Functional linear model with missing values in the covariate and the response

Christophe Crambes, Chayma Daayeb, Ali Gannoun, Yousri Henchiri

Functional linear model with missing values in the covariate and the response.

Christophe Crambes
Université de Montpellier, IMAG, France.
and
Chayma Daayeb
Université de Montpellier, IMAG, France.
Université de Tunis El Manar, ENIT-LAMSIN, Tunisie.
and
Ali Gannoun
Université de Montpellier, IMAG, France.
and
Yousri Henchiri*
Université de la Manouba, ISAMM, Tunisie.
Université de Tunis El Manar, ENIT-LAMSIN, Tunisie.

December 18, 2020

Abstract

Dealing with missing values is an important issue in data observation or data recording process. In this paper, we consider a functional linear regression model, where some observations of the real response and the functional covariate are affected by missing data. We use a reconstruction operator that aims to recover the missing parts of the explanatory curves, then we are interested in regression imputation method of missing data on the response variable, using functional principal component regression to estimate the functional coefficient of the model. We study the asymptotic behavior of the prediction error we commit when missing data are replaced by the imputed values in the original dataset. The practical behavior of the method is also studied on simulated data and a real dataset.

Keywords: Functional linear model, Missing data, Functional Principal Components, Missing At Random, Missing Completely At Random, Regression imputation.

*Principal corresponding author. E-mail: christophe.crambes@umontpellier.fr (Christophe Crambes), chayma.daayeb@etu.umontpellier.fr (Chayma Daayeb), ali.gannoun@umontpellier.fr (Ali Gannoun), yousri.henchiri@umontpellier.fr (Yousri Henchiri).

1 Introduction

The analysis of the functional data has grown very significantly in recent years, as evidenced by the numerous literatures on the subject: Ramsay and Silverman (2005), Ferraty and Vieu (2006), Hsing and Eubank (2015), Horváth and Kokoszka (2012) provide a non-exhaustive list of monographs giving an overview of this topic.

One of the most popular model in functional data analysis is the functional linear model, when one is interested to consider a relationship between a real-valued variable Y and a covariate $X = (X(t), t \in [a, b])$ valued in a real separable Hilbert space H of functions defined on a compact interval $[a, b]$ of $\mathbb{R}$. In the following, we consider the space $H = L^2([a, b])$ of square integrable functions defined on $[a, b]$, endowed with its usual inner product defined by $\langle u, v\rangle = \int_a^b u(t)v(t)\mathrm{d}t$ for all functions $u, v \in H$, and its associated norm $\|.\|$. This model, studied by many authors as for instance Cardot et al. (1999), Cai and Hall (2006), Hall and Horowitz (2007), Crambes et al. (2009), is defined by

$$Y = \theta_0 + \int_a^b \theta(t)X(t)\mathrm{d}t + \varepsilon, \tag{1}$$

where $\theta_0 \in \mathbb{R}$ and θ is a square integrable function defined on $[a, b]$ modeling the relationship between the real random variable Y and the square integrable random function X. The error of the model ε is a centered real random variable independent of X with finite variance $\mathbb{E}(\varepsilon^2) = \sigma_\varepsilon^2$. We can also write the functional linear regression model (1) as

$$Y = \theta_0 + \Theta X + \varepsilon, \tag{2}$$

where $\Theta : H \to \mathbb{R}$ is a linear continuous operator defined by $\Theta u = \langle \theta, u\rangle$ for any function $u \in H$. The existence and unicity of this regression function θ is discussed in Cardot et al. (2003). A smooth version of the functional principal components regression (SPCR) is introduced. It consists in considering the empirical covariance operator of the predictor X and diagonalizing it to select the eigenfunctions associated to the highest eigenvalues. Then, a least squares regression is performed with the response Y and the coordinates of the functional covariate X projection on the space spanned by the selected eigenfunctions.

Considering a sample $(X_i, Y_i)_{i=1,\dots,n}$ of independent and identically distributed couples with the same distribution as (X, Y), we define the empirical cross covariance operator $\widehat{\Delta}_n$ given by $\widehat{\Delta}_n u = \frac{1}{n}\sum_{i=1}^n \langle X_i, u\rangle Y_i$ for all $u \in H$, the empirical covariance operator $\widehat{\Gamma}_n$ given by $\widehat{\Gamma}_n u = \frac{1}{n}\sum_{i=1}^n \langle X_i, u\rangle X_i$ for all $u \in H$. Denoting $(\widehat{\phi}_j)_{j=1,\dots,k_n}$ the eigenfunctions associated to $\widehat{\Gamma}_n$ corresponding to the k_n highest eigenvalues $\widehat{\lambda}_1 > \ldots > \widehat{\lambda}_{k_n} > 0$, we define the orthogonal projection operator $\widehat{\Pi}_{k_n}$ onto the subspace $\mathrm{Span}(\widehat{\phi}_1, \dots, \widehat{\phi}_{k_n})$ by $\widehat{\Pi}_{k_n} u = \sum_{j=1}^{k_n} \langle \widehat{\phi}_j, u\rangle \widehat{\phi}_j$ for all $u \in H$. Then, the functional principal component regression estimator $\widehat{\Theta}$ of Θ is defined by

$$\widehat{\Theta} = \langle \widehat{\theta}, . \rangle = \widehat{\Pi}_{k_n} \widehat{\Delta}_n (\widehat{\Pi}_{k_n} \widehat{\Gamma}_n \widehat{\Pi}_{k_n})^{-1}. \tag{3}$$

The corresponding estimator of θ is given by

$$\widehat{\theta} = \frac{1}{n}\sum_{i=1}^n \sum_{j=1}^{k_n} \frac{\langle X_i, \widehat{\phi}_j\rangle Y_i}{\widehat{\lambda}_j} \widehat{\phi}_j = \sum_{j=1}^{k_n} \widehat{s}_j \widehat{\phi}_j, \tag{4}$$

with $\widehat{s}_j = \frac{1}{n\widehat{\lambda}_j}\sum_{i=1}^n \langle X_i, \widehat{\phi}_j\rangle Y_i$. In addition, the estimator of $\theta_0 = \mathbb{E}(Y) - \int_a^b \theta(t)\mathbb{E}(X)(t)\mathrm{d}t$ is written as follows: $\widehat{\theta}_0 = \overline{Y} - \int_a^b \widehat{\theta}(t)\overline{X}(t)\mathrm{d}t$ with $\overline{Y} = \frac{1}{n}\sum_{i=1}^n Y_i$ and $\overline{X} = \frac{1}{n}\sum_{i=1}^n X_i$. Now, given $\widehat{\theta}_0$ and $\widehat{\theta}$, it is easy to obtain the residuals of the fit, given by $\widehat{\varepsilon}_{i,k_n} = Y_i - \widehat{\theta}_0 - \langle X_i, \widehat{\theta}\rangle$, for $i = 1, \dots, n$, that can be used to estimate the error variance, σ_ε^2, through

$$\widehat{\sigma}^2_{\varepsilon,k_n} = \frac{1}{n - k_n - 1}\sum_{i=1}^n \widehat{\varepsilon}^2_{i,k_n}.$$

In the previously cited works on the functional linear model, data is fully observed. This may not always be the case, and missing data appear in many situations, for example when the measuring device breaks down. This topic has to be studied a lot in the multivariate framework, for example we refer the reader to Little and Rubin (2002) and Graham (2012). For functional data, the literature is not so well developed. In functional linear regression, the work of Crambes and Henchiri (2019) considers a missing data mechanism on the response Y while the functional covariate is completely observed. A regression imputation methodology for the missing data is proposed and the authors propose an estimation of the functional parameter θ with the reconstructed dataset, as well as the prediction of new values. The method consistency is studied both from a theoretical and a practical point

of view. The same problematic is studied in another paper Febrero-Bande et al. (2019), although not exploring theoretical results. In functional nonparametric setting, Ferraty et al. (2013) consider the estimation of the mean of a real response with a functional covariate, in a nonparametric regression model where the functional covariate is completely observed and the real response is missing at random. In our work, we want to consider the functional linear model where both the response and the covariate are affected with missing data, which is an unexplored topic as far as we know.

For the missing data mechanism in the response, we consider a dichotomous random variable $\delta^{[Y]}$ leading to the sample $(\delta_i^{[Y]})_{i=1,\dots,n}$ such that $\delta_i^{[Y]} = 1$ if the value Y_i is available and $\delta_i^{[Y]} = 0$ if the value Y_i is missing, for all $i = 1, \dots, n$. Here, we consider that the data in the response is missing at random (MAR): the fact that the value Y is missing does not depend on the response of the model, but can possibly depend on the covariate, that is,

$$\mathbb{P}(\delta^{[Y]} = 1 \mid X, Y) = \mathbb{P}(\delta^{[Y]} = 1 \mid X).$$

As a consequence of this MAR assumption, the variable $\delta^{[Y]}$ (the fact that an observation is missing) is independent of the error of the model ε. In the following, the number of missing values among $Y_1, \dots, Y_n$ is denoted

$$m_n^{[Y]} = \sum_{i=1}^{n} \mathbb{1}_{\{\delta_i^{[Y]}=0\}}.$$

In Crambes and Henchiri (2019), a regression imputation methodology is given for the missing values on the response, under this MAR assumption, but the covariate is supposed to be completely observed, which is no more the case here. We consider a functional variable $\delta^{[X]}$ leading to the sample $(\delta_i^{[X]})_{i=1,\dots,n}$ such that, for all $t \in [a, b]$, $\delta_i^{[X]}(t) = 1$ if the value $X(t)$ is available and $\delta_i^{[X]}(t) = 0$ if the value $X(t)$ is missing. We consider that the data in the covariate are missing completely at random (MCAR): the fact that X contains missing data does not depend on the covariate of the model, neither on the response, that is, for any $t \in [a, b]$

$$\mathbb{P}(\delta^{[X]}(t) = 1 \mid X, Y) = \mathbb{P}(\delta^{[X]}(t) = 1).$$

Finally, $\mathscr{O}^{[Y]}$ and $\mathscr{O}^{[X]}$ correspond to the sets of indexes for which data Y and X are completely observed, in other words

$$\mathscr{O}^{[Y]} = \left\{ i \in \{1, \ldots, n\} \quad / \quad \delta_i^{[Y]} = 1 \right\},$$

and

$$\mathscr{O}^{[X]} = \left\{ i \in \{1, \ldots, n\} \quad / \quad \exists t \in [a, b], \delta_i^{[X]}(t) = 1 \right\}.$$

We denote the number of curves where missing values appear

$$m_n^{[X]} = n - Card\left(\mathscr{O}^{[X]}\right).$$

The objective of this paper is: (I) to reconstruct missing X curves and to impute missing data to Y, (II) to estimate either θ or Θ with the reconstructed dataset, (III) predict a new value of the response Y given a new test observation on the explanatory variable X. Moreover, we want to explore the interest of the imputation methodology compared to the naive method which would consist in simply ignoring the missing data and only using the observations when both X and Y are observed.

In the following, we give in section 2 theoretical results when missing data appear on the covariate. Then, in section 3, we extend these results when missing data appear both on the covariate and the response. In section 4, we present some simulation results to show the behaviour of the method in practice. Section 5 is devoted to a real dataset application. Finally, all the proofs are postponed to section 6.

2 Missing data on the covariate

2.1 Curve reconstruction

Some works dealing with curve reconstruction use the data structure through functional principal components analysis. Recent contributions on this topic are Kraus (2015), Delaigle and Hall (2016), Descary and Panaretos (2019), Kneip and Liebl (2020), Kraus and

Stefanucci (2020), Delaigle et al. (2020), Lin et al. (2020), Lin and Wang (2020). In this part, we are first going to reconstruct the missing parts of the functional covariate using the methodology from Kneip and Liebl (2020). Let $(O_i)_{i=1,\dots,n}$ be the sample of the observation parts of the curves, in other words $O_i = \left\{ t \in [a,b], \quad / \quad \delta_i^{[X]}(t) = 1 \right\}$, for all $i = 1, \dots, n$. On the contrary, we denote $M_i = [a,b] \setminus O_i$, for all $i = 1, \dots, n$. In the following, we use "O" and "M" to denote a given production of O_i and M_i. In addition, we denote the observed and missing parts of X_i by X_i^O and X_i^M. We write the Karhunen-Loève (KL) decomposition of X_i^O in $\mathbb{L}^2(O)$

$$X_i^O(t) = \sum_{k=1}^{+\infty} \xi_{ik}^O \phi_k^O(t), \tag{5}$$

where $t \in O$. In this decomposition, the principal component scores are defined for all $i = 1, \dots, n$ and $k \geq 1$ by $\xi_{ik}^O = \langle \phi_k^O, X_i^O \rangle$, where $\mathbb{E}(\xi_{ik}^O) = 0$ and $\mathbb{E}(\xi_{ik}^O \xi_{i\ell}^O) = \lambda_k^O$ for all $k = \ell$ and zero for all $k \neq \ell$. The decreasing eigenvalues $\lambda_1^O > \lambda_2^O > \ldots > 0$ are tending to zero and $(\widehat{\phi}_k)_{k \geq 1}$ are the eigenfunctions of the covariance operator of X_i^O.

The missing part of the curves are assumed to be related to the observed parts, writing

$$X_i^M(s) = L(X_i^O(t)) + Z_i(s), \tag{6}$$

for all $t \in O$ and $s \in M$, where $L : \mathbb{L}^2(O) \to \mathbb{L}^2(M)$ is a linear reconstruction operator and $Z_i \in \mathrm{L}^2(M)$ is the reconstruction error. This reconstruction estimator is estimated in Kneip and Liebl (2020) by

$$\mathcal{L}(X_i^O)(s) = \sum_{k=1}^{+\infty} \xi_{ik}^O \widetilde{\phi_k^O}(s) = \sum_{k=1}^{+\infty} \xi_{ik}^O \frac{\langle \phi_k^O, \gamma_s \rangle}{\lambda_k^O}, \tag{7}$$

for all $s \in M$, where $\gamma_s(t) = \mathbb{E}\left(X_i^M(s) X_i^O(t)\right)$ for all $t \in O$ and $s \in M$. In particular, it is shown in Kneip and Liebl (2020) that $\mathcal{L}(X_i^O)$ has a continuous and finite variance function and is unbiaised.

2.2 Estimation of the reconstruction in practice

We consider a discretization without measurement errors, that is $((W_{i1}, t_{i1}), \ldots, (W_{ip}, t_{ip}))$ denote the observable data pairs of the function X_i^O, namely

$$W_{ij} = X_i^O(t_{ij}), \tag{8}$$

for $i = 1, \ldots, n$ and $j = 1, \ldots, p$, where $t_{ij} \in O_i$. In order to estimate the curve X_i^O and the covariance function γ_s, a nonparametric curve estimation by local polynomials smoothers is used. For the curve X_i^O, the kernel is denoted κ_1 and the bandwidth h_X, and for the covariance function γ_s, the kernel is denoted κ_2 and the bandwidth h_γ. More precisely, we consider

$$\sum_{j=1}^{p} (W_{ij} - \beta_0 - \beta_1(t_{ij} - t))^2 \kappa_1 \left(\frac{t_{ij} - t}{h_X} \right), \tag{9}$$

which we minimize with respect to β_0, β_1 for all $t \in O$. The local linear smoother of the curve X_i^O is defined by $\widehat{X}_i^O(t; h_X) = \widehat{\beta}_0$. Similarly, we consider

$$\sum_{i=1}^{n} \sum_{j,\ell=1}^{p} (C_{ij\ell} - \tau_0 - \tau_1(t_{ij} - t) - \tau_2(t_{i\ell} - s))^2 \kappa_2 \left(\frac{t_{ij} - t}{h_\gamma}, \frac{t_{i\ell} - s}{h_\gamma} \right), \tag{10}$$

which we minimize with respect to τ_0, τ_1, τ_2 for all $t \in O, s \in M$, where $C_{ij\ell} = W_{ij} W_{i\ell}$ are the raw covariance points. The local linear smoother of the covariance function γ is defined by $\widehat{\gamma}(t, s; h_\gamma) = \widehat{\tau}_0$.

For estimating the eigenvalues λ_k^O and the eigenfunctions ϕ_k^O, we use the Fredholm integral equation

$$\int_O \widehat{\gamma}(t, u; h_\gamma) \widehat{\phi}_k^O(u) \mathrm{d}u = \widehat{\lambda}_k^O \widehat{\phi}_k^O(t),$$

for all $t \in O$. For the functional principal component scores $\xi_{ik}^O = \int_O X_i^O(t) \phi_k(t) \mathrm{d}t$, the estimator is defined by

$$\widehat{\xi}_{ik}^O = \sum_{j=1}^{p} \widehat{\phi}_k^O(t_{ij}) W_{ij} (t_{ij} - t_{i,j-1}), \quad \text{with } t_{i0} = a.$$

Finally, to estimate $\mathcal{L}(X_i^O)$ in (7), considering a positive integer k_n, we define

$$\widehat{\mathcal{L}}_{k_n}(X_i^O)(s) = \sum_{k=1}^{k_n} \widehat{\xi}_{ik}^O \frac{\langle \widehat{\phi}_k^O, \widehat{\gamma}_s \rangle}{\widehat{\lambda}_k^O}, \tag{11}$$

where $\widehat{\gamma}_s = \widehat{\gamma}(., s; h_\gamma)$. In this step we are able to find the estimator of the missing parts of X_i^O

$$\widehat{X}_i^M(s) = \widehat{\mathcal{L}}((X_i^O)(t)), \tag{12}$$

for all $t \in O$ and $s \in M$. A boundary problem is highlighted in Kneip and Liebl (2020), due to the fact that the nonparametric smoothing of X on the observed interval may not coincide with the estimation of X on the missing interval at the boundary. Consequently, the authors consider a corrected version of the estimation of $\mathcal{L}(X_i^O)$. Let V_s be the boundary point closest to $s \in M$, the corrected estimator of $\mathcal{L}(X_i^O)$ is written in the following form

$$\widehat{\mathcal{L}}_{k_n}^\star(X_i^O)(s) = \widehat{X}_i^O(V_s; h_X) + \sum_{k=1}^{k_n} \widehat{\xi}_{ik}^O \left(\frac{\langle \widehat{\phi}_k^O, \widehat{\gamma}_s \rangle}{\widehat{\lambda}_k^O} - \frac{\langle \widehat{\phi}_k^O, \widehat{\gamma}_{V_s} \rangle}{\widehat{\lambda}_k^O} \right). \tag{13}$$

In the following, we denote

$$\widetilde{X}_i(t) = \begin{cases} X_i^O(t) \text{ if } t \in O, \\ \widehat{\mathcal{L}}_{k_n}(X_i^O)(t) \text{ if } t \in M. \end{cases} \tag{14}$$

2.3 Estimation of θ and prediction

Let $X^\star$ be defined by

$$X^\star = \delta^{[X]} X + (1 - \delta^{[X]}) \widetilde{X}, \tag{15}$$

with $\widetilde{X}$ defined by (14).

Remark 2.1 *The gap between $X^\star$ and the true function leads us to the gap between the true function and its reconstruction with the proportion of curves completely observed.*

$$\begin{aligned} X_i^\star - X_i &= \delta_i^{[X]} X_i + (1 - \delta_i^{[X]}) \widetilde{X}_i - X_i \\ &= (1 - \delta_i^{[X]}) \widetilde{X}_i - (1 - \delta_i^{[X]}) X_i = (1 - \delta_i^{[X]})(\widetilde{X}_i - X_i). \end{aligned}$$

Concerning the estimation of θ, we define

$$\widehat{\widehat{\theta}} = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{k_n}\frac{\langle X_i^\star, \widehat{\phi}_{j,rec}\rangle Y_i}{\widehat{\lambda}_{j,rec}}\widehat{\phi}_{j,rec} = \sum_{j=1}^{k_n}\widehat{\widehat{s}}_j\widehat{\phi}_{j,rec}, \tag{16}$$

with $\widehat{\widehat{s}}_j = \frac{1}{n\widehat{\lambda}_{j,rec}}\sum_{i=1}^{n}\langle X_i^\star, \widehat{\phi}_{j,rec}\rangle Y_i$. The estimation of the operator Θ is given by

$$\widehat{\widehat{\Theta}} = \langle\widehat{\widehat{\theta}}, .\rangle = \widehat{\Pi}_{k_n,rec}\widehat{\Delta}_{n,rec}(\widehat{\Pi}_{k_n,rec}\widehat{\Gamma}_{n,rec}\widehat{\Pi}_{k_n,rec})^{-1}, \tag{17}$$

where $\widehat{\Delta}_{n,rec}$ is the reconstructed cross covariance operator given by $\widehat{\Delta}_{n,rec} = \frac{1}{n}\sum_{i=1}^{n}\langle X_i^\star, .\rangle Y_i$, $\widehat{\Gamma}_{n,rec}$ is the reconstructed covariance operator given by $\widehat{\Gamma}_{n,rec} = \frac{1}{n}\sum_{i=1}^{n}\langle X_i^\star, .\rangle X_i^\star$, and $\widehat{\Pi}_{k_n,rec}$ is the projection operator onto the subspace $\mathrm{Span}(\widehat{\phi}_{1,rec},\ldots,\widehat{\phi}_{k_n,rec})$, that is the subspace spanned by the k_n first eigenfunctions of the covariance operator $\widehat{\Gamma}_{n,rec}$. The eigenvalues of the covariance operator $\widehat{\Gamma}_{n,rec}$ are denoted $\widehat{\lambda}_{1,rec},\ldots,\widehat{\lambda}_{k_n,rec}$.

Moreover, the estimator of θ_0 is defined by

$$\widehat{\widehat{\theta}}_0 = \overline{Y} - \int_a^b \widehat{\widehat{\theta}}(t)\overline{X}^\star(t)\mathrm{d}t, \tag{18}$$

where $\overline{Y}$ and $\overline{X}^\star$ represent the respective empirical means of $Y_1,\ldots,Y_n$ and $X_1^\star,\ldots,X_n^\star$. Given $\widehat{\widehat{\theta}}_0$ and $\widehat{\widehat{\theta}}$, the residuals of the fit, $\widehat{\widehat{\varepsilon}}_{i,k_n} = Y_i - \widehat{\widehat{\theta}}_0 - \langle X_i^\star, \widehat{\widehat{\theta}}\rangle$, for $i = 1,\ldots,n$, can be used to estimate the error variance as following

$$\widehat{\widehat{\sigma}}^2_{\varepsilon,k_n} = \frac{1}{n-k_n-1}\sum_{i=1}^{n}\widehat{\widehat{\varepsilon}}^2_{i,k_n}.$$

Finally, given a new observation of the covariate X, denoted X_{new}, with possibly missing data, we predict the corresponding value of the response Y by

$$\widehat{Y}_{new} = \widehat{\widehat{\theta}}_0 + \langle\widehat{\widehat{\theta}}, X_{new}^\star\rangle. \tag{19}$$

2.4 Assumptions

We present in this part the assumptions needed for our results. These assumptions are used in Kneip and Liebl (2020) in order to control the curve reconstruction for the covariate.

(A.1) The variable X has a finite four moment order, that is $\mathbb{E}\left(\|X\|^4\right) < \infty$.

(A.2) Let $np \to \infty$ when $n \to \infty$ and $p = p(n)$. We assume $p = n^{\eta_1}$ with $0 < \eta_1 < \infty$ in the following.

(A.3) The bandwidth h_X satisfies $h_X \to 0$ and $(ph_X) \to \infty$ as $p \to \infty$. For instance, we assume that $h_X = \frac{1}{n^{\eta_2}}$ with $0 < \eta_2 < \eta_1$. The bandwidth h_γ satisfies $h_\gamma \to 0$ and $(n(p^2 - p)h_\gamma) \to \infty$ as $n(p^2 - p) \to \infty$. For example, we can take $h_\gamma = \frac{1}{n^{\eta_3}}$ with $0 < \eta_3 < 2\eta_1 + 1$.

(A.4) Let κ_1 and κ_2 be nonnegative, second order univariate and bivariate kernel functions with support $[-1, 1]$. For example, we can use univariate and bivariate Epanechnikov kernel functions with compact support $[-1, 1]$, namely $\kappa_1(x) = \frac{3}{4}(1 - x^2)\mathbb{1}_{[-1,1]}(x)$ and $\kappa_2(x, y) = \frac{9}{16}(1 - x^2)(1 - y^2)\mathbb{1}_{[-1,1]}(x)\mathbb{1}_{[-1,1]}(y)$.

(A.5) For any subinterval $O \subseteq [a, b]$, we assume that the eigenvalues $\lambda_1 > \lambda_2 > \ldots > 0$ have multiplicity one. Moreover, we assume that there exist $a_O > 1$ and $0 < c_O < \infty$ such that `(i)` $\lambda_k^O - \lambda_{k+1}^O \geq c_O k^{-a_O - 1}$, `(ii)` $\lambda_k^O = \mathcal{O}(k^{-a_O})$, `(iii)` $1/\lambda_k^O = \mathcal{O}(k^{a_O})$ as $k \to \infty$.

(A.6) For any subinterval $O \subseteq [a, b]$, we assume that there exists $0 < D_O < \infty$ such that the eigenfunctions satisfy $\sup_{t \in [a,b]} \sup_{k \geq 1} \left| \widetilde{\phi}_k^O(t) \right| \leq D_O$.

Assumption ***(A.1)*** holds for many processes X (Gaussian processes, bounded processes). Assumption ***(A.2)*** is mild and can be satisfied even if the number of observation points p does not go fast to infinity. As in Kneip and Liebl (2020), we assume that $p = n^{\eta_1}$ with $0 < \eta_1 < \infty$. Assumptions ***(A.3)*** and ***(A.4)*** are classic in the context of local polynomials smoothers. Assumptions ***(A.5)*** and ***(A.6)***, related to eigenvalues and eigenfunctions of the covariance operator of X, are given in Kneip and Liebl (2020). In particular, a polynomial decrease of the eigenvalues is required, allowing a large class of eigenvalues for the covariance operator of X.

2.5 Asymptotic results

Under assumptions ***(A.1)-(A.6)***, it is proved in Kneip and Liebl (2020) that, in the case where $p \sim n^{\eta_1}$ with $\eta_1 \leq 1/2$, we have for any $t \in [a, b]$

$$\left|\widetilde{X}_i(t) - X_i(t)\right| = \mathcal{O}_p\big(p^{-(a_O-1)/(2(a_O+2))}\big)\,, \tag{20}$$

from which we deduce, for any $t \in [a, b]$

$$|X_i^\star(t) - X_i(t)| = \mathcal{O}_p\big(p^{-(a_O-1)/(2(a_O+2))}\big)\,. \tag{21}$$

The previous result allows to obtain some bounds between quantities related to functional principal components analysis with the constructed curves and with the original curves. These bounds are given in the following proposition. For any linear continuous operator $T : H \to H$ or any linear continuous operator $S : H \to \mathbb{R}$, we define the operator norm of T as $\|T\|_\infty = \sup_{\|x\|=1} \|Tx\|$, and the operator norm of S as $\|S\|_\infty = \sup_{\|x\|=1} |Sx|$.

Proposition 2.2 *Under assumptions* ***(A.1)-(A.6)****, we have*

$$\begin{aligned}
&(i) \quad \left\|\widehat{\Gamma}_{n,rec} - \widehat{\Gamma}_n\right\|_\infty = \mathcal{O}_p\big(p^{-(a_O-1)/(2(a_O+2))}\big)\,,\\
&(ii) \quad \left\|\widehat{\Delta}_{n,rec} - \widehat{\Delta}_n\right\|_\infty = \mathcal{O}_p\big(p^{-(a_O-1)/(2(a_O+2))}\big)\,,\\
&(iii) \quad \forall k \geq 1, \quad \left\|\widehat{\phi}_{k,rec} - \widehat{\phi}_k\right\| = \mathcal{O}_p\big(\widehat{\alpha}_k^{-1} p^{-(a_O-1)/(2(a_O+2))}\big)\,,\\
&(iv) \quad \forall k \geq 1, \quad \left|\widehat{\lambda}_{k,rec} - \widehat{\lambda}_k\right| = \mathcal{O}_p\big(p^{-(a_O-1)/(2(a_O+2))}\big)\,,
\end{aligned}$$

where we set $\widehat{\alpha}_1 = \widehat{\lambda}_1 - \widehat{\lambda}_2$ *and* $\widehat{\alpha}_k = \min\left(\widehat{\lambda}_{k-1} - \widehat{\lambda}_k; \widehat{\lambda}_k - \widehat{\lambda}_{k+1}\right)$ *for all* $k \geq 2$.

We finish this section about missing data on the covariate with the main result giving a bound for the prediction error of Y_{new} with a new value of the covariate X_{new}.

Theorem 2.3 *Under assumptions* ***(A.1)-(A.6)****, if we take* $k_n \sim p^{1/(a_O+2)}$ *and* $p \sim n^{\eta_1}$ *with* $\eta_1 \leq 1/2$*, the prediction error is*

$$\mathbb{E}\left(\widehat{\widehat{\theta}}_0 + \langle\widehat{\widehat{\theta}}, X_{new}^\star\rangle - \theta_0 - \langle\theta, X_{new}^\star\rangle\right)^2 = \mathcal{O}_p\big(n^{-\eta_1(a_O-1)/(2(a_O+2))}\big)\,.$$

This prediction error rate $\mathcal{O}_p\big(n^{-\eta_1(a_O-1)/(2(a_O+2))}\big)$ is related to the rate given in Corollary 4.1 in Kneip and Liebl (2020) (in the particular case where $\eta_1 = 1/2$). This means that, provided with some conditions on the number of observation points p and the number of principal components k_n are fulfilled, the prediction error rate has the same order as the curve reconstruction error rate. In other words, this means that, when reconstructing missing parts of the explanatory curves in a functional linear model and then predicting a new value of the response, the most important step is the curve reconstruction. This step is going to fix the convergence rate of the prediction.

Remark 2.4 *Due to the bound (21), the result of Theorem 2.3 remains valid if we replace* $X^\star_{new}$ *with* X_{new}.

Corollary 2.5 *Under the hypotheses of Theorem 2.3, in the favorable situation where* $\eta_1 = 1/2$*, the prediction error is*

$$\mathbb{E}\left(\widehat{\widehat{\theta}}_0 + \langle\widehat{\widehat{\theta}}, X^\star_{new}\rangle - \theta_0 - \langle\theta, X^\star_{new}\rangle\right)^2 = \mathcal{O}_p\big(n^{-(a_O-1)/(4(a_O+2))}\big)\,.$$

3 Missing data on the covariate and the response

In this section, we are interested in the most general case of missing data in functional linear regression: when both the covariate and the response are affected by missing data. We have seen in the previous section the methodology for reconstructing the missing parts of the explanatory curves. Concerning missing data on the response, we are going to apply the methodology presented in Crambes and Henchiri (2019), imputing missing values on the response using a regression imputation. Next, once the initial sample is completed, we will present the estimation of the functional parameter θ and predict new values for the response.

3.1 Regression imputation on the response

In this subsection, we use the methodology to impute a missing value of Y as in Crambes and Henchiri (2019). We consider the whole data, possibly with reconstructed explanatory

curves, except the ones for which the value of Y is not available. We define the covariance operator with the reconstructed curves

$$\widehat{\Gamma}_{n,rec}^{obs} = \frac{1}{n - m_n^{[Y]}} \sum_{i=1}^{n} \langle X_i^\star, . \rangle \delta_i^{[Y]} X_i^\star.$$

Let $\widehat{\Pi}_{k_n,rec}^{obs}$ be the projection operator onto the subspace $\mathrm{Span}(\widehat{\phi}_{1,rec}^{obs}, \ldots, \widehat{\phi}_{k_n,rec}^{obs})$ where $\widehat{\phi}_{1,rec}^{obs}, \ldots, \widehat{\phi}_{k_n,rec}^{obs}$ are the k_n first eigenfunctions of the covariance operator $\widehat{\Gamma}_{n,rec}^{obs}$. With analogous notations, $\widehat{\lambda}_{1,rec}^{obs}, \ldots, \widehat{\lambda}_{k_n,rec}^{obs}$ represent the k_n first eigenvalues of $\widehat{\Gamma}_{n,rec}^{obs}$. We first estimate θ with the observed responses and the observed or reconstructed covariates

$$\widetilde{\theta} = \frac{1}{n - m_n^{[Y]}} \sum_{i=1}^{n - m_n^{[Y]}} \sum_{j=1}^{k_n} \frac{\langle X_i^\star, \widehat{\phi}_{j,rec}^{obs} \rangle \delta_i^{[Y]} Y_i}{\widehat{\lambda}_{j,rec}^{obs}} \widehat{\phi}_{j,rec}^{obs} = \sum_{j=1}^{k_n} \widetilde{s}_j \widehat{\phi}_{j,rec}^{obs}, \tag{22}$$

with $\widetilde{s}_j = \frac{1}{(n - m_n^{[Y]}) \widehat{\lambda}_{j,rec}^{obs}} \sum_{i=1}^{n - m_n^{[Y]}} \langle X_i^\star, \widehat{\phi}_{j,rec}^{obs} \rangle \delta_i^{[Y]} Y_i$. We also estimate the intercept θ_0 with

$$\widetilde{\theta}_0 = \overline{Y}_{obs} - \int_a^b \widetilde{\theta}(t) \overline{X}^\star(t) \mathrm{d}t,$$

where $\overline{Y}_{obs} = \frac{1}{n - m_n^{[Y]}} \sum_{i=1}^{n} \delta_i^{[Y]} Y_i$. Now, the residuals of the fit, $\widetilde{\varepsilon}_{i,k_n} = \delta_i^{[Y]} Y_i - \widetilde{\theta}_0 - \langle X_i^\star, \widetilde{\theta} \rangle$, for $i = 1, \ldots, n - m_n^{[Y]}$, can be used to estimate the error variance as following

$$\widetilde{\sigma}_{\varepsilon,k_n}^2 = \frac{1}{n - m_n^{[Y]} - k_n - 1} \sum_{i=1}^{n - m_n^{[Y]}} \widetilde{\varepsilon}_{i,k_n}^2.$$

Then, considering a missing value on the response, say Y_ℓ such that $\delta_\ell^{[Y]} = 0$, we define the imputed value $Y_{\ell,imp}$ by

$$Y_{\ell,imp} = \widetilde{\theta}_0 + \langle \widetilde{\theta}, X_\ell^\star \rangle = \widetilde{\theta}_0 + \sum_{j=1}^{k_n} \widetilde{\widetilde{s}}_j \langle X_\ell^\star, \widehat{\phi}_{j,rec}^{obs} \rangle, \tag{23}$$

with $\widetilde{\widetilde{s}}_j = \frac{1}{(n - m_n^{[Y]}) \widehat{\lambda}_{j,rec}^{obs}} \sum_{\substack{i=1 \\ i \neq \ell}}^{n} \langle X_i^\star, \widehat{\phi}_{j,rec}^{obs} \rangle \delta_i^{[Y]} Y_i$. Let us remark that the imputation $Y_{\ell,imp}$ can also be written

$$Y_{\ell,imp} = \widehat{\Pi}_{k_n,rec}^{obs} \widehat{\Delta}_{n,rec}^{obs} \left(\widehat{\Pi}_{k_n,rec}^{obs} \widehat{\Gamma}_{k_n,rec}^{obs} \widehat{\Pi}_{k_n,rec}^{obs} \right)^{-1} X_\ell^\star, \tag{24}$$

where $\widehat{\Delta}_{n,rec}^{obs} = \frac{1}{n - m_n^{[Y]}} \sum_{i=1}^{n} \langle X_i^\star, . \rangle \delta_i^{[Y]} Y_i$.

3.2 Estimation of θ and prediction

Once the whole database has been reconstructed, we estimate the functional coefficient θ with

$$\widehat{\theta}^{\star} = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{k_n}\frac{\langle X_i^{\star}, \widehat{\phi}_{j,rec}^{\star}\rangle Y_i^{\star}}{\widehat{\lambda}_{j,rec}^{\star}}\widehat{\phi}_{j,rec}^{\star} = \sum_{j=1}^{k_n}\widehat{s}_j^{\star}\widehat{\phi}_{j,rec}^{\star}, \tag{25}$$

where $\widehat{s}_j^{\star} = \frac{1}{n\widehat{\lambda}_{j,rec}^{\star}}\sum_{i=1}^{n}\langle X_i^{\star}, \widehat{\phi}_{j,rec}^{\star}\rangle Y_i^{\star}$ and $Y_i^{\star} = Y_i\delta_i^{[Y]} + Y_{i,imp}(1-\delta_i^{[Y]})$ for all $i = 1, \ldots, n$. The estimation of the operator Θ is similarly given by

$$\widehat{\Theta}^{\star} = \langle\widehat{\theta}^{\star}, .\rangle = \widehat{\Pi}_{k_n,rec}^{\star}\widehat{\Delta}_{n,rec}^{\star}\left(\widehat{\Pi}_{k_n,rec}^{\star}\widehat{\Gamma}_{n,rec}^{\star}\widehat{\Pi}_{k_n,rec}^{\star}\right)^{-1}, \tag{26}$$

where the cross covariance operator is $\widehat{\Delta}_{n,rec}^{\star} = \frac{1}{n}\sum_{i=1}^{n}\langle X_i^{\star}, .\rangle Y_i^{\star}$, the covariance operator is $\widehat{\Gamma}_{n,rec}^{\star} = \frac{1}{n}\sum_{i=1}^{n}\langle X_i^{\star}, .\rangle X_i^{\star}$, and $\widehat{\phi}_{1,rec}^{\star}, \ldots, \widehat{\phi}_{k_n,rec}^{\star}$ and $\widehat{\lambda}_{1,rec}^{\star}, \ldots, \widehat{\lambda}_{k_n,rec}^{\star}$ represent respectively the k_n first eigenfunctions and eigenvalues of the operator $\widehat{\Gamma}_{n,rec}^{\star}$. We use this estimation to predict a new value of the response Y when a new explanatory curve X_{new} is given

$$\begin{aligned}\widehat{Y}_{new} = \widehat{\theta}_0^{\star} + \langle\widehat{\theta}^{\star}, X_{new}^{\star}\rangle &= \widehat{\theta}_0^{\star} + \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{k_n}\frac{\langle X_i^{\star}, \widehat{\phi}_{j,rec}^{\star}\rangle\langle X_{new}^{\star}, \widehat{\phi}_{j,rec}^{\star}\rangle Y_i^{\star}}{\widehat{\lambda}_{j,rec}^{\star}} \\ &= \widehat{\theta}_0^{\star} + \sum_{j=1}^{k_n}\widehat{s}_j^{\star}\langle X_{new}^{\star}, \widehat{\phi}_{j,rec}^{\star}\rangle,\end{aligned} \tag{27}$$

where $\widehat{\theta}_0^{\star} = \overline{Y}^{\star} - \int_a^b\widehat{\theta}^{\star}(t)\overline{X}^{\star}(t)\mathrm{d}t$ and $\overline{Y}^{\star} = \frac{1}{n}\sum_{i=1}^{n}Y_i^{\star}$. Then, the residuals of the fit, $\widehat{\varepsilon}_{i,k_n}^{\star} = Y_i^{\star} - \widehat{\theta}_0^{\star} - \langle X_i^{\star}, \widehat{\theta}^{\star}\rangle$, for $i = 1, \ldots, n$, can be used to estimate the error variance through

$$(\widehat{\sigma}_{\varepsilon,k_n}^{\star})^2 = \frac{1}{n-k_n-1}\sum_{i=1}^{n}(\widehat{\varepsilon}_{i,k_n}^{\star})^2.$$

3.3 Asymptotic results

The first result gives an error rate of the imputed values.

Theorem 3.1 *Under assumptions **(A.1)-(A.6)**, if we take $k_n \sim p^{1/(a_O+2)}$ and $p \sim n^{\eta_1}$ with $\eta_1 \leq 1/2$, we have*

$$\mathbb{E}\left(Y_{\ell,imp} - \theta_0 - \langle \theta, X_\ell^\star \rangle\right)^2 = \mathcal{O}_p\left(n^{-\eta_1(a_O-1)/(2(a_O+2))} + \frac{n^{\eta_1/(a_O+2)}}{n - m_n^{[Y]}}\right).$$

Moreover, the aggregate error for all the imputed values is given by

$$\sum_{\ell=1}^{n}(1-\delta_\ell^{[Y]})\mathbb{E}\left(Y_{\ell,imp} - \theta_0 - \langle \theta, X_\ell^\star \rangle\right)^2 = \mathcal{O}_p\left(m_n^{[Y]} n^{-\eta_1(a_O-1)/(2(a_O+2))} + \frac{m_n^{[Y]} n^{\eta_1/(a_O+2)}}{n - m_n^{[Y]}}\right).$$

The following corollary explores some specific cases of the above error rates. The given results simply come from a comparison between the convergence rates of the above result, hence the proof is ommited.

Corollary 3.2 *We consider cases where the number of missing values on the covariate are* (i) *negligeable with respect to the sample size,* (ii) *proportional to the sample size,* (iii) *of the same order than the sample size. More precisely*

(i) $m_n^{[Y]} = \lfloor a_n n \rfloor$ *where* a_n *goes to zero when* n *goes to infinity,*

(ii) $m_n^{[Y]} \sim \lfloor \rho n \rfloor$ *with* $0 < \rho < 1$,

(iii) $n - m_n^{[Y]} = \lfloor n^\gamma \rfloor$ *with* $0 < \gamma < 1$.

We summarize the error rate for a single imputed value and the aggregate error in Table 1.

Table 1: Single and aggregate imputation mean square error convergence rates.

		single error	aggregate error
(i) $m_n^{[Y]} = \lfloor a_n n \rfloor$		$\mathcal{O}_p\big(n^{-\eta_1(a_O-1)/(2(a_O+2))}\big)$	$\mathcal{O}_p\big(a_n n^{1-\eta_1(a_O-1)/(2(a_O+2))}\big)$
(ii) $m_n^{[Y]} \sim \lfloor \rho n \rfloor$		$\mathcal{O}_p\big(n^{-\eta_1(a_O-1)/(2(a_O+2))}\big)$	$\mathcal{O}_p\big(n^{1-\eta_1(a_O-1)/(2(a_O+2))}\big)$
(iii) $n - m_n^{[Y]} = \lfloor n^\gamma \rfloor$	$\gamma \geq \frac{\eta_1(a_O+1)}{2(a_O+2)}$	$\mathcal{O}_p\big(n^{-\eta_1(a_O-1)/(2(a_O+2))}\big)$	$\mathcal{O}_p\big(n^{1-\eta_1(a_O-1)/(2(a_O+2))}\big)$
	$\gamma < \frac{\eta_1(a_O+1)}{2(a_O+2)}$	$\mathcal{O}_p\big(n^{\eta_1/(a_O+2)-\gamma}\big)$	$\mathcal{O}_p\big(n^{1+\eta_1/(a_O+2)-\gamma}\big)$

We finish the theoretical results with the prediction error of Y_{new} with a new value of the covariate X_{new}. The proof of this result is ommited as it uses previous results of Theorems 2.3 and 3.1 and follows exactly the same lines as the proof of Theorem 2.3.

Theorem 3.3 *Under assumptions* ***(A.1)-(A.6)****, and* $k_n \sim p^{1/(a_O+2)}$ *and* $p \sim n^{\eta_1}$ *with* $\eta_1 \leq 1/2$*, the prediction error is*

$$\mathbb{E}\left(\widehat{\theta}_0^\star + \langle \widehat{\theta}^\star, X_{new}^\star \rangle - \theta_0 - \langle \theta, X_{new}^\star \rangle\right)^2 = \mathcal{O}_p\left(n^{-\eta_1(a_O-1)/(2(a_O+2))} + \frac{n^{\eta_1/(a_O+2)}}{n - m_n^{[Y]}}\right).$$

In the particular case where $\eta_1 = 1/2$*, the first term in the convergence rate is* $\mathcal{O}_p\big(n^{-(a_O-1)/(4(a_O+2))}\big)$.

As before, we consider cases in the corollary below where the number of missing values on the covariate are (i) negligeable with respect to the sample size, (ii) proportional to the sample size, (iii) of the same order than the sample size.

Corollary 3.4 *In the cases* (i)*,* (ii) *and* (iii) *with* $\gamma \geq \frac{\eta_1(a_O+1)}{2(a_O+2)}$*, the prediction error of a new value of the response is*

$$\mathbb{E}\left(\widehat{\theta}_0^\star + \langle \widehat{\theta}^\star, X_{new}^\star \rangle - \theta_0 - \langle \theta, X_{new}^\star \rangle\right)^2 = \mathcal{O}_p\big(n^{-\eta_1(a_O-1)/(2(a_O+2))}\big).$$

In the case (iii) *with* $\gamma < \frac{\eta_1(a_O+1)}{2(a_O+2)}$*, the prediction error of a new value of the response is*

$$\mathbb{E}\left(\widehat{\theta}_0^\star + \langle \widehat{\theta}^\star, X_{new}^\star \rangle - \theta_0 - \langle \theta, X_{new}^\star \rangle\right)^2 = \mathcal{O}_p\big(n^{\eta_1/(a_O+2)-\gamma}\big).$$

In other words, in situations where the number of missing values on the response is negligeable or moderate with respect to the sample size, the convergence rate of the prediction error is given by the convergence rate obtained in Kneip and Liebl (2020) for the curve reconstruction.

Remark 3.5 *As noticed at the end of the previous section, all the results obtained in this section remain valid if we replace* $X^\star$ *with* X*.*

4 Simulations

4.1 Models and samples

All the procedures described below were implemented with the R software. In the simulations, we deal with functions defined on the interval $[0,1]$. We consider the model

$$Y = \theta_0 + \langle \theta, X \rangle + \varepsilon, \tag{28}$$

where the error ε is a Gaussian noise: $\varepsilon \sim N(0, \sigma_\varepsilon^2)$ with $\sigma_\varepsilon = 0.2$. We derive three different models from (28).

First model

In this model called $Model_1$, as in Hall and Horowitz (2007), the functional covariate X is generated by a set of cosine basis functions $\phi_1 \equiv 1$ and $\phi_{j+1} = \sqrt{2}\cos(j\pi t)$ for $j > 1$, such that $X(t) = \sum_{j=1}^{150} \varrho_j \zeta_j \phi_j(t)$ for all $t \in [0,1]$, where the ζ_j's are independently sampled from the uniform distribution on $[-\sqrt{3}, \sqrt{3}]$ and the ϱ_j's are defined by $\varrho_j = (-1)^{j+1}(j)^{-\beta/2}$ with $\beta = 4$. The covariance function writes

$$cov(X(t), X(s)) = \sum_{j=1}^{150} \frac{2}{j^\beta} \cos(j\pi t)\cos(j\pi s).$$

The true parameters of the model are $\theta_0 = 3$ and $\theta := \theta_1(t)$ defined for all $t \in [0,1]$ by

$$\theta_1(t) = \sum_{j=1}^{50} b_j \phi_j(t),$$

with $b_1 = 0.3$ and $b_j = 4(-1)^{j+1} j^{-2}$ for all $j > 1$.

Second model

In this model called $Model_2$, we consider the same model as $Model_1$, with $\beta = 1$. This corresponds to a less smooth process X.

Third model

In this model called $Model_3$, even if a polynomial decrease of the eigenvalues of the covariance operator of X is required in our theoretical results (see assumption ***(A.5)***), we want to see how the method works in practice if this assumption is no more satisfied. The covariance function is defined, for $t, s \in [0,1]$ by

$$cov(X(s), X(t)) = \frac{\sigma^2 \exp(-|t-s|^{\alpha})}{\varsigma},$$

with $\sigma > 0$, $\alpha > 0$ and $\varsigma > 0$. In our case, we take $\sigma = 1$, $\alpha = 2$, and $\varsigma = 0.2$. The true slope function of the model is $\theta := \theta_2(t)$ defined for all $t \in [0,1]$ by

$$\theta_2(t) = \ln(15t^2 + 10) + \cos(4\pi t).$$

Samples

The trajectories of X_i for $i = 1, \ldots, N$ are discretized in $p = 100$ equidistant points. To comprehend the effect of sample size, we consider $n = \frac{4}{5}N$ for the training sets $(X_1, Y_1), \ldots, (X_n, Y_n)$ and $n_1 = \frac{1}{5}N$ for the test sets $(X_{n+1}, Y_{n+1}), \ldots, (X_{n+n_1}, Y_{n+n_1})$ where $N = 90$ and 1440. In each simulation, we replicated $\boldsymbol{S} = 400$ samples.

4.2 Criteria

The criteria we used are the following. Criteria 1 and 2 are related to the imputation step with the training samples, criteria 3 and 4 are related to the prediction step with the test samples.

- Criterion 1: the mean square errors (MSE) averaged over $\boldsymbol{S}$ samples

$$\overline{MSE} = \frac{1}{\boldsymbol{S}} \sum_{j=1}^{\boldsymbol{S}} MSE(j)$$

where $MSE(j) = \frac{1}{m_n^{[Y]}} \sum_{\ell=1}^{n} \left(Y_{\ell,imp}^{j} - \theta_0 - \langle \theta, X_\ell^j \rangle\right)^2 (1 - \delta_\ell)$ is the mean square error computed on the j^{th} simulated sample, $j \in \{1, \ldots, \boldsymbol{S}\}$.

- Criterion 2: the ratio respect to truth between the mean square prediction error and the mean square prediction error when the true mean is known averaged over $\boldsymbol{S}$ samples

$$\overline{RT} = \frac{1}{\boldsymbol{S}} \sum_{j=1}^{\boldsymbol{S}} RT(j),$$

where $RT(j) = \frac{\sum_{\ell=1}^{n} \left(Y_{\ell,imp}^{j} - \theta_0 - \langle \theta, X_{\ell}^{j} \rangle\right)^2 (1-\delta_\ell)}{\sum_{\ell=1}^{n} \left(\epsilon_{\ell}^{j}\right)^2 (1-\delta_\ell)}$ is the ratio between the mean square prediction error and the mean square prediction error when the true mean is known, computed on the j^{th} simulated sample.

- Criterion 3: the mean square errors (MSE') averaged over $\boldsymbol{S}$ samples

$$\overline{MSE'} = \frac{1}{\boldsymbol{S}} \sum_{j=1}^{\boldsymbol{S}} MSE'(j),$$

where $MSE'(j) = \frac{1}{n_1} \sum_{\ell'=n+1}^{n+n_1} \left(Y_{\ell'}^{j} - \theta_0 - \langle \theta, X_{\ell'}^{j} \rangle\right)^2$ is the mean square error computed on the j^{th} simulated sample, $j \in \{1, \ldots, \boldsymbol{S}\}$.

- Criterion 4: the ratio respect to truth between the mean square prediction error and the mean square prediction error when the true mean is known averaged over $\boldsymbol{S}$ samples

$$\overline{RT'} = \frac{1}{\boldsymbol{S}} \sum_{j=1}^{\boldsymbol{S}} RT'(j),$$

where $RT'(j) = \frac{\sum_{\ell'=n+1}^{n+n_1} \left(Y_{\ell'}^{j} - \theta_0 - \langle \theta, X_{\ell'}^{j} \rangle\right)^2}{\sum_{\ell'=n+1}^{n+n_1} \left(\epsilon_{\ell'}^{j}\right)^2}$ is the ratio between the mean square prediction error and the mean square prediction error when the true mean is known, computed on the j^{th} simulated sample.

Notice that all the criteria tend to zero when the sample size tends to infinity. Criteria RT and RT' are rescaled versions of MSE and MSE' if we substitute the denominator by its limit (specifically, $MSE(j) = RT(j)\sigma_\epsilon^2$).

4.3 Methodology

As in Crambes and Henchiri (2019), we use a smoothed version of the estimator (4) based on the SPCR. We use a regression spline basis with parameters: the number κ of knots of the spline functions, the degree q of spline functions and the order m of derivative. Let us remark that, with appropriate conditions, all the theoretical results obtained in our work will also apply to the SPCR estimation. Here, we have fixed $\kappa = 20$, $q = 3$ and $m = 2$. The choice of these parameters is not crucial in our study, especially in comparison with the choice of the number of principal components (see Crambes and Henchiri (2019) for more details). In this subsection, we firstly present the missing data simulation scenarios for the response and functional covariate. Secondly, we give a procedure to choose the optimal tuning parameter on a growing sequence of dimension $k_n = 2, \ldots, 22$.

Missing data simulation scenario

In our simulations, we have adopted the following scenario to determine the number of missing data on the response Y as in Crambes and Henchiri (2019): we simulate δ according to the logistic functional regression. The variable δ follows the Bernoulli law with parameter $p(X)$ such that

$$\log\big(\frac{p(X)}{1-p(X)}\big) = \big\langle \alpha_0, X \big\rangle + c,$$

where $\alpha_0 = \sin(2\pi t)$ for all $t \in [0,1]$ and c is a constant allowing to take different levels of missing data. For exemple $c = 2$ for around 12.02% of missing data, $c = 1$ for around 26.91% of missing data and $c = 0.2$ for around 45.08% of missing data.

To generate missing data on the covariate, we have adopted the missing data simulation scenario as in Kneip and Liebl (2020) such that

- 85% (respectively 70%) of the curves are fully observed on $[0,1]$,
- and for the 15% (respectively 30%) of curves affected by missing data, the curve X_i is fully observed on $[A_i, B_i] \subset [0,1]$ with A_i drawn with uniform law on the interval $[0, A]$ and $B_i = A_i + B$, in the two following cases

– A = 1/8 and B = 7/8 (case corresponding to $Model_1$, $Model_2$ and $Model_3$),

– A = 3/8 and B = 6/8 (case corresponding to $Model'_1$, $Model'_2$ and $Model'_3$).

Choice of the optimal parameter

Theoretical results are generally obtained under assumptions concerning the rate of convergence of the integer k_n. In practice, this integer is selected by minimizing a certain empirical criterion, for example the Generalized Cross Validation (GCV) criterion, the Cross Validation (CV) criterion and the K-fold Cross Validation (K-fold CV) criterion (see Crambes and Henchiri (2019)). In our simulations, we chose the GCV procedure, known to be computationally fast. The GCV criterion is given below for imputation

$$\mathrm{GCV}(k_n) = \frac{(n - m_n^{[Y]}) \sum_{i=1}^{n} (\widehat{Y}_i - \theta_0 - \langle \theta, X_i \rangle)^2 \delta_i}{((n - m_n^{[Y]}) - k_n)^2},$$

and the analogous criterion for prediction

$$\mathrm{GCV}(k_n) = \frac{n \sum_{i=1}^{n} (\widehat{Y}_i - \theta_0 - \langle \theta, X_i \rangle)^2}{(n - k_n)^2}.$$

4.4 Analysis of results

The complete results can be found in supplementary material. Tables 1-12 in (A1) give the mean and standard deviation errors for the imputed values on training samples for different models. Tables 1-12 in (A2) give the mean and standard deviation errors for the predicted values on test samples for different models. These errors are computed according to the different cases listed below.

- Case 1: X and Y are fully observed, this corresponds to the complete reference dataset.

- Case 2: X is fully observed and Y is affected with missing values, which will be imputed according to the method presented in Crambes and Henchiri (2019).

- Case 3: X is affected with missing values and Y is fully observed, the missing parts of X will be reconstructed according to the method from Kneip and Liebl (2020) presented in section 2 of this paper.

- Case 4: X and Y are affected with missing values, the missing parts of X are reconstructed and the missing values of Y are imputed, according to the method presented in this paper.

- Case 5: X and Y are affected with missing values, these missing data are removed from the sample.

As it can be expected, the errors decrease as the sample size increases. Moreover, the regularity of the process X does not seem to have a crucial impact on the results. The main point we want to discuss is related to the level of missing data in the sample, in particular with respect to cases 4 and 5. The most favorable situation for our method (case 4) seems to be when there is a quite small sample size, and when the missing part of the curves is not so much important (see for example Table 1 in (A2)). In this situation, our method (case 4) behaves better than the naive method (case 5). Our method show a real advantage in reconstructing the missing parts of the curves and imputing the missing values of the response. It is particularly clear when the percentage of missing data on Y increases. The difference between cases 4 and 5 narrows (generally still in favor of case 4, though) when the sample size increases (see for example Table 2 in (A2)). In this situation, even if we have important percentages of missing data on Y, there are enough remaining data in the sample. Finally, when the missing part of the curves is more important (see for example Tables 3 and 4 in (A2)), the curve reconstruction from Kneip and Liebl (2020) is more difficult, and the difference between cases 4 and 5 also narrows in this situation.

5 Real dataset study

In order to illustrate our contribution in a prediction setting when the functional covariate and the real response are affected with missing data, we present in this section an energy consumption and economic dataset application.

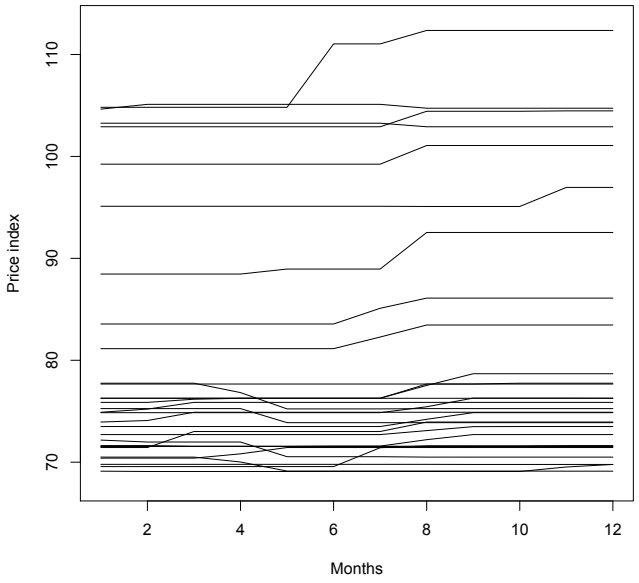

Figure 1: Plot of the 30 electricity price index curves.

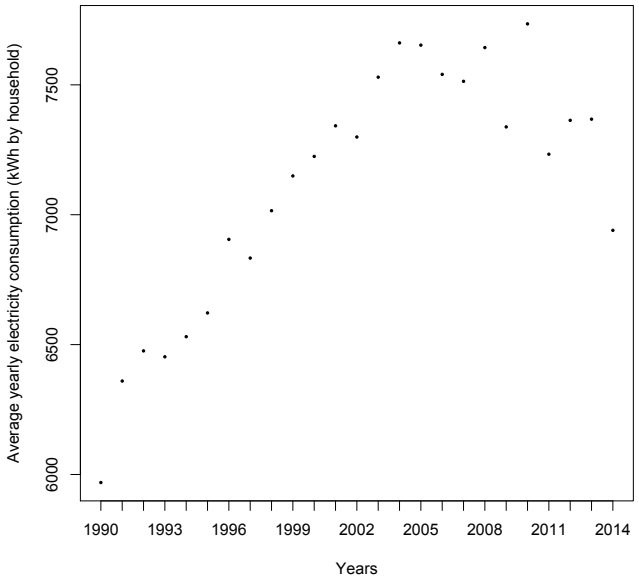

Figure 2: Plot of the average yearly electricity consumption (kWh by household).

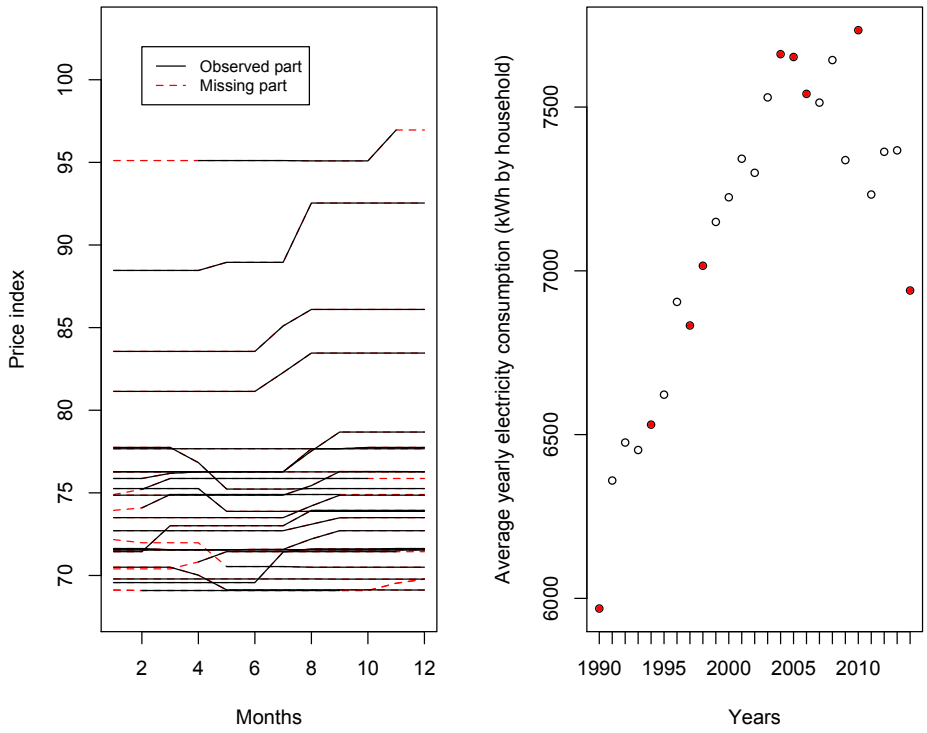

Figure 3: Example of missing data simulation scenario on X and Y: 24% of missing data on X and 36% of missing data on Y.

We start by describing the dataset. The functional covariate X is the electricity price index curve in France from January 1990 to December 2019[1]. This yearly curve is observed at $p = 12$ discretization points corresponding to the monthly values. The graphical display of the electricity price index curves can be observed in Figure 1. The response variable Y is the average yearly electricity consumption in France[2]. The response variable goes up to 2014.

In the following, we consider missing data in this sample, created as in the simulations. From the sample of 25 pairs $(Y_i, X_i)_{i=1,\ldots,25}$, we want to forecast the yearly electricity consumption for the years 2015, 2016, 2017, 2018 and 2019.

[1]http://www.insee.fr/fr/statistiques/serie/001763554

[2]https://donnees.banquemondiale.org/indicateur/EG.USE.ELEC.KH.PC?locations=FR

We have fixed the number of knots $\kappa = 12$, the degree of splines $q = 3$ and the order of derivative $m = 2$. Then, we use the GCV criterion to find the best parameter of projection dimension k_n trying growing sequences: $k_n = 2, 3, \ldots, 11, 12$.

In order to see the impact of missing data on this dataset, we have randomly drawn 700 tests samples in the initial sample and computed prediction errors on these tests samples, using the remaining of the sample as training sample. Results are given in Table 2. To measure the performance of the prediction, we consider mean absolute errors (MAE) averaged over $\boldsymbol{S} = 700$ samples. The first mean absolute error is defined by

$$\overline{MAE} = \frac{1}{\boldsymbol{S}} \sum_{i=1}^{\boldsymbol{S}} MAE(i),$$

where $MAE(i) = \frac{1}{5} \sum_{j=2015}^{2019} \left| \widehat{Y}_j^i - \widehat{Y}_j^{\star,i} \right|$ is the mean absolute error computed on the i^{th} simulated testing sample, for $i \in \{1, \ldots, \boldsymbol{S}\}$. This error corresponds to our method (case 4 presented in the simulations) when we reconstruct the missing parts of the curves and we imput missing data on the response. The second mean absolute error is defined by

$$\overline{MAE'} = \frac{1}{\boldsymbol{S}} \sum_{i=1}^{\boldsymbol{S}} MAE'(i),$$

where $MAE'(i) = \frac{1}{5} \sum_{j=2015}^{2019} \left| \widehat{Y}_j^i - \widehat{Y}_j^{suppr,i} \right|$ is the mean absolute error computed on the i^{th} simulated testing sample $i \in \{1, \ldots, \boldsymbol{S}\}$. This error corresponds to the naive method (case 5 presented in the simulations): $\widehat{Y}_j^{suppr,i}$ is the prediction of Y_j for the i^{th} simulated testing sample, when we remove the missing data from the sample. We can see that, in this illustrating example, our method (case 4) favorably compares again to the naive method (case 5).

Finally, the predicted values of yearly electricity consumption for the years 2015, 2016, 2017, 2018 and 2019 are given in Table 3

Table 2: Real dataset: prediction errors over 700 drawn tests samples.

Rate of missing data (%) of Y	42.6
Rate of missing data (%) of X	20.06
$\overline{MAE}$	858.9762
	(756.4337)
$\overline{MAE'}$	1932.134
	(2514.136)

Table 3: Predicted values of yearly electricity consumption.

year	2015	2016	2017	2018	2019
case 1	9006.685	10284.094	9454.371	10009.066	10112.296
case 4	9190.074	10198.675	9616.474	10293.353	9502.263
case 5	9451.357	10204.158	9842.176	10167.635	8420.580

6 Proofs

Proof of Proposition 2.2

For any $x \in H$ such that $\|x\| = 1$, we have

$$\widehat{\Gamma}_{n,rec}x - \widehat{\Gamma}_n x = \frac{1}{n}\sum_{i=1}^{n}\langle X_i^\star - X_i, x\rangle X_i + \frac{1}{n}\sum_{i=1}^{n}\langle X_i, x\rangle\,(X_i^\star - X_i) + \frac{1}{n}\sum_{i=1}^{n}\langle X_i^\star - X_i, x\rangle\,(X_i^\star - X_i)\,.$$

Using the Cauchy-Schwarz inequality, we get

$$\|\langle X_i^\star - X_i, x\rangle X_i\| \leq \|X_i^\star - X_i\|\,\|x\|\,\|X_i\|\,,$$

from which we deduce with (21) that

$$\|\langle X_i^\star - X_i, x\rangle X_i\| = \mathcal{O}_p\big(p^{-(a_O-1)/(2(a_O+2))}\big)\,.$$

We prove in the same way that

$$\|\langle X_i, x\rangle\,(X_i^\star - X_i)\| = \mathcal{O}_p\big(p^{-(a_O-1)/(2(a_O+2))}\big)\,,$$

and

$$\|\langle X_i^\star - X_i, x\rangle\,(X_i^\star - X_i)\| = \mathcal{O}_p\big(p^{-(a_O-1)/(a_O+2)}\big)\,,$$

which gives the first result (i). The result (ii) can be shown exactly the same way. Concerning results (iii) and (iv), they are directly deduced from (i) and respectively Lemma 2.3 and Lemma 2.2 in Horváth and Kokoszka (2012).

Proof of Theorem 2.3

We start with the decomposition

$$\begin{aligned}
&\mathbb{E}\left(\langle\widehat{\widehat{\theta}}, X_{new}^\star\rangle - \langle\theta, X_{new}^\star\rangle\right)^2 \\
&= \mathbb{E}\left(\widehat{\Pi}_{k_n,rec}\widehat{\Delta}_{n,rec}\left(\widehat{\Pi}_{k_n,rec}\widehat{\Gamma}_{n,rec}\widehat{\Pi}_{k_n,rec}\right)^{-1} X_{new}^\star - \Theta X_{new}^\star\right)^2 \\
&\leq 2\mathbb{E}\left(\widehat{\Pi}_{k_n,rec}\Theta\widehat{\Gamma}_{n,rec}\left(\widehat{\Pi}_{k_n,rec}\widehat{\Gamma}_{n,rec}\widehat{\Pi}_{k_n,rec}\right)^{-1} X_{new}^\star - \Theta X_{new}^\star\right)^2 \\
&+ 2\mathbb{E}\left(\widehat{\Pi}_{k_n,rec}\left(\frac{1}{n}\sum_{i=1}^n \langle X_i^\star, .\rangle\varepsilon_i\right)\left(\widehat{\Pi}_{k_n,rec}\widehat{\Gamma}_{n,rec}\widehat{\Pi}_{k_n,rec}\right)^{-1} X_{new}^\star - \Theta X_{new}^\star\right)^2.
\end{aligned}$$

Applying several times the identity $(a+b)^2 \leq 2a^2 + 2b^2$ for any $a, b \in \mathbb{R}$, we get

$$\begin{aligned}
\mathbb{E}\left(\langle\widehat{\widehat{\theta}}, X_{new}^\star\rangle - \langle\theta, X_{new}^\star\rangle\right)^2 \leq\ & 32\mathbb{E}\left(\Theta\widehat{\Pi}_{k_n,rec}X_{new}^\star - \Theta\widehat{\Pi}_{k_n}X_{new}^\star\right)^2 \\
&+ 32\mathbb{E}\left(\Theta\widehat{\Pi}_{k_n}X_{new}^\star - \Theta\widehat{\Pi}_{k_n}X_{new}\right)^2 \\
&+ 16\mathbb{E}\left(\Theta\widehat{\Pi}_{k_n}X_{new} - \Theta\Pi_{k_n}X_{new}\right)^2 \\
&+ 8\mathbb{E}\left(\Theta\Pi_{k_n}X_{new} - \Theta X_{new}\right)^2 \\
&+ 4\mathbb{E}\left(\Theta X_{new} - \Theta X_{new}^\star\right)^2 \\
&+ 2\mathbb{E}\left(\frac{1}{n}\sum_{i=1}^n \langle X_i^\star, \left(\widehat{\Pi}_{k_n,rec}\widehat{\Gamma}_{n,rec}\widehat{\Pi}_{k_n,rec}\right)^{-1} X_{new}^\star\rangle\varepsilon_i\right)^2.
\end{aligned}$$

We start with the first term in the above decomposition $A_1 := 32\mathbb{E}\left(\Theta\widehat{\Pi}_{k_n,rec}X_{new}^\star - \Theta\widehat{\Pi}_{k_n}X_{new}^\star\right)^2$. Applying Lemma 5.1 in Crambes and Henchiri (2019), we obtain

$$A_1 = o\left(\frac{\widehat{\lambda}_{k_n} k_n^2}{n} + \frac{k_n}{n}\right).$$

With Lemma 2.2 in Horváth and Kokoszka (2012), we get

$$A_1 = o\left(\frac{\lambda_{k_n} k_n^2}{n} + \frac{k_n}{n}\right).$$

Now, we use (21) to obtain

$$A_2 := 32\mathbb{E}\left(\Theta\widehat{\Pi}_{k_n} X^\star_{new} - \Theta\widehat{\Pi}_{k_n} X_{new}\right)^2 = \mathcal{O}_p\big(p^{-(a_O-1)/(2(a_O+2))}\big).$$

Moreover, again with Lemma 5.1 in Crambes and Henchiri (2019), we obtain

$$A_3 := 16\mathbb{E}\left(\Theta\widehat{\Pi}_{k_n} X_{new} - \Theta\Pi_{k_n} X_{new}\right)^2 = o\left(\frac{\lambda_{k_n} k_n^2}{n} + \frac{k_n}{n}\right).$$

We go on with $A_4 := 8\mathbb{E}\left(\Theta\Pi_{k_n} X_{new} - \Theta X_{new}\right)^2$. With Lemma 5.3 in Crambes and Henchiri (2019), we get

$$A_4 = 8\sum_{j=k_n+1}^{+\infty}\left(\Theta\Gamma^{1/2}\phi_j\right)^2.$$

Next, using again (21), we can write

$$A_5 := 4\mathbb{E}\left(\Theta X_{new} - \Theta X^\star_{new}\right)^2 = \mathcal{O}_p\big(p^{-(a_O-1)/(2(a_O+2))}\big).$$

Finally, the last term of the decomposition comes from Lemma 5.2 in Crambes and Henchiri (2019) and gives

$$A_6 := 2\mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n}\langle X_i^\star, \left(\widehat{\Pi}_{k_n,rec}\widehat{\Gamma}_{n,rec}\widehat{\Pi}_{k_n,rec}\right)^{-1} X^\star_{new}\rangle\varepsilon_i\right)^2 = \frac{2\sigma_\varepsilon^2 k_n}{n} + o\left(\frac{k_n}{n}\right).$$

We can now conclude the proof of Theorem 2.3. The decomposition from the beginning of the proof gives

$$\mathbb{E}\left(\langle\widehat{\widehat{\theta}}, X^\star_{new}\rangle - \langle\theta, X^\star_{new}\rangle\right)^2 = \mathcal{O}_p\left(\sum_{j=k_n+1}^{+\infty}\left(\Theta\Gamma^{1/2}\phi_j\right)^2 + p^{-(a_O-1)/(2(a_O+2))} + \frac{\sigma_\varepsilon^2 k_n}{n}\right) + o\left(\frac{\lambda_{k_n}k_n^2}{n} + \frac{k_n}{n}\right).$$

The first term in the convergence rate is

$$\sum_{j=k_n+1}^{+\infty}\left(\Theta\Gamma^{1/2}\phi_j\right)^2 = \sum_{j=k_n+1}^{+\infty}\lambda_j\left(\Theta\phi_j\right)^2 \leq \sum_{j=k_n+1}^{+\infty} j^{-a_O}.$$

Comparing the latter sum to an integral, we get

$$\sum_{j=k_n+1}^{+\infty}\left(\Theta\Gamma^{1/2}\phi_j\right)^2 = \mathcal{O}\big(k_n^{-(a_O+1)}\big) = \mathcal{O}\big(p^{-(a_O+1)/(a_O+2)}\big) = \mathcal{O}\big(n^{-\eta_1(a_O+1)/(a_O+2)}\big).$$

The second term in the convergence rate is

$$p^{-(a_O-1)/(2(a_O+2))} \sim n^{-\eta_1(a_O-1)/(2(a_O+2))},$$

and the third term in the convergence rate is

$$\frac{\sigma_\varepsilon^2 k_n}{n} \sim \frac{\sigma_\varepsilon^2 n^{\eta_1/(a_O+2)}}{n} = \sigma_\varepsilon^2 n^{\eta_1/(a_O+2)-1}.$$

If we compare the different rates, with the condition $\eta_1 \leq 1/2$, we get

$$\mathbb{E}\left(\langle\widehat{\widehat{\theta}}, X^\star_{new}\rangle - \langle\theta, X^\star_{new}\rangle\right)^2 = \mathcal{O}_p\big(n^{-\eta_1(a_O-1)/(2(a_O+2))}\big).$$

Finally, we can write

$$\begin{aligned}\mathbb{E}\left(\widehat{\widehat{\theta}}_0 + \langle\widehat{\widehat{\theta}}, X^\star_{new}\rangle - \theta_0 - \langle\theta, X^\star_{new}\rangle\right)^2 &= \mathbb{E}\left(\overline{Y} - \theta_0 - \langle\theta, \overline{X^\star}\rangle + \langle\widehat{\widehat{\theta}}, X^\star_{new} - \overline{X^\star}\rangle - \langle\theta, X^\star_{new} - \overline{X^\star}\rangle\right)^2 \\ &\leq 2\mathbb{E}\left(\overline{Y} - \mathbb{E}(Y)\right)^2 + 2\mathbb{E}\left(\langle\widehat{\widehat{\theta}}, X^\star_{new} - \overline{X^\star}\rangle - \langle\theta, X^\star_{new} - \overline{X^\star}\rangle\right)^2.\end{aligned}$$

The first term of the right-hand side is given by $\mathbb{E}\left(\overline{Y} - \mathbb{E}(Y)\right)^2 \leq 2\overline{\varepsilon}^2 + 2\langle\theta, \mathbb{E}(X)\rangle^2 = \mathcal{O}_p(n^{-1})$ (with Bienaymé-Tchebychev inequality), and the second term of the right-hand

side gives a convergence rate in probability of $n^{-\eta_1(a_O-1)/(2(a_O+2))}$, which gives the desired result

$$\mathbb{E}\left(\widehat{\widehat{\theta}}_0 + \langle\widehat{\widehat{\theta}}, X^\star_{new}\rangle - \theta_0 - \langle\theta, X^\star_{new}\rangle\right)^2 = \mathcal{O}_p\big(n^{-\eta_1(a_O-1)/(2(a_O+2))}\big)\,.$$

Proof of Theorem 3.1

This proof follows the same lines as the proof of Theorem 2.3. We write the decomposition

$$\begin{aligned}
\mathbb{E}\left(\langle\widetilde{\theta}, X^\star_\ell\rangle - \langle\theta, X^\star_\ell\rangle\right)^2 \leq\ & 32\mathbb{E}\left(\Theta\widehat{\Pi}^{obs}_{k_n,rec}X^\star_\ell - \Theta\widehat{\Pi}_{k_n}X^\star_\ell\right)^2 \\
& + 32\mathbb{E}\left(\Theta\widehat{\Pi}^{obs}_{k_n}X^\star_\ell - \Theta\widehat{\Pi}_{k_n}X_\ell\right)^2 \\
& + 16\mathbb{E}\left(\Theta\widehat{\Pi}_{k_n}X_\ell - \Theta\Pi_{k_n}X_{new}\right)^2 \\
& + 8\mathbb{E}\left(\Theta\Pi_{k_n}X_\ell - \Theta X_\ell\right)^2 \\
& + 4\mathbb{E}\left(\Theta X_\ell - \Theta X^\star_\ell\right)^2 \\
& + 2\mathbb{E}\left(\frac{1}{n}\sum_{i=1}^n \langle X^\star_i, \left(\widehat{\Pi}^{obs}_{k_n,rec}\widehat{\Gamma}^{obs}_{n,rec}\widehat{\Pi}^{obs}_{k_n,rec}\right)^{-1} X^\star_\ell\rangle \delta^{[Y]}_i \varepsilon_i\right)^2 .
\end{aligned}$$

The first term in the above decomposition $B_1 := 32\mathbb{E}\left(\Theta\widehat{\Pi}^{obs}_{k_n,rec}X^\star_\ell - \Theta\widehat{\Pi}_{k_n}X^\star_\ell\right)^2$. Applying Lemma 5.1 in Crambes and Henchiri (2019) and Lemma 2.2 in Horváth and Kokoszka (2012), we get

$$B_1 = o\left(\frac{\lambda_{k_n}k_n^2}{n - m_n^{[Y]}} + \frac{k_n}{n - m_n^{[Y]}}\right).$$

Now, we use (21) to obtain

$$B_2 := 32\mathbb{E}\left(\Theta\widehat{\Pi}^{obs}_{k_n}X^\star_\ell - \Theta\widehat{\Pi}_{k_n}X_\ell\right)^2 = \mathcal{O}_p\big(p^{-(a_O-1)/(2(a_O+2))}\big)\,.$$

Again with Lemma 5.1 in Crambes and Henchiri (2019), we obtain

$$B_3 := 16\mathbb{E}\left(\Theta\widehat{\Pi}_{k_n}X_\ell - \Theta\Pi_{k_n}X_{new}\right)^2 = o\left(\frac{\lambda_{k_n}k_n^2}{n} + \frac{k_n}{n}\right).$$

The next term is $B_4 := 8\mathbb{E}\left(\Theta\Pi_{k_n}X_\ell - \Theta X_\ell\right)^2$. With Lemma 5.3 in Crambes and Henchiri (2019), we get

$$B_4 = 8\sum_{j=k_n+1}^{+\infty}\left(\Theta\Gamma^{1/2}\phi_j\right)^2.$$

Then, using again (21), we can write

$$B_5 := 4\mathbb{E}\left(\Theta X_\ell - \Theta X_\ell^\star\right)^2 = \mathcal{O}_p\big(p^{-(a_O-1)/(2(a_O+2))}\big).$$

Finally, the last term of the decomposition comes from Lemma 5.2 in Crambes and Henchiri (2019) and gives

$$B_6 := 2\mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n}\langle X_i^\star, \left(\widehat{\Pi}_{k_n,rec}^{obs}\widehat{\Gamma}_{n,rec}^{obs}\widehat{\Pi}_{k_n,rec}^{obs}\right)^{-1}X_\ell^\star\rangle\delta_i^{[Y]}\varepsilon_i\right)^2 = \frac{2\sigma_\varepsilon^2 k_n}{n-m_n^{[Y]}} + o\left(\frac{k_n}{n-m_n^{[Y]}}\right).$$

We can now conclude the proof of Theorem 3.1. Coming back to the decomposition from the beginning, we get

$$\begin{aligned}\mathbb{E}\left(\langle\widetilde{\theta}, X_\ell^\star\rangle - \langle\theta, X_\ell^\star\rangle\right)^2 = \mathcal{O}_p\left(\sum_{j=k_n+1}^{+\infty}\left(\Theta\Gamma^{1/2}\phi_j\right)^2 + p^{-(a_O-1)/(2(a_O+2))} + \frac{\sigma_\varepsilon^2 k_n}{n-m_n^{[Y]}}\right)\\ + o\left(\frac{\lambda_{k_n}k_n^2}{n-m_n^{[Y]}} + \frac{k_n}{n-m_n^{[Y]}}\right).\end{aligned}$$

Comparing the convergence rates, we obtain

$$\mathbb{E}\left(\langle\widetilde{\theta}, X_\ell^\star\rangle - \langle\theta, X_\ell^\star\rangle\right)^2 = \mathcal{O}_p\left(n^{-\eta_1(a_O-1)/(2(a_O+2))} + \frac{n^{\eta_1/(a_O+2)}}{n-m_n^{[Y]}}\right).$$

Finally, we can get the desired result including the intercept. We follow the end of the proof of Theorem 2.3 to write

$$\begin{aligned}\mathbb{E}\left(\widetilde{\theta}_0 + \langle\widetilde{\theta}, X_\ell^\star\rangle - \theta_0 - \langle\theta, X_\ell^\star\rangle\right)^2 &= \mathbb{E}\left(\overline{Y}_{obs} - \theta_0 - \langle\theta, \overline{X}_{obs}^\star\rangle + \langle\widehat{\widetilde{\theta}}, X_\ell^\star - \overline{X}_{obs}^\star\rangle - \langle\theta, X_\ell^\star - \overline{X}_{obs}^\star\rangle\right)^2\\ &\leq 2\mathbb{E}\left(\overline{Y}_{obs} - \mathbb{E}(Y)\right)^2 + 2\mathbb{E}\left(\langle\widetilde{\theta}, X_\ell^\star - \overline{X}_{obs}^\star\rangle - \langle\theta, X_\ell^\star - \overline{X}_{obs}^\star\rangle\right)^2.\end{aligned}$$

The first term of the right-hand side is given by $\mathbb{E}\left(\overline{Y}_{obs} - \mathbb{E}(Y)\right)^2 \leq 2\bar{\varepsilon}^2_{obs} + 2\langle\theta, \mathbb{E}(X)\rangle^2 = \mathcal{O}_p\Big((n - m_n^{[Y]})^{-1}\Big)$ (with Bienaymé-Tchebychev inequality), and the second term of the right-hand side gives a convergence rate in probability of $n^{-\eta_1(a_O-1)/(2(a_O+2))} + \frac{n^{\eta_1/(a_O+2)}}{n-m_n^{[Y]}}$, which gives

$$\mathbb{E}\left(\widetilde{\theta}_0 + \langle\widetilde{\theta}, X_\ell^\star\rangle - \theta_0 - \langle\theta, X_\ell^\star\rangle\right)^2 = \mathcal{O}_p\left(n^{-\eta_1(a_O-1)/(2(a_O+2))} + \frac{n^{\eta_1/(a_O+2)}}{n - m_n^{[Y]}}\right).$$

Proof of Theorem 3.3

Following the same lines of previous proofs but first we write the cross covariance operator as

$$\begin{aligned}
\widehat{\Delta}^\star_{n,rec} =& \frac{1}{n}\sum_{i=1}^n \langle X_i^\star, .\rangle Y_i^\star \\
=& \frac{1}{n}\sum_{i=1}^n \langle X_i^\star, .\rangle \left(Y_i\delta_i^{[Y]} + Y_{i,imp}(1-\delta_i^{[Y]})\right) \\
=& \frac{1}{n}\sum_{i=1}^n \langle X_i^\star, .\rangle \delta_i^{[Y]} Y_i + \frac{1}{n}\sum_{i=1}^n \langle X_i^\star, .\rangle (1-\delta_i^{[Y]}) Y_{i,imp}.
\end{aligned}$$

Next, we observe that

$$\begin{aligned}
&\mathbb{E}\left(\langle\widehat{\theta}^\star, X^\star_{new}\rangle - \langle\theta, X^\star_{new}\rangle\right)^2 \\
&= \mathbb{E}\left(\widehat{\Pi}_{k_n,rec}\widehat{\Delta}^\star_{n,rec}\left(\widehat{\Pi}_{k_n,rec}\widehat{\Gamma}_{n,rec}\widehat{\Pi}_{k_n,rec}\right)^{-1} X^\star_{new} - \Theta X^\star_{new}\right)^2 \\
&\leq 2\mathbb{E}\left(\widehat{\Pi}_{k_n,rec}\frac{1}{n}\sum_{i=1}^n \langle X_i^\star, .\rangle \delta_i^{[Y]} Y_i \left(\widehat{\Pi}_{k_n,rec}\widehat{\Gamma}_{n,rec}\widehat{\Pi}_{k_n,rec}\right)^{-1} X^\star_{new} - \Theta X^\star_{new}\right)^2 \\
&+ 2\mathbb{E}\left(\widehat{\Pi}_{k_n,rec}\left(\frac{1}{n}\sum_{i=1}^n \langle X_i^\star, .\rangle Y_{i,imp}(1-\delta_i^{[Y]})\right)\left(\widehat{\Pi}_{k_n,rec}\widehat{\Gamma}_{n,rec}\widehat{\Pi}_{k_n,rec}\right)^{-1} X^\star_{new} - \Theta X^\star_{new}\right)^2.
\end{aligned}$$

The first term is given by the result of Theorem 2.3. For the second term

$$\mathbb{E}\left(\widehat{\Pi}_{k_n,rec}\left(\frac{1}{n}\sum_{i=1}^{n}\langle X_i^\star,.\rangle Y_{i,imp}(1-\delta_i^{[Y]})\right)\left(\widehat{\Pi}_{k_n,rec}\widehat{\Gamma}_{n,rec}\widehat{\Pi}_{k_n,rec}\right)^{-1}X_{new}^\star-\Theta X_{new}^\star\right)^2$$
$$\leq 2\mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n}\langle X_i^\star,\left(\widehat{\Gamma}_{n,rec}\widehat{\Pi}_{k_n,rec}\right)^{-1}X_{new}^\star(Y_{i,imp}-Y_i)(1-\delta_i^{[Y]})\right)$$
$$+2\mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n}\langle X_i^\star,\left(\widehat{\Gamma}_{n,rec}\widehat{\Pi}_{k_n,rec}\right)^{-1}X_{new}^\star\rangle Y_i(1-\delta_i^{[Y]})-\Theta X_{new}^\star\right)^2.$$

We notice that the first term above is exactly the same as in Theorem 3.1 and the second term is directly the result of the Theorem 2.3. So, comparing the convergence rates, we get

$$\mathbb{E}\left(\langle\widehat{\theta^\star},X_{new}^\star\rangle-\langle\theta,X_{new}^\star\rangle\right)^2=\mathcal{O}_p\left(n^{-\eta_1(a_O-1)/(2(a_O+2))}+\frac{n^{\eta_1/(a_O+2)}}{n-m_n^{[Y]}}\right),$$

which gives the desired result.

SUPPLEMENTARY MATERIAL

Title: Imputation values (A1).

Title: Predicted values (A2).

References

Cai, T. and Hall, P. (2006), "Prediction in functional linear regression," *The Annals of Statistics*, 34, 2159–2179.

Cardot, H., Ferraty, F., and Sarda, P. (1999), "Functional linear model," *Statistics and Probability Letters*, 45, 11–22.

— (2003), "Spline estimators for the functional linear model," *Statistica Sinica*, 13, 571–591.

Crambes, C. and Henchiri, Y. (2019), "Regression imputation in the functional linear model with missing values in the response," *Journal of Statistical Planning and Inference*, 201, 103–119.

Crambes, C., Kneip, A., and Sarda, P. (2009), "Smoothing splines estimators for functional linear regression," *The Annals of statistics*, 37, 35–72.

Delaigle, A. and Hall, P. (2016), "Approximating fragmented functional data by segments of Markov chains," *Biometrika*, 103, 779–799.

Delaigle, A., Hall, P., Huang, W., and Kneip, A. (2020), "Estimating the covariance of fragmented and other related types of functional data," *Journal of the American Statistical Association*, to appear.

Descary, M.-H. and Panaretos, V. (2019), "Recovering covariance from functional fragments," *Biometrika*, 106, 145–160.

Febrero-Bande, M., Galeano, P., and Gonzalez-Manteiga, W. (2019), "Estimation, imputation and prediction for the functional linear model with scalar response with responses missing at random," *Computational Statistics and Data Analysis*, 131, 91–103.

Ferraty, F., Sued, M., and Vieu, P. (2013), "Mean estimation with data missing at random for functional covariables," *Statistics*, 47, 688–706.

Ferraty, F. and Vieu, P. (2006), *Nonparametric functional data analysis: Theory and practice*, Springer-Verlag, New York.

Graham, J. (2012), *Missing data analysis and design*, Springer-Verlag, New York.

Hall, P. and Horowitz, J. L. (2007), "Methodology and convergence rates for functional linear regression," *The Annals of Statistics*, 35, 70–91.

Horváth, L. and Kokoszka, P. (2012), *Inference for functional data with applications*, Springer-Verlag, New York.

Hsing, T. and Eubank, R. (2015), *Theoretical foundations of functional data analysis, with an introduction to linear operators*, Wiley series in probability and statistics, John Wiley & Sons.

Kneip, A. and Liebl, D. (2020), "On the optimal reconstruction of partially observed functional data," *The Annals of Statistics*, 4, 1692–1717.

Kraus, D. (2015), "Components and completion of partially observed functional data," *Journal of the Royal Statistical Society: Series B*, 77, 777–801.

Kraus, D. and Stefanucci, M. (2020), "Ridge reconstruction of partially observed functional data is asymptotically optimal," *Statistics and Probability Letters*, 165, DOI: 10.1016/j.spl.2020.108813.

Lin, Z. and Wang, J. L. (2020), "Mean and covariance estimation for functional snippets," *Technical report, University of California, Davis*.

Lin, Z., Wang, J.-L., and Zhong, Q. (2020), "Basis Expansions for Functional Snippets," *Biometrika*, asaa088, 1–18.

Little, R. and Rubin, D. B. (2002), *Statistical analysis with missing data (Second edition)*, John Wiley, New York.

Ramsay, J. O. and Silverman, B. W. (2005), *Statistical analysis with missing data (Second edition)*, Springer-Verlag, New York.