j - 0.5)\pi t), \quad \forall t \in [0, 1], \quad j = 1, 2, \dots,$$

with all eigenvalues of Γ strictly positive.

From Tables 4, 5, and 6, we see that $\hat{\theta}_{k_n}^{\alpha_n}$ gives better estimates for θ than $\hat{\theta}_{k_n}$, according with the theoretical results. Again, simulations confirm the effect

Table 1. Case A: Fixed sample size $n = 100$ and percentage of outliers $out = 10\%$. Bold entries highlight the smallest errors for each selection method.

<i>error</i>	<i>r</i>	<i>GCV</i>			<i>opt_{R(Y)}</i>		<i>opt_{R(θ)}</i>	
		$\hat{\theta}_{PS}$	$\hat{\theta}_{k_n}$	$\hat{\theta}_{k_n}^{\alpha_n}$	$\hat{\theta}_{k_n}$	$\hat{\theta}_{k_n}^{\alpha_n}$	$\hat{\theta}_{k_n}$	$\hat{\theta}_{k_n}^{\alpha_n}$
<i>R(Y)</i>	0.02	4.98	4.99	4.96	4.94	4.88	4.96	4.92
	0.2	4.99	5.01	4.97	4.96	4.89	4.98	4.93
	0.5	5.05	5.07	5.05	5.02	4.95	5.03	5.00
	1	5.27	5.28	5.25	5.22	5.14	5.24	5.19
	2	6.01	6.03	5.99	5.96	5.84	5.98	5.92
<i>R(θ)</i>	0.02	6.84	4.42	3.64	5.50	3.65	4.03	3.31
	0.2	6.76	4.57	3.65	5.38	3.69	4.02	3.32
	0.5	6.75	4.54	3.67	5.57	3.66	4.04	3.31
	1	6.84	4.60	3.65	5.55	3.70	4.05	3.32
	2	7.25	4.69	3.72	6.11	3.72	4.14	3.32

Table 2. Case A: Fixed ratio $r = 0.2$ and percentage of outliers $out = 10\%$. Bold entries highlight the smallest errors for each selection method.

<i>error</i>	<i>n</i>	<i>GCV</i>			<i>opt_{R(Y)}</i>		<i>opt_{R(θ)}</i>	
		$\hat{\theta}_{PS}$	$\hat{\theta}_{k_n}$	$\hat{\theta}_{k_n}^{\alpha_n}$	$\hat{\theta}_{k_n}$	$\hat{\theta}_{k_n}^{\alpha_n}$	$\hat{\theta}_{k_n}$	$\hat{\theta}_{k_n}^{\alpha_n}$
<i>R(Y)</i>	25	4.26	4.27	4.09	4.10	3.87	4.18	3.97
	50	5.08	5.09	4.99	4.99	4.86	5.03	4.94
	100	4.99	5.01	4.97	4.96	4.89	4.98	4.93
	200	4.95	4.95	4.94	4.93	4.90	4.94	4.93
<i>R(θ)</i>	25	13.24	8.50	3.86	12.05	3.90	6.30	3.37
	50	9.52	6.02	3.82	8.70	3.83	4.78	3.35
	100	6.76	4.57	3.65	5.38	3.69	4.02	3.32
	200	5.75	3.89	3.48	4.35	3.53	3.63	3.29

Table 3. Case A: Fixed sample size $n = 100$ and ratio $r = 0.2$. Bold entries highlight the smallest errors for each selection method.

<i>error</i>	<i>out</i>	<i>GCV</i>			<i>opt_{R(Y)}</i>		<i>opt_{R(θ)}</i>	
		$\hat{\theta}_{PS}$	$\hat{\theta}_{k_n}$	$\hat{\theta}_{k_n}^{\alpha_n}$	$\hat{\theta}_{k_n}$	$\hat{\theta}_{k_n}^{\alpha_n}$	$\hat{\theta}_{k_n}$	$\hat{\theta}_{k_n}^{\alpha_n}$
<i>R(Y)</i>	0%	0.01	0.01	0.01	0.01	0.01	0.01	0.01
	10%	4.99	5.01	4.97	4.96	4.89	4.98	4.93
	20%	9.97	9.99	9.92	9.90	9.78	9.94	9.85
<i>R(θ)</i>	0%	0.39	0.24	0.28	0.99	0.25	0.24	0.10
	10%	6.76	4.57	3.65	5.38	3.69	4.02	3.32
	20%	8.49	5.46	3.82	7.58	3.79	4.61	3.35

of the sample size and noise on the expected reduction for estimation of θ (see Remark 7): more improvement when n is small (see Table 5) when the “noise level” is large (see Tables 4 and 6). As happens in Case A, $\hat{\theta}_{k_n}$ and $\hat{\theta}_{k_n}^{\alpha_n}$ give

Table 4. Case B: Fixed sample size $n = 100$ and percentage of outliers $out = 10\%$. Bold entries highlight the smallest errors for each selection method.

error	r	GCV			$opt_{R(Y)}$		$opt_{R(\theta)}$	
		$\hat{\theta}_{PS}$	$\hat{\theta}_{k_n}$	$\hat{\theta}_{k_n}^{\alpha_n}$	$\hat{\theta}_{k_n}$	$\hat{\theta}_{k_n}^{\alpha_n}$	$\hat{\theta}_{k_n}$	$\hat{\theta}_{k_n}^{\alpha_n}$
$R(Y)$	0.02	5.00	5.01	4.99	4.93	4.88	4.97	4.94
	0.2	5.13	5.16	5.13	5.04	4.99	5.09	5.07
	0.5	5.70	5.71	5.68	5.59	5.52	5.64	5.61
	1	7.66	7.70	7.65	7.52	7.42	7.59	7.53
	2	15.43	15.45	15.32	15.00	14.81	15.17	15.06
$R(\theta)$	0.02	2.42	2.40	2.31	3.11	2.06	1.84	1.33
	0.2	2.45	2.55	2.37	3.15	2.17	1.83	1.34
	0.5	2.75	2.67	2.37	3.34	2.12	1.92	1.33
	1	3.53	3.40	2.67	4.29	2.36	2.15	1.44
	2	6.27	5.55	3.32	7.20	3.15	3.18	1.54

Table 5. Case B: Fixed ratio $r = 0.2$ and percentage of outliers $out = 10\%$. Bold entries highlight the smallest errors for each selection method.

error	n	GCV			$opt_{R(Y)}$		$opt_{R(\theta)}$	
		$\hat{\theta}_{PS}$	$\hat{\theta}_{k_n}$	$\hat{\theta}_{k_n}^{\alpha_n}$	$\hat{\theta}_{k_n}$	$\hat{\theta}_{k_n}^{\alpha_n}$	$\hat{\theta}_{k_n}$	$\hat{\theta}_{k_n}^{\alpha_n}$
$R(Y)$	25	4.50	4.48	4.38	4.15	3.95	4.29	4.14
	50	5.24	5.28	5.23	5.08	4.96	5.14	5.09
	100	5.13	5.16	5.13	5.04	4.99	5.09	5.07
	200	5.07	5.08	5.07	5.03	5.00	5.05	5.04
$R(\theta)$	25	9.56	8.07	4.64	9.09	3.49	3.49	1.67
	50	4.96	4.39	3.23	4.88	2.64	2.60	1.51
	100	2.45	2.55	2.37	3.15	2.17	1.83	1.34
	200	1.49	1.83	1.84	2.05	1.68	1.49	1.23

similar results in terms of $R(Y)$.

Comparing our presmoothed estimator with $\hat{\theta}_{PS}$, behaviour is similar as far as prediction error is concerned, while $\hat{\theta}_{k_n}^{\alpha_n}$ gives smaller estimation errors than the penalized B-spline estimator when n is small or the noise is large.

These simulations suggest that the presmoothing estimate improves the standard FPCA linear estimate, and especially when the sample size is small. The choice of the parameters of the presmoothed estimate is of course a key point. A general practical guideline is to choose these parameters by cross-validation techniques. Even if the results above might point to a new way for selecting these parameters that could give even more improvement (see results for the “optimal” choice with respect $R(\theta)$), CV gives good data-driven results.

Table 6. Case B: Fixed sample size $n = 100$ and ratio $r = 0.2$. Bold entries highlight the smallest errors for each selection method.

error	out	GCV			$opt_{R(Y)}$		$opt_{R(\theta)}$	
		$\hat{\theta}_{PS}$	$\hat{\theta}_{k_n}$	$\hat{\theta}_{k_n}^{\alpha_n}$	$\hat{\theta}_{k_n}$	$\hat{\theta}_{k_n}^{\alpha_n}$	$\hat{\theta}_{k_n}$	$\hat{\theta}_{k_n}^{\alpha_n}$
$R(Y)$	0%	0.11	0.11	0.11	0.11	0.10	0.11	0.11
	10%	5.13	5.16	5.13	5.04	4.99	5.09	5.07
	20%	10.16	10.19	10.13	10.00	9.86	10.06	9.99
$R(\theta)$	0%	0.56	1.21	1.14	1.14	1.05	0.87	0.66
	10%	2.45	2.55	2.37	3.15	2.17	1.83	1.34
	20%	4.62	3.83	3.02	4.65	2.55	2.47	1.46

5. Data Applications

In order to demonstrate the improved performance of our methodology with respect to the FPCA estimator in applications, we chose three functional datasets with different sized n : Canadian weather data ($n = 35$), spectrometric data ($n = 215$), and atmospheric pollution data ($n = 1,000$).

Since the variables involved are not centered, we used the regression model (1.2), and proceeded through the following steps.

- Step 1. Calculate the sample mean of curves (\bar{X}) and scalar responses (\bar{Y}).
- Step 2. Split the sample into a learning sample $\{(X_i, Y_i)\}_{i \in I_{LS}}$, and a testing sample $\{(X_i, Y_i)\}_{i \in I_{TS}}$.
- Step 3. Use the centered learning sample $\{(X_i - \bar{X}, Y_i - \bar{Y})\}_{i \in I_{LS}}$ to build $\hat{\theta}$ ($\hat{\theta}_{PS}$, $\hat{\theta}_{k_n}$ or $\hat{\theta}_{k_n}^{\alpha_n}$), and estimate the intercept term by $\hat{\theta}_0 = \bar{Y} - \langle \hat{\theta}, \bar{X} \rangle$.
- Step 4. Compute the predicted responses for the testing sample, $\hat{Y}_i = \hat{\theta}_0 + \langle \hat{\theta}, X_i \rangle$, $\forall i \in I_{TS}$.
- Step 5. Obtain $R(Y) = (1/\#I_{TS}) \sum_{i \in I_{TS}} (Y_i - \hat{Y}_i)^2$, with $\#I_{TS}$ the testing sample size.

To avoid the effect of sample selection, this procedure was iterated 200 times, and the mean, the median, and the standard deviation of $R(Y)$ over replications were calculated.

We followed the general guidelines of the simulations study to build the different estimates of θ . In particular, the involved parameters were selected by cross-validation.

5.1. Canadian weather data

The Canadian weather data is available in the R package *fda*. This dataset contains daily temperature and precipitation at 35 different locations in Canada averaged over 1960 to 1994. Here Y is the logarithm of total annual precipitation

Table 7. Canadian weather data: mean, median, and standard deviation of $R(Y)$.

<i>error</i>	$\hat{\theta}_{PS}$	$\hat{\theta}_{k_n}$	$\hat{\theta}_{k_n}^{\alpha_n}$
<i>mean</i> ($R(Y)$)	0.037	0.03566	0.03293
<i>median</i> ($R(Y)$)	0.036	0.03466	0.03230
<i>sd</i> ($R(Y)$)	0.018	0.01579	0.01436

Table 8. Spectrometric data: mean, median, and standard deviation of $R(Y)$.

<i>error</i>	$\hat{\theta}_{PS}$	$\hat{\theta}_{k_n}$	$\hat{\theta}_{k_n}^{\alpha_n}$
<i>mean</i> ($R(Y)$)	6.656	6.926	6.604
<i>median</i> ($R(Y)$)	6.313	6.524	6.088
<i>sd</i> ($R(Y)$)	1.935	2.153	1.939

at each weather station, and X is the daily temperature curve (see Ramsay and Silverman (2005, Chap. 15) for a Fourier basis approach to this case). The original sample was split into two subsamples: a learning sample (25 stations) and a testing one (10 stations). Table 7 shows the mean, median, and standard deviation of $R(Y)$. Bold entries highlight the smallest values for the mean of squared error of prediction; these were generated by $\hat{\theta}_{k_n}^{\alpha_n}$.

5.2. Spectrometric data

The data contain 215 spectrometric curves obtained from pieces of finely chopped meat, and a scalar value corresponding to the fat content of each of them (this dataset is available at <http://www.math.univ-toulouse.fr/staph/npfd>). In the chemometric community, it is well-known that derivatives of NIR spectra are more informative than the original ones. Therefore, in our regression model, X_i is the second derivative of the spectrometric curve for each meat piece, and Y_i is its fat content (Ferraty and Vieu (2006) also chose the second derivatives in order to get better predictive results in the functional nonparametric regression context). We split the sample 200 times into a learning sample composed of 160 pieces and a test sample composed of 55 pieces. Table 8 contains the results for this example (bold entries correspond to the smallest errors).

5.3. Atmospheric pollution data

The data here correspond to hourly averaged NO_x concentrations measured in the neighbourhood of a power station of ENDESA, located in As Pontes in the Northwest of Spain, from 2007 to 2009. During unfavorable meteorological conditions, NO_x levels can rise quickly and cause an air-quality episode. The aim is to forecast NO_x with half an hour horizon to allow the power plant staff to preclude NO_x concentrations reaching the limits fixed by the current environmental

Table 9. Atmospheric pollution data: mean, median, and standard deviation of $R(Y)$.

<i>error</i>	$\hat{\theta}_{PS}$	$\hat{\theta}_{k_n}$	$\hat{\theta}_{k_n}^{\alpha_n}$
<i>mean</i> ($R(Y)$)	2.865	3.238	3.207
<i>median</i> ($R(Y)$)	2.780	3.074	3.062
<i>sd</i> ($R(Y)$)	0.966	0.866	0.833

legislation (a functional kernel and linear autoregressive approach to this problem with SO_2 levels is in Fernández de Castro, Guillas and González-Manteiga (2005)).

Here we have a time series $Z(t)$ that contains the hourly averaged NO_x concentrations: for each minute t , $Z(t)$ is the average of the NO_x measurements recorded over the last hour. Each observation (X_i, Y_i) is built as follows. The curve X_i consists of a 4-hour period of the time series Z , and the response Y_i is the value of the time series half an hour ahead, that is, $X_i(t) = Z(240 \times (i-1) + t)$ with $t \in \{1, \dots, 240\}$, and $Y_i = Z(240 \times i + 30)$. As in the previous examples, we have considered 200 different pairs of learning/testing samples composed of 750 and 250 observations, respectively, and the results are compiled in Table 9, where the smallest errors are in boldface.

Remark 9. In Subsection 3.2, the second order efficiency of our estimator $\hat{\theta}_{k_n}^\alpha$ was stated. We also indicated the existence of an inverse relation between the improvement of $\hat{\theta}_{k_n}^\alpha$ with respect to $\hat{\theta}_{k_n}$, and the sample size n . This fact is reflected in the real data applications: the gain of our estimator decreases from small to large sample size (see Table 7, Table 8 and Table 9).

6. Comments

We have proposed a new FPCA-estimator for the linear model parameter θ . Our methodology can be seen as an extension of the ordinary multivariate ridge regression estimator to general Hilbert spaces: we slightly perturbate the eigenvalues of the second moment operator in order to avoid ill-conditioned problems.

We have shown that our estimator conserves the consistency properties of the standard FPCA estimator (see Theorem 2), and we have obtained expressions for conditional mean square errors for prediction and estimation (see Theorem 3). The proof of Theorem 3 is derived from a general result (see Lemma A.1, Appendix), that seems certain to be of use in further research on functional linear regression. Remark 6 highlights the effect of the bias term in the conditional error of prediction: we can only obtain clear efficiency when the bias is negligible. As far as the conditional error of estimation is concerned, we are able to

get improvement over the FPCA estimate, especially if the model noise is large and/or the sample size is small.

We have tested the effectiveness of the presmoothed estimator relative to the standard FPCA estimator and the penalized B-spline estimator by means of simulation studies and data applications. In terms of conditional error of estimation, our proposal gives better results than the FPCA estimator, and it is a serious rival to the spline estimator when the sample size is small or the noise level is large. Second order efficiency only generates a clear improvement for small sample sizes, whereas for n large enough, our estimator and the FPCA estimator have similar behaviour.

Acknowledgement

The authors would like to thank the two reviewers and an Associate Editor for their fruitful and interesting comments. The first and fourth authors wish to thank the participants in the working group STAPH in Toulouse for continuous support and comments. The work of the second and the third authors was supported by Ministerio de Ciencia e Innovación (grant MTM2008-03010) and by Consellería de Innovación e Industria (regional grant PGIDIT07PXIB207031PR), and Consellería de Economía e Industria (regional grant 10MDS207015PR), Xunta de Galicia.

Appendix

We present the proofs of the main results in the paper. We also give a general result (Lemma A.1) that covers the standard FPCA estimator and the presmoothed FPCA estimator discussed in this paper (and useful for other purposes in the functional linear regression context), and some auxiliary lemmas.

A.1. Proof of Theorem 1

Theorem 1 is easily shown using Lemma A.1 with $\gamma_j = \hat{\lambda}_j^{-1}$ and $w_j = \hat{v}_j$. In this case, $R_{k_n}^{(\gamma, \omega)} = \theta - \sum_{j=1}^{k_n} \hat{\lambda}_j^{-1} \langle \Gamma_n \hat{v}_j, \theta \rangle \hat{v}_j = \hat{R}_{k_n}$ where \hat{R}_{k_n} is defined at (2.4). Then, for the conditional error of prediction,

$$\begin{aligned} E_{\mathcal{X}^{n+1}}(Y_{n+1} - \langle X_{n+1}, \hat{\theta}_{k_n} \rangle)^2 &= \sigma^2 \\ &+ \frac{\sigma^2}{n} \sum_{j_1=1}^{k_n} \sum_{j_2=1}^{k_n} \frac{\langle \Gamma_n \hat{v}_{j_1}, \hat{v}_{j_2} \rangle}{\hat{\lambda}_{j_1} \hat{\lambda}_{j_2}} \langle X_{n+1}, \hat{v}_{j_1} \rangle \langle X_{n+1}, \hat{v}_{j_2} \rangle + \langle X_{n+1}, \hat{R}_{k_n} \rangle^2 \\ &= \sigma^2 + \frac{\sigma^2}{n} \sum_{j=1}^{k_n} \frac{\langle X_{n+1}, \hat{v}_j \rangle^2}{\hat{\lambda}_j} + \langle X_{n+1}, \hat{R}_{k_n} \rangle^2, \end{aligned}$$

and for the conditional error of estimation,

$$\begin{aligned} E_{\mathcal{X}^n}(\|\theta - \hat{\theta}_{k_n}\|^2) &= \frac{\sigma^2}{n} \sum_{j_1=1}^{k_n} \sum_{j_2=1}^{k_n} \frac{\langle \Gamma_n \hat{v}_{j_1}, \hat{v}_{j_2} \rangle}{\hat{\lambda}_{j_1} \hat{\lambda}_{j_2}} \langle \hat{v}_{j_1}, \hat{v}_{j_2} \rangle + \|\hat{R}_{k_n}\|^2 \\ &= \frac{\sigma^2}{n} \sum_{j=1}^{k_n} \frac{1}{\hat{\lambda}_j} + \|\hat{R}_{k_n}\|^2. \end{aligned}$$

A.2. Proof of Theorem 2

This proof is similar to that of Theorem 3.2 in Cardot, Ferraty, and Sarda (1999). First, take $m_{k_n} = \Delta \Pi_{k_n} (\Pi_{k_n} \Gamma \Pi_{k_n})^{-1}$, where Π_{k_n} is the orthogonal projection onto the space spanned by the first k_n eigenfunctions of Γ . It is clear that

$$\|m - \hat{m}_{k_n}^{\alpha_n}\|_{H'} \leq \|m - m_{k_n}\|_{H'} + \|m_{k_n} - \hat{m}_{k_n}^{\alpha_n}\|_{H'}.$$

Cardot, Ferraty, and Sarda (1999) showed that $\|m - m_{k_n}\|_{H'} \rightarrow 0$, so we just need to show that $\|m_{k_n} - \hat{m}_{k_n}^{\alpha_n}\|_{H'} \rightarrow 0$.

Let $E_n = \{\lambda_{k_n}/2 < \hat{\lambda}_{k_n} < 3\lambda_{k_n}/2\}$. In E_n , Lemma A.2 ensures that

$$\|m_{k_n} - \hat{m}_{k_n}^{\alpha_n}\|_{H'} \leq \delta_n \|\Delta\|_{H'} \|\Gamma - \Gamma_n\|_{\infty} + 2 \frac{\|\Delta - \Delta_n\|_{H'}}{\lambda_{k_n}} + 2\alpha_n \frac{\|\Delta\|_{H'}}{\lambda_{k_n}^2},$$

with $\delta_n = 2/\lambda_{k_n}^2 + 6 \sum_{j=1}^{k_n} a_j/\lambda_{k_n}$. Therefore

$$\begin{aligned} &P(\|m_{k_n} - \hat{m}_{k_n}^{\alpha_n}\|_{H'} > \eta) \\ &\leq P\left(\|\Gamma - \Gamma_n\|_{\infty} > \frac{\eta}{(3\delta_n \|\Delta\|_{H'})}\right) \\ &\quad + P\left(\|\Delta - \Delta_n\|_{H'} > \frac{\lambda_{k_n} \eta}{6}\right) + 1_{\{\alpha_n > \eta \lambda_{k_n}^2 / (6 \|\Delta\|_{H'})\}} + P(\bar{E}_n). \end{aligned} \quad (\text{A.1})$$

It can be shown that

$$P(\bar{E}_n) \leq P(\|\Gamma - \Gamma_n\|_{\infty} > \lambda_{k_n}/2) \leq 2 \exp(-Cn\lambda_{k_n}^2), \quad (\text{A.2})$$

where C is a positive constant independent of n , and the last inequality is derived from Lemma 5.3 in Cardot, Ferraty, and Sarda (1999) Furthermore,

$$P\left(\|\Gamma - \Gamma_n\|_{\infty} > \frac{\eta}{3\delta_n \|\Delta\|_{H'}}\right) \leq 2 \exp\left(-\frac{A_{\eta} n}{\delta_n^2}\right), \quad (\text{A.3})$$

$$P\left(\|\Delta - \Delta_n\|_{H'} > \frac{\lambda_{k_n} \eta}{6}\right) \leq 2 \exp(-B_{\eta} n \lambda_{k_n}^2), \quad (\text{A.4})$$

where A_η and B_η are positive constants independent of n . These two inequalities are obtained using Lemma 5.3 in Cardot, Ferraty, and Sarda (1999), and Lemma A.3. Using (A.2), (A.3), and (A.4) in (A.1), we get

$$\begin{aligned} & P(\|m_{k_n} - \hat{m}_{k_n}^{\alpha_n}\|_{H'} > \eta) \\ & \leq 2 \exp\left(-\frac{A_\eta n}{\delta_n^2}\right) + 2 \exp(-B_\eta n \lambda_{k_n}^2) + 1_{\{\alpha_n > \eta \lambda_{k_n}^2 / (6\|\Delta\|_{H'})\}} + 2 \exp(-Cn \lambda_{k_n}^2). \end{aligned}$$

Cardot, Ferraty, and Sarda (1999) showed $\exp(-A_\eta n / \delta_n^2)$, $\exp(-B_\eta n \lambda_{k_n}^2)$, and $\exp(-Cn \lambda_{k_n}^2)$ are general terms of convergent series under hypotheses (H.4). On the other hand, under (H.5), there is an n_0 such that $\alpha_n < \eta \lambda_{k_n}^2 / (6\|\Delta\|_{H'})$ for all $n > n_0$. Therefore $\sum_{n \in \mathbb{N}^*} P(\|m_{k_n} - \hat{m}_{k_n}^{\alpha_n}\|_{H'} > \eta) < \infty$, and Borel-Cantelli Lemma give us $\|m_{k_n} - \hat{m}_{k_n}^{\alpha_n}\|_{H'} \rightarrow 0$ a.s..

A.3. Proof of Theorem 3

Consider Lemma A.1, with $\gamma_j = (\hat{\lambda}_j + \alpha_n)^{-1}$ and $w_j = \hat{v}_j$. Then

$$R_{k_n}^{(\gamma, \omega)} = \theta - \sum_{j=1}^{k_n} \frac{\hat{\lambda}_j}{\hat{\lambda}_j + \alpha_n} \langle \hat{v}_j, \theta \rangle \hat{v}_j = \hat{R}_{k_n} + \sum_{j=1}^{k_n} \frac{\alpha_n}{\hat{\lambda}_j + \alpha_n} \langle \hat{v}_j, \theta \rangle \hat{v}_j,$$

with \hat{R}_{k_n} defined at (2.4). We then obtain for the conditional error of prediction,

$$\begin{aligned} E_{\mathcal{X}^{n+1}}(Y_{n+1} - \langle X_{n+1}, \hat{\theta}_{k_n}^{\alpha_n} \rangle)^2 &= \sigma^2 + \frac{\sigma^2}{n} \sum_{j=1}^{k_n} \frac{\hat{\lambda}_j}{(\hat{\lambda}_j + \alpha_n)^2} \langle X_{n+1}, \hat{v}_j \rangle^2 \\ &\quad + \langle X_{n+1}, \hat{R}_{k_n} + \sum_{j=1}^{k_n} \frac{\alpha_n}{\hat{\lambda}_j + \alpha_n} \langle \hat{v}_j, \theta \rangle \hat{v}_j \rangle^2. \end{aligned} \quad (\text{A.5})$$

Some calculations and (H.6) allow us to obtain

$$\begin{aligned} \frac{\hat{\lambda}_j}{(\hat{\lambda}_j + \alpha_n)^2} &= \frac{1}{\hat{\lambda}_j} - 2 \frac{\alpha_n}{\hat{\lambda}_j^2} + \frac{\alpha_n^2}{\hat{\lambda}_j(\hat{\lambda}_j + \alpha_n)^2} \left(3 + 2 \frac{\alpha_n}{\hat{\lambda}_j}\right) \\ &= \frac{1}{\hat{\lambda}_j} - 2 \frac{\alpha_n}{\hat{\lambda}_j^2} (1 + o_{a.s.}(1)), \end{aligned} \quad (\text{A.6})$$

$$\frac{\alpha_n}{\hat{\lambda}_j + \alpha_n} = \frac{\alpha_n}{\hat{\lambda}_j} (1 - \alpha_n (\hat{\lambda}_j + \alpha_n)^{-1}) = \frac{\alpha_n}{\hat{\lambda}_j} (1 + o_{a.s.}(1)). \quad (\text{A.7})$$

Using (A.6) and (A.7) in (A.5), we get

$$\begin{aligned} & E_{\mathcal{X}^{n+1}}((Y_{n+1} - \langle X_{n+1}, \hat{\theta}_{k_n}^{\alpha_n} \rangle)^2) \\ &= \sigma^2 + \frac{\sigma^2}{n} \sum_{j=1}^{k_n} \frac{\langle X_{n+1}, \hat{v}_j \rangle^2}{\hat{\lambda}_j} - 2\alpha_n \frac{\sigma^2}{n} \sum_{j=1}^{k_n} \frac{\langle X_{n+1}, \hat{v}_j \rangle^2}{\hat{\lambda}_j^2} (1 + o_{a.s.}(1)) + \langle X_{n+1}, \hat{R}_{k_n} \rangle^2 \\ &\quad + 2\alpha_n \langle X_{n+1}, \hat{R}_{k_n} \rangle \langle X_{n+1}, \hat{T}_{k_n} \rangle (1 + o_{a.s.}(1)) + \alpha_n^2 \langle X_{n+1}, \hat{T}_{k_n} \rangle^2 (1 + o_{a.s.}(1)), \end{aligned}$$

where $\hat{T}_{k_n} = \sum_{j=1}^{k_n} \hat{\lambda}_j^{-1} \langle \theta, \hat{v}_j \rangle \hat{v}_j$. Comparing this expression with the conditional error of prediction for $\hat{\theta}_{k_n}$ given in Theorem 1, we get

$$\begin{aligned} & E_{\mathcal{X}^{n+1}}((Y_{n+1} - \langle X_{n+1}, \hat{\theta}_{k_n}^{\alpha_n} \rangle)^2) - E_{\mathcal{X}^{n+1}}((Y_{n+1} - \langle X_{n+1}, \hat{\theta}_{k_n} \rangle)^2) \\ &= -2\alpha_n \frac{\sigma^2}{n} \sum_{j=1}^{k_n} \frac{\langle X_{n+1}, \hat{v}_j \rangle^2}{\hat{\lambda}_j^2} (1 + o_{a.s.}(1)) \\ &\quad + 2\alpha_n \langle X_{n+1}, \hat{R}_{k_n} \rangle \langle X_{n+1}, \hat{T}_{k_n} \rangle (1 + o_{a.s.}(1)) + \alpha_n^2 \langle X_{n+1}, \hat{T}_{k_n} \rangle^2 (1 + o_{a.s.}(1)). \end{aligned}$$

On the other hand, for the conditional error of estimation Lemma A.1 implies

$$E_{\mathcal{X}^n}(\|\theta - \hat{\theta}_{k_n}^{\alpha_n}\|^2) = \frac{\sigma^2}{n} \sum_{j=1}^{k_n} \frac{\hat{\lambda}_j}{(\hat{\lambda}_j + \alpha_n)^2} + \|\hat{R}_{k_n}\|^2 + \left\| \sum_{j=1}^{k_n} \frac{\alpha_n}{\hat{\lambda}_j + \alpha_n} \langle \hat{v}_j, \theta \rangle \hat{v}_j \right\|^2. \quad (\text{A.8})$$

Using (A.6) and (A.7) in (A.8), we have

$$\begin{aligned} E_{\mathcal{X}^n}(\|\theta - \hat{\theta}_{k_n}^{\alpha_n}\|^2) &= \frac{\sigma^2}{n} \sum_{j=1}^{k_n} \frac{1}{\hat{\lambda}_j} - 2\alpha_n \frac{\sigma^2}{n} \sum_{j=1}^{k_n} \frac{1}{\hat{\lambda}_j^2} (1 + o_{a.s.}(1)) + \|\hat{R}_{k_n}\|^2 \\ &\quad + \alpha_n^2 \|\hat{T}_{k_n}\|^2 (1 + o_{a.s.}(1)), \end{aligned}$$

with \hat{T}_{k_n} defined as before. Bearing in mind Theorem 1,

$$\begin{aligned} & E_{\mathcal{X}^n}(\|\theta - \hat{\theta}_{k_n}^{\alpha_n}\|^2) - E_{\mathcal{X}^n}(\|\theta - \hat{\theta}_{k_n}\|^2) \\ &= -2\alpha_n \frac{\sigma^2}{n} \sum_{j=1}^{k_n} \frac{1}{\hat{\lambda}_j^2} (1 + o_{a.s.}(1)) + \alpha_n^2 \|\hat{T}_{k_n}\|^2 (1 + o_{a.s.}(1)). \end{aligned}$$

A.4. Proof of Corollary 1

Note that

$$\begin{aligned} \langle X_{n+1}, \hat{T}_{k_n} \rangle^2 &= \left(\sum_{j=1}^{k_n} \frac{\langle \theta, \hat{v}_j \rangle}{\hat{\lambda}_j} \langle X_{n+1}, \hat{v}_j \rangle \right)^2 \leq \left(\sum_{j=1}^{k_n} \langle \theta, \hat{v}_j \rangle^2 \right) \left(\sum_{j=1}^{k_n} \frac{\langle X_{n+1}, \hat{v}_j \rangle^2}{\hat{\lambda}_j^2} \right) \\ &\leq \|\theta\|^2 \sum_{j=1}^{k_n} \frac{\langle X_{n+1}, \hat{v}_j \rangle^2}{\hat{\lambda}_j^2}, \end{aligned}$$

and so $\alpha_n \langle X_{n+1}, \hat{T}_{k_n} \rangle^2 (n^{-1} \sum_{j=1}^{k_n} \hat{\lambda}_j^{-2} \langle X_{n+1}, \hat{v}_j \rangle^2)^{-1} \leq n\alpha_n \|\theta\|^2$. Hence, using Theorem 3 and (H.7), we have

$$\begin{aligned} & E_{\mathcal{X}^{n+1}}(Y_{n+1} - \langle X_{n+1}, \hat{\theta}_{k_n}^{\alpha_n} \rangle)^2 - E_{\mathcal{X}^{n+1}}(Y_{n+1} - \langle X_{n+1}, \hat{\theta}_{k_n} \rangle)^2 \\ &= -2\alpha_n \frac{\sigma^2}{n} \sum_{j=1}^{k_n} \frac{\langle X_{n+1}, \hat{v}_j \rangle^2}{\hat{\lambda}_j^2} (1 + o_{a.s.}(1)) + 2\alpha_n \langle X_{n+1}, \hat{R}_{k_n} \rangle \langle X_{n+1}, \hat{T}_{k_n} \rangle (1 + o_{a.s.}(1)). \end{aligned}$$

Furthermore, $\|\hat{T}_{k_n}\|^2 = \sum_{j=1}^{k_n} \hat{\lambda}_j^{-2} \langle \theta, \hat{v}_j \rangle^2 \leq \|\theta\|^2 \sum_{j=1}^{k_n} \hat{\lambda}_j^{-2}$. Then $\alpha_n \|\hat{T}_{k_n}\|^2 ((1/n) \sum_{j=1}^{k_n} \hat{\lambda}_j^{-2}) \leq n\alpha_n \|\theta\|^2$. Applying this inequality and (H.7) to Theorem 3, we get

$$E_{\mathcal{X}^n}(\|\theta - \hat{\theta}_{k_n}^{\alpha_n}\|^2) - E_{\mathcal{X}^n}(\|\theta - \hat{\theta}_{k_n}\|^2) = -2\alpha_n \frac{\sigma^2}{n} \sum_{j=1}^{k_n} \frac{1}{\hat{\lambda}_j^2} (1 + o_{a.s.}(1)).$$

A.5. Technical lemmas

Lemma A.1. *For the regression model (1.1), if*

$$\hat{\theta} = \sum_{j=1}^{k_n} \gamma_j \Delta_n w_j w_j,$$

where $\{(\gamma_j, w_j)\}_j \subset \mathbb{R} \times H$ only depend on $\mathcal{X}^n = \{X_1, \dots, X_n\}$, then

$$\begin{aligned} & E_{\mathcal{X}^{n+1}}((Y_{n+1} - \langle X_{n+1}, \hat{\theta} \rangle)^2) \\ &= \sigma^2 + \frac{\sigma^2}{n} \sum_{j_1=1}^{k_n} \sum_{j_2=1}^{k_n} \gamma_{j_1} \gamma_{j_2} \langle \Gamma_n w_{j_1}, w_{j_2} \rangle \langle X_{n+1}, w_{j_1} \rangle \langle X_{n+1}, w_{j_2} \rangle + \langle X_{n+1}, R_{k_n}^{(\gamma, \omega)} \rangle^2, \\ & E_{\mathcal{X}^n}(\|\theta - \hat{\theta}\|^2) = \frac{\sigma^2}{n} \sum_{j_1=1}^{k_n} \sum_{j_2=1}^{k_n} \gamma_{j_1} \gamma_{j_2} \langle \Gamma_n w_{j_1}, w_{j_2} \rangle \langle w_{j_1}, w_{j_2} \rangle + \|R_{k_n}^{(\gamma, \omega)}\|^2, \end{aligned}$$

where $R_{k_n}^{(\gamma, \omega)} = \theta - \sum_{j=1}^{k_n} \gamma_j \langle \Gamma_n w_j, \theta \rangle w_j$.

Proof. Observe that $\Delta_n x = \langle \Gamma_n x, \theta \rangle + \Delta_n^\epsilon x$ for all $x \in H$, with $\Delta_n^\epsilon = n^{-1} \sum_{i=1}^n \langle X_i, \cdot \rangle \epsilon_i$. Consequently,

$$\theta - \hat{\theta} = R_{k_n}^{(\gamma, \omega)} - \sum_{j=1}^{k_n} \gamma_j \Delta_n^\epsilon w_j w_j, \quad \text{with} \quad R_{k_n}^{(\gamma, \omega)} = \theta - \sum_{j=1}^{k_n} \gamma_j \langle \Gamma_n w_j, \theta \rangle w_j. \quad (\text{A.9})$$

Then the regression model (1.1), (A.9), and conditions on ϵ imply that

$$\begin{aligned} & E_{\mathcal{X}^{n+1}}(Y_{n+1} - \langle X_{n+1}, \hat{\theta} \rangle)^2 \\ &= E_{\mathcal{X}^{n+1}}(\epsilon_{n+1} - \langle X_{n+1}, R_{k_n}^{(\gamma, \omega)} - \sum_{j=1}^{k_n} \gamma_j \Delta_n^\epsilon w_j w_j \rangle)^2 \\ &= \sigma^2 + \sum_{j_1=1}^{k_n} \sum_{j_2=1}^{k_n} \gamma_{j_1} \gamma_{j_2} E_{\mathcal{X}^n}(\Delta_n^\epsilon w_{j_1} \Delta_n^\epsilon w_{j_2}) \langle X_{n+1}, w_{j_1} \rangle \langle X_{n+1}, w_{j_2} \rangle \\ &\quad - 2 \langle X_{n+1}, R_{k_n}^{(\gamma, \omega)} \rangle \langle X_{n+1}, \sum_{j=1}^{k_n} \gamma_j E_{\mathcal{X}^n}(\Delta_n^\epsilon w_j) w_j \rangle + \langle X_{n+1}, R_{k_n}^{(\gamma, \omega)} \rangle^2. \end{aligned}$$

Now, note that

$$E_{\mathcal{X}^n}(\Delta_n^\epsilon w_j) = 0, \quad E_{\mathcal{X}^n}(\Delta_n^\epsilon w_{j_1} \Delta_n^\epsilon w_{j_2}) = \frac{\sigma^2}{n} \langle \Gamma_n w_{j_1}, w_{j_2} \rangle. \quad (\text{A.10})$$

Therefore, using (A.10), we have

$$\begin{aligned} & E_{\mathcal{X}^{n+1}}(Y_{n+1} - \langle X_{n+1}, \hat{\theta} \rangle)^2 \\ &= \sigma^2 + \frac{\sigma^2}{n} \sum_{j_1=1}^{k_n} \sum_{j_2=1}^{k_n} \gamma_{j_1} \gamma_{j_2} \langle \Gamma_n w_{j_1}, w_{j_2} \rangle \langle X_{n+1}, w_{j_1} \rangle \langle X_{n+1}, w_{j_2} \rangle + \langle X_{n+1}, R_{k_n}^{(\gamma, \omega)} \rangle^2. \end{aligned}$$

Analogously, for the conditional error of estimation, (A.9) gives

$$\begin{aligned} E_{\mathcal{X}^n}(\|\theta - \hat{\theta}\|^2) &= \sum_{j_1=1}^{k_n} \sum_{j_2=1}^{k_n} \gamma_{j_1} \gamma_{j_2} E_{\mathcal{X}^n}(\Delta_n^\epsilon w_{j_1} \Delta_n^\epsilon w_{j_2}) \langle w_{j_1}, w_{j_2} \rangle \\ &\quad - 2 \langle R_{k_n}^{(\gamma, \omega)} \rangle, \sum_{j=1}^{k_n} \gamma_j E_{\mathcal{X}^n}(\Delta_n^\epsilon w_j) w_j + \|R_{k_n}^{(\gamma, \omega)}\|^2 \end{aligned}$$

and, applying (A.10) again,

$$E_{\mathcal{X}^n}(\|\theta - \hat{\theta}\|^2) = \frac{\sigma^2}{n} \sum_{j_1=1}^{k_n} \sum_{j_2=1}^{k_n} \gamma_{j_1} \gamma_{j_2} \langle \Gamma_n w_{j_1}, w_{j_2} \rangle \langle w_{j_1}, w_{j_2} \rangle + \|R_{k_n}^{(\gamma, \omega)}\|^2.$$

Lemma A.2. *With $\gamma_n = \|\Delta\|_{H'} \{1/(\lambda_{k_n} \hat{\lambda}_{k_n}) + 2(1/\lambda_{k_n} + 1/\hat{\lambda}_{k_n}) \sum_{j=1}^{k_n} a_j\}$,*

$$\|m_{k_n} - \hat{m}_{k_n}^{\alpha_n}\|_{H'} \leq \gamma_n \|\Gamma - \Gamma_n\|_\infty + \frac{\|\Delta - \Delta_n\|_{H'}}{\hat{\lambda}_{k_n}} + \alpha_n \frac{\|\Delta\|_{H'}}{\lambda_{k_n} \hat{\lambda}_{k_n}}.$$

Proof. Take $\tilde{\Gamma}_{k_n} = \sum_{j=1}^{k_n} \lambda_j \langle \hat{v}_j, \cdot \rangle \hat{v}_j$. The first step is to write

$$\begin{aligned} \|m_{k_n} - \hat{m}_{k_n}^{\alpha_n}\|_{H'} &\leq \|\Delta \Pi_{k_n}\|_{H'} \|(\Pi_{k_n} \Gamma \Pi_{k_n})^{-1} - \tilde{\Gamma}_{k_n}^{-1}\|_\infty \\ &\quad + \|\Delta \Pi_{k_n}\|_{H'} \|\tilde{\Gamma}_{k_n}^{-1} - (\hat{\Pi}_{k_n}(\Gamma_n + \alpha_n I) \hat{\Pi}_{k_n})^{-1}\|_\infty \\ &\quad + \|\Delta \Pi_{k_n} - \Delta_n \hat{\Pi}_{k_n}\|_{H'} \|(\hat{\Pi}_{k_n}(\Gamma_n + \alpha_n I) \hat{\Pi}_{k_n})^{-1}\|_\infty. \quad (\text{A.11}) \end{aligned}$$

From (11) and (14) in the proof of Lemma 5.1 in Cardot, Ferraty, and Sarda (1999),

$$\|\Delta \Pi_{k_n}\|_{H'} \|(\Pi_{k_n} \Gamma \Pi_{k_n})^{-1} - \tilde{\Gamma}_{k_n}^{-1}\|_\infty \leq 2 \frac{\|\Delta\|_{H'}}{\lambda_{k_n}} \|\Gamma - \Gamma_n\|_\infty \sum_{j=1}^{k_n} a_j, \quad (\text{A.12})$$

$$\|\Delta \Pi_{k_n} - \Delta_n \hat{\Pi}_{k_n}\|_{H'} \leq 2 \|\Delta\|_{H'} \|\Gamma - \Gamma_n\|_\infty \sum_{j=1}^{k_n} a_j + \|\Delta - \Delta_n\|_{H'}. \quad (\text{A.13})$$

Moreover, with arguments as in the proof of Lemma 5.1 in Cardot, Ferraty, and Sarda (1999), we have

$$\|\Delta \Pi_{k_n}\|_{H'} \|\tilde{\Gamma}_{k_n}^{-1} - (\hat{\Pi}_{k_n}(\Gamma_n + \alpha_n I) \hat{\Pi}_{k_n})^{-1}\|_\infty \leq \frac{\|\Delta\|_{H'}}{\lambda_{k_n} \hat{\lambda}_{k_n}} (\|\Gamma - \Gamma_n\|_\infty + \alpha_n), \quad (\text{A.14})$$

$$\|(\hat{\Pi}_{k_n}(\Gamma_n + \alpha_n I) \hat{\Pi}_{k_n})^{-1}\|_\infty \leq \frac{1}{\hat{\lambda}_{k_n}}. \quad (\text{A.15})$$

Hence, using (A.12), (A.13), (A.14), and (A.15) in (A.11), we get

$$\begin{aligned} \|m_{k_n} - \hat{m}_{k_n}^{\alpha_n}\|_{H'} &\leq \|\Delta\|_{H'} \left(\frac{1}{\lambda_{k_n} \hat{\lambda}_{k_n}} + 2 \left(\frac{1}{\lambda_{k_n}} + \frac{1}{\hat{\lambda}_{k_n}} \right) \sum_{j=1}^{k_n} a_j \right) \|\Gamma - \Gamma_n\|_\infty \\ &\quad + \frac{\|\Delta - \Delta_n\|_{H'}}{\hat{\lambda}_{k_n}} + \alpha_n \frac{\|\Delta\|_{H'}}{\lambda_{k_n} \hat{\lambda}_{k_n}}. \end{aligned}$$

Lemma A.3. *Under (H.2) and (H.3), we have*

$$P(\|\Delta_n - \Delta\|_{H'} > \xi) \leq 2 \exp\left(-\frac{\xi^2 n}{2c_3'(c_3' + c_4'\xi)}\right).$$

Proof. This result adapts Lemma 5.3 in Cardot, Ferraty, and Sarda (1999) to the weaker assumption (H.3) (see Remark 4). Take

$$W_i = \langle X_i, \cdot \rangle Y_i - E(\langle X, \cdot \rangle Y), \quad i = 1, \dots, n,$$

with $E(W_i) = 0$. Note that $W_i = \langle X_i, \cdot \rangle m(X_i) - E(\langle X, \cdot \rangle m(X)) + \langle X_i, \cdot \rangle \epsilon_i$, and consequently

$$\|W_i\|_{H'} \leq \|m\|_{H'} (\|X_i\| + E\|X_i\|) + |\epsilon_i| \|X_i\|.$$

As a by-product, $\forall l \geq 2$, one has

$$\|W_i\|_{H'}^l \leq \sum_{k=0}^l C_l^k \|m\|_{H'}^k (\|X_i\| + E\|X_i\|)^k |\epsilon_i|^{l-k} \|X_i\|^{l-k}.$$

From this inequality, (H.2), (H.3), and $\|m\|_{H'} < c_0 < +\infty$ imply

$$E(\|W_i\|_{H'}^l) \leq l! \sum_{k=0}^l C_l^k c_0^k (2c_1)^k c_2 c_1^{l-k} = l! c_2 \sum_{k=0}^l C_l^k (2c_0 c_1)^k c_1^{l-k} = l! c_2 (c_1 + 2c_0 c_1)^l.$$

Then $E(\|W_i\|_{H'}^l) \leq (l/2) b_i^2 h^{l-2}$, where $b_i = \sqrt{2c_2}(c_1 + 2c_0 c_1)$ and $h = c_1 + 2c_0 c_1$. We can then apply the Yurinskii exponential inequality (see Yurinskii (1976)) to obtain

$$P\left(\left\|\sum_{i=1}^n W_i\right\|_{H'} > x B_n\right) \leq 2 \exp\left(-\frac{x^2}{2(1 + 1.62xh/B_n)}\right),$$

where $B_n = (\sum_{i=1}^n b_i^2)^{1/2} = n^{1/2}c'_3$ with $c'_3 = \sqrt{2c_2}(c_1 + 2c_0c_1)$. Hence,

$$P(\|\Delta_n - \Delta\|_{H'} > \xi) = P(\|\sum_{i=1}^n W_i\|_{H'} > \frac{\sqrt{n}\xi}{c'_3}B_n) \leq 2 \exp\left(-\frac{\xi^2 n}{2c'_3(c'_3 + c'_4\xi)}\right),$$

where c'_4 is a positive finite constant.

References

- Aerts, M., Hens, N. and Simonoff, J. S. (2010). Model selection in regression based on pre-smoothing. *J. Appl. Statist.* **37**, 1455-1472.
- Akritis, M. G. (1996). On the use of nonparametric regression techniques for fitting parametric regression models. *Biometrics* **52**, 1342-1362.
- Cai, T. T. and Hall, P. (2006). Prediction in functional linear regression. *Ann. Statist.* **34**, 2159-2179.
- Cao-Abad, R., López-de-Ullibarri, I., Janssen, P. and Veraverbeke, N. (2005). Presmoothed Kaplan-Meier and Nelson-Aalen estimators. *J. Nonparametr. Stat.* **17**, 31-56.
- Cardot, H., Ferraty, F. and Sarda, P. (1999). Functional linear model. *Statist. Probab. Lett.* **45**, 11-22.
- Cardot, H., Ferraty, F. and Sarda, P. (2003). Spline estimators for the functional linear model. *Statist. Sinica* **13**, 571-591.
- Cardot, H., Mas, A. and Sarda, P. (2007). CLT in functional linear regression models. *Probab. Theory Related Fields* **138**, 325-361.
- Crambes, C., Kneip, A. and Sarda, P. (2009). Smoothing splines estimators for functional linear regression. *Ann. Statist.* **37**, 35-72.
- Cristóbal-Cristóbal, J. A., Faraldo-Roca, P. and González-Manteiga, W. (1987). A class of linear regression parameter estimators constructed by nonparametric estimation. *Ann. Statist.* **15**, 603-609.
- Faraldo-Roca, P. and González-Manteiga, W. (1987). On efficiency of a new class of linear regression estimates obtained by preliminary non-parametric estimation. In *New Perspectives in Theoretical and Applied Statistics* (Edited by M.L. Puri, J. Perez-Vilaplana and W. Wertz), 229-242. Wiley, New York.
- Ferraty, F. and Vieu, P. (2006). *Nonparametric Functional Data Analysis: Theory and Practice*. Springer, New York.
- Hall, P. and Horowitz, J. L. (2007). Methodology and convergence rates for functional linear regression. *Ann. Statist.* **35**, 70-91.
- Hall, P. and Hosseini-Nasab, M. (2006). On properties of functional principal components analysis. *J. R. Statist. Soc. Ser. B* **68**, 109-126.
- Hitchcock, D. B., Casella, G. and Booth, J. G. (2006). Improved estimation of dissimilarities by presmoothing functional data. *J. Amer. Statist. Assoc.* **101**, 211-222.
- Janssen, P., Swanepoel, J. and Veraverbeke, N. (2001). Efficiency of linear regression estimators based on presmoothing. *Comm. Statist. Theory Methods* **30**, 2079-2097.
- Jácome, M., Gijbels, I. and Cao, R. (2008). Comparison of presmoothing methods in kernel density estimation under censoring. *Comput. Statist.* **23**, 381-406.

- Jácome, M. A. and Iglesias-Pérez, M. C. (2008). Presmoothed estimation with left-truncated and right-censored data. *Comm. Statist. Theory Methods* **37**, 2964-2983.
- Ramsay, J. O. and Silverman, B. W. (2002). *Applied Functional Data Analysis. Methods and Case Studies*. Springer, New York.
- Ramsay, J. O. and Silverman, B. W. (2005). *Functional Data Analysis*. 2nd edition. Springer, New York.
- Yurinskii, V. V. (1976). Exponential inequalities for sums of random vectors. *J. Multivariate Anal.* **6**, 473-499.
- Zhang, J. T. and Chen, J. (2007). Statistical inferences for functional data. *Ann. Statist.* **35**, 1052-1079.

Laboratoire de Statistique et Probabilités, Institut de Mathématiques, Université Paul Sabatier, 118, Route de Narbonne, 31062 Toulouse Cédex, France.

E-mail: ferraty@cict.fr

Departamento de Estatística e I.O., Facultade de Matemáticas, Universidade de Santiago de Compostela, c/ Lope Gómez de Marzoa, Campus Vida, 15782 Santiago de Compostela, Spain.

E-mail: wenceslao.gonzalez@usc.es

Departamento de Estatística e I.O., Facultade de Matemáticas, Universidade de Santiago de Compostela, c/ Lope Gómez de Marzoa, Campus Vida, 15782 Santiago de Compostela, Spain.

E-mail: adela.martinez@usc.es

Laboratoire de Statistique et Probabilités, Institut de Mathématiques, Université Paul Sabatier, 118, Route de Narbonne, 31062 Toulouse Cédex, France.

E-mail: vieu@cict.fr

(Received April 2010; accepted January 2011)