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

Christophe Crambes, Chayma Daayeb, Ali Gannoun, Yousri Henchiri

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

Christophe Crambes [1], Chayma Daayeb [1, 2], Ali Gannoun [1], and Yousri Henchiri [2, 3]

[1] Institut Montpelliérain Alexander Grothendieck (IMAG), Université de Montpellier, France.

[2] Université de Tunis El Manar, Laboratoire de Modélisation Mathématique et Numérique dans les Sciences de l'Ingénieur (ENIT-LAMSIN), Tunisie.

[3] Université de la Manouba, Institut Supérieur des Arts Multimédia de la Manouba (ISAMM), Tunisie.

Address for correspondence: Yousri Henchiri, Université de la Manouba, Institut Supérieur des Arts Multimédia de la Manouba (ISAMM), Tunisie.

E-mail: `yousri.henchiri@umontpellier.fr`.

Phone: (+216) 56 272 229.

Fax: (+216) 71 603 450.

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 with partially observed covariate and missing values in the response. We use a reconstruction operator that aims at recovering 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 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.

Key words: Functional linear model; Functional Principal Components; Missing data; Missing At Random; Missing Completely At Random; Regression imputation.

1 Introduction

The analysis of 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 in considering 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 $\mathbb{R}$. We assume that X is centered, that is $\mathbb{E}(X(t)) = 0$ for all $t \in [a, b]$. 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

$$Y = \theta_0 + \int_a^b \theta(t)X(t)\mathrm{d}t + \varepsilon, \tag{1.1}$$

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 (1.1) as

$$Y = \theta_0 + \Theta X + \varepsilon, \tag{1.2}$$

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 $(X_i, Y_i)_{i=1,\ldots,n}$ 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 $(\widehat{\phi}_j)_{j=1,\ldots,k_n}$ the eigenfunctions associated to $\widehat{\Gamma}_n$ corresponding to the k_n highest eigenvalues $\widehat{\lambda}_1 > \ldots > \widehat{\lambda}_{k_n} > 0$ (where k_n is an integer depending on n), we define the orthogonal projection operator $\widehat{\Pi}_{k_n}$ onto the subspace $\mathrm{Span}(\widehat{\phi}_1, \ldots, \widehat{\phi}_{k_n})$ 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

$$\widehat{\Theta} = \langle \widehat{\theta}, . \rangle = \widehat{\Pi}_{k_n} \widehat{\Delta}_n (\widehat{\Pi}_{k_n} \widehat{\Gamma}_n \widehat{\Pi}_{k_n})^{-1}. \tag{1.3}$$

The corresponding estimator of θ is given by

$$\widehat{\theta} = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{k_n} \frac{\langle X_i, \widehat{\phi}_j \rangle Y_i}{\widehat{\lambda}_j} \widehat{\phi}_j = \sum_{j=1}^{k_n} \widehat{s}_j \widehat{\phi}_j, \tag{1.4}$$

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 θ_0 is $\widehat{\theta}_0 = \overline{Y} = \frac{1}{n} \sum_{i=1}^{n} Y_i$. 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 $i = 1, \ldots, n$, 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 or when an observation interval is not available. 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 only starts developing. 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. Other works explore the context of missing data in the response while the response is missing at random in a nonparametric setting (see Ferraty et al. (2013), Ling et al. (2015)) or in a functional partial linear regression setting (see Ling et al. (2019), Zhou and Peng (2020)) or while the response is not missing at random (see Li et al. (2018)). In our work, we want to consider the functional linear model where some observations of the real response are affected with missing data and the covariate is partially observed, 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 $(\delta_i^{[Y]})_{i=1,\ldots,n}$ 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 $i = 1, \ldots, n$. 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 $Y_1, \ldots, Y_n$ is denoted

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

For the missing data mechanism of the functional covariate, we adopt the paradigm of partially observed functions as in Kneip and Liebl (2020) or Kraus (2015). We also refer the reader to Delaigle et al. (2020) or Kraus and Stefanucci (2020) for recent contributions on this topic. More precisely, for each curve X_i, $i = 1, \dots, n$, we consider the observed part $O_i \subseteq [a,b]$ of X_i and the missing part $M_i = [a,b] \setminus O_i$. The observed part O_i refers to an interval (or several intervals) where the curve X_i is observed at some measure points of O_i. Based on the punctual observations, the whole curve can be reconstructed on O_i with usual methods (e.g. smoothing splines, regression splines, local polynomial smoothing, ...). On the contrary, no information is available on the missing part M_i. An example of such partially observed functions is given in section 5 of the paper.

The objective of this paper is to predict a new value of the response Y given a new test observation on the explanatory variable X once the partially observed curves X have been reconstructed and the missing data Y have been imputed. 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 the covariate is partially observed. Then, in section 3, we extend these results when the covariate is partially observed and some observations of the real response are affected with missing data. 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 in supplementary material.

2 Partially observed covariate

2.1 Curve reconstruction

We write "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. As noticed in Kneip and Liebl (2020, p. 7) all the following remains valid if we consider the more general case of several observed subintervals, that is $O_i = \cup_{j=1}^J O_i^J$ where $O_i^1, \ldots, O_i^J$ are J disjoint intervals where the curve X_i is observed. For the sake of simplicity, we will take $J = 1$ and $O_i = O_i^1$. We write the Karhunen-Loève (KL) decomposition of X_i^O in $\mathbb{L}^2(O)$

$$X_i^O(t) = \sum_{k=1}^{+\infty} \xi_{ik}^O \phi_k^O(t), \tag{2.1}$$

where $t \in O$. In this decomposition, the principal component scores are defined for all $i = 1, \ldots, n$ 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 $k \neq \ell$. Moreover, the eigenfunctions satisfy

$$\phi_k^O(t) = \frac{\langle \phi_k^O, \gamma_t^O \rangle}{\lambda_k^O}, \tag{2.2}$$

for all $t \in O$ and $k \geq 1$, where $\gamma_t^O(s) = \gamma^O(t,s) = \mathbb{E}\left(X_i^O(t) X_i^O(s)\right)$, and the decreasing eigenvalues $\lambda_1^O > \lambda_2^O > \ldots > 0$ are tending to zero.

We consider a reconstruction problem relating the missing part of the curves to the observed part, writing

$$X_i^M(s) = L(X_i^O(t)) + Z_i(s), \tag{2.3}$$

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

$$\mathcal{L}(X_i^O)(s) = \sum_{k=1}^{+\infty} \xi_{ik}^O \widetilde{\phi}_k^O(s) = \sum_{k=1}^{+\infty} \xi_{ik}^O \frac{\langle \phi_k^O, \gamma_s \rangle}{\lambda_k^O}, \tag{2.4}$$

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 $s \in M$. The definition of $\widetilde{\phi}_k^O$ is a way to extend the relation (2.2) to the missing parts of the curves. 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

$$W_{ij} = X_i^O(t_{ij}), \tag{2.5}$$

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. The latter is similar to the procedure in Yao et al. (2005) or Hall et al. (2006). 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

$$\sum_{j=1}^{p} (W_{ij} - \beta_0 - \beta_1(t_{ij} - t))^2 \kappa_1 \left(\frac{t_{ij} - t}{h_X} \right), \tag{2.6}$$

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

$$\sum_{i=1}^{n} \sum_{j,\ell=1}^{p} (C_{ij\ell} - \tau_0 - \tau_1(t_{ij} - t) - \tau_2(t_{i\ell} - s))^2 \kappa_2 \left(\frac{t_{ij} - t}{h_\gamma}, \frac{t_{i\ell} - s}{h_\gamma} \right), \tag{2.7}$$

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 (2.4), considering a positive integer k_n, we define

$$\widehat{\mathcal{L}}_{k_n}(X_i^O)(s) = \sum_{k=1}^{k_n} \widehat{\xi}_{ik}^O \frac{\langle \widehat{\phi}_k^O, \widehat{\gamma}_s \rangle}{\widehat{\lambda}_k^O}, \tag{2.8}$$

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

$$\widehat{X}_i^M(s) = \widehat{\mathcal{L}}((X_i^O)(t)), \tag{2.9}$$

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

$$\widehat{\mathcal{L}}_{k_n}^{\star}(X_i^O)(s) = \widehat{X}_i^O(V_s; h_X) + \sum_{k=1}^{k_n} \widehat{\xi}_{ik}^O \left(\frac{\langle \widehat{\phi}_k^O, \widehat{\gamma}_s \rangle}{\widehat{\lambda}_k^O} - \frac{\langle \widehat{\phi}_k^O, \widehat{\gamma}_{V_s} \rangle}{\widehat{\lambda}_k^O} \right). \tag{2.10}$$

In the following, we denote

$$X_i^\star(t) = \begin{cases} X_i^O(t) \text{ if } t \in O, \\ \widehat{\mathcal{L}}_{k_n}(X_i^O)(t) \text{ if } t \in M. \end{cases} \tag{2.11}$$

2.3 Estimation of θ and prediction

For estimating θ, we set

$$\widehat{\widehat{\theta}} = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{k_n}\frac{\langle X_i^\star, \widehat{\phi}_{j,rec}\rangle Y_i}{\widehat{\lambda}_{j,rec}}\widehat{\phi}_{j,rec} = \sum_{j=1}^{k_n}\widehat{\widehat{s}}_j\widehat{\phi}_{j,rec}, \tag{2.12}$$

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

$$\widehat{\widehat{\Theta}} = \langle\widehat{\widehat{\theta}}, .\rangle = \widehat{\Pi}_{k_n,rec}\widehat{\Delta}_{n,rec}(\widehat{\Pi}_{k_n,rec}\widehat{\Gamma}_{n,rec}\widehat{\Pi}_{k_n,rec})^{-1}, \tag{2.13}$$

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 $\text{Span}(\widehat{\phi}_{1,rec}, \ldots, \widehat{\phi}_{k_n,rec})$, 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 $\widehat{\widehat{\theta}}_0 = \overline{Y}$. 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 follows

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

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

$$\widehat{Y}_{new} = \widehat{\widehat{\theta}}_0 + \langle\widehat{\widehat{\theta}}, X^\star_{new}\rangle. \tag{2.14}$$

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 $t \in [a, b]$

$$|X_i^\star(t) - X_i(t)| = \mathcal{O}_p\left(p^{-(a_O-1)/(2(a_O+2))}\right). \tag{2.15}$$

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

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

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 with the main result giving a bound for the prediction error of Y_{new} with a new value of the covariate X_{new}.

Theorem 2.2 *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*

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

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 1 Due to the bound (2.15), the result of Theorem 2.2 remains valid if we replace $X^{\star}_{new}$ with X_{new}.

Corollary 2.3 *Under the hypotheses of Theorem 2.2, 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 Partially observed covariate and missing data on the response

In this section, we are interested in the most general case of missing data in functional linear regression: when the covariate is partially observed and when the response is 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

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

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

$$\widetilde{\theta} = \frac{1}{n - m_n^{[Y]}} \sum_{i=1}^{n-m_n^{[Y]}} \sum_{j=1}^{k_n} \frac{\langle X_i^\star, \widehat{\phi}^{obs}_{j,rec} \rangle \delta_i^{[Y]} Y_i}{\widehat{\lambda}^{obs}_{j,rec}} \widehat{\phi}^{obs}_{j,rec} = \sum_{j=1}^{k_n} \widetilde{s}_j \widehat{\phi}^{obs}_{j,rec}, \tag{3.1}$$

with $\widetilde{s}_j = \frac{1}{(n-m_n^{[Y]})\widehat{\lambda}^{obs}_{j,rec}} \sum_{i=1}^{n-m_n^{[Y]}} \langle X_i^\star, \widehat{\phi}^{obs}_{j,rec} \rangle \delta_i^{[Y]} Y_i$. We also estimate the intercept θ_0 with $\widetilde{\theta}_0 = \overline{Y}_{obs} = \frac{1}{n-m_n^{[Y]}} \sum_{i=1}^{n} \delta_i^{[Y]} Y_i$. Now, the residuals of the fit, $\widetilde{\varepsilon}_{i,k_n} = Y_i - \widetilde{\theta}_0 - \langle X_i^\star, \widetilde{\theta} \rangle$ for $i = 1, \ldots, n$, can be used to estimate the error variance as follows

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

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

$$Y_{\ell,imp} = \widetilde{\theta}_0 + \langle \widetilde{\theta}, X_\ell^\star \rangle = \widetilde{\theta}_0 + \sum_{j=1}^{k_n} \widetilde{\widetilde{s}}_j \langle X_\ell^\star, \widehat{\phi}^{obs}_{j,rec} \rangle, \tag{3.2}$$

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

$$Y_{\ell,imp} = \widehat{\Pi}^{obs}_{k_n,rec} \widehat{\Delta}^{obs}_{n,rec} \left(\widehat{\Pi}^{obs}_{k_n,rec} \widehat{\Gamma}^{obs}_{k_n,rec} \widehat{\Pi}^{obs}_{k_n,rec} \right)^{-1} X_\ell^\star, \tag{3.3}$$

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

3.2 Estimation of θ and prediction

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

$$\widehat{\theta}^\star = \frac{1}{n} \sum_{i=1}^n \sum_{j=1}^{k_n} \frac{\langle X_i^\star, \widehat{\phi}^\star_{j,rec}\rangle Y_i^\star}{\widehat{\lambda}^\star_{j,rec}} \widehat{\phi}^\star_{j,rec} = \sum_{j=1}^{k_n} \widehat{s}_j^\star \widehat{\phi}^\star_{j,rec}, \tag{3.4}$$

where $\widehat{s}_j^\star = \frac{1}{n\widehat{\lambda}^\star_{j,rec}} \sum_{i=1}^n \langle X_i^\star, \widehat{\phi}^\star_{j,rec}\rangle Y_i^\star$ 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

$$\widehat{\Theta}^\star = \langle \widehat{\theta}^\star, .\rangle = \widehat{\Pi}^\star_{k_n,rec} \widehat{\Delta}^\star_{n,rec} \left(\widehat{\Pi}^\star_{k_n,rec} \widehat{\Gamma}^\star_{n,rec} \widehat{\Pi}^\star_{k_n,rec} \right)^{-1}, \tag{3.5}$$

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

$$\begin{aligned} \widehat{Y}_{new} = \widehat{\theta}_0^\star + \langle \widehat{\theta}^\star, X^\star_{new}\rangle &= \widehat{\theta}_0^\star + \frac{1}{n} \sum_{i=1}^n \sum_{j=1}^{k_n} \frac{\langle X_i^\star, \widehat{\phi}^\star_{j,rec}\rangle \langle X^\star_{new}, \widehat{\phi}^\star_{j,rec}\rangle Y_i^\star}{\widehat{\lambda}^\star_{j,rec}} \\ &= \widehat{\theta}_0^\star + \sum_{j=1}^{k_n} \widehat{s}_j^\star \langle X^\star_{new}, \widehat{\phi}^\star_{j,rec}\rangle, \end{aligned} \tag{3.6}$$

where $\widehat{\theta}_0^\star = \overline{Y}^\star = \frac{1}{n}\sum_{i=1}^n Y_i^\star$. Then, the residuals of the fit, $\widehat{\varepsilon}^\star_{i,k_n} = Y_i^\star - \widehat{\theta}_0^\star - \langle X_i^\star, \widehat{\theta}^\star\rangle$ for $i = 1, \ldots, n$, allow to estimate the error variance writing

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

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 response 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 2.2 and 3.1 and follows exactly the same lines as the proof of Theorem 2.2.

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 response 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*

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

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*

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

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 2 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 Model 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

$$Y = \theta_0 + \langle \theta, X \rangle + \varepsilon, \tag{4.1}$$

where the error ε is a Gaussian noise: $\varepsilon \sim N(0, \sigma_\varepsilon^2)$ with $\sigma_\varepsilon = 0.2$. We derived different models from (4.1), simulating more or less smooth processes X. For the sake of concision, we only give the results for the model presented below. Results for the other models are available on demand to the authors.

In this model, 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 $j > 1$.

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

We used the following criteria, 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 $MSE(j) = \frac{1}{n_1} \sum_{\ell=n+1}^{n+n_1} \left(Y_\ell^j - \widehat{\theta}_0 - \langle \widehat{\theta}, X_\ell^j \rangle \right)^2$ is the mean square error computed on the j^{th} simulated sample, $j \in \{1, \ldots, \boldsymbol{S}\}$.

- Criterion 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 $RT(j) = \frac{\sum_{\ell=n+1}^{n+n_1} \left(Y_\ell^j - \widehat{\theta}_0 - \langle \widehat{\theta}, X_\ell^j \rangle \right)^2}{\sum_{\ell=n+1}^{n+n_1} \left(\epsilon_\ell^j \right)^2}$ 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. Criterion RT is a rescaled version of 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 (1.4) 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 when using the SPCR estimation. We take $\kappa = 20$, $q = 3$ 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 $\delta^{[Y]}$ 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 deal with partially observed curves for the covariate, we adopted the missing data simulation scenario from Kneip and Liebl (2020) such that

- 85% (respectively 70%) of the curves are fully observed on $[0,1]$,
- for the 15% (respectively 30%) of partially observed curves, 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,
  - A = 3/8 and B = 6/8.

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 criteria were 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 partially observed and Y is 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 3: X is partially observed and Y is affected with missing values, the individuals presenting either a partially observed curve or a missing response are removed from the sample.

Other intermediate cases have been examined (when X is fully observed and Y is affected by missing values, or when X is partially observed and Y is not affected by missing values). Complete results are available on demand to the authors.

Table 2: Mean and standard deviation errors for the predicted values based on 400 simulation replications with different levels of missing data and a sample size $N = 90$. Partially observed curves are fully observed on $[1/8, 7/8]$.

Rate of missing data in X in %	15.10 (3.63)	14.96 (3.72)	15.21 (4.02)	29.64 (4.93)	29.96 (4.64)	29.84 (4.63)
Rate of missing data in Y in %	12.06 (3.99)	27.61 (5.24)	44.96 (5.59)	11.64 (3.97)	27.25 (5.38)	44.85 (6.04)
(Case 1) $\overline{MSE} \times 10^2$	26.42 (29.38)	27.13 (29.56)	27.91 (30.68)	25.89 (27.50)	26.57 (27.80)	25.10 (29.51)
(Case 1) $\overline{RT}$	8.99 (10.37)	9.06 (10.01)	8.93 (10.11)	7.83 (7.41)	8.29 (8.23)	8.20 (9.31)
(Case 2) $\overline{MSE} \times 10^2$	33.63 (37.40)	45.17 (50.88)	65.83 (72.27)	32.14 (34.24)	44.78 (46.99)	57.00 (60.82)
(Case 2) $\overline{RT}$	11.10 (12.77)	14.50 (17.61)	19.72 (23.65)	9.46 (9.13)	13.34 (14.04)	17.18 (19.00)
(Case 3) $\overline{MSE} \times 10^2$	36.26 (41.51)	45.16 (52.42)	73.32 (85.45)	42.78 (50.63)	60.34 (71.71)	102.60 (410.11)
(Case 3) $\overline{RT}$	12.02 (14.07)	14.68 (18.27)	23.08 (31.69)	13.25 (15.53)	23.77 (104.03)	30.13 (104.58)

As it can be expected, the errors decrease as the sample size increases. The main

Table 3: Mean and standard deviation errors for the predicted values based on 400 simulation replications with different levels of missing data and a sample size $N = 1440$. Partially observed curves are fully observed on $[1/8, 7/8]$.

Rate of missing	15.05	14.92	14.90	29.84	30.05	30.05
data in X in %	(0.92)	(0.92)	(0.94)	(1.19)	(1.20)	(1.21)
Rate of missing	12.02	26.92	45.09	11.98	26.94	44.89
data in Y in %	(0.91)	(1.31)	(1.44)	(0.94)	(1.34)	(1.33)
(Case 1) $\overline{MSE} \times 10^2$	1.74	1.88	1.89	1.74	1.89	1.89
	(1.57)	(1.72)	(1.62)	(1.56)	(1.77)	(1.77)
(Case 1) $\overline{RT}$	1.43	1.46	1.48	1.44	1.47	1.47
	(0.40)	(0.42)	(0.42)	(0.39)	(0.45)	(0.45)
(Case 2) $\overline{MSE} \times 10^2$	2.16	3.06	4.85	2.18	3.10	4.94
	(1.99)	(2.85)	(4.56)	(2.00)	(2.95)	(4.39)
(Case 2) $\overline{RT}$	1.54	1.76	2.22	1.54	1.78	2.24
	(0.51)	(0.70)	(1.16)	(0.50)	(0.75)	(1.10)
(Case 3) $\overline{MSE} \times 10^2$	2.70	4.26	7.82	2.94	4.69	8.65
	(2.51)	(3.75)	(6.35)	(2.86)	(4.20)	(7.33)
(Case 3) $\overline{RT}$	1.68	2.08	2.96	1.73	2.19	3.19
	(0.63)	(0.94)	(1.62)	(0.72)	(1.08)	(1.85)

point we want to discuss is related to the level of missing data in the sample, in particular with respect to cases 2 and 3. The most favorable situation for our method (case 2) appears when there is a quite small sample size, and when the missing part of the curves is not so much important (see Table 2). In this situation, our method

Table 4: Mean and standard deviation errors for the predicted values based on 400 simulation replications with different levels of missing data and a sample size $N = 90$. Partially observed curves are fully observed on $[3/8, 6/8]$.

Rate of missing data in X in %	14.49 (3.91)	15.15 (3.77)	15.16 (3.80)	29.76 (4.90)	30.12 (4.91)	29.86 (4.71)
Rate of missing data in Y in %	12.05 (3.88)	26.75 (5.00)	45.10 (5.45)	12.24 (3.66)	26.77 (5.32)	44.92 (5.67)
(Case 1) $\overline{MSE} \times 10^2$	27.51 (30.74)	24.74 (26.44)	26.67 (27.89)	24.57 (25.51)	28.76 (30.39)	25.10 (26.77)
(Case 1) $\overline{RT}$	8.96 (10.10)	7.87 (9.65)	8.35 (8.72)	7.72 (7.29)	9.02 (9.31)	8.39 (8.54)
(Case 2) $\overline{MSE} \times 10^2$	34.89 (38.63)	40.96 (44.75)	67.29 (77.38)	32.59 (34.47)	48.76 (51.67)	60.81 (66.21)
(Case 2) $\overline{RT}$	11.12 (12.53)	12.33 (16.36)	19.78 (25.39)	9.88 (9.50)	14.80 (16.17)	18.90 (21.43)
(Case 3) $\overline{MSE} \times 10^2$	36.48 (41.58)	41.88 (46.59)	66.24 (74.63)	39.89 (44.10)	62.83 (73.75)	514.92 (469.21)
(Case 3) $\overline{RT}$	11.85 (14.03)	12.76 (15.52)	20.12 (26.41)	12.95 (14.59)	20.32 (27.62)	127.83 (1072.52)

(case 2) behaves better than the naive method (case 3). It shows a real advantage in reconstructing the unobserved 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 2 and 3 narrows (generally still in favor of case 2, though) when the sample size increases (see Table 3). 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 Tables 4), the curve reconstruction from Kneip and Liebl (2020) is more difficult, and the difference between cases 2 and 3 also narrows in this situation.

5 Real dataset study

In this section, we are interested in a model involving electricity production, demand and prices of the German power market. Kneip and Liebl (2020) were already interested in the curve reconstruction problem of electricity prices curves (function of the demand). These data are provided from three different publicly available sources: The European Power Exange (www.epexspot.com), the European Network of Transmission System Operators for Electricity (www.entsoe.eu) and the European Energy Exchange (www.eex-transparency.com). The observation period corresponds to $n = 241$ working days from March 15, 2012 to March 14, 2013. The dataset consists in $n = 241$ daily electricity prices curves in Germany (measured every hour) in function of the residual electricity demand, which is the relevant value for considering electricity demand. It corresponds to germany's gross electricity demand minus infeeds from renewable energy sources plus net-imports from foreign countries. Some prices greater than 120 EUR/MWh have to be treated as outliers since they cannot be explained by the model and were set to the value 120. Negative prices are not impossible in this situation: electricity producers prefer to sell electricity at negative prices (meaning that they are paying for delivering electricity), it is sometimes more profitable than shutting off and restarting a central plant. Figure 1 shows the prices curves (in EUR/MWh) in function of the residual demand (in MWh), and Figure 2 shows the reconstructed curves with the method from Kneip and Liebl (2020). Price curves can be seen as partially observed curves, as some prices cannot be observed with respect to some residual demand values.

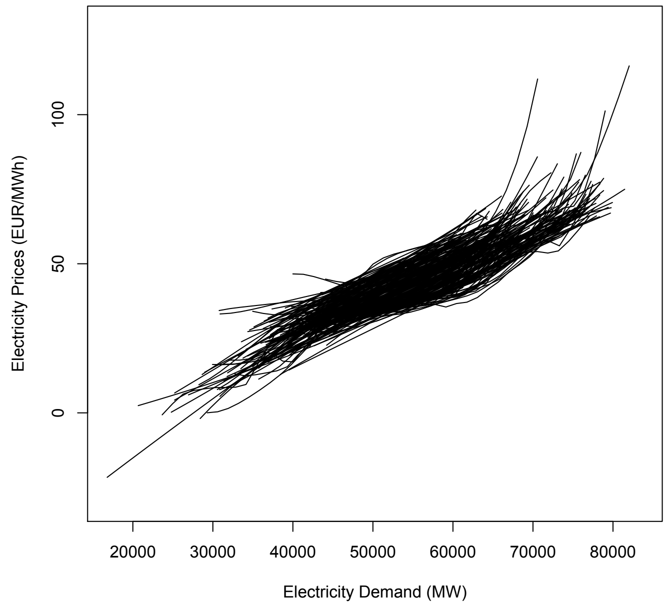

Figure 1: Daily electricity price curves in function of the residual demand.

Here, the price-demand functions are observed on different domains. This distinguishes our functional data set from classical functional data sets, where all functions are observed on a common domain. We consider a standardized domain where the standardization can be achieved as follows: for $i = 1, \ldots, n$, we consider a sequence from $\min_{1\leq j\leq p} t_{ij}$ to $\max_{1\leq j\leq p} t_{ij}$ with a regular step $(b-a)/p$, where $a := \min_{1\leq i\leq n} \min_{1\leq j\leq p} t_{ij}$ and $b := \max_{1\leq i\leq n} \max_{1\leq j\leq p} t_{ij}$.

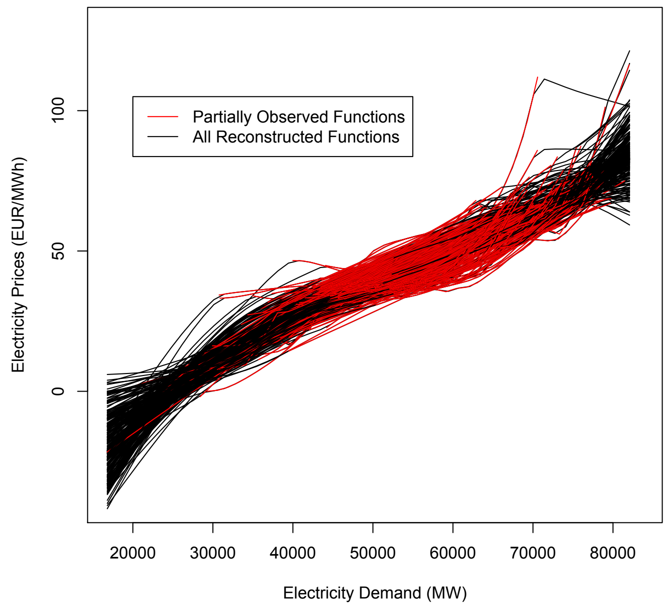

Figure 2: Reconstructed daily electricity price curves in function of the residual demand.

Our experimental study is based on two steps. In the first treatment step, we do not observe the price-demand functions directly but we have to estimate each price-demand function by a local polynomial smoother estimator. Here, we choose the Gaussian kernel and we consider a cross validation criterion to select the optimal tuning bandwidth parameter from a grid of parameter values in the interval [1070,35000]. In the second step, we reconstructed the missing parts of the differents curves.

We introduce now the model

$$Y_i = \theta_0 + \langle \theta, X_i \rangle + \varepsilon_i, \tag{5.1}$$

for $i = 1, \ldots, 241$, where X_i is the daily electricity price curve on day i (function of the residual demand), and Y_i is the value of electricity production (in MWh) on day i. The production data come from https://www.agora-energiewende.de[1]. Only a graphic (with numerical values marked at the observation points) was available on this website to collect a data (neither a table nor an Excel file). It can be possible to use a software to get numerical values from a graphic (see https://automeris.io[2]). However, this software is not completely reliable and some numerical values, being not possible, can be considered as missing data for the response variable. In our case, the percentage of missing data is 13.26%.

We split the initial sample into a learning sample (the index set is denoted I_L) with size 181 and a test sample with size 60 (the index set is denoted I_T). Firstly, we reconstructed the missing parts of the differents curves (see Figure 2) and, on the learning sample, we imputed the missing values on the response. Then, on the test sample, we computed the prediction values for the response. In order to evaluate the quality of the prediction, we calculated the mean square error $MSE = \frac{1}{60}\sum_{i\in I_T}(Y_i - \widehat{Y}_i)^2 = 40.44$ and the mean absolute error $MAE = \frac{1}{60}\sum_{i\in I_T}|Y_i - \widehat{Y}_i| = 5.35$. In this situation, we can see that the naive method (corresponding to case 3 in the simulation) would not be possible since all the curves are partially observed and this would cause removing all individuals in the sample.

References

Cai, T. and Hall, P. (2006). Prediction in functional linear regression. *The Annals of*

[1]https://www.agora-energiewende.de/en/service/recent-electricity-data/chart/power_generation/15.03.2012/14.03.2013/

[2]https://automeris.io/WebPlotDigitizer/

Statistics, **34**, 2159–2179.

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

Cardot, H., Ferraty, F., and Sarda, P. (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., 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**.

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

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

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.

Hall, P., Müller, H.-G., and Wang, J.-L. (2006). Properties of principal component methods for functional and longitudinal data analysis. *The Annals of Statistics*, **34**, 1493–1517.

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.

Li, T., Xie, F., Feng, X., Ibrahim, J., Zhu, H., and the Alzheimers Disease Neuroimaging Initiative (2018). Functional linear regression models for nonignorable missing scalar responses. *Statistica Sinica*, **28**, 1867–1886.

Ling, N., Liang, L., and Vieu, P. (2015). Nonparametric regression estimation for functional stationary ergodic data with missing at random. *Journal of Statistical Planning and Inference*, **162**, 75–87.

Ling, N., Kan, R., Vieu, P., and Meng, S. (2019). Semi-functional partially linear regression model with responses missing at random. *Metrika*, **82**, 39–70.

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.

Yao, F., Müller, H.-G., and Wang, J.-L. (2005). Functional data analysis for sparse longitudinal data. *Journal of the American Statistical Association*, **100**, 577–590.

Zhou, J. and Peng, Q. (2020). Estimation for functional partial linear models with missing responses. *Statistics and Probability Letters*, **156**.

Supplementary material: Functional linear model with partially observed covariate and missing values in the response.

Christophe Crambes, Chayma Daayeb, Ali Gannoun, and Yousri Henchiri.

Proofs

Proof of Proposition 2.1

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

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

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 (2.15) 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 $\|\langle X_i, x\rangle\left(X_i^\star - X_i\right)\| = \mathcal{O}_p\big(p^{-(a_O-1)/(2(a_O+2))}\big)$, and $\|\langle X_i^\star - X_i, x\rangle\left(X_i^\star - X_i\right)\| = \mathcal{O}_p\big(p^{-(a_O-1)/(a_O+2)}\big)$, 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 2.2

We start with the decomposition

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

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

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

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

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

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

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

Now, we use (2.15) to obtain

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

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 (2.15), we can write

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

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 2.2. The decomposition from the beginning of the proof gives

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

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

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

The first term of the right-hand side is given by $\mathbb{E}\left(\overline{Y} - \mathbb{E}(Y)\right)^2 = \mathcal{O}_p(n^{-1})$ (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 2.2. We write the decomposition

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

The first term in the above decomposition $B_1 := 32\mathbb{E}\left(\Theta\widehat{\Pi}_{k_n,rec}^{obs}X_\ell^\star - \Theta\widehat{\Pi}_{k_n}X_\ell^\star\right)^2$. 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 (2.15) to obtain

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

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 (2.15), 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 2.2 to write

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

The first term of the right-hand side is given by $\mathbb{E}\left(\overline{Y}_{obs} - \mathbb{E}(Y)\right)^2 = \mathcal{O}_p\Big((n-m_n^{[Y]})^{-1}\Big)$ (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

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

Next, we observe that

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

The first term is given by the result of Theorem 2.2. For the second term

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

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 2.2. So, comparing the convergence rates, we get

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

which gives the desired result.

References

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.

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