New perspectives in smoothing: minimax estimation of the mean and principal components of discretized functional data.

Angelina Roche

Minimax estimation of the mean and principal components of discretized functional data: some thoughts and perspectives in smoothing

Angelina Roche roche@ceremade.dauphine.fr

September 16, 2022

Abstract

Functional data analysis has been the subject of increasing interest over the past decades. Most existing theoretical contributions assume that the curves are fully observed, whereas in practice the data are observed on a finite grid and may be affected by noise. To account for the presence of noise and discretization, it is common to smooth the data. The purpose of this paper is to review some of the recent works studying the influence of the observation scheme for estimating the mean and principal components. Some of this work questions the need to smooth the data when the observation grid is fixed.

1 Introduction

In recent decades, the analysis of functional data has received increasing attention. We refer to Ramsay and Silverman (2005); Ferraty and Vieu (2006); Ferraty and Romain (2011); Hsing and Eubank (2015) for monographs and reviews on the subject.

In the following, let $(\Omega, \mathcal{A}, \mathbb{P})$ be a probability space. A *functional random variable* is a random variable taking values in a function space $\mathcal{F}$. In this article, we suppose that $\mathcal{F} = \mathbb{L}^2([0,1])$ and equip it with its usual scalar product $\langle f, g\rangle = \int_0^1 f(t)g(t)dt$ and associated norm $\|\cdot\|$ and Borel σ-field $\mathcal{B}$. A *functional random variable* is then a measurable map,

$$X : (\Omega, \mathcal{A}) \to (\mathcal{F}, \mathcal{B}).$$

If we assume that $\mathbb{E}[\|X\|^2] < +\infty$, one of the most common approach to understand the variability of a functional random variable X is its Karhunen-Loève decomposition (Grenander, 1950). Let $\mu(t) = \mathbb{E}[X(t)]$ the mean function of X then we can write

$$X = \mu + \sum_{j\geq 1} \sqrt{\lambda_j}\xi_j\eta_j,$$

where $(\lambda_j)_{j\geq 1}$ is a summable sequence of non-negative real numbers, $(\xi_j)_{j\geq 1}$ a sequence of uncorrelated random variables and $(\eta_j)_{j\geq 1}$ an orthonormal basis of $\mathcal{F}$, the *principal components*. The sum converges in norm, and also uniformly if the bi-variate function $K(s,t) = \mathrm{Cov}(X(s), X(t))$ is assumed to be a continuous function.

Finding estimations of the mean μ, and of the sequence $(\lambda_j)_{j\geq 1}$ and $(\eta_j)_{j\geq 1}$ from a sample $X_1, \ldots, X_n$ of functional data is then of particular interest. The estimation of principal components are indeed useful to vizualise the data (Principal Components Analysis, see Ramsay and

Silverman (2005)) or as a dimension reduction tool to perform regression tasks Brunel et al. 2016; Hall and Horowitz 2007; Zhou et al. 2022 or classification tasks Escabias et al. 2014; Preda et al. 2007; Jacques and Preda 2014, see also Müller 2005 and references therein.

Most theoretical contributions on the subject focus on the case where the curves $X_i(t)$ are fully observed, i.e. $X_i(t)$ is observed for all $t \in [0,1]$. However, in practice, the data is observed on a finite grid. The grid can be either fixed and regular or random and different from one individual to another. The latter observation scheme is often considered to be in the domain of Longitudinal Data Analysis. However, these two fields are very close and inference methods for functional data may in many cases be successfully applied to longitudinal data (see Müller 2005; Hall et al. 2006).

Moreover, the presence of noise in the observations has, therefore, naturally led to the consideration that a preliminary data smoothing step is generally essential to process functional data. However, despite its importance for practitioners, the impact of the smoothing step has rarely been studied from a theoretical point of view.

In this article, we first describe the usual smoothing and reconstruction techniques for functional data in Section 2. Then in Section 3 we attempt to summarize the main results on the convergence rates for the mean estimation. The results concerning the estimation of Principal Components are devoted to Section 4.

2 Usual approaches for smoothing and reconstruction of functional data

2.1 Fixed and random designs models

Suppose that $X_1, \dots, X_n \sim_{i.i.d.} X$ is a sample of functional data, and suppose also that X is continuous a.s. In most theoretical works (see e.g. Mas and Ruymgaart (2015); Bosq (2000)), it is assumed that the data is fully observed, without noise, that is to say that we observe $X_i(t)$ for all $t \in [0,1]$ and for all $i = 1, \dots, n$. We will call this setting *the ideal case*. However, in practice, the X_i's are only observed on a grid, which can be considered either as fixed, or random.

As in Cai and Yuan (2011), we consider two observation settings.

Setting (FG) We observe $\{Y_{i,j}), i = 1, \dots, n; j = 1, \dots, p\}$ where $\{t_1, \dots, t_p\}$ is a fixed grid of $[0,1]$ such that

$$\max_{j,k=1,\dots,p, j\neq k} |t_j - t_k| \leq C p^{-1}$$

and

$$Y_{i,j} = X_i(t_j) + \varepsilon_{i,j},$$

where $\{\varepsilon_{i,j}\}_{i=1,\dots,n; j=1,\dots,p}$ is a centered noise independent of $X_1, \dots, X_n$.

Setting (RG) We observe $\{Y_{i,j}, i = 1, \dots, n; j = 1, \dots, p_i\}$ where $\{T_{i,j}, i = 1, \dots, n; j = 1, \dots, p_i\}$ is an i.i.d. sequence following the uniform distribution on $[0,1]$ and

$$Y_{i,j} = X_i(T_{i,j}) + \varepsilon_{i,j},$$

where $\{\varepsilon_{i,j}\}_{i=1,\dots,n; j=1,\dots,p}$ is a centered noise independent of the X_i's and of the $T_{i,j}$'s. The grid $\{T_{i,j}, i = 1, \dots, n; j = 1, \dots, t_p\}$ also is independent of the X_i's.

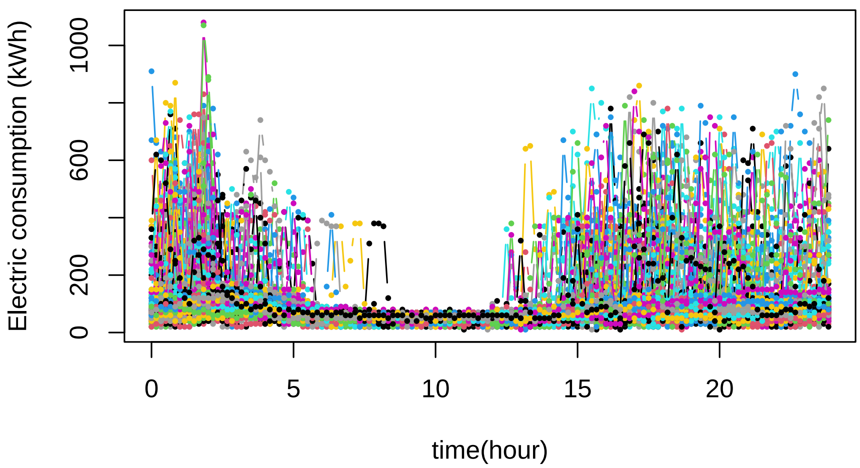

Figure 1: Electric consumption of appliances in a low energy house (Candanedo et al., 2017).

An example of functional data measured on a fixed grid, and which we can potentially reasonably model by the assumptions of the framework (FG) is given in Figure 1. This setting is typical of time-dependent phenomena (such as temperature, power consumption,...) that are measured automatically. The example of data that can be reasonably modeled by (RG)-type assumptions is given in Figure 2. This framework is natural for example for phenomena that are measured manually. Figure 2 is an example of a growth simulation curve, simulated from setting (RG). The framework (RG) has some similarities to that of longitudinal data analysis (LDA) (see Yu et al. 2022; Wong et al. 2022 for a recent contribution on the subject).

To avoid cumbersome notations, we will note hereafter $T_{i,j} = t_j$ and $p_i = p$ in the definitions and properties covering both cases (FG) and (RG).

The objective is to reconstruct, from the discrete observations, the unobserved random functions $X_i(t)$ for all $t \in [0, 1]$. This can be done in different ways with two objectives that can be distinct.

- If we consider that the impact of noise is negligible or has no importance on the data processing, it is enough to find, for all i, functions $\widetilde{X}_i : [0, 1] \to \mathbb{R}$ such that

$$\widetilde{X}_i(T_{i,j}) = Y_{i,j}, \qquad j = 1, \ldots, p_i. \tag{1}$$

  The curves $X_1, \ldots, X_n$ are then simply reconstructed by interpolating them.

- If we want to remove the effect of noise from the data as accurately as possible, we can smooth the data. In this case, Eq. (1) is no longer true.

The next two subsections detail some reconstruction and smoothing methods.

2.2 Kernel smoothing

One of the usual approaches to simultaneously reconstruct and smooth data is to use kernels. Let $K : \mathbb{R} \to \mathbb{R}$ be a kernel function, i.e. an integrable function which satisfies $\int_{\mathbb{R}} K(t)dt = 1$,

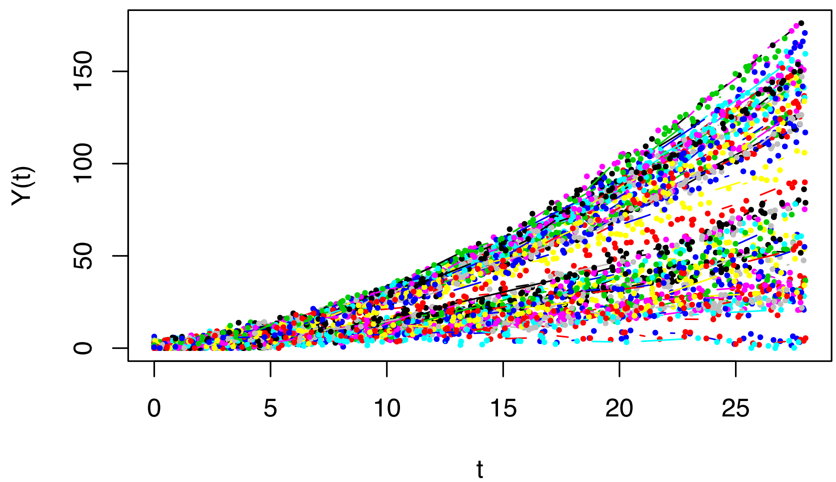

Figure 2: Simulated growth curves.

and $h > 0$ a parameter called *bandwidth*. We note

$$\widetilde{X}_i^{(KS,h)}(t) = \frac{\sum_{j=1}^{p_i} Y_{i,j} K\left(\frac{T_{i,j}-t}{h}\right)}{\sum_{j=1}^{p_i} K\left(\frac{T_{i,j}-t}{h}\right)}.$$

Figure 3 represents the reconstructed versions of the first curves for the two samples in figures 1 and 2. We observe that the parameter h plays a crucial role in the smoothing step. When h is small, there is almost no smoothing, and we simply reconstruct the Y_i's. On the contrary, when h is too large, the data are too smoothed and do not look like the observations anymore. In all cases, we introduce a bias that should be taken into account from a theoretical point of view.

2.3 Reconstruction with an orthonormal system of functions

An alternative to kernel smoothing is to reconstruct the data on a system of D orthonormal functions $\{\phi_1, \ldots, \phi_D\}$. We can always complete the orthonormal system $\{\phi_1, \ldots, \phi_D\}$ to obtain a Hilbert basis $(\phi_d)_{d \geq 1}$ of $\mathbb{L}^2([0,1])$. In that case, we have, if D is large,

$$X_i = \sum_{d \geq 1} \langle X_i, \phi_j \rangle \phi_d \approx \sum_{d=1}^{D} \langle X_i, \phi_d \rangle \phi_d,$$

the convergence of the series is in $\mathbb{L}^2([0,1])$.

Since the X_i's are unobserved, the scalar products $\langle X_i, \phi_d \rangle$ are not calculable in practice. We define an approximation of $\langle X_i, \phi_d \rangle$ e.g. by the trapezoidal rule

$$\widetilde{x_{i,d}} := \frac{1}{2} \sum_{d=2}^{p_i} (T_{i,(d)} - T_{i,(d-1)})(Y_{i,(d)} \phi(T_{i,(d)}) + Y_{i,(d-1)} \phi(T_{i,(d-1)})$$

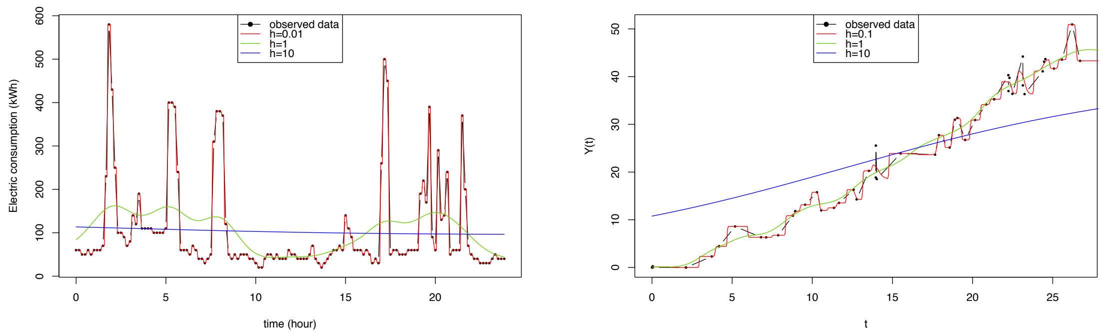

Figure 3: Kernel reconstruction of the first observation X_1 for the electric consumption data (left) and the growth data (right).

that gives us

$$\widetilde{X}_i^{(basis,D)}(t) = \sum_{d=1}^{D} \widetilde{x_{i,d}}\phi_d(t).$$

There are a large number of orthonormal systems that can be used to reconstruct functional data (B-splines, wavelets, Fourier basis, histogram system,...). We refer to Ramsay and Silverman (2005) for a description of the most common orthonormal systems and to Härdle et al. (1998) for a detailed reference on wavelet systems. We focus here on two classical orthonormal systems that are widely employed.

Histograms Let $I_1, \ldots, I_D$ be a partition of $[0,1]$ into D nonempty subintervals, we define

$$\phi_d(t) = \sqrt{|I_d|}^{-1}\mathbf{1}_{I_d}(t), \qquad d = 1, \ldots, D.$$

where $|I_d|$ denotes the length of I_d.

Fourier basis We suppose here that there exists $D' \in \mathbb{N}$ such that $D = 2D' + 1$ and set

$$\phi_1 \equiv 1, \quad \phi_{2d}(t) = \sqrt{2}\cos(2\pi dt) \text{ and } \phi_{2d+1}(t) = \sqrt{2}\sin(2\pi dt), \qquad d = 1, \ldots, D'.$$

We can see from figures 4 and 5 that the quality of the reconstruction strongly depends on the parameter D. Moreover, assuming that the power consumption peaks are of interest in the data we study, we see that when the window h is chosen too large or the number of functions D too small, the power consumption peaks disappear. This suggests that these power consumption data should not be too smoothed or that the two previous approaches are not well suited. We also observe that the Fourier basis is not a good choice to reconstruct the data, although they can be considered as periodic because a value of D large enough to make the peaks appear gives a reconstruction of X that is too oscillatory.

3 Minimax rates for mean estimation

In the ideal case, where we recall that we assume $X_i(t)$ is observed for all $i = 1, \ldots, n$ and $t \in [0,1]$, one would define a natural moment estimator of μ as follows

$$\widehat{\mu}(t) = \frac{1}{n}\sum_{i=1}^{n} X_i(t). \tag{2}$$

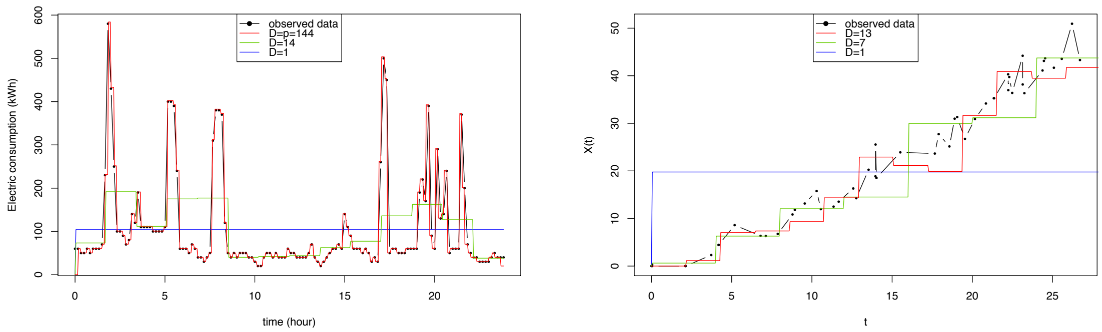

Figure 4: Histogram reconstruction of the first observation X_1 for the electric consumption data (left) and the growth data (right).

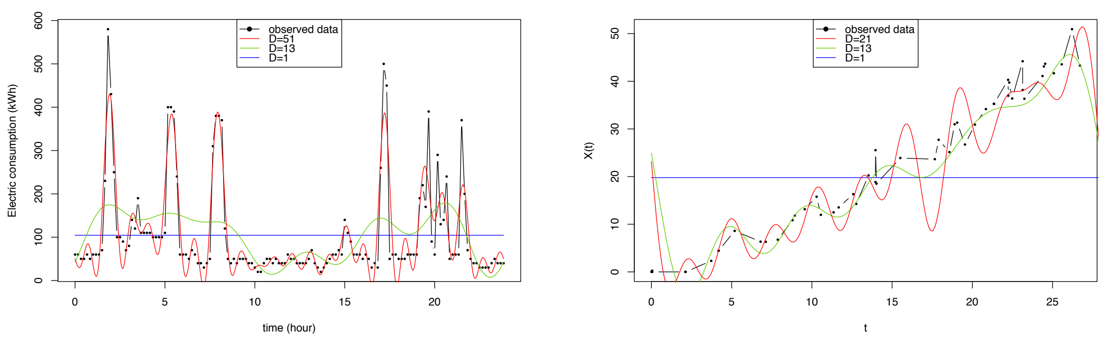

Figure 5: Fourier reconstruction of the first observation X_1 for the electric consumption data (left) and the growth data (right).

It can be proven easily that, if $\int_0^1 \mathrm{Var}(X(t))dt < +\infty$, $\widehat{\mu}$ converge to μ in $\mathbb{L}^2$-norm at a rate n^{-1} since

$$\mathbb{E}\left[\|\widehat{\mu}-\mu\|^2\right] = \int_0^1 \mathbb{E}[(\widehat{\mu}(t)-\mu(t))^2]dt = \int_0^1 \mathrm{Var}(\widehat{\mu}(t))dt = \frac{1}{n}\int_0^1 \mathrm{Var}(X(t))dt.$$

However, this assumption is not necessary. Bosq (2000) proves a law of large numbers in the case $\mathbb{E}[\|X\|] < +\infty$ (Theorem 2.4). Large deviation inequalities as well as Bernstein type inequalities for the estimator $\widehat{\mu}$ are also proven under stronger assumptions.

In the case where the data is observed on a grid, Cai and Yuan (2011) have proven minimax rates in both fixed design (FD) and random design (RD) settings. In the fixed design (FD) setting, they obtain the following lower bound for the rate (see Cai and Yuan (2011, Theorem 2.1)): define, for $r \in \mathbb{N}^*$ and $M_0 > 0$ $\mathcal{P}(r; M_0)$ the set of probability measures for a random function X such that X is a.s. r-times differentiable and verifies

$$\mathbb{E}\left[\|X^{(r)}\|^2\right] \leq M_0.$$

They prove that there exists a constant $d = d(M_0) > 0$ such that for any estimator $\tilde{\mu}$ of μ

$$\limsup_{n\to\infty} \sup_{\mathcal{L}(X)\in\mathcal{P}(r;M_0)} \mathbb{P}(\|\tilde{\mu}-\mu\|^2 > d(p^{-2r}+n^{-1})) > 0.$$

Comparing this rate with the rate n^{-1} obtained in the ideal case, we see here explicitly the influence of the discretization through the addition of the term p^{-2r}, which depends on the regularity of the curves of the sample. They prove that the smoothing splines estimator defined by Rice and Silverman (1991)

$$\widehat{\mu}_\lambda = \arg\min_{g\in\mathcal{W}_2^r}\left\{\frac{1}{n}\sum_{i=1}^n \frac{1}{p_i}\sum_{j=1}^{p_i}(Y_{ij}-g(T_j))^2 + \lambda\|g^{(r)}\|^2\right\}$$

with $\mathcal{W}_2^r$ the space of r-times differentiable absolutely continuous functions $g : [0,1] \to \mathbb{R}$ such that, for all $j = 1, \ldots, r-1$, $g^{(j)}$ is absolutely continuous and $g^{(r)} \in \mathbb{L}^2([0,1])$, attains the lower bound if $\lambda = O(p^{-2r}+n^{-1})$.

The study of these convergence rates give us some precious information, both from a theoretical and a practical point of view.

- First the estimator $\widehat{g}_\lambda$ attains the minimax rate even in the case $\lambda = 0$ corresponding to splines interpolation (in that case $\widehat{\mu}_0(T_{ij}) = Y_{ij}$, for all i and j). The regularization is therefore not necessary from a theoretical point of view and can even be detrimental because we do not know in practice the exact value of the regularity parameter appearing in the penalty.

- Second, note that, in the case of a fixed design, consistent estimation is not possible when the number of discretization points p is bounded (it must tend to infinity as n tends to infinity).

It is worth noting that the rates differs completely in the random design (RD) setting. Cai and Yuan (2011, Theorem 3.1) prove that, in the case where the T_{ij}'s are i.i.d. from a density η such that $\inf_{t\in[0,1]}\eta(t) > 0$, there exists a constant $d = d(M_0) > 0$ such that for any estimator $\tilde{\mu}$ of μ

$$\limsup_{n\to\infty} \sup_{\mathcal{L}(X)\in\mathcal{P}(r;M_0),\sum_{i=1}^n p_i^{-1}=p^{-1}} \mathbb{P}(\|\tilde{\mu}-\mu\|^2 > d((np)^{-2r/(2r+1)}+n^{-1})) > 0.$$

Here, the discretization adds a term of the order of $(np)^{-2r/(2r+1)}$. As in the (FD) setting, the smoothing splines estimate of Rice and Silverman (1991) also attain the rate given by the lower bound but in that case, the parameter λ is of order $(np)^{-2r/(2r+1)}$. Hence, we can remark that, contrary to the (FD) setting, it is possible to consistently estimate the mean even in the case where p_i is bounded by a constant. However regularization is needed to achieve the optimal rate. We also refer to Li and Hsing (2010) for a similar discussion on the local linear kernel smoother.

4 Minimax rates for the estimation of Principal Components

We first define the *covariance operator* as follows

$$\begin{aligned}\Gamma: \ & \mathcal{F} \ \to \mathcal{F} \\ & f \ \mapsto \mathbb{E}[\langle X-\mu, f\rangle (X-\mu)].\end{aligned}$$

In our case, where the space $\mathcal{F} = \mathbb{L}^2([0,1])$, Γ is an integral operator with kernel

$$K(s,t) = \mathrm{Cov}(X(s), X(t)), \qquad s,t \in [0,1]$$

in the sense that, for all $f \in \mathcal{F}$,

$$\Gamma f(t) = \int_0^1 f(s)K(s,t)ds, \qquad t \in [0,1].$$

Classical results of operator theory allows us to write (see e.g. Eq. (4) p. 6 of Bosq 2000)

$$\Gamma f = \sum_{j\geq 1} \lambda_j \langle f, \eta_j\rangle \eta_j, \qquad f \in \mathcal{F}$$

where $(\eta_j)_{j\geq 1}$ is an orthonormal basis of $\mathcal{F}$, the eigenfunctions of Γ, that we call *principal components basis* (PC basis) and $(\lambda_j)_{j\geq 1}$, the associated eigenvalues, is a summable sequence of non-negative real numbers. Usually, we assume without loss of generality that $(\lambda_j)_{j\geq 1}$ is a non-increasing sequence. The estimation of the PC basis and its associated eigenvalues is important in the functional data analysis field. In particular, since $(\eta_j)_{j\geq 1}$ is an orthonormal basis of X, we can write

$$X = \mu + \sum_{j\geq 1} \langle X, \eta_j\rangle \eta_j = \mu + \sum_{j\geq 1} \sqrt{\lambda_j}\xi_j\eta_j, \tag{3}$$

where $\xi_j = \langle X, \eta_j\rangle/\sqrt{\lambda_j}$ if $\lambda_j \neq 0$ and ξ_j is any standard random variable otherwise. The sequence $(\xi_j)_{j\geq 1}$ is a sequence of uncorrelated standard random variable, called *principal components scores*. We retrieve then here the *Karhunen-Loève decomposition* of X.

4.1 Estimation of principal components in the ideal case

A natural estimator of Γ is the *empirical covariance operator* defined as follows

$$\begin{aligned}\widehat{\Gamma}: \ & \mathcal{F} \ \to \mathcal{F} \\ & f \ \mapsto \frac{1}{n}\sum_{i=1}^n \langle X_i - \widehat{\mu}, f\rangle (X_i - \widehat{\mu}),\end{aligned}$$

with $\widehat{\mu}$ the natural estimator defined in equation (2).

The operator $\widehat{\Gamma}$ is a finite-rank operator, hence it is a compact operator. It is also self-adjoint. Then, the diagonalisation theorem for compact self-adjoint operators, ensures the existence of an orthonormal basis $(\widehat{\eta}_j)_{j\geq 1}$ of $\mathcal{F}$ of eigenfunctions of $\widehat{\Gamma}$. Denoting by $(\widehat{\lambda}_j)_{j\geq 1}$ the associated eigenvalues, sorted in decreasing order, it is natural to consider $\widehat{\eta}_j$ (resp. $\widehat{\lambda}_j$) as an estimator of η_j (resp. λ_j), for $j \geq 1$.

The risk of the estimators of the PC basis defined above and eigenfunctions can be related to the risk of the empirical covariance operator via the Bosq Inequalities (Bosq, 2000), in the case where the λ_j's are distinct,

$$\begin{aligned} \|\widehat{\eta}_j - \eta_{\pm,j}\|^2 &\leq b_j \|\widehat{\Gamma} - \Gamma\|_\infty^2 \qquad (4) \\ \sup_{j\geq 1} |\widehat{\lambda}_j - \lambda_j| &\leq \|\widehat{\Gamma} - \Gamma\|_\infty^2, \qquad (5) \end{aligned}$$

where $b_j = 8\min\{\lambda_j - \lambda_{j+1}; \lambda_{j-1} - \lambda_j\}^{-2}$ for $j \geq 2$, $b_1 = 8(\lambda_1 - \lambda_2)^{-2}$ and $\|\cdot\|$ is the usual operator norm defined by $\|T\|_\infty = \sup_{f\in\mathcal{F}} \|Tf\|/\|f\|$ for a linear operator T on $\mathcal{F}$ and $\eta_{\pm,j} = \mathrm{sign}(\langle \eta_j, \widehat{\eta}_j \rangle \eta_j$. Using a Bernstein type inequality for Banach random variables (Bosq, 2000, Corollary 4.1) we can prove the following upper-bound

$$\|\widehat{\Gamma} - \Gamma\|_\infty^2 = O\left(\frac{\log^2(n)}{n}\right) \; a.s.$$

that implies, up to a log term, a parametric convergence rate of the order n^{-1} for both the eigenfunctions and the eigenvectors. Under moment assumptions on the scores $(\xi_j)_{j\geq}$, that are verified e.g. if X is a Gaussian or a bounded process, and assuming that there exists $c > 0$ and $\alpha > 0$ such that $\lambda_j = cj^{-1-\alpha}$ or $\lambda_j = ce^{-\alpha j}$, the order of the quantity b_j appearing on the upper-bound on the eigenfunctions can be improved. Indeed, it can be deduced from the proof of Mas and Ruymgaart (2015, Corollary 2) that

$$\mathbb{E}[\|\widehat{\eta}_j - \eta_{\pm,j}\|^2] \leq \frac{c_2 j^2 \log^2(n)}{n} + e^{-c_1 \log^2 n}, \qquad (6)$$

where c_1 and c_2 are positive constants.

4.2 Estimation of principal components in the fixed design (FD) model

Consider the following Hölder regularity class for X. Let, $\alpha \in]0,1]$, M_0 and $\mathcal{P}(\alpha; M_0)$ the set of probability measures for a random function X such that

$$\mathbb{E}[(X(t) - X(s))^2] \leq M_0|t-s|^\alpha.$$

In the fixed design (FD) model, Belhakem et al. (2021) obtained a minimax lower bound for the estimation of the first eigenfunction η_1

$$\min_{\widehat{\eta}} \max_{\mathcal{L}(X)\in\mathcal{P}(\alpha;M_0)} \mathbb{E}[\|\widehat{\eta} - \eta_{\pm,1}\|^2] \geq c(p^{-2\alpha} + n^{-1}),$$

where $c > 0$ depending on M_0.

This rate is similar to the one obtained for the estimation of the mean function μ by Cai and Yuan (2011). They prove that the estimator obtained by projecting the data into the histogram basis with $D = p$ bins achieves the minimax rate. Descary and Panaretos (2016) obtain a similar decomposition under a more general assumption on the noise and a different assumption on the covariance operator of X.

4.3 Estimation of principal components in the random design (RD) model

This case has been investigated by Hall et al. (2006). They first estimate the covariance kernel $C(s,t) = \mathrm{Cov}(X(s), X(t))$ by taking the minimizer $\widehat{C}_h(s,t)$ in a_0 of the criterion

$$\sum_{i=1}^{n} \sum_{j \neq k} (Y_{i,j}Y_{i,k} - a_0 - b_1(s - T_{i,j}) - b_2(t - T_{i,k}))^2 \, K((T_{i,j} - s)/h)K((T_{i,k} - t)/h),$$

where $h > 0$ is the bandwidth and K is a kernel function. Once the kernel function is estimated, one can take the eigenfunctions of the associated covariance operator estimator $\widehat{\Gamma}_h : f \mapsto \int_0^1 \widehat{C}_h(\cdot, t) f(t) dt$ as an estimator of the eigenfunction. Under the condition that p is bounded, they prove that, for an optimal choice of the bandwidth $h \sim n^{-1/5}$,

$$\lim_{C \to \infty} \limsup_{n \to \infty} \max_{j=1,\dots,r} \sup_{(\eta_1,\dots,\eta_r) \in \Psi} \mathbb{P}(\|\widehat{\eta}_{j,h} - \eta_j\| > Cn^{-1/5}),$$

where Ψ is a class of r-uplet of orthonormal functions which are two times differentiable with uniformly bounded first and second derivatives. This rate is proven to be asymptotically minimax on the class Ψ.

More recently Zhou et al. (2022) have obtained a convergence rate of the order of

$$\|\widehat{\eta}_{j,h} - \eta_j\|^2 \leq C_j \left(\frac{1}{n} + \frac{1}{n^{4/5} p^{4/5}} \right)$$

under more restrictive assumptions on the eigenvalue sequence, without making the assumption that p is bounded and with a similar estimator. The minimal risk is obtained for $h \sim n^{1/5} p^{1/5}$. We do not know at this time if these rates are optimal but we can notice that they coincide with the rates obtained for the estimation of the mean by Cai and Yuan (2011). The constant C_j appearing in the upper bound depends on the rank of the eigenfunction that is estimated.

Concluding remarks

In view of the recent results of Cai and Yuan (2011); Belhakem et al. (2021), the question of the necessity of smoothing the functional data seems legitimately debatable, at least in the case where the observation grid is fixed, and even in the presence of noise. On the contrary, in the case where the grid is random, a regularization or a smoothing is necessary to obtain the best convergence rate. Particular attention must then be paid to the choice of the smoothing or regularization parameter whose optimal value depends on the unknown regularity of the data.

However, the convergence rates of the estimation of the eigenvalues $(\lambda_j)_{j \geq 1}$ remain, to our knowledge, an open question. Inequality (5) allows us to obtain an upper bound on this rate but it is not clear if this upper bound is precise enough or if it can be improved.

Another issue, which is of particular importance when PCA is considered as a dimension reduction tool, is to obtain the most accurate constants possible in the upper bounds of the eigenfunctions. In the ideal case where all the curves are observed Mas and Ruymgaart (2015) have obtained in Eq. (6) a constant equal to $c_2 j^2$. In the case where the data are observed on a grid (fixed or random and with or without noise), Inequality (4) allows us to obtain a constant of the order of the squared inverse of the spectral gap. Using the tools of perturbation theory, Zhou et al. (2022) have improved this constant in the case of a random observation grid (RD) but it remains larger than the constant of the ideal case, raising the question of whether or not it can be improved.

References

R. Belhakem, F. Picard, V. Rivoirard, and A. Roche. Minimax estimation of functional principal components from noisy discretized functional data, 2021. URL `https://arxiv.org/abs/2110.12739`.

D. Bosq. *Linear processes in function spaces*, volume 149 of *Lecture Notes in Statistics*. Springer-Verlag, New York, 2000. Theory and applications.

E. Brunel, A. Mas, and A. Roche. Non-asymptotic adaptive prediction in functional linear models. *J. Multivariate Anal.*, 143:208–232, 2016.

T. T. Cai and M. Yuan. Optimal estimation of the mean function based on discretely sampled functional data: phase transition. *Ann. Statist.*, 39(5):2330–2355, 2011.

L. M. Candanedo, V. Feldheim, and D. Deramaix. Data driven prediction models of energy use of appliances in a low-energy house. *Energy and Buildings*, 140, 2017.

M.-H. Descary and V. M. Panaretos. Functional data analysis by matrix completion. *Ann. Statist.*, 2016.

M. Escabias, A. M. Aguilera, and M. C. Aguilera-Morillo. Functional pca and base-line logit models. *J. Classif.*, 31(3):296–324, 2014.

F. Ferraty and Y. Romain, editors. *The Oxford handbook of functional data analysis*. Oxford University Press, Oxford, 2011.

F. Ferraty and P. Vieu. *Nonparametric functional data analysis*. Springer Series in Statistics. Springer, New York, 2006. Theory and practice.

U. Grenander. Stochastic processes and statistical inference. *Arkiv för matematik*, 1(3):195–277, 1950.

P. Hall and J. L. Horowitz. Methodology and convergence rates for functional linear regression. *Ann. Statist.*, 35(1):70 – 91, 2007.

P. Hall, H. Müller, Hans-Georg, and J.-L. Wang. Properties of principal component methods for functional and longitudinal data analysis. *Ann. Statist.*, 34(3):1493–1517, 2006.

W. Härdle, G. Kerkyacharian, D. Picard, and A. Tsybakov. *Wavelets, approximation, and statistical applications*, volume 129 of *Lecture Notes in Statistics*. Springer-Verlag, New York, 1998.

T. Hsing and R. Eubank. *Theoretical Foundations of Functional Data Analysis, with an Introduction to Linear Operators*. Wiley, May 2015.

J. Jacques and C. Preda. Functional data clustering: a survey. *Advances in Data Analysis and Classification*, 8(3):231–255, 2014.

Y. Li and T. Hsing. Uniform convergence rates for nonparametric regression and principal component analysis in functional/longitudinal data. *Ann. Statist.*, 38(6):3321 – 3351, 2010.

A. Mas and F. Ruymgaart. High-dimensional principal projections. *Complex Anal. Oper. Theory*, 9(1):35–63, 2015.

H.-G. Müller. Functional modelling and classification of longitudinal data. *Scand. J. Stat.*, 32 (2):223–240, 2005.

C. Preda, G. Saporta, and C. Lévéder. PLS classification of functional data. *Comput. Stat.*, 22 (2):223–235, 2007.

J. O. Ramsay and B. W. Silverman. *Functional data analysis*. Springer Series in Statistics. Springer, New York, second edition, 2005.

J. A. Rice and B. W. Silverman. Estimating the mean and covariance structure nonparametrically when the data are curves. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 53(1):233–243, 1991.

K. Y. Wong, D. Zeng, and D. Y. Lin. Semiparametric latent-class models for multivariate longitudinal and survival data. *Ann. Statist.*, 50(1):487–510, 2022.

J. Yu, T. Nummi, and J. Pan. Mixture regression for longitudinal data based on joint mean-covariance model. *J. Multivariate Anal.*, 190, 2022.

H. Zhou, F. Yao, and H. Zhang. Functional linear regression for discretely observed data: From ideal to reality. https://www.math.pku.edu.cn/teachers/yaof/papers/2022-flr.pdf, 2022.