Regression imputation in the functional linear model with missing values in the response

Christophe Crambes, Yousri Henchiri

Regression imputation in the functional linear model with missing values in the response.

Christophe CRAMBES[a], Yousri HENCHIRI[b,c,*]

[a] *Institut Montpelliérain Alexander Grothendieck (IMAG), Université de Montpellier, France.*
[b] *Université de la Manouba, Institut Supérieur des Arts Multimédia de la Manouba (ISAMM), Tunisie.*
[c] *Université de Tunis El Manar, Laboratoire de Modélisation Mathématique et Numérique dans les Sciences de l'Ingénieur (ENIT-LAMSIN), Tunisie.*

Abstract

We are interested in functional linear regression when some observations of the real response are missing, while the functional covariate is completely observed. A complete case regression imputation method of missing data is presented, using functional principal component regression to estimate the functional coefficient of the model. We study the asymptotic behaviour of the error when the missing data are replaced by the regression imputed value, in a 'missing at random' framework. The completed database can be used to estimate the functional coefficient of the model and to predict new values of the response. The practical behaviour of the method is also studied on simulated data sets. A real dataset illustration is performed in the environmental context of air quality.

Keywords: Functional linear model, Missing data, Missing at random, Principal components regression, Mean square error prediction.
2010 MSC: 62G20, 62J05, 62F12, 62P12.

*Principal corresponding author
Email addresses: crambes.christophe@univ-montp2.fr (Christophe CRAMBES), yousri.henchiri@univ-montp2.fr (Yousri HENCHIRI)

1. Introduction

Literature on functional data is really wide, as attested by the numerous books on this subject these last years. The estimation and forecasting theories of linear processes in function spaces are developed in [1]. A comprehensive introduction to functional data analysis can be found in [26]. In the focus of [13] are nonparametric approaches. Computational issues are explained in [27]. Nonparametric statistical methods for functional regression analysis, specifically the methods based on a Gaussian process prior in a functional space are discussed in [28]. In [18] inferential procedures based on functional principal components are considered. [32] mainly focuses on hypothesis testing problems about functional data. Among this, the functional linear model has received a special attention (see [25, 4, 5, 3, 16, 10, 6, 31] for main references).

In this paper, we are interested in the functional linear model

$$Y = \langle \theta, X \rangle + \varepsilon, \tag{1}$$

where θ is the unknown function of the model, Y is a real variable of interest, ε is a centered real random variable representing the error of the model, with finite variance $\mathbb{E}(\varepsilon^2) = \sigma_\varepsilon^2$, and X is a functional covariate belonging to some functional space H endowed with an inner product $\langle ., . \rangle$ and its associated norm $\|.\|$. Usually, H is the space $L^2([a, b])$ of square integrable functions defined on some real compact $[a, b]$ and the corresponding inner product is defined by $\langle f, g \rangle = \int_a^b f(t)g(t)\,\mathrm{d}t$ for functions $f, g \in L^2([a, b])$. Without loss of generality, we consider our work on $[0, 1]$. Moreover, we assume that X and ε are independent.

All the previously cited works are devoted to analyse complete data, however, this is not the case in many interesting applications including for example survival data analysis. For this reason, we focus in this work on the problem

of missing data (see [20, 15] for a wide introduction in the multivariate framework). This subject has been widely studied, in particular the way to impute missing data and the accuracy of this imputation according to the types of missing data: Missing Completely At Random (MCAR), Missing At Random (MAR) and Missing Not At Random (MNAR). Even if this problematic has received a lot of attention in a multivariate framework, it is not the case for the functional data framework. Our objective is to study the problem of combining regression imputation, missing data mechanisms and functional data analysis. As far as we know, few results are available for the moment. In MAR setting, [17] have explored this area by developing a functional multiple imputation approach modeling missing longitudinal response under a functional mixed effects model. They developed a Gibbs sampling algorithm to draw model parameters and imputations for missing values. Besides, [14] have considered two kinds of mean estimates of a scalar outcome, based on a sample in which an explanatory variable is observed for every subject while responses are missing (which is the closest to our context). A weak convergence result was proved. In MCAR setting, [24] have adapted a methodology based on the NIPALS (Nonlinear Iterative Partial Least Squares) algorithm, which provides an imputation method for missing data, which have affected the functional covariates. In MNAR setting, [2] adapts a specification test for functional data with the presence of missing observations. His method is able to extract the information available in the observed portion of the data while being agnostic about the nature of the missing observations. In MAR and MCAR setting, [9] have recently proposed a nonparametric approach to missing value imputation and outlier detection for functional data. To our knowledge, there is no existing theoretical result in the case of functional linear model under missing assumption operating on the response variable, this problem only being until now the subject of studies in the multivariate framework (see for instance [21], [22]).

We carefully distinguish the missing data problem from a simple prediction problem. Indeed, the missing data mechanism involves a random variable (which

indicates whether the response is missing or not) which plays a central role when obtaining our asymptotic results. This random variable and the variable X are dependent in the MAR case. This is also highlighted in [14]. In this paper, we first propose an imputation method, based on the completely observed cases, to replace missing values in the response of the functional linear model. We get mean square error rates for these imputed values. Secondly, once the database is completed, we are able to estimate the unknown function θ of the model with the whole sample. This estimator can then be used for predicting other values of the response on a test set.

Combining missing data and functional variables offers a very large field of applications. Among all possible applications, environment is a core issue interesting many people for the future of our planet, in particular in the study of pollution indexes. The dataset we study here deals with temperature curves in some French cities to predict a specific pollution atmospheric index. The atmospheric index is missing in some cities in the northwest of France, for which the corresponding temperature curves (the explanatory variable) are mild, and leads to consider MAR data. The main objective is to get a map of the atmospheric index on the whole French territory.

The rest of the paper is organized as follows. Section 2 introduces the problem of functional linear model under missing assumption operating on the response variable and formulates our main results of the imputation method and of the mean square error for prediction of a new observation using the complete dataset. A simulation study is performed in Section 3. An environmental data illustration is presented in Section 4. Some preliminary lemmas, which are used in the proofs of the main results, are collected in Section 5.

2. Imputation of a missing value of the response

2.1. Functional principal components regression

Let us consider a sample $(X_i, Y_i)_{i=1,\dots,n}$ independent and identically distributed with the same distribution as (X, Y). An estimation of θ based on

principal components analysis of the curves $X_1, \ldots, X_n$ has been studied in many papers, see for instance [4]. We recall below the construction of this estimator. Considering the covariance operator of X defined under the condition $\mathbb{E}\left(\|X\|^2\right) < +\infty$ (which is supposed to be satisfied in the following) by

$$\Gamma u = \mathbb{E}\Big(\langle X, u\rangle X\Big),$$

for all $u \in H$ and its empirical version

$$\widehat{\Gamma}_n u = \frac{1}{n}\sum_{i=1}^{n}\langle X_i, u\rangle X_i,$$

we call $(\lambda_j)_{j\geq 1}$ $\Big(\text{resp. } \left(\widehat{\lambda}_j\right)_{j\geq 1}\Big)$ the sequence of eigenvalues of Γ $\Big(\text{resp. } \widehat{\Gamma}_n\Big)$ and $(v_j)_{j\geq 1}$ $\Big(\text{resp. } (\widehat{v}_j)_{j\geq 1}\Big)$ the sequence of eigenfunctions of Γ $\Big(\text{resp. } \widehat{\Gamma}_n\Big)$. The identifiability of model (1) is ensured as long as we assume that $\lambda_1 > \lambda_2 > \ldots > 0$ (see [4]). Moreover, assuming that $\widehat{\lambda}_1 > \ldots > \widehat{\lambda}_{k_n} > 0$ for some integer k_n depending on n, the estimator of θ is defined by

$$\widehat{\theta} = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{k_n}\frac{\langle X_i, \widehat{v}_j\rangle Y_i}{\widehat{\lambda}_j}\,\widehat{v}_j. \tag{2}$$

A consistency result of this estimator is given in [4], while more recent results can be found in [3, 16]. In particular, [4] give technical conditions on the decreasing rate to zero of the eigenvalues λ_j's in order to ensure the consistency of the estimator.

2.2. Operatorial point of view

We notice in this subsection that the model (1) can be seen from an operatorial point of view. Indeed, we can write the model

$$Y = \Theta X + \varepsilon, \tag{3}$$

where $\Theta : H \longrightarrow \mathbb{R}$ is a linear continuous operator defined by $\Theta u = \langle \theta, u\rangle$ for any function $u \in H$. Let us consider $\widehat{\Delta}_n$ the cross covariance operator defined by

$\widehat{\Delta}_n u = \frac{1}{n}\sum_{i=1}^n \langle X_i, u\rangle Y_i$, for all $u \in H$. Then, it is easily seen that an estimator $\widehat{\Theta}$ of Θ, satisfying $\widehat{\Theta} = \langle\widehat{\theta}, .\rangle$, is given by

$$\widehat{\Theta} = \langle\widehat{\theta}, .\rangle = \widehat{\Pi}_{k_n}\widehat{\Delta}_n\left(\widehat{\Pi}_{k_n}\widehat{\Gamma}_n\right)^{-1}, \tag{4}$$

where $\widehat{\Pi}_{k_n}$ is the projection operator onto the subspace $Span(\widehat{v}_1, \ldots, \widehat{v}_{k_n})$.

2.3. Imputation principle

Now, we present the context of missing data. There can be many reasons for which missing data can appear: breakdown in a measurement process, a person who is not willing to answer to some question of a questionnaire, ... We consider that some of the observations $Y_1, \ldots, Y_n$ are not available. We define the real variable δ and we consider the sample $(\delta_i)_{i=1,\ldots,n}$ such that $\delta_i = 1$ if the value Y_i is available and $\delta_i = 0$ if the value Y_i is missing, for all $i = 1, \ldots, n$. The data we observe are

$$\{(Y_i, \delta_i, X_i)\}_{i=1}^n.$$

We consider that the missing values are MAR. The MAR assumption implies that δ and Y are conditionally independent given X. That is,

$$\mathrm{P}\left(\delta = 1 \mid X, Y\right) = \mathrm{P}\left(\delta = 1 \mid X\right). \tag{5}$$

Note that the MAR assumption is much weaker than MCAR (for which $\mathrm{P}\left(\delta = 1 \mid X, Y\right) = \mathrm{P}\left(\delta = 1\right)$), as it allows the missing data to possibly depend on the observed data and may be reasonable for many practical problems. As a consequence of this MAR assumption, the variable δ (the fact that an observation is missing) is independent of the error of the model ϵ, conditionally on X. In the following, the number of missing values in the sample is denoted

$$m_n = \sum_{i=1}^n \mathbb{1}_{\{\delta_i = 0\}}. \tag{6}$$

Then, to impute a missing value, say Y_ℓ (where ℓ is a given integer between 1 and n), a simple way is to consider complete case analysis (see for instance

[20, 7, 30, 23, 29]). This regression imputation method uses the pairs of observed data to define the estimator of the model coefficient. More precisely, we define

$$Y_{\ell,imp} = \frac{1}{n-m_n} \sum_{\substack{i=1 \\ i\neq\ell}}^{n} \sum_{j=1}^{k_n} \frac{\langle X_i, \widehat{v}_j\rangle \langle X_\ell, \widehat{v}_j\rangle \delta_i Y_i}{\widehat{\lambda}_j}. \tag{7}$$

From the operatorial point of view, the imputation of the missing value Y_ℓ comes back to

$$Y_{\ell,imp} = \widehat{\Pi}_{k_n,obs} \widehat{\Delta}_{n,obs} \left(\widehat{\Pi}_{k_n,obs} \widehat{\Gamma}_{n,obs} \right)^{-1} X_\ell, \tag{8}$$

where $\widehat{\Gamma}_{n,obs} = \frac{1}{n-m_n} \sum_{i=1}^{n} \langle X_i, . \rangle \delta_i X_i$, $\widehat{\Delta}_{n,obs} = \frac{1}{n-m_n} \sum_{i=1}^{n} \langle X_i, . \rangle \delta_i Y_i$ and $\widehat{\Pi}_{k_n,obs}$ is the projection operator onto the subspace $\text{span}(\widehat{v}_{1,obs}, \ldots, \widehat{v}_{k_n,obs})$ where $\widehat{v}_{1,obs}$, $\ldots, \widehat{v}_{k_n,obs}$ are the k_n first eigenfunctions of the covariance operator $\widehat{\Gamma}_{n,obs}$.

Now we give our main results. We consider the following assumptions.

$(A.1)$ We assume that there exists a convex function λ such that $\lambda(j) = \lambda_j$ for all $j \geq 1$ that continuously interpolates the λ_j's between j and $j+1$.

$(A.2)$ There exists a positive constant C such that

$$\mathbb{E}\left(\|X\|^4 \right) \leq C.$$

Our assumptions are quite classic in this context. Assumption $(A.1)$ is similar to an assumption from [11]. It is a mild condition that allows a large class of decreasing rate of eigenvalues for the covariance operator Γ, for example polynomial decay or exponential decay (see example 1 below, in page 7, for more details). Assumption $(A.2)$ holds for many processes X (Gaussian processes, bounded processes) and can also be found for example in [4]. Then, we give our main results.

Remark 1. *Notice that the assumptions* $(A.1)$ *and* $(A.2)$ *are just needed to obtain a convergence rate, whether there are missing data on the response or not. The only assumption needed on missing data is actually the MAR model.*

Theorem 2.1. *Assume* $(A.1)$ *and* $(A.2)$ *are satisfied, if, moreover* $\lambda_{k_n} k_n$ *goes to zero as* n *goes to infinity, we have the mean square error*

$$\mathbb{E}\Big(Y_{\ell,imp} - \langle \theta, X_\ell \rangle\Big)^2 = \sum_{j=k_n+1}^{+\infty} \Big(\Theta \Gamma^{1/2} v_j\Big)^2 + \frac{\sigma_\varepsilon^2 k_n}{n - m_n} + o\left(\frac{k_n}{n - m_n}\right).$$

Moreover, for the aggregate mean square error of all the imputed values, we have

$$\sum_{\ell=1}^{n} (1-\delta_\ell) \mathbb{E}\Big(Y_{\ell,imp} - \langle \theta, X_\ell \rangle\Big)^2 = m_n \sum_{j=k_n+1}^{+\infty} \Big(\Theta \Gamma^{1/2} v_j\Big)^2 + \frac{\sigma_\varepsilon^2 k_n m_n}{n - m_n} + o\left(\frac{k_n m_n}{n - m_n}\right).$$

In order to precise the convergence rate of the imputed value $Y_{\ell,imp}$ to the real one $\langle \theta, X_\ell \rangle$, we need an additional notation. For a function $\varphi : \mathbb{R}_+^\star \longrightarrow \mathbb{R}_+^\star$ and a positive real number L, we define

$$\mathcal{C}(\varphi, L) = \Big\{ T : H \longrightarrow \mathbb{R} \ / \ \forall j \geq 1, T v_j \leq L \sqrt{\varphi(j)} \Big\}.$$

Note that simple cases satisfy the fact that $\Theta \Gamma^{1/2}$ belongs to $\mathcal{C}(\varphi, L)$. For example, consider the operator Θ expressed in the eigenfunctions basis $(v_j)_{j \geq 1}$ such that $\Theta u = \sum_{j=1}^{+\infty} \theta_j \langle v_j, u \rangle$ for any $u \in H$, with θ_j going to zero as j goes to infinity. Hence there exists a bound L such that $\theta_j \leq L$ for any $j \geq 1$ and $\Theta \Gamma^{1/2} v_j = \theta_j \sqrt{\lambda_j} \leq L \sqrt{\lambda_j}$.

Remark 2. *We introduce two notations to compare the magnitudes of two functions* $\tilde{u}(x)$ *and* $\tilde{v}(x)$ *as the argument* x *tends to a limit* $\tilde{\ell}$ *(not necessarily finite). The notation* $\tilde{u}(x) \underset{x \to \tilde{\ell}}{\sim} \tilde{v}(x)$*, stands for*

$$\lim_{x \to \tilde{\ell}} \frac{\tilde{u}(x)}{\tilde{v}(x)} = 1,$$

and the notation $\tilde{u}(x) \underset{x \to \tilde{\ell}}{\lesssim} \tilde{v}(x)$ *denotes that* $|\tilde{u}(x)/\tilde{v}(x)|$ *remains bounded as* $x \to \tilde{\ell}$.

Theorem 2.2. *Let* $L = \big\|\Theta \Gamma^{1/2}\big\|_\infty$ *and* φ *the function defined by* $\varphi(j) = \frac{\left(\Theta \Gamma^{1/2} v_j\right)^2}{L^2}$ *for all* $j \geq 1$ *that continuously interpolates the* $\varphi(j)$*'s between* j *and*

$j+1$. *Under assumptions* (A.1)-(A.2), *the operator* $\Theta\Gamma^{1/2}$ *belongs to* $\mathcal{C}(\varphi, L)$ *and*

$$\mathbb{E}\left(Y_{\ell,imp} - \langle \theta, X_\ell \rangle\right)^2 \underset{n\to+\infty}{\sim} 2\sigma_\varepsilon^2 \frac{k_n^\star}{n - m_n},$$

where $k_n^\star$ *is the solution of the equation in* x

$$\int_x^{+\infty} \varphi(t)\,\mathrm{d}t = \frac{\sigma_\varepsilon^2}{L^2(n-m_n)} x. \tag{9}$$

Again, for the aggregate mean square error of all the imputed values, we have

$$\sum_{\ell=1}^{n} (1-\delta_\ell)\mathbb{E}\left(Y_{\ell,imp} - \langle \theta, X_\ell \rangle\right)^2 \underset{n\to+\infty}{\sim} 2\sigma_\varepsilon^2 \frac{k_n^\star m_n}{n - m_n}.$$

Remark 3. *Notice that the equation (9) has a unique solution (the left and right hand sides are descreasing and increasing in* x*, respectively). The practical resolution of equation (9) to get* $k_n^\star$ *seems quite complicated due to the computation of* L*. In order to solve this problem, we will use other ways to select the optimal number of principal components (see Section 3 below).*

The last result giving the convergence rate of the imputed value $Y_{\ell,imp}$ is similar to the convergence rate obtained in [11] (who considered the case of a completely observed functional response). The rate is simply affected by the number m_n of missing values. We precise the resulting rate of convergence in the following examples.

Example 1. *We consider two different functions* φ *such that* $\varphi_{pol}(j) = C_\alpha j^{-(2+\alpha)}$ *and* $\varphi_{exp}(j) = D_\alpha \exp(-\alpha j)$ *where* C_α *and* D_α *are positive constants and* $\alpha > 0$*. Then the solution of equation (9) is*

$$\begin{cases} k_{n,pol}^\star \underset{n\to+\infty}{\sim} \left(\frac{C_\alpha L^2}{(1+\alpha)\sigma_\varepsilon^2}\right)^{1/(2+\alpha)} n^{1/(2+\alpha)}, & \text{if } \varphi = \varphi_{pol}, \\ k_{n,exp}^\star \underset{n\to+\infty}{\lesssim} \frac{\log n}{\alpha}, & \text{if } \varphi = \varphi_{exp}. \end{cases}$$

For $\varphi = \varphi_{pol}$, the result of Theorem 2.2 becomes

$$\mathbb{E}\left(Y_{\ell,imp} - \langle \theta, X_\ell \rangle\right)^2 \underset{n\to+\infty}{\sim} 2\left(\sigma_\varepsilon^2\right)^{(1+\alpha)/(2+\alpha)} \left(\frac{C_\alpha L^2}{1+\alpha}\right)^{1/(2+\alpha)} \frac{n^{1/(2+\alpha)}}{n-m_n},$$

for a single imputation and

$$\sum_{\ell=1}^{n}(1-\delta_\ell)\mathbb{E}\left(Y_{\ell,imp} - \langle \theta, X_\ell \rangle\right)^2 \underset{n\to+\infty}{\sim} 2\left(\sigma_\varepsilon^2\right)^{(1+\alpha)/(2+\alpha)} \left(\frac{C_\alpha L^2}{1+\alpha}\right)^{1/(2+\alpha)} \frac{n^{1/(2+\alpha)} m_n}{n-m_n},$$

for the aggregate error of all the imputed values.

For $\varphi = \varphi_{exp}$, the result of Theorem 2.2 becomes

$$\mathbb{E}\left(Y_{\ell,imp} - \langle \theta, X_\ell \rangle\right)^2 \underset{n\to+\infty}{\lesssim} \frac{2\sigma_\varepsilon^2 \log n}{\alpha(n-m_n)},$$

for a single imputation and

$$\sum_{\ell=1}^{n}(1-\delta_\ell)\mathbb{E}\left(Y_{\ell,imp} - \langle \theta, X_\ell \rangle\right)^2 \underset{n\to+\infty}{\lesssim} \frac{2\sigma_\varepsilon^2 m_n \log n}{\alpha(n-m_n)},$$

for the aggregate error of all the imputed values.

Example 2. *To precise in more specific cases our convergence rates, we consider three different levels of missing data: (i) when the number of missing data m_n is negligeable compared to the sample size, that is $m_n = a_n n$ with a_n going to zero as n goes to infinity, (ii) when the number of missing values is proportional to the sample size, that is $m_n = \rho n$ with $0 < \rho < 1$, and (iii) when the number of observed values is negligeable compared to the sample size, that is $u_n := n - m_n = o(n)$. We can sum up all the rates of convergence for the single imputation mean square error (Table 1) and for the aggregate mean square error (Table 2).*

We can see that missing data do not affect the convergence rate for a single imputed value when there are not too many missing values ($m_n = o(n)$ or $m_n = \rho n$). The rate $1/n^{(1+\alpha)/(2+\alpha)}$ matches the usual optimal rates in this context. The rate $\log n/\alpha n$ is not exact but obviously sharp since parametric up to a logarithm. It is no more the case when the number of missing values is high ($m_n \sim n$), the convergence rate is affected. For the aggregate error of several

Table 1: Single imputation mean square error convergence rates, where $K_\alpha := 2\left(\sigma_\varepsilon^2\right)^{(1+\alpha)/(2+\alpha)}\left(\frac{C_\alpha L^2}{1+\alpha}\right)^{1/(2+\alpha)}$.

	$\varphi = \varphi_{\text{pol}}$	$\varphi = \varphi_{\exp}$
$m_n := a_n n = o(n)$	$\underset{n\to+\infty}{\sim} K_\alpha n^{-(1+\alpha)/(2+\alpha)}$	$\underset{n\to+\infty}{\lesssim} \frac{2\sigma_\varepsilon^2 \log n}{\alpha n}$
$m_n = \rho n$	$\underset{n\to+\infty}{\sim} K_\alpha (1-\rho)^{1/(2+\alpha)} n^{-(1+\alpha)/(2+\alpha)}$	$\underset{n\to+\infty}{\lesssim} \frac{2\sigma_\varepsilon^2 \log n}{\alpha(1-\rho)n}$
$u_n := n - m_n = o(n)$	$\underset{n\to+\infty}{\sim} K_\alpha u_n^{-(1+\alpha)/(2+\alpha)}$	$\underset{n\to+\infty}{\lesssim} \frac{2\sigma_\varepsilon^2 \log n}{\alpha u_n}$

Table 2: Aggregate imputation mean square error convergence rates, where $K_\alpha := 2\left(\sigma_\varepsilon^2\right)^{(1+\alpha)/(2+\alpha)}\left(\frac{C_\alpha L^2}{1+\alpha}\right)^{1/(2+\alpha)}$.

	$\varphi = \varphi_{\text{pol}}$	$\varphi = \varphi_{\exp}$
$m_n := a_n n = o(n)$	$\underset{n\to+\infty}{\sim} K_\alpha a_n n^{1/(2+\alpha)}$	$\underset{n\to+\infty}{\lesssim} \frac{2\sigma_\varepsilon^2 a_n \log n}{\alpha}$
$m_n = \rho n$	$\underset{n\to+\infty}{\sim} K_\alpha \rho(1-\rho)^{1/(2+\alpha)} n^{1/(2+\alpha)}$	$\underset{n\to+\infty}{\lesssim} \frac{2\sigma_\varepsilon^2 \rho \log n}{\alpha(1-\rho)}$
$u_n := n - m_n = o(n)$	$\underset{n\to+\infty}{\sim} K_\alpha n u_n^{-(1+\alpha)/(2+\alpha)}$	$\underset{n\to+\infty}{\lesssim} \frac{2\sigma_\varepsilon^2 n \log n}{\alpha u_n}$

imputed values, when there are not too many missing values ($m_n = o(n)$), the number of missing values plays a crucial role, since the convergence depends on the fact that $a_n n^{1/(2+\alpha)}$ or $a_n \log n$ go to zero as n goes to infinity. In other cases ($m_n = \rho n$ or $m_n \sim n$), missing data affect the convergence of the aggregate error term for several imputed values, since it cannot converge to zero.

2.4. Estimation of θ and prediction of future values

Once the database being reconstructed, we can use the full database to estimate the functional coefficient θ of the model (directly inspired from (2)) (see also [8]), namely

$$\widetilde{\theta} = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{k_n} \frac{\langle X_i, \widehat{v}_j\rangle Y_i^\star}{\widehat{\lambda}_j}\widehat{v}_j, \tag{10}$$

where $Y_i^\star = Y_i\delta_i + Y_{i,imp}(1-\delta_i)$ for all $i = 1,\ldots,n$. Then this estimator of θ can be used to predict new values of the response Y on a test sample. Indeed,

if X_{new} is a new curve, the corresponding predicted response value is

$$\widehat{Y}_{new} = \langle X_{new}, \widetilde{\theta} \rangle = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{k_n} \frac{\langle X_i, \widehat{v}_j \rangle \langle X_{new}, \widehat{v}_j \rangle Y_i^\star}{\widehat{\lambda}_j}. \tag{11}$$

We give below a result allowing to control the mean square prediction error of $\widehat{Y}_{new}$.

Theorem 2.3. *Under the assumptions of Theorem 2.1, if we additionally assume that $m_n = o(n)$ (that is $m_n = a_n n$ with a_n going to zero as n goes to infinity) and $a_n^2 n = o(1)$, then*

$$\mathbb{E}\Big(\widehat{Y}_{new} - \langle \theta, X_{new} \rangle\Big)^2 = \sum_{j=k_n+1}^{+\infty} \Big(\Theta \Gamma^{1/2} v_j\Big)^2 + O\left(\frac{k_n}{n}\right).$$

Remark 4. *This result shows that, under the condition that there are not too many missing values, the convergence rate of the mean square error prediction of a new value of the covariate remains the same compared to the non missing values case.*

3. Simulations

To observe the behavior of our estimator in practice, this section considers a simulation study.

3.1. Models

Two models are considered:

$$Model_1 \quad : \quad Y = \int_0^1 \sin(4\pi t) X_t \, \mathrm{d}t \; + \epsilon, \tag{12}$$

$$Model_2 \quad : \quad Y = \int_0^1 \big(\log(15t^2 + 10) + \cos(4\pi t)\big) \; X_t \, \mathrm{d}t \; + \epsilon, \tag{13}$$

where the error ϵ is a Gaussian noise : $\epsilon \sim \mathcal{N}(0, \sigma_\epsilon)$ and

- in equation (12), $X := \{X_t\}_{t \in [0,1]}$ is the standard Brownian motion.

- In equation (13), $X := \{X_t\}_{t\in[0,1]}$ is a Gaussian process where the covariance function is defined as

$$cov(X_t, X_{t'}) = \exp(-\frac{|t-t'|^2}{0.2}).$$

The simulation aims at considering processes X with different regularities (the standard Brownian motion being the case of the less smooth) in order to see if it has an impact on the results.

All the procedures described below were implemented by using the R software:

- ⋆ the trajectories of X_i, $1 \leq i \leq n$, in the two models are discretized in $p = 100$ equidistant points,
- ⋆ values of Y are computed using integration by rectangular interpolation,
- ⋆ the variability of noise is such that $\sigma_\epsilon = \tau * \mathrm{Var}\Big(\int_0^1 \theta(t)X(t)\mathrm{d}t\Big) \approx 0.2$. Note that some Monte Carlo experiments are achieved to determine the values of τ: $\tau \approx 21.726$ for the $model_1$ (low level of noise) and $\tau \approx 0.048$ for the $model_2$ (high level of noise),
- ⋆ the sample sizes are respectively $n = 100, 300$ and 1200 for the training sets $(X_1, Y_1), \ldots, (X_n, Y_n)$ and $n_1 = 50, 150$ and 600 for the test sets $(X_{n+1}, Y_{n+1}), \ldots, (X_{n+n_1}, Y_{n+n_1})$.

3.2. 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, and criteria 5 is related to the estimation step with the reconstructed database.

- 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 $MSE(j) = \frac{1}{m_n} \sum_{\ell=1}^{n} \left(Y^j_{\ell,imp} - \langle \theta, X^j_\ell \rangle\right)^2 (1 - \delta_\ell)$ 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 $RT(j) = \dfrac{\sum_{\ell=1}^{n} \left(Y^j_{\ell,imp} - \langle \theta, X^j_\ell \rangle\right)^2 (1 - \delta_\ell)}{\sum_{\ell=1}^{n} \left(\epsilon^j_\ell\right)^2 (1 - \delta_\ell)}$ 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 $MSE'(j) = \frac{1}{n_1} \sum_{\ell'=n+1}^{n+n_1} \left(Y^j_{\ell'} - \langle \theta, X^j_{\ell'} \rangle\right)^2$ 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 $RT'(j) = \dfrac{\sum_{\ell'=n+1}^{n+n_1} \left(Y^j_{\ell'} - \langle \theta, X^j_{\ell'} \rangle\right)^2}{\sum_{\ell'=n+1}^{n+n_1} \left(\epsilon^j_{\ell'}\right)^2}$ 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.

- Criterion 5: the mean square errors (MSE'') averaged over $\boldsymbol{S}$ samples

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

where $MSE''(j) = \left\| \widetilde{\theta}^j - \theta \right\|^2$ is the square error of estimation computed on the j^{th} simulated sample. The MSE'' criterion is decomposed into variance and square bias in our results.

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

3.3. Methodology

We use a smoothed version of the estimator (2) based on the Smooth Principal Components Regression (SPCR) [5]. 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 number m of derivatives. Let us remark that, with appropriate conditions, all the theoretical results obtained in section 2 will also apply to the SPCR estimation. For example, we assume that the estimator $\widetilde{\theta}$ has r' derivatives for some integer r' and $\widetilde{\theta}^{(r')}$ satisfies, for some $\nu \in]0, 1]$

$$\left| \widetilde{\theta}^{(r')}(t_1) - \widetilde{\theta}^{(r')}(t_2) \right| \leq C \left| t_1 - t_2 \right|^{\nu},$$

for all $t_1, t_2 \in [0, 1]$. If we denote $r = r' + \nu$ and if we assume that the degree q of the splines is such that $q \geq r$, then

$$\sup_{t \in [0,1]} \left| \widetilde{\theta}(t) - S_{\kappa,q}(\widetilde{\theta})(t) \right| = O\left(\kappa^{-r}\right),$$

where $S_{\kappa,q}(\widetilde{\theta})$ is the spline approximation of $\widetilde{\theta}$ (see [12]). In other words, any of the convergence results obtained in Section 2 can be transposed to the smoothed version of the estimators.

Here, we have fixed the number of knots to be 20, the degree has been chosen to be 3 and the number of derivatives was fixed to the moderate value of 2. The choice of these parameters is not the most important in our study, especially in comparison with the choice of the number of principal components.

In this subsection, we show firstly how to determine the number of missing data. Secondly, we present a procedure to choose the optimal tuning parameter (the best dimension k_n^* of the projection space for the SPCR).

3.3.1. Missing data simulation scenario

To determine the number of missing data in our simulations, we have adopted the following scenario. In the MAR case, we simulate δ according to the logistic functional regression. The variable δ follows the Bernoulli law with parameter $p(X)$ such that

$$\log\left(\frac{p(X)}{1-p(X)}\right) = \langle \alpha_0, X\rangle + ct,$$

where $\alpha_0(t) = \sin(2\pi t)$ for all $t \in [0,1]$ and ct is a constant allowing to take different levels of missing data. We take $ct = 2$ for around 12.5% of missing data, $ct = 1$ for around 27.4% of missing data and $ct = 0.2$ for around 44.9% of missing data. Notice that, in the MCAR case, we simulate δ with the Bernoulli law with parameter $p(X) := p = 0.9$ (10% of missing data), $p(X) := p = 0.75$ (25% of missing data) or $p(X) := p = 0.6$ (40% of missing data).

3.3.2. Criteria for optimal parameter selection

We focus on the procedure allowing to select the optimal tuning parameter. We consider a Generalized Cross Validation (GCV) criterion versus a Cross Validation (CV) criterion and K-fold Cross Validation (K-fold CV) criterion and we select the optimal tuning parameter k_n^* by minimizing these criteria.

The GCV procedure is known to be computationally fast. The CV, K-fold CV and GCV criteria are respectively given as follows for imputation

$$\begin{aligned}
\mathrm{CV}(k_n) &= \frac{1}{n-m_n}\sum_{i=1}^{n}(\hat{Y}_i^{[-i]} - \langle\theta, X_i\rangle)^2\delta_i, \\
\text{K-fold CV}(k_n) &= \frac{1}{\mathrm{K}}\sum_{k=1}^{\mathrm{K}}|B_k|^{-1}\sum_{i\in B_k}(\hat{Y}_i^{[-B_k]} - \langle\theta, X_i\rangle)^2\delta_i, \\
\mathrm{GCV}(k_n) &= \frac{(n-m_n)\sum_{i=1}^{n}(\hat{Y}_i - \langle\theta, X_i\rangle)^2\delta_i}{((n-m_n)-k_n)^2}.
\end{aligned}$$

The analogous criteria are given as follows for prediction

$$\begin{aligned}
\mathrm{CV}(k_n) &= \frac{1}{n}\sum_{i=1}^{n}(\hat{Y}_i^{*^{[-i]}} - \langle\theta, X_i\rangle)^2, \\
\text{K-fold CV}(k_n) &= \frac{1}{\mathrm{K}}\sum_{k=1}^{\mathrm{K}}|B_k|^{-1}\sum_{i\in B_k}(\hat{Y}_i^{*^{[-B_k]}} - \langle\theta, X_i\rangle)^2, \\
\mathrm{GCV}(k_n) &= \frac{n\sum_{i=1}^{n}(\hat{Y}_i^{*} - \langle\theta, X_i\rangle)^2}{(n-k_n)^2},
\end{aligned}$$

where $\hat{Y}_i^{[-i]}$ and $\hat{Y}_i^{[-B_k]}$ respectively mean that the value of Y_i is predicted using the whole sample except the i^{th} observation or except the set of observations indexed in B_k. In the same way $\hat{Y}_i^{*^{[-i]}}$ and $\hat{Y}_i^{*^{[-B_k]}}$ respectively mean that the value of Y_i is predicted using the whole sample except the i^{th} observation or except the set of observations indexed in B_k. The data set is randomly partitioned into K equally sized (as equal as possible) subsets $\cup_{k=1}^{\mathrm{K}}B_k$ such that $B_j \cap B_k = \emptyset$ $(j \neq k)$. In practice, often $\mathrm{K} = 5$ or $\mathrm{K} = 10$ are used. In our case, the K-fold CV splits are chosen in a special deterministic way. For imputation, we consider

$$\text{K-fold CV}(k_n) = \frac{1}{\mathrm{K}}\sum_{k=1}^{\mathrm{K}}((n-m_n)/K)^{-1}\sum_{i=(n(k-1))/K+1}^{nk/K}(\hat{Y}_i^{[-k]} - \langle\theta, X_i\rangle)^2\delta_i.$$

The analogous criterion is given as follows for prediction

$$\text{K-fold CV}(k_n) = \frac{1}{\mathrm{K}}\sum_{k=1}^{\mathrm{K}}(n/K)^{-1}\sum_{i=(n(k-1))/K+1}^{nk/K}(\hat{Y}_i^{*^{[-k]}} - \langle\theta, X_i\rangle)^2.$$

In order to illustrate the advantage of the GCV criterion, we compared the computational times to obtain the tuning parameter with the three criteria on a growing sequence of dimension $k_n = 2, \ldots, 22$. The characteristics of the computer used to perform these computations were McBook pro: Processor 2.66 GHz intel core 2 Duo, Memory 4 Gb 1067 MHz DDR3. The computational times are displayed in Table A.11 in the appendix. The GCV criterion shows a clear advantage with regard to computational time compared with the CV and K-fold criteria. In addition, we see that the three criteria behave in the same way and select the same optimal projection dimension (see Fig. 1 and 2) for both models (under $n = 1000$ and $p = 100$). Notice that the GCV criterion (faster to compute) has been used in different simulations.

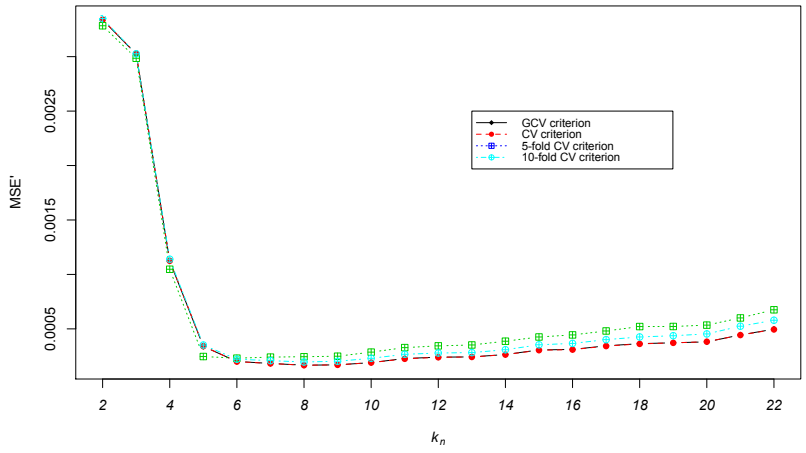

Figure 1: GCV, CV and K-fold criteria for different values of dimension k_n in $model_1$: best dimension $k_n^* = 8$ and MSE' $(\times 10^4) = 1.6640$ (in GCV criterion case), best dimension $k_n^* = 6$ and MSE' $(\times 10^4) = 2.3081$ (in 5-fold CV criterion case), best dimension $k_n^* = 8$ and MSE' $(\times 10^4) = 1.9584$ (in 10-fold CV criterion case), best dimension $k_n^* = 8$ and MSE' $(\times 10^4) = 1.6598$ (in CV criterion case), for the $model_1$.

We show on Fig. 3 and Fig. 4 different estimates of the slope function of the $Model_1$ and $Model_2$ (under $n = 1000$ and $p = 100$) with different values of

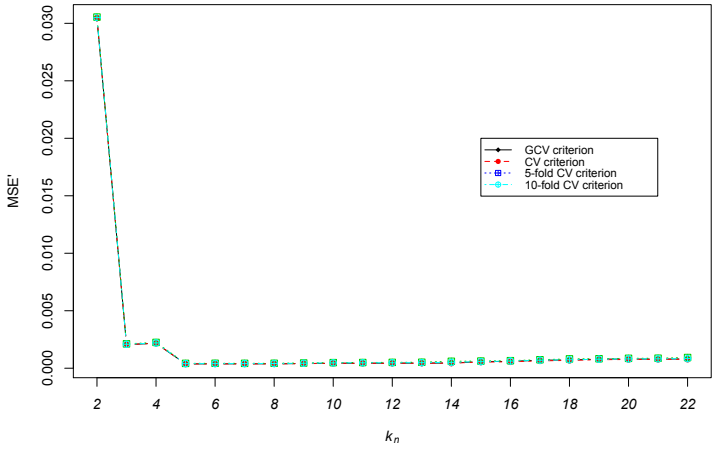

Figure 2: GCV, CV and K-fold criteria for different values of dimension k_n in $model_2$: best dimension $k_n^* = 5$ and MSE' $(\times 10^4) = 3.7589$ (in GCV criterion case), best dimension $k_n^* = 5$ and MSE' $(\times 10^4) = 4.2132$ (in 5-fold CV criterion case), best dimension $k_n^* = 5$ and MSE' $(\times 10^4) = 3.9758$ (in 10-fold CV criterion case), best dimension $k_n^* = 5$ and MSE' $(\times 10^4) = 3.7270$ (in CV criterion case), for the $model_2$.

dimension ($k_n = 4, 6, 8, 12, 16$) and ($k_n = 2, 3, 5, 7, 8$), respectively, by using the GCV criterion (used for its computational efficiency). We have chosen a percentage of missing values equal to 45.8518% for $model_1$ and equal to 46.8888% for $model_2$ (we obtain this rate with $ct = 1$ for both models).

325

3.4. Analysis of results

In this subsection, we analyse the results of the criteria presented in the previous subsection. Both MAR and MCAR context were considered. We only show the results for MAR and the results for MCAR are available on demand.
330 The different results given in Appendix A. Tables A.5, A.6 give the mean and standard deviation errors for the imputed values on training samples for both models. Tables A.7, A.8 give the mean and standard deviation errors for the predicted values on test samples for both models. Tables A.9, A.10 give the

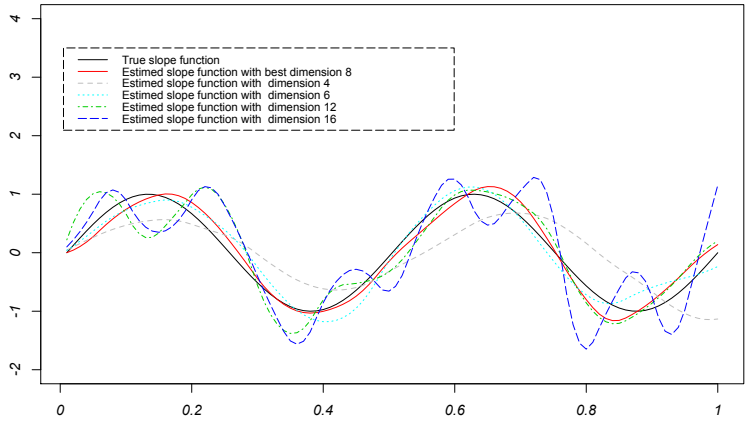

Figure 3: Plots of the true slope function (solid black) and estimates with different values of dimension k_n in $model_1$. The plots of estimates slope function with best dimension $k_n^* = 8$ (solid red), with dimension $k_n = 4$ (dotted), with dimension $k_n = 6$ (dashed), with dimension $k_n = 12$ (dotdashed), with dimension $k_n = 16$ (twodash).

mean and standard deviation errors for the estimation of θ using the fullfilled database with imputed values for both models. We can see that the errors increase when the rate of missing data increases. Similarly, the errors decrease as the size of the sample increases. When we compare the case of MAR and MCAR, we see that the error in case of MAR is slightly higher that in the MCAR case. Moreover, we can see that the regularity of the process X does not have a crucial impact on the results at least on these simulated examples. All the results in these simulations are in accordance with what we can expect and confirm the theoretical results obtained in the previous section.

4. Illustration

In order to illustrate the contribution of our approach in functional prediction setting when the covariates are functions and some observations of the

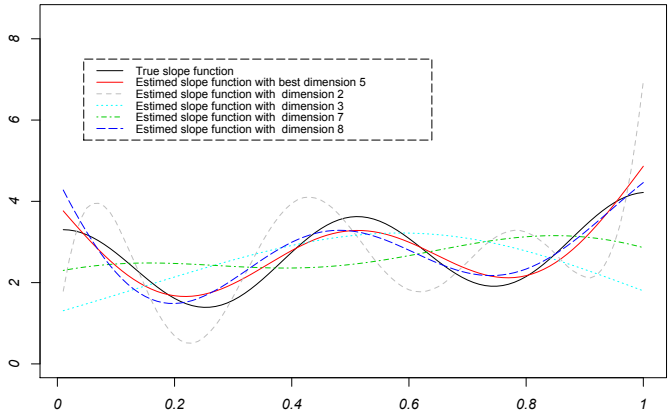

Figure 4: Plots of the true slope function (solid black) and estimates with different values of dimension k_n in $model_2$. The plots of estimates slope function with best dimension $k_n^* = 5$ (solid red), with dimension $k_n = 2$ (dotted), with dimension $k_n = 3$ (dashed), with dimension $k_n = 7$ (dotdashed), with dimension $k_n = 8$ (twodash).

real response are missing, we present in this section an environmental dataset application.

We start by describing the dataset. The functional covariate X is a daily temperature curve in some cities in France (from May 7, 2015 at 4 pm up to May 8, 2015 at 3 pm) obtained from www.meteociel.fr. This daily continuous curve is observed at some discretization points (here, at 24 discretization points, every hour). The graphical display of this daily temperature curves can be observed in Fig 5. The response variable Y is an atmospheric index of air quality called ATMO (for a detailed description of this atmospheric index, see www.atmo-france.org). Its values range from 1 (very good quality of air) to 10 (very bad quality of air). Though these values are discrete, we will consider that Y is a continuous approximation. We obtained the values of the atmospheric index on May 8, 2015, for these same cities, from www2.prevair.org. Furthermore, we added some cities for which the temperature curve is available but

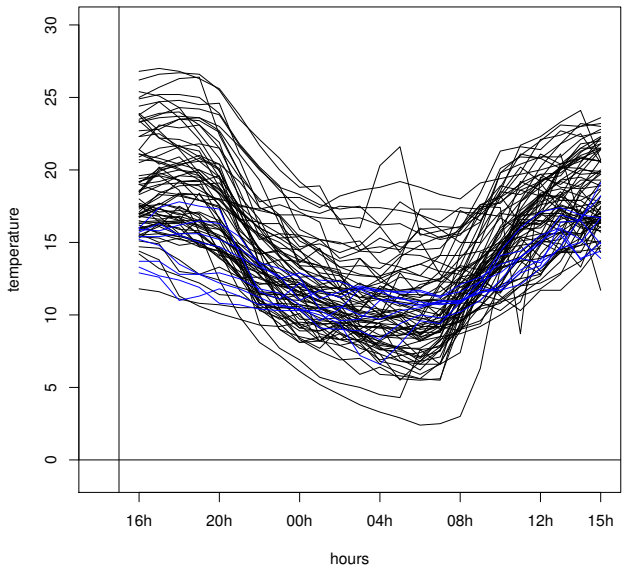

Figure 5: Plot of the 78 daily temperature curves (the blue curves are given when the response variable Y is missing).

the atmospheric index is missing. Notice that the response is missing for mild temperature curves cities: the fact that the value of the response variable Y is missing for these cities depends on the temperature curve X, and thus we consider the MAR case. We also refer the reader to the paper [19] for more discussions about missing data mechanism when dealing with air quality data. In particular, this paper highlights the fact that air quality missing data can be considered as MAR. Fig 6 illustrates the selected cities in our study, the blue cities are given when the response variable Y is missing and the red cities are given when the response variable Y is observed. It is of primary importance to get a map of the atmospheric index on the whole French territory, and thus to impute missing data.

We have built a sample of 78 pairs $\{(Y_i, X_i)\}_{i=1}^{78}$, where we have 8 missing

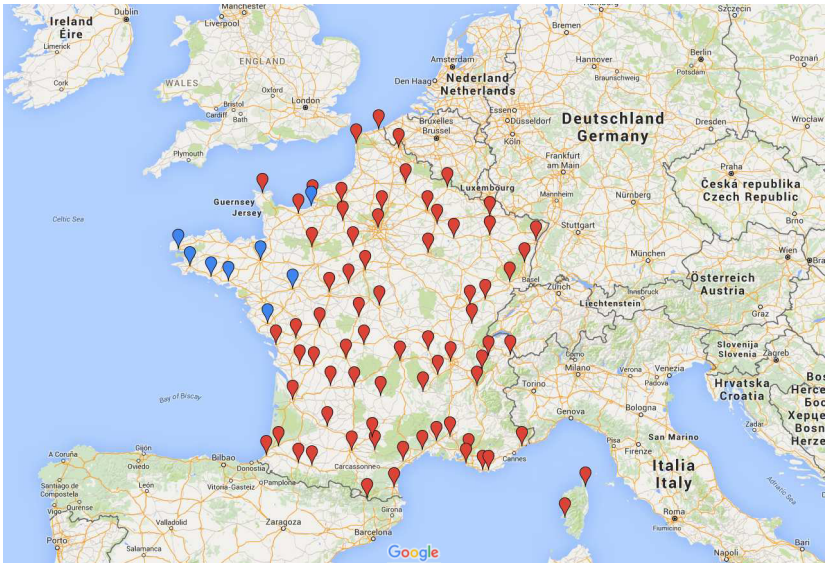

Figure 6: Map of France locating the selected cities of our study: the cities are red when the variable Y is observed and the cities are blue when the variable Y is missing.

values of the variable Y (the $Y_i's, i = 71, \ldots, 78$, are missing). Our goal is to impute these missing values $\{Y_i\}_{i=71}^{78}$.

We have fixed the number of knots to be 20, the degree of splines has been chosen equal to 3 and the number of derivatives was fixed to the moderate value of 2. Then, we use the GCV criterion to find the best parameter of projection dimension k_n trying growing sequences: $k_n = 2, 3, \ldots, 21, 22$. 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 3. Here again, the more we have missing data in the training set, the more the prediction error on the test sample is.

Now, we come back to the initial goal, imputing the missing data. The

Table 3: Real data set: prediction errors over 700 drawn samples.

	$n = 78$, 8 missing data, 70 observed data		
Test sets	$n/4$	$n/3$	$n/2$
Rate of missing data (%)	13	15	20
$\overline{MSE'} \times 10^2$	24.5650 (8.4750)	25.5172 (8.1444)	29.7827 (15.0889)

minimum value of the GCV criterion is reached for $k_n^* = 5$ and MSE' $(\times 10^2) = 20.791$. Table 4 gives the imputed values of the missing data. We see imputed values mainly around 4, which is a moderate value for the atmospheric index corresponding to a good quality of air. It is in accordance with the fact that these cities have moderate temperature curves. We can mention two particular cases. The highest imputed value (4.161) corresponds to the city of Angers, and in parallel, we can see that the temperature curve of this city becomes high at the end of May 8. On the contrary, the lowest imputed value (3.491) corresponds to the city of Quimper, and the temperature curve of this city presents few variations along the 24 hours.

Table 4: Imputed values of the missing response variable.

Missing values of Y	Y_{71}	Y_{72}	Y_{73}	Y_{74}	Y_{75}	Y_{76}	Y_{77}	Y_{78}
Imputed values	4.161	3.496	3.850	3.758	3.590	3.491	3.990	3.821

5. Proof of the results

5.1. Proof of Theorem 2.1

We begin with the following decomposition

$$\widehat{\Delta}_{n,obs} = \frac{1}{n-m_n}\sum_{i=1}^{n}\langle X_i,.\rangle\delta_i\Theta X_i + \frac{1}{n-m_n}\sum_{i=1}^{n}\langle X_i,.\rangle\delta_i\varepsilon_i = \Theta\widehat{\Gamma}_{n,obs} + U_{n,obs},$$

with $U_{n,obs} = \frac{1}{n-m_n}\sum_{i=1}^{n}\langle X_i,.\rangle\delta_i\varepsilon_i$. Then, ε being independent from X and δ (MAR assumption), we deduce

$$\begin{aligned}
\mathbb{E}\left(Y_{\ell,imp} - \langle\theta, X_\ell\rangle\right)^2 &= \mathbb{E}\left(\Theta\widehat{\Pi}_{k_n,obs}X_\ell - \Theta X_\ell\right)^2 \\
&\quad + \mathbb{E}\left(\frac{1}{n-m_n}\sum_{i=1}^{n}\langle X_i, \left(\widehat{\Pi}_{k_n,obs}\widehat{\Gamma}_{n,obs}\right)^{-1}X_\ell\rangle\delta_i\varepsilon_i\right)^2 \\
&\leq 2\mathbb{E}\left(\Theta\widehat{\Pi}_{k_n,obs}X_\ell - \Theta\Pi_{k_n,obs}X_\ell\right)^2 \\
&\quad + 2\mathbb{E}\left(\Theta\Pi_{k_n,obs}X_\ell - \Theta X_\ell\right)^2 \\
&\quad + \mathbb{E}\left(\frac{1}{n-m_n}\sum_{i=1}^{n}\langle X_i, \left(\widehat{\Pi}_{k_n,obs}\widehat{\Gamma}_{n,obs}\right)^{-1}X_\ell\rangle\delta_i\varepsilon_i\right)^2,
\end{aligned}$$

where $\Pi_{k_n,obs}$ is the projection onto the subspace $\mathrm{span}(v_{1,obs},\ldots,v_{k_n,obs})$ where $v_{1,obs},\ldots,v_{k_n,obs}$ are the k_n first eigenfunctions of the covariance operator $\Gamma_{n,obs}$. For a single imputation, the end of the proof of Theorem 2.1 is based on the following lemmas. For the aggregate error term of m_n imputed values, it is just a sum of m_n terms that behave like the term for single imputation.

Lemma 5.1. *We have*

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

Lemma 5.2. *We have*

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

Lemma 5.3. *We have*

$$\mathbb{E}\left(\Theta\Pi_{k_n,obs}X_\ell - \Theta X_\ell\right)^2 = \sum_{j=k_n+1}^{+\infty}\left(\Theta\Gamma^{1/2}v_j\right)^2.$$

5.2. Proof of Lemma 5.1

Writing X_ℓ in the basis $(v_j)_{j\geq 1}$, we obtain

$$
\begin{aligned}
& \mathbb{E}\left(\Theta\widehat{\Pi}_{k_n,obs}X_\ell - \Theta\Pi_{k_n,obs}X_\ell\right)^2 \\
= & \sum_{j=1}^{+\infty}\sum_{j'=1}^{+\infty}\mathbb{E}\left[\langle X_\ell, v_j\rangle\langle X_\ell, v_{j'}\rangle\Theta\left(\widehat{\Pi}_{k_n,obs}-\Pi_{k_n,obs}\right)v_j\Theta\left(\widehat{\Pi}_{k_n,obs}-\Pi_{k_n,obs}\right)v_{j'}\right].
\end{aligned}
$$

Noticing that the variable X_ℓ corresponds to the missing data Y_ℓ hence independent of $\widehat{\Pi}_{k_n,obs}$, we get

$$
\begin{aligned}
& \mathbb{E}\left(\Theta\widehat{\Pi}_{k_n,obs}X_\ell - \Theta\Pi_{k_n,obs}X_\ell\right)^2 \\
= & \sum_{j=1}^{+\infty}\sum_{j'=1}^{+\infty}\langle\Gamma v_j, v_{j'}\rangle\mathbb{E}\left[\Theta\left(\widehat{\Pi}_{k_n,obs}-\Pi_{k_n,obs}\right)v_j\Theta\left(\widehat{\Pi}_{k_n,obs}-\Pi_{k_n,obs}\right)v_{j'}\right] \\
= & \sum_{j=1}^{+\infty}\lambda_j\mathbb{E}\left[\Theta\left(\widehat{\Pi}_{k_n,obs}-\Pi_{k_n,obs}\right)v_j\right]^2.
\end{aligned}
$$

Now, following the proof of Proposition 15 in [11], for any $m \geq 1$ we denote $\mathcal{B}_m$ the oriented circle of the complex plane with center λ_m and radius $\rho_m/2$ where $\rho_m = \min\left(\lambda_m - \lambda_{m+1}, \lambda_{m-1} - \lambda_m\right)$ for $m \geq 2$ and $\rho_1 = \lambda_2 - \lambda_1$. With the convexity assumption $(A.1)$, we actually have $\rho_m = \lambda_m - \lambda_{m+1}$ for all $m \geq 1$. With these notations, denoting by ι the complex number such that $\iota^2 = -1$, the difference between the projection operators $\widehat{\Pi}_{k_n,obs}$ and $\Pi_{k_n,obs}$ can be written

$$
\widehat{\Pi}_{k_n,obs} - \Pi_{k_n,obs} = \frac{1}{2\pi\iota}\sum_{m=1}^{k_n}\int_{\mathcal{B}_m}\Lambda(z)\left(\Gamma - \widehat{\Gamma}_{n,obs}\right)\Lambda(z)\mathrm{d}z,
$$

where $\Lambda(z) = (zI - \Gamma)^{-1}$. Noticing that $\Lambda(z)v_j = \frac{1}{z-\lambda_j}v_j$, we deduce

$$
\begin{aligned}
& \Theta\left(\widehat{\Pi}_{k_n,obs}-\Pi_{k_n,obs}\right)v_j \\
= & \frac{1}{2\pi\iota}\sum_{m=1}^{k_n}\Theta\int_{\mathcal{B}_m}\Lambda(z)\left(\Gamma - \widehat{\Gamma}_{n,obs}\right)\frac{dz}{z-\lambda_j} \\
= & \frac{1}{2\pi\iota}\sum_{m=1}^{k_n}\Theta\int_{\mathcal{B}_m}\sum_{j'=1}^{+\infty}\frac{\langle\left(\Gamma - \widehat{\Gamma}_{n,obs}\right)v_j, v_{j'}\rangle v_{j'}}{(z-\lambda_{j'})(z-\lambda_j)}\mathrm{d}z.
\end{aligned}
$$

Still using the results from [11], we have

$$\sum_{m=1}^{k_n}\int_{\mathcal{B}_m}\frac{dz}{(z-\lambda_{j'})(z-\lambda_j)}=\begin{cases}0, & \text{if } j,j'>k_n,\\ 0, & \text{if } j,j'\leq k_n,\\ (\lambda_j-\lambda_{j'})^{-1}, & \text{if } j\leq k_n<j',\\ (\lambda_{j'}-\lambda_j)^{-1}, & \text{if } j'\leq k_n<j.\end{cases}$$

hence we deduce

$$\begin{aligned}&\mathbb{E}\left(\Theta\widehat{\Pi}_{k_n,obs}X_\ell-\Theta\Pi_{k_n,obs}X_\ell\right)^2\\ =\;&\mathbb{E}\left[\frac{1}{4\pi^2}\sum_{j=1}^{k_n}\lambda_j\left(\sum_{j'=k_n+1}^{+\infty}\frac{\langle\left(\Gamma-\widehat{\Gamma}_{n,obs}\right)v_j,v_{j'}\rangle}{\lambda_j-\lambda_{j'}}\Theta v_{j'}\right)^2\right]\\ &+\mathbb{E}\left[\frac{1}{4\pi^2}\sum_{j=k_n+1}^{+\infty}\lambda_j\left(\sum_{j'=1}^{k_n}\frac{\langle\left(\Gamma-\widehat{\Gamma}_{n,obs}\right)v_j,v_{j'}\rangle}{\lambda_{j'}-\lambda_j}\Theta v_{j'}\right)^2\right].\end{aligned}$$

In the following, C corresponds to a generic constant. We denote $\mathbb{E}(A)$ and $\mathbb{E}(B)$ the above two terms. We start with the computation of $\mathbb{E}(A)$. Using the same technique as in [11], we get the following bound

$$\mathbb{E}\left(\langle\left(\Gamma-\widehat{\Gamma}_{n,obs}\right)v_j,v_{j'}\rangle\langle\left(\Gamma-\widehat{\Gamma}_{n,obs}\right)v_j,v_r\rangle\right)\leq\frac{C}{n-m_n}\lambda_j\sqrt{\lambda_{j'}}\sqrt{\lambda_r},$$

noticing that the n rate of convergence given in [11] is here transformed into the $n-m_m$ rate because we use $\widehat{\Gamma}_{n,obs}$ with $n-m_m$ observed data. Hence we deduce

$$\begin{aligned}&\mathbb{E}\left(\frac{\langle\left(\Gamma-\widehat{\Gamma}_{n,obs}\right)v_j,v_{j'}\rangle}{\lambda_j-\lambda_{j'}}\Theta v_{j'}\right)^2\\ =\;&\sum_{j'=k_n+1}^{+\infty}\sum_{r=k_n+1}^{+\infty}\frac{\mathbb{E}\left(\langle\left(\Gamma-\widehat{\Gamma}_{n,obs}\right)v_j,v_{j'}\rangle\langle\left(\Gamma-\widehat{\Gamma}_{n,obs}\right)v_j,v_r\rangle\right)}{(\lambda_j-\lambda_{j'})(\lambda_j-\lambda_r)}\Theta v_{j'}\Theta v_r\\ \leq\;&\frac{C\lambda_j}{n-m_n}\left(\sum_{j'=k_n+1}^{+\infty}\frac{\sqrt{\lambda_j}}{\lambda_j-\lambda_{j'}}\Theta v_{j'}\right)^2.\end{aligned}$$

Coming back to the computation of $\mathbb{E}(A)$, we can write (using Lemma 12 in [11])

$$\begin{aligned}
\mathbb{E}(A) &\leq \frac{C}{n-m_n}\sum_{j=1}^{k_n}\frac{\lambda_j^2\lambda_{k_n+1}}{(\lambda_j-\lambda_{k_n+1})^2}\left(\sum_{j'=k_n+1}^{+\infty}\Theta v_{j'}\right)^2 \\
&\leq \frac{C\lambda_{k_n+1}}{n-m_n}\sum_{j=1}^{k_n}\frac{(k_n+1)^2}{(k_n+1-j)^2}\left(\sum_{j'=k_n+1}^{+\infty}\Theta v_{j'}\right)^2 \\
&\leq \frac{C\lambda_{k_n+1}(k_n+1)^2}{n-m_n}\sum_{j=1}^{k_n}\frac{1}{j^2}\left(\sum_{j'=k_n+1}^{+\infty}\Theta v_{j'}\right)^2.
\end{aligned}$$

As $\theta \in L^2([0,1])$ (hence θ is integrable), we finally get

$$\mathbb{E}(A) \leq \frac{C\lambda_{k_n}k_n^2}{n-m_n}a_n,$$

where $(a_n)_{n\geq 1}$ is a sequence of real numbers going to zero as n goes to infinity. We are now interested in the computation of $\mathbb{E}(B)$. Beginning in the same way as $\mathbb{E}(A)$ and still using Lemma 12 in [11], we get

$$\begin{aligned}
\mathbb{E}(B) &\leq \frac{C}{n-m_n}\sum_{j=k_n+1}^{+\infty}\lambda_j^2\left(\sum_{j'=1}^{k_n}\frac{\sqrt{\lambda_{j'}}}{\lambda_{j'}-\lambda_j}\Theta v_{j'}\right)^2 \\
&\leq \frac{C}{n-m_n}\sum_{j=k_n+1}^{+\infty}\lambda_j\left(\sum_{j'=1}^{k_n}\frac{\lambda_{j'}}{\lambda_{j'}-\lambda_j}\Theta v_{j'}\right)^2 \\
&\leq \frac{C}{n-m_n}\sum_{j=k_n+1}^{+\infty}\lambda_j\left(\frac{j}{j-k_n}\right)^2\left(\sum_{j'=1}^{k_n}\Theta v_{j'}\right)^2.
\end{aligned}$$

Now, again with the integrability of θ and the fact that

$$\sum_{j=k_n+1}^{+\infty}\lambda_j\left(\frac{j}{j-k_n}\right)^2 \leq Ck_nb_n,$$

with $(b_n)_{n\geq 1}$ going to zero as n goes to infinity (see [11] p.19 in the proof of Proposition 15), we conclude

$$\mathbb{E}(B) \leq \frac{Ck_n}{n-m_n} b_n,$$

and this achieves the proof of Lemma 5.1.

5.3. Proof of Lemma 5.2

Let us denote

$$T_n = \frac{1}{n-m_n} \sum_{i=1}^{n} \langle X_i, \left(\widehat{\Pi}_{k_n,obs}\widehat{\Gamma}_{n,obs}\right)^{-1} X_\ell \rangle \delta_i \varepsilon_i.$$

We can write

$$\begin{aligned} T_n^2 &= \frac{1}{(n-m_n)^2} \sum_{i=1}^{n} \langle X_i, \left(\widehat{\Pi}_{k_n,obs}\widehat{\Gamma}_{n,obs}\right)^{-1} X_\ell \rangle^2 \delta_i^2 \varepsilon_i^2 \\ &\quad + \frac{1}{(n-m_n)^2} \sum_{i=1}^{n} \sum_{\substack{i'=1 \\ i' \neq i}}^{n} \langle X_i, \left(\widehat{\Pi}_{k_n,obs}\widehat{\Gamma}_{n,obs}\right)^{-1} X_\ell \rangle \langle X_{i'}, \left(\widehat{\Pi}_{k_n,obs}\widehat{\Gamma}_{n,obs}\right)^{-1} X_\ell \rangle \delta_i \delta_{i'} \varepsilon_i \varepsilon_{i'}. \end{aligned}$$

From the independence between ε and X and the MAR assumption, the expectation of the second term above is zero, hence

$$\begin{aligned} \mathbb{E}\left(T_n^2\right) &= \frac{1}{n-m_n} \mathbb{E}\left[\langle X_i, \left(\widehat{\Pi}_{k_n,obs}\widehat{\Gamma}_{n,obs}\right)^{-1} X_\ell \rangle^2 \delta_i^2 \varepsilon_i^2\right] \\ &= \frac{\sigma_\varepsilon^2}{n-m_n} \mathbb{E}\left[\langle X_i, \left(\widehat{\Pi}_{k_n,obs}\widehat{\Gamma}_{n,obs}\right)^{-1} X_\ell \rangle^2 \delta_i^2\right], \end{aligned}$$

the index i corresponding to an observed data in the sample (and consequently $\delta_i = 1$ for this observation). We finally get

$$\mathbb{E}\left(T_n^2\right) = \frac{\sigma_\varepsilon^2}{n-m_n} \mathbb{E}\left[\langle X_i, \left(\widehat{\Pi}_{k_n,obs}\widehat{\Gamma}_{n,obs}\right)^{-1} X_\ell \rangle^2\right].$$

Following the same lines of the proof of Proposition 17 and Lemma 19 in [11], we obtain

$$\mathbb{E}\left[\langle X_i, \left(\widehat{\Pi}_{k_n,obs}\widehat{\Gamma}_{n,obs}\right)^{-1} X_\ell \rangle^2\right] = k_n + o\left(k_n\right),$$

which achieves the proof of the Lemma.

5.4. Proof of Lemma 5.3

The proof of this lemma is quite immediate, noticing that

$$\begin{aligned}\mathbb{E}\left(\Theta\Pi_{k_n,obs}X_\ell - \Theta X_\ell\right)^2 &= \mathbb{E}\left(\langle(\Pi_{k_n,obs} - I)\, X_\ell, \theta\rangle^2\right)\\ &= \langle(\Pi_{k_n,obs} - I)\,\Gamma\theta, \theta\rangle\\ &= \sum_{j=k_n+1}^{+\infty}\left(\Theta\Gamma^{1/2}v_j\right)^2.\end{aligned}$$

5.5. Proof of Theorem 2.2

From Theorem 2.1, the last term in the asymptotic development is negligible, so we just have to achieve the usual trade-off between the square bias and the variance. Given that

$$\sum_{j=k_n+1}^{+\infty}\left(\Theta\Gamma^{1/2}v_j\right)^2 = \sum_{j=k_n+1}^{+\infty} L^2\varphi(j),$$

we approximate this sum with the integral $\int_x^{+\infty} L^2\varphi(t)\,\mathrm{d}t$, which gives the desired result.

5.6. Proof of Theorem 2.3

First, if we follow the same lines of the proof of Lemmas 5.1 and 5.3 in Theorem 2.1 but with all the sample $X_1, \ldots, X_n$, we get

$$\mathbb{E}\left(\Theta\widehat{\Pi}_{k_n}X_{new} - \Theta\Pi_{k_n}X_{new}\right)^2 = o\left(\frac{\lambda_{k_n}k_n^2}{n} + \frac{k_n}{n}\right), \tag{14}$$

and

$$\mathbb{E}\left(\Theta\Pi_{k_n}X_{new} - \Theta X_{new}\right)^2 = \sum_{j=k_n+1}^{+\infty}\left(\Theta\Gamma^{1/2}v_j\right)^2. \tag{15}$$

Now, let us denote, for $i = 1, \ldots, n$,

$$\varepsilon_{i,imp} = Y_{i,imp} - \langle\theta, X_i\rangle,$$

and

$$\varepsilon_i^\star = \delta_i\varepsilon_i + (1-\delta_i)\varepsilon_{i,imp}.$$

We immediately can write

$$\varepsilon_{i,imp} = \varepsilon_i + Y_{i,imp} - Y_i,$$

and

$$\varepsilon_i^\star = \varepsilon_i + (1-\delta_i)(Y_{i,imp} - Y_i).$$

Then, following the proof of Lemma 5.2 in Theorem 2.1, we denote

$$S_n = \frac{1}{n}\sum_{i=1}^n \langle X_i, \left(\widehat{\Pi}_{k_n}\widehat{\Gamma}_n\right)^{-1} X_{new}\rangle \varepsilon_i^\star,$$

whence,

$$\begin{aligned} S_n^2 &= \frac{1}{n^2}\sum_{i=1}^n \langle X_i, \left(\widehat{\Pi}_{k_n}\widehat{\Gamma}_n\right)^{-1} X_{new}\rangle^2 \left(\varepsilon_i^\star\right)^2 \\ &\quad + \frac{1}{n^2}\sum_{i=1}^n \sum_{\substack{i'=1 \\ i'\neq i}}^n \langle X_i, \left(\widehat{\Pi}_{k_n}\widehat{\Gamma}_n\right)^{-1} X_{new}\rangle \langle X_{i'}, \left(\widehat{\Pi}_{k_n}\widehat{\Gamma}_n\right)^{-1} X_{new}\rangle \varepsilon_i^\star \varepsilon_{i'}^\star. \end{aligned}$$

We notice that, for $i \neq i'$, we have

$$\mathbb{E}\left(\varepsilon_i^\star \varepsilon_{i'}^\star\right) \leq 4\mathbb{E}(Y_{i,imp} - Y_i)^2 \leq 8\left[\mathbb{E}(Y_{i,imp} - \langle \theta, X_i\rangle)^2 + \sigma_\varepsilon^2\right].$$

This bound and the lines of the proof of Lemma 5.2 give

$$\mathbb{E}\left(\frac{1}{n}\sum_{i=1}^n \langle X_i, \left(\widehat{\Pi}_{k_n}\widehat{\Gamma}_n\right)^{-1} X_{new}\rangle \varepsilon_i^\star\right)^2 = O\left(\frac{(n-m_n)k_n}{n^2} + \frac{m_n^2 k_n^2}{n^2}\right). \quad (16)$$

Now, combining relations (14), (15) and (16) and the fact that $m_n = o(n)$ and $m_n^2 k_n = O(n)$ (due to $a_n^2 n = o(1)$), we get the desired result.

Appendix A. Appendix

References

[1] Bosq, D., Linear Processes in Function Spaces: Theory and Applications (First edition). NY: Springer, New York, 2000.

[2] Bugni, F. A., Specification test for missing functional data, Econometric Theory, 28 (2012) 959–1002.

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

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

[5] Cardot, H., Ferraty, F. and Sarda, P., Spline estimators for the functional linear model, Statistica Sinica, 13 (2003) 571–591.

[6] Cardot, H. and Johannes, J., Thresholding projection estimators in functional linear models, Journal of Multivariate Analysis, 101 (2010) 5395–408.

[7] Cheng, P. E., Nonparametric Estimation of Mean Functionals with Data Missing at Random, Journal of the American Statistical Association, 89 (1994) 81–87.

[8] Chu, C. K. Cheng, P. E., Nonparametric regression estimation with missing data, Journal of Statistical Planning and Inference, 48 (1995) 85–99.

[9] Chiou, J-M., Zhang, Y-C., Chen, W-H. and Chang, C-W., A functional data approach to missing value imputation and outlier detection for traffic flow data, Transportmetrica B: Transport Dynamics, 2 (2014) 106–129.

[10] Crambes, C., Kneip, A. and Sarda, P., Smoothing splines estimators for functional linear regression, The Annals of Statistics, 37 (2009) 35–72.

[11] Crambes, C. and Mas, A., Asymptotics of prediction in functional linear regression with functional outputs, Bernoulli, 19 (2013) 2627–2651.

[12] De Boor, C., A practical guide to splines. Applied Mathematical Sciences, Springer, New York, 1978.

[13] Ferraty, F. and Vieu, P., Nonparametric functional data analysis: Theory and practice. NY: Springer-Verlag, New York, 2006.

[14] Ferraty, F., Sued, M. and Vieu, P., Mean estimation with data missing at random for functional covariables, Statistics: A Journal of Theoretical and Applied Statistics, 47 (2013) 688–706.

[15] Graham, J. W., Missing data analysis and design. NY: Springer, New York, 2012.

[16] Hall, P. and Horowitz, J. L., Methodology and Convergence Rates for Functional Linear Regression, The Annals of Statistics, 35 (2007) 70–91.

[17] He, Y., Yucel, R. and Raghunathan, T. E., A functional multiple imputation approach to incomplete longitudinal data, Statistics in Medicine, 30 (2011) 1137–1156.

[18] Horváth, L. and Kokoszka, P., Inference for Functional Data with Applications. NY: Springer-Verlag, New York, 2012.

[19] Junninen, H., Niska, H., Tuppurainen, K., Ruuskanen, J. and Kolehmainen, M., Methods for imputation of missing values in air quality datasets. Atmospheric environment, 38 (2004) 2895–2907.

[20] Little, R. J. A. and Rubin, D. B., Statistical analysis with missing data (Second edition). NY: John Wiley, New York, 2002.

[21] Manski, C.F., Identification problems in the social sciences. Harvard University Press, 1995.

[22] Manski, C.F., Partial identification of probability distributions. Springer-Verlag, 2003.

[23] Mojirsheibani, M., Nonparametric curve estimation with missing data: A general empirical process approach, Journal of Statistical Planning and Inference, 137 (2007) 2733–2758.

[24] Preda, C., Saporta, G. and Hadj M. M. H., The NIPALS Algorithm for Functional Data, Revue Roumaine de Mathématique Pures et Appliquées, 55 (2010) 315–326.

[25] Ramsay, J. O. and Dalzell, C., Some tools for functional data analysis, Journal Royal Statistical Society B, 53 (1991) 539–572.

[26] Ramsay, J. O. and Silverman, B. W., Functional Data Analysis (Second edition). NY: Springer-Verlag, New York, 2005.

[27] Ramsay, J. O., Hooker, G. and Graves, S., Functional Data Analysis with R and MATLAB (Fisrt edition). NY: Springer Publishing Company, New York, 2009.

[28] Shi, J. Q. and Choi, T., Gaussian Process Regression Analysis for Functional Data. Chapman and Hall (CRC Press), London, 2011.

[29] Van Buuren, S., Flexible Imputation of Missing Data. NJ: Chapman and Hall (CRC Press), Hoboken, 2012.

[30] Wang, Q., Linton, O. and Härdle, W., Semiparametric Regression Analysis with Missing Response at Random, Journal of the American Statistical Association, 99 (2004) 334–345.

[31] Yuan, M. and Cai, T. T., A reproducing kernel Hilbert space approach to functional linear regression, The Annals of Statistics, 38 (2010) 412–444.

[32] Zhang, J. T., Analysis of Variance for Functional Data. NY: Chapman and Hall, New York, 2014.

Table A.5: MAR ($Model_1$): Imputed values mean errors and standard deviations for samples with different sizes discretized in $p = 100$ equidistant points based on 500 simulation replications.

	$n + n_1 = 150$		
Rate of missing data (%)			
Mean	12.520	27.420	44.882
Median	13	27	45
SD	3.307	4.515	5.038
Criterion 1: $[\overline{MSE} \times 10^3]$	2.3592	2.7845	3.2821
	(1.8375)	(2.0370)	(2.0679)
Criterion 2: $[\overline{RT} \times 10^2]$	7.0001	7.5194	8.6148
	(6.6216)	(5.7701)	(5.7158)

	$n + n_1 = 450$		
Rate of missing data (%)			
Mean	12.433	27.456	45.209
Median	12.333	27.333	45.333
SD	1.877	2.487	3.041
Criterion 1: $[\overline{MSE} \times 10^3]$	0.8349	1.0048	1.3364
	(0.5728)	(0.6843)	(0.9037)
Criterion 2: $[\overline{RT} \times 10^2]$	2.2327	2.5724	3.4547
	(1.5754)	(1.7245)	(2.3383)

	$n + n_1 = 1800$		
Rate of missing data (%)			
Mean	12.529	27.536	45.213
Median	12.500	27.500	45.250
SD	0.934	1.280	1.355
Criterion 1: $[\overline{MSE} \times 10^3]$	0.2326	0.2759	0.3521
	(0.1321)	(0.1519)	(0.2018)
Criterion 2: $[\overline{RT} \times 10^2]$	0.5933	0.6962	0.8822
	(0.3492)	(0.3891)	(0.5036)

Table A.6: MAR ($Model_2$): Imputed values mean errors and standard deviations for samples with different sizes discretized in $p = 100$ equidistant points based on 500 simulation replications.

	$n + n_1 = 150$		
Rate of missing data (%)			
Mean	12.912	28.026	45.472
Median	13	28	45
SD	3.524	4.493	5.118
Criterion 1: $[\overline{MSE} \times 10^3]$	2.4786	2.9537	3.7448
	(2.0871)	(2.2814)	(2.8036)
Criterion 2: $[\overline{RT} \times 10^2]$	7.5424	7.7867	9.7596
	(8.0437)	(5.7674)	(7.1366)

	$n + n_1 = 450$		
Rate of missing data (%)			
Mean	12.924	28.018	45.277
Median	13	28	45.33
SD	1.871	2.533	2.844
Criterion 1: $[\overline{MSE} \times 10^3]$	0.8594	1.0189	1.2727
	(0.6156)	(0.6901)	(0.8227)
Criterion 2: $[\overline{RT} \times 10^2]$	2.2861	2.6008	3.2415
	(1.6605)	(1.7465)	(2.0856)

	$n + n_1 = 1800$		
Rate of missing data (%)			
Mean	13.010	28.081	45.289
Median	13	28.083	45.250
SD	0.970	1.330	1.456
Criterion 1: $[\overline{MSE} \times 10^2]$	0.1958	0.2420	0.2977
	(0.1262)	(0.1610)	(0.1852)
Criterion 2: $[\overline{RT} \times 10^2]$	0.5023	0.6193	0.7618
	(0.3284)	(0.4157)	(0.4776)

Table A.7: MAR ($Model_1$): Predicted values mean errors and standard deviations for samples with different sizes discretized in $p = 100$ equidistant points based on 500 simulation replications.

	$n + n_1 = 150$		
Rate of missing data (%)			
Mean	12.520	27.420	44.882
Median	13	27	45
SD	3.307	4.515	5.038
Criterion 3: $[\overline{MSE'} \times 10^3]$	2.3383	2.7173	3.1939
	(1.4987)	(1.8390)	(2.0391)
Criterion 4: $[\overline{RT'} \times 10^2]$	5.9523	6.9769	8.2677
	(3.7338)	(4.9933)	(5.6516)

	$n + n_1 = 450$		
Rate of missing data (%)			
Mean	12.433	27.456	45.209
Median	12.333	27.333	45.333
SD	1.877	2.487	3.041
Criterion 3: $[\overline{MSE'} \times 10^3]$	0.8453	0.9984	1.3046
	(0.5530)	(0.6729)	(0.8897)
Criterion 4: $[\overline{RT'} \times 10^2]$	2.1534	2.5348	3.3255
	(1.3984)	(1.6629)	(2.2417)

	$n + n_1 = 1800$		
Rate of missing data (%)			
Mean	12.529	27.536	45.213
Median	12.500	27.500	45.250
SD	0.934	1.280	1.355
Criterion 3: $[\overline{MSE'} \times 10^3]$	0.2295	0.2746	0.3474
	(0.1282)	(0.1512)	(0.1982)
Criterion 4: $[\overline{RT'} \times 10^2]$	0.5756	0.6887	0.8699
	(0.3165)	(0.3753)	(0.4888)

Table A.8: MAR ($Model_2$): Predicted values mean errors and standard deviations for samples with different sizes discretized in $p = 100$ equidistant points based on 500 simulation replications.

	$n + n_1 = 150$		
Rate of missing data (%)			
Mean	12.912	28.026	45.472
Median	13	28	45
SD	3.524	4.493	5.118
Criterion 3: $[\overline{MSE'} \times 10^3]$	2.3556	2.9148	3.6204
	(1.6157)	(2.2111)	(2.7093)
Criterion 4: $[\overline{RT'} \times 10^2]$	6.0704	7.4692	9.2007
	(4.1999)	(5.6623)	(6.5708)

	$n + n_1 = 450$		
Rate of missing data (%)			
Mean	12.924	28.018	45.277
Median	13	28	45.33
SD	1.871	2.533	2.844
Criterion 3: $[\overline{MSE'} \times 10^3]$	0.8183	0.9882	1.2666
	(0.5391)	(0.6270)	(0.8146)
Criterion 4: $[\overline{RT'} \times 10^2]$	2.0977	2.5322	3.2364
	(1.3686)	(1.5836)	(2.0620)

	$n + n_1 = 1800$		
Rate of missing data (%)			
Mean	13.010	28.081	45.289
Median	13	28.083	45.250
SD	0.970	1.330	1.456
Criterion 3: $[\overline{MSE'} \times 10^2]$	0.1896	0.2360	0.2935
	(0.1216)	(0.1531)	(0.1812)
Criterion 4: $[\overline{RT'} \times 10^2]$	0.4856	0.6035	0.7492
	(0.3148)	(0.3959)	(0.4618)

Table A.9: MAR ($Model_1$): Estimation of θ mean square errors, variance and square bias for samples with different sizes discretized in $p = 100$ equidistant points based on 500 simulation replications.

	$n + n_1 = 150$		
Rate of missing data (%)			
Mean	12.520	27.420	44.882
Median	13	27	45
SD	3.307	4.515	5.038
Criterion 5: $\overline{MSE''} \times 10^2$	20.33993	22.84329	25.59843
$\overline{Variance} \times 10^2$	16.42143	17.02001	17.58919
$\overline{Bias^2} \times 10^2$	3.918497	5.823277	8.009239

	$n + n_1 = 450$		
Rate of missing data (%)			
Mean	12.433	27.456	45.209
Median	12.333	27.333	45.333
SD	1.877	2.487	3.041
Criterion 5: $\overline{MSE''} \times 10^2$	8.923099	10.01299	12.37846
$\overline{Variance} \times 10^2$	7.636041	8.680379	10.64885
$\overline{Bias^2} \times 10^2$	1.287058	1.332613	1.729613

	$n + n_1 = 1800$		
Rate of missing data (%)			
Mean	12.529	27.536	45.213
Median	12.500	27.500	45.250
SD	0.934	1.280	1.355
Criterion 5: $\overline{MSE''} \times 10^2$	3.268755	3.663376	4.294925
$\overline{Variance} \times 10^2$	2.517848	2.870331	3.410527
$\overline{Bias^2} \times 10^2$	0.7509066	0.793045	0.884398

Table A.10: MAR ($Model_2$): Estimation of θ mean square errors, variance and square bias for samples with different sizes discretized in $p = 100$ equidistant points based on 500 simulation replications.

	$n + n_1 = 150$		
Rate of missing data (%)			
Mean	12.912	28.026	45.472
Median	13	28	45
SD	3.524	4.493	5.118
Criterion 5: $\overline{MSE''} \times 10^2$	25.77594	30.94147	35.58789
$\overline{Variance} \times 10^2$	17.87099	20.83862	21.5734
$\overline{Bias^2} \times 10^2$	7.904949	10.10285	14.01449

	$n + n_1 = 450$		
Rate of missing data (%)			
Mean	12.924	28.018	45.277
Median	13	28	45.33
SD	1.871	2.533	2.844
Criterion 5: $\overline{MSE''} \times 10^2$	12.80462	14.15714	16.64587
$\overline{Variance} \times 10^2$	6.696352	8.047992	10.44823
$\overline{Bias^2} \times 10^2$	6.108267	6.109149	6.197638

	$n + n_1 = 1800$		
Rate of missing data (%)			
Mean	13.010	28.081	45.289
Median	13	28.083	45.250
SD	0.970	1.330	1.456
Criterion 5: $\overline{MSE''} \times 10^2$	7.50709	8.091252	8.477034
$\overline{Variance} \times 10^2$	1.746334	2.096911	2.495418
$\overline{Bias^2} \times 10^2$	5.760756	5.994341	5.981616

Table A.11: MAR ($Model_1$): Computation times and selected dimensions of the CV, GCV and K-fold criteria for samples with different sizes discretized in $p = 100$ equidistant points.

$n + n_1$	150	450	1800
CV			
Computational times (sec.)	10.5928	74.1095	1158.8180
Best dimension k_n^* (For imputation)	5	5	6
Best dimension k_n^{**} (For prediction)	5	5	6
5-fold CV			
Computational times (sec.)	0.7885	1.3610	4.6047
Best dimension k_n^* (For imputation)	5	5	6
Best dimension k_n^{**} (For prediction)	5	5	6
10-fold CV			
Computational times (sec.)	1.2671	2.6702	9.9181
Best dimension k_n^* (For imputation)	5	5	6
Best dimension k_n^{**} (For prediction)	5	5	6
GCV			
Computational times (sec.)	0.3235	0.4065	1.3558
Best dimension k_n^* (For imputation)	5	5	6
Best dimension k_n^{**} (For prediction)	5	5	6