Bernoulli **19**(5B), 2013, 2627–2651
DOI: 10.3150/12-BEJ469

Asymptotics of prediction in functional linear regression with functional outputs

CHRISTOPHE CRAMBES and ANDRÉ MAS[*]

Institut de Mathématiques et de Modélisation de Montpellier, UMR CNRS 5149, Equipe de Probabilités et Statistique, Université Montpellier II, CC 051, Place Eugène Bataillon, 34095 Montpellier Cedex 5, France. E-mail: [*]*mas@math.univ-montp2.fr*

We study prediction in the functional linear model with functional outputs, $Y = SX + \varepsilon$, where the covariates X and Y belong to some functional space and S is a linear operator. We provide the asymptotic mean square prediction error for a random input with exact constants for our estimator which is based on the functional PCA of X. As a consequence we derive the optimal choice of the dimension k_n of the projection space. The rates we obtain are optimal in minimax sense and generalize those found when the output is real. Our main results hold for class of inputs $X(\cdot)$ that may be either very irregular or very smooth. We also prove a central limit theorem for the predictor. We show that, due to the underlying inverse problem, the bare estimate cannot converge in distribution for the norm of the function space.

Keywords: functional data; functional output; linear regression model; optimality; prediction mean square error; weak convergence

1. Introduction

1.1. The model

Functional data analysis has become these last years an important field in statistical research, showing a lot of possibilities of applications in many domains (climatology, teledetection, linguistics, economics, ...). When one is interested on a phenomenon continuously indexed by time, for instance, it seems appropriate to consider this phenomenon as a whole curve. Practical aspects also go in this direction, since actual technologies allow to collect data on thin discretized grids. The paper by Ramsay and Dalzell [18] began to pave the way in favour of this idea of taking into account the functional nature of these data, and highlighted the drawbacks of considering a multivariate point of view. A major reference in this domain is the monograph by Ramsay and Silverman [19] which gives an overview about the philosophy and the basic models involving functional data. Important nonparametric issues are treated in the monograph by Ferraty and Vieu [13].

A particular problem in statistics is to predict the value of an interest variable Y knowing a covariate X. An underlying model can then write:

$$Y = r(X) + \varepsilon,$$

where r is an operator representing the link between the variables X and Y and ε is a noise random variable. In our functional data context, we want to consider that both variables X and

1350-7265 © 2013 ISI/BS

Y are of functional nature, that is, are random functions taking values on an interval $I = [a, b]$ of $\mathbb{R}$. We assume that X and Y take values in the space $L^2(I)$ of square integrable on I. In the following and in order to simplify, we assume that $I = [0, 1]$, which is not restrictive since the simple transformation $x \longmapsto (x - a)/(b - a)$ allows to come back to that case.

We assume as well that X and Y are centered. The issue of estimating the means $\mathbb{E}(X)$ and $\mathbb{E}(Y)$ in order to center the data was exhaustively treated in the literature and is of minor interest in our setting. The objective of this paper is to consider the model with functional input and output:

$$Y(t) = \int_0^1 \mathcal{S}(s, t)X(s)\,\mathrm{d}s + \varepsilon(t), \qquad \mathbb{E}(\varepsilon|X) = 0, \qquad \mathbb{E}(\varepsilon \otimes \varepsilon|X) = \mathbb{E}(\varepsilon \otimes \varepsilon) = \Gamma_\varepsilon, \quad (1)$$

where $\mathcal{S}(\cdot, \cdot)$ is an integrable kernel: $\iint |\mathcal{S}(s, t)|\,\mathrm{d}s\,\mathrm{d}t < +\infty$ and Γ_ε is the covariance operator of ε (the homoskedasticity of the errors is assumed). The kernel $\mathcal{S}$ may be represented on a $3D$-plot by a surface. The functional historical model (Malfait and Ramsay [17]) is

$$Y(t) = \int_0^t \mathcal{S}_{\text{hist}}(s, t)X(s)\,\mathrm{d}s + \varepsilon(t)$$

and may be recovered from the first model be setting $\mathcal{S}(s, t) = \mathcal{S}_{\text{hist}}(s, t)\mathbb{1}_{\{s \le t\}}$, the surface defining $\mathcal{S}$ being null when (s, t) is located in the triangle above the first diagonal of the unit square.

Model (1) may be viewed as a random Fredholm equation where both the input an the output are random (or noisy). This model has already been the subject of some studies, as, for instance, Chiou, Müller and Wang [8] or Yao, Müller and Wang [23], which propose an estimation of the functional parameter $\mathcal{S}$ using functional PCAs of the curves X and Y. One of the first studies about this model is due to Cuevas, Febrero and Fraiman [10] which considered the case of a fixed design. In this somewhat different context, they study an estimation of the functional coefficient of the model and give consistency results for this estimator. Recently, Antoch *et al.* [2] proposed a spline estimator of the functional coefficient in the functional linear model with a functional response, while Aguilera, Ocaña and Valderrama [1] proposed a wavelet estimation of this coefficient.

We start with a sample $(Y_i, X_i)_{1\le i\le n}$ with the same law as (Y, X), and we consider a new observation X_{n+1}. In all the paper, our goal will be to predict the value of the conditional expectation evaluated at a new random input $\mathbb{E}[Y|X_{n+1}]$.

The model (1) may be revisited if one acknowledges that $\int_0^1 \mathcal{S}(s, t)X(s)\,\mathrm{d}s$ is the image of X through a general linear integral operator. Denoting S the operator defined on and with values in $L^2([0, 1])$ by $(Sf)(t) = \int_0^1 \mathcal{S}(s, t)f(s)\,\mathrm{d}s$ we obtain from (1) that $Y(t) = S(X)(t) + \varepsilon(t)$ or

$$Y = SX + \varepsilon \qquad \text{where } S(X)(t) = \int \mathcal{S}(s, t)X(s)\,\mathrm{d}s.$$

This fact motivates a more general framework: it may be interesting to consider Sobolev spaces $W^{m,p}$ instead of $L^2([0, 1])$ in order to allow some intrinsic smoothness for the data. It turns out that, amongst this class of spaces, we choose to work with Hilbert spaces. Indeed the unknown parameter is a linear operator and spectral theory of these operators acting on Hilbert space allows enough generality, intuitive approaches and easier practical implementation. That is why in all the sequel we consider a sample $(Y_i, X_i)_{1\le i\le n}$ where Y and X are independent, identically distributed and take values in the same Hilbert space H endowed with inner product $\langle\cdot,\cdot\rangle$ and associated norm $\|\cdot\|$.

Obviously the model we consider generalizes the regression model with a real output y:

$$y = \int_0^1 \beta(s)X(s)\,\mathrm{d}s + \varepsilon = \langle\beta, X\rangle + \varepsilon, \tag{2}$$

and all our results hold in this direction. The literature is wide about (2) but we picked articles which are close to our present concerns and will be cited again later in this work: Yao, Müller and Wang [23], Hall and Horowitz [14], Crambes, Kneip and Sarda [9].

Since the unknown parameter is here an operator, the infinite-dimensional equivalence of a matrix, it is worth giving some basic information about operator theory on Hilbert spaces. The interested reader can find basics and complements about this topic in the monograph Dunford and Schwartz [11]. We denote by $\mathcal{L}$ the space of bounded – hence continuous – operators on a Hilbert space H. For our statistical or probabilistic purposes, we restrain this space to the space of compact operators $\mathcal{L}_c$. Then, any compact and symmetric operator T belonging to $\mathcal{L}_c$ admits a unique Schmidt decomposition of the form $T = \sum_{j\in\mathbb{N}}\mu_j\phi_j\otimes\phi_j$ where the (μ_j, ϕ_j)'s are called the eigenelements of T, and the tensor product notation $\otimes$ is defined in the following way: for any function f, g and h belonging to H, we define $f\otimes g = \langle g,\cdot\rangle f$. Finally, we mention two subclasses of $\mathcal{L}_c$ one of which will be our parameter space. The space of Hilbert–Schmidt operators and trace class operators are defined, respectively, by $\mathcal{L}_2 = \{T\in\mathcal{L}_c : \sum_{j\in\mathbb{N}}\mu_j^2 < +\infty\}$ and $\mathcal{L}_1 = \{T\in\mathcal{L}_c : \sum_{j\in\mathbb{N}}\mu_j < +\infty\}$. It is well known that if S is the linear operator associated to the kernel $\mathcal{S}$ like in line (1) then if $\iint|\mathcal{S}(s,t)|\,\mathrm{d}s\,\mathrm{d}t < +\infty$, S is Hilbert–Schmidt and S is trace class if $\mathcal{S}(s,t)$ is continuous as a function of (s,t). The Hilbert–Schmidt norm is denoted $\|\cdot\|_2$.

1.2. Estimation

Our purpose here is first to introduce the estimator. This estimate looks basically like the one studied in Yao, Müller and Wang [23]. Our second goal is to justify from a more theoretical position the choice of such a candidate.

Two strategies may be carried out to propose an estimate of S. They join finally, like in the finite-dimensional framework. One could consider the theoretical mean square program (convex in S) that minimize $\mathbb{E}\|Y - SX\|^2$ over $S\in\mathcal{L}_2$, whose solution S_* is defined by the equation $\mathbb{E}[Y\otimes X] = S_*\mathbb{E}[X\otimes X]$. On the other hand it is plain that the moment equation:

$$\mathbb{E}[Y\otimes X] = \mathbb{E}\big[S(X)\otimes X\big] + \mathbb{E}[\varepsilon\otimes X]$$

leads to the same solution. Finally denoting $\Delta = \mathbb{E}[Y\otimes X]$, $\Gamma = \mathbb{E}[X\otimes X]$ we get $\Delta = S\Gamma$. Turning to empirical counterparts with

$$\Delta_n = \frac{1}{n}\sum_{i=1}^n Y_i\otimes X_i, \qquad \Gamma_n = \frac{1}{n}\sum_{i=1}^n X_i\otimes X_i,$$

the estimate $\widehat{S}_n$ of S should naturally be defined by $\Delta_n = \widehat{S}_n \Gamma_n$. Once again the moment method and the minimization of the mean square program coincide. By the way, note that $\Delta_n = S\Gamma_n + U_n$ with $U_n = \frac{1}{n}\sum_{i=1}^{n} \varepsilon_i \otimes X_i$. The trouble is that, from $\Delta_n = S_n \Gamma_n$ we cannot directly derive an explicit form for S_n. Indeed Γ_n is not invertible on the whole H since it has finite rank. The next section proposes solutions to solve this inverse problem by classical methods.

As a last point, we note that if $\widehat{S}_n$ is an estimate of S, a statistical predictor given a new input X_{n+1} is:

$$\widehat{Y}_{n+1}(t) = \widehat{S}_n(X_{n+1})(t) = \int \widehat{\mathcal{S}}(s,t) X_{n+1}(s)\,\mathrm{d}s. \tag{3}$$

1.3. Identifiability, inverse problem and regularization issues

We turn again to the equation which defines the operator S: $\Delta = S\Gamma$. Taking a one-to-one Γ is a first and basic requirement for identifiability. It is simple to check that if $v \in \ker\Gamma \neq \{0\}$, $\Delta = S\Gamma = (S + v \otimes v)\Gamma$, for instance, and the unicity of S is no more ensured. More precisely, the inference based on the equation $\Delta = S\Gamma$ does not ensure the identifiability of the model. From now on, we assume that $\ker\Gamma = \{0\}$. At this point, some more theoretical concerns should be mentioned. Indeed, writing $S = \Delta\Gamma^{-1}$ is untrue. The operator Γ^{-1} exists whenever $\ker\Gamma = \{0\}$ but is unbounded, that is, not continuous. We refer once again to Dunford and Schwartz [11], for instance, for developments on unbounded operators. It turns out that Γ^{-1} is a linear mapping defined on a dense domain $\mathcal{D}$ of H which is measurable but continuous at no point of its domain. Let us denote (λ_j, e_j) the eigenelements of Γ. Elementary facts of functional analysis show that $S_{|\mathcal{D}} = \Delta\Gamma^{-1}$ where $\mathcal{D}$ is the domain of Γ^{-1}, that is, the range of Γ and is defined by $\mathcal{D} = \{x = \sum_j x_j e_j \in H : \sum_j \frac{x_j^2}{\lambda_j^2} < +\infty\}$.

An illustrative example is the Gaussian case. If Γ is the covariance operator of a Gaussian random element X on H (a process, a random function, etc.), then the Reproducing Kernel Hilbert Space of X coincides with the domain of $\Gamma^{-1/2}$ and the range of $\Gamma^{1/2}$: $\{x \in H : \sum_j x_j^2/\lambda_j < +\infty\}$.

The last stumbling stone comes from switching population parameters to empirical ones. We construct our estimate from the equation $\Delta_n = S\Gamma_n + U_n$ as seen above and setting $\Delta_n = \widehat{S}_n\Gamma_n$. Here the inverse of Γ_n does not even exist since this covariance operator is finite-rank. If Γ_n was invertible, we could set $S_n = \Delta_n\Gamma_n^{-1}$ but we have to regularize Γ_n first. We carry out techniques which are classical in inverse problems theory. Indeed, the spectral decomposition of Γ_n is $\Gamma_n = \sum_j \widehat{\lambda}_j(\widehat{e}_j \otimes \widehat{e}_j)$ where $(\widehat{\lambda}_j, \widehat{e}_j)$ are the empirical eigenelements of Γ_n (the $\widehat{\lambda}_j$'s are sorted in a decreasing order and some of them may be null) derived from the functional PCA. The spectral cut regularized inverse is given for some integer k by

$$\Gamma_n^{\dagger} = \sum_{j=1}^{k} \widehat{\lambda}_j^{-1}(\widehat{e}_j \otimes \widehat{e}_j). \tag{4}$$

The choice of $k = k_n$ is crucial; all the $\widehat{\lambda}_j$'s cannot be null and one should stress that $\widehat{\lambda}_j^{-1}$ tends to infinity when j increases. The reader will note that we could define equivalently

$\Gamma^{\dagger} = \sum_{j=1}^{k} \lambda_j^{-1}(e_j \otimes e_j)$. From the definition of the regularized inverse above, we can derive a useful equation. Indeed, let $\widehat{\Pi}_k$ denote the projection of the k first eigenvectors of Γ_n, that is the projection on span$(\widehat{e}_1, \ldots, \widehat{e}_k)$. Then $\Gamma_n^{\dagger}\Gamma_n = \Gamma_n\Gamma_n^{\dagger} = \widehat{\Pi}_k$. For further purpose, we define as well Π_k to be the projection operator on (the space spanned by) the k first eigenvectors of Γ.

Other regularizations are possible by replacing $\widehat{\lambda}_j^{-1}$ in (4) by a smooth function of it, which converges to $\widehat{\lambda}_j^{-1}$. See Section 3 of Cardot, Mas and Sarda [7] for more details and the books by Tikhonov and Arsenin [22] and Engl, Hanke and Neubauer [12] on the general topic of inverse problems.

1.4. Assumptions

The assumptions we need are classically of three types: regularity of the regression parameter S, moment assumptions on X and regularity assumptions on X which are often expressed in terms of spectral properties of Γ (especially the rate of decrease to zero of its eigenvalues).

Assumption on S. We assume that S is Hilbert–Schmidt which may be rewritten: for any basis $(\phi_j)_{j\in\mathbb{N}}$ of H

$$\sum_{j,\ell} \langle S(\phi_\ell), \phi_j\rangle^2 < +\infty. \tag{5}$$

This assumption finally echoes assumption $\sum_j \beta_j^2 < +\infty$ in the functional linear model (2) with real outputs. We already underlined that (5) is equivalent to assuming that $\mathcal{S}$ is doubly integrable if H is $L^2([0, T])$. Finally, no continuity or smoothness is required for the kernel $\mathcal{S}$ at this point.

Moment assumptions on X. In order to better understand the moment assumptions on X, we recall the Karhunen–Loeve development, which is nothing but the decomposition of X in the basis of the eigenvectors of Γ, $X = \sum_{j=1}^{+\infty} \sqrt{\lambda_j}\xi_j e_j$ a.s. where the ξ_j's are independent centered real random variables with unit variance. We need higher moment assumptions because we need to apply Bernstein's exponential inequality to functionals of $\Gamma - \Gamma_n$. We assume that for all $j, \ell \in \mathbb{N}$ there exists a constant b such that

$$\mathbb{E}(|\xi_j|^\ell) \le \frac{\ell!}{2} b^{\ell-2} \cdot \mathbb{E}(|\xi_j|^2), \tag{6}$$

which echoes the assumption (2.19), page 49, in Bosq [4]. As a consequence, we see that

$$\mathbb{E}\langle X, e_j\rangle^4 \le C\big(\mathbb{E}\langle X, e_j\rangle^2\big)^2. \tag{7}$$

This requirement already appears in several papers. It assesses that the sequence of the fourth moment of the margins of X tends to 0 quickly enough. The assumptions above always hold for a Gaussian X. These assumptions are close to the moment assumptions usually required when rates of convergence are addressed.

Assumptions on the spectrum of Γ. The covariance operator Γ is assumed to be injective hence with *strictly* positive eigenvalues arranged in a decreasing order. Let the function $\lambda : \mathbb{R}^+ \to \mathbb{R}^{+*}$

be defined by $\lambda(j) = \lambda_j$ for any $j \in \mathbb{N}$ (the λ_j's are continuously interpolated between j and $j+1$. From the assumption above, we already know that $\sum_j \lambda_j < +\infty$. Indeed the summability of the eigenvalues of Γ is ensured whenever $\mathbb{E}\|X\|^2 < +\infty$. Besides, assume that for x large enough

$$x \to \lambda(x) \text{ is convex.} \tag{8}$$

These last conditions are mild and match a very large class of eigenvalues: with arithmetic decay $\lambda_j = Cj^{-1-\alpha}$ where $\alpha > 0$ (like in Hall and Horowitz [14]), with exponential decay $\lambda_j = Cj^{-\beta}\exp(-\alpha j)$, Laurent series $\lambda_j = Cj^{-1-\alpha}(\log j)^{-\beta}$ or even $\lambda_j = Cj^{-1}(\log j)^{-1-\alpha}$. Such a rate of decay occurs for extremely irregular processes, even more irregular than the Brownian motion for which $\lambda_j = Cj^{-2}$. In fact, our framework initially relaxes prior assumptions on the rate of decay of the eigenvalues, hence on the regularity of X. These eigenvalues play the role of Fourier coefficients in Fourier analysis (and are indeed these coefficients when the eigenbasis is cosine). A regular or smoothly reconstructed X will feature rapidly decaying λ's. It will be seen later that exact risk and optimality are obtained when considering specific classes of eigenvalues. Assumption (8) is crucial however since the most general lemmas rely on convex inequalities for the eigenvalues.

2. Asymptotic results

We are now in a position to introduce our estimate.

Definition 1. *The estimate $\widehat{S}_n$ of S is defined by: $\widehat{S}_n = \Delta_n \Gamma_n^{\dagger}$, the associated predictor is $\widehat{Y}_{n+1} = \widehat{S}_n(X_{n+1}) = \Delta_n \Gamma_n^{\dagger}(X_{n+1})$. It is possible to provide a kernel form. We deduce from $S_n = \Delta_n \Gamma_n^{\dagger}$ that*

$$\mathcal{S}_n(s,t) = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{k}\frac{\int X_i \widehat{e}_j}{\widehat{\lambda}_j}\cdot Y_i(t)\widehat{e}_j(s).$$

Though distinct, this estimate remains close from the one proposed in Yao, Müller and Wang [23], the difference consisting in the fact that we do not consider a Karhunen–Loeve development of Y. In the sequel, our main results are usually given in term of $\widehat{S}_n$ but we frequently switch to the 'kernel' viewpoint since it may be sometimes more illustrative. Then we implicitly assume that $H = L^2([0, 1])$.

2.1. Mean square prediction error and optimality

From now on, all our results are stated when assumptions of the Section 1.4 hold. We start with an upper bound from which we deduce, as a corollary, the exact asymptotic risk of the predictor. What is considered here is the prediction of the regression function $\mathbb{E}(Y_{n+1}|X_{n+1}) = S(X_{n+1})$ based on the estimate $\widehat{S}_n$. Evaluating $\widehat{S}_n$ at a random input with the same distribution as X averages in a way the predictor and unfortunately prevents from considering specific design points. But conversely prediction at a fixed x will converge at a rate depending on x. If x is outside the support of X one may be mistaken on the validity and the sharpness of the result. This approach may gain some kind of robustness and lose some kind of pliancy. In the following, we denote $\sigma_\varepsilon^2 = \operatorname{tr} \Gamma_\varepsilon$, which we assume to be finite.

Theorem 2. *The mean square prediction error of our estimate has the following exact asymptotic development*:

$$\mathbb{E}\left\| \widehat{S}_n(X_{n+1}) - S(X_{n+1}) \right\|^2 = \sigma_\varepsilon^2 \frac{k}{n} + \sum_{j=k+1}^{+\infty} \lambda_j \left\| S(e_j) \right\|^2 + A_n + B_n, \tag{9}$$

where $A_n \le C_A \|S\|_2 k^2 \lambda_k / n$ and $B_n \le C_B k^2 \log^2 k / n^2$ where C_A and C_B are constants which do not depend on k, n or S.

The two first terms determine the convergence rate: the variance effect appears through $\sigma_\varepsilon^2 k/n$ and the bias (related to smoothness) through $\sum_{j=k+1}^{+\infty} \lambda_j \|S(e_j)\|^2$. Several comments are needed at this point. The term A_n comes from bias decomposition and B_n is a residue from variance. Both are negligible with respect to the first two terms. Indeed, $k\lambda_k \to 0$ since $\sum_k \lambda_k < +\infty$ and $A_n = \mathrm{o}(k/n)$. Turning to B_n is a little bit more tricky. It can be seen that necessarily $(k \log k)^2/n \to 0$ which ensures that $B_n = \mathrm{o}(k/n)$. A second interesting property arises from Theorem 2. Rewriting $\lambda_j \|S(e_j)\|^2 = \|S\Gamma^{1/2}(e_j)\|^2$ we see that the only regularity assumptions needed may be made from the spectral decomposition of the operator $S\Gamma^{1/2}$ itself and not from X (or Γ as well) and S separately.

Before turning to optimality, we introduce the class of parameters S over which optimality will be obtained.

Definition 3. *Let $\varphi : \mathbb{R}^+ \to \mathbb{R}^+$ be a C^1 decreasing function such that $\sum_{j=1}^{+\infty} \varphi(j) = 1$ and set $\mathcal{L}_2(\varphi, L)$ be the class of linear operator from H to H be defined by*

$$\mathcal{L}_2(\varphi, L) = \left\{ T \in \mathcal{L}_2, \|T\|_2 \le L : \left\| T(e_j) \right\| \le L\sqrt{\varphi(j)} \right\}.$$

The set $\mathcal{L}_2(\varphi, L)$ is entirely determined by the bounding constant L and the function φ. Hall and Horowitz [14] consider the case when $\varphi(j) = Cj^{-(\alpha+2\beta)}$ where $\alpha > 1$ and $\beta > 1/2$. As mentioned earlier, we are free here to take any φ such that $\int^{+\infty} \varphi(s)\,\mathrm{d}s < +\infty$ and which leaves assumption (8) unchanged.

As an easy consequence, we derive the uniform bound with exact constants below.

Theorem 4. *Set $L = \|S\Gamma^{1/2}\|_2$, $\varphi(j) = \lambda_j \|S(e_j)\|^2 / L^2$ and k_n^* as the integer part of the unique solution of the integral equation* (*in x*):

$$\frac{1}{x} \int_x^{+\infty} \varphi(x)\,\mathrm{d}x = \frac{1}{n} \frac{\sigma_\varepsilon^2}{L^2}. \tag{10}$$

Let $\mathcal{R}_n(\varphi, L)$ be the uniform prediction risk of the estimate $\widehat{S}_n$ over the class $\mathcal{L}_2(\varphi, L)$:

$$\mathcal{R}_n(\varphi, L) = \sup_{S\Gamma^{1/2}\in\mathcal{L}_2(\varphi,L)} \mathbb{E}\big\|\widehat{S}_n(X_{n+1}) - S(X_{n+1})\big\|^2,$$

then

$$\lim\sup_{n\to+\infty} \frac{n}{k_n^*}\mathcal{R}_n(\varphi, L) = 2\sigma_\varepsilon^2.$$

Equation (10) has a unique solution because the function of x on the left-hand side is strictly decreasing. The integer k_n^* is the optimal dimension: the parameter which minimizes the prediction risk. It plays the same role as the optimal bandwidth in nonparametric regression. The upper bound in the line above is obvious from (9). This upper bound is attained when taking for S the diagonal operator defined in the basis of eigenvectors by $Se_j = L\varphi^{1/2}(j)\lambda_j^{-1/2}e_j$. The proof of this theorem is an easy consequence of Theorem 2, hence omitted.

The next corollary is an attempt to illustrate the consequences of the previous theorem by taking explicit sequences $(\varphi(j))_{j\in\mathbb{N}}$. We chose to treat the case of general Laurent series (including very irregular input and parameter when $\alpha = 0$) and the case of exponential decay.

Corollary 5. *Set $\varphi_a(j) = C_{\alpha,\beta}(j^{2+\alpha}(\log j)^\beta)^{-1}$ and $\varphi_b(j) = C'_\alpha \exp(-\alpha j)$ where either $\alpha > 0$ and $\beta \in \mathbb{R}$ or $\alpha = 0$ and $\beta > 1$, $C_{\alpha,\beta}$ and C'_α are normalizing constants, then*

$$\mathcal{R}_n(\varphi_a, L) \sim \frac{(\log n)^{\beta/(2+\alpha)}}{n^{(1+\alpha)/(2+\alpha)}}\left(\frac{C_{\alpha,\beta}L^2}{2\sigma_\varepsilon^2}\right)^{1/(2+\alpha)}, \qquad \mathcal{R}_n(\varphi_b, L) \le \frac{\log n}{\alpha n}.$$

For $\mathcal{R}_n(\varphi_b, L)$ we could not compute an exact bound because equation (10) has no explicit solution. But the term $(\log n)/\alpha n$ is obviously sharp since parametric up to $\log n$. The special case $\beta = 0$ and $\alpha > 1$ matches the optimal rate derived in Hall and Horowitz [14] with a slight damage due to the fact that the model shows more complexity ($\mathcal{S}$ is a function of two variables whereas β the slope parameter in the latter article and in model (2) was a function of a single variable). We also refer the reader to Stone [21] who underlines this effect of dimension on the convergence rates in order to check that our result matches the ones announced by Stone.

In our setting, the data Y are infinite dimensional. Obtaining lower bound for optimality in minimax version is slightly different than in the case studied in Hall and Horowitz [14], Crambes, Kneip and Sarda [9]. In order to get a lower bound, our method is close to the one carried out by Cardot and Johannes [6], based on a variant of Assouad's lemma. We consider Gaussian observations under 2^{k_n} distinct models.

Theorem 6. *The following bound on the minimax asymptotic risk up to constants proves that our estimator is optimal in minimax sense*:

$$\inf_{\widehat{S}_n} \sup_{S\in\mathcal{L}_2(\varphi,L)} \mathbb{E}\big\|\widehat{S}_n(X_{n+1}) - S(X_{n+1})\big\|^2 \asymp \frac{k_n^*}{n}.$$

Remark 7. The bound above holds with highly irregular data (e.g., $\lambda_j \asymp Cj^{-1}(\log j)^{-1-\alpha}$ with $\alpha > 0$) or with very regular data featuring a flat spectrum with $\lambda_j \asymp Cj^{-\gamma}\exp(-\alpha j)$ or even the intermediate situation like $\lambda_j \asymp Cj^{-1-\beta}(\log j)^{1+\alpha}$. The same remarks are valid when turning to the regularity of the kernel $\mathcal{S}$ or of the operator S expressed through the sequence $\|S(e_j)\|^2$. Our method of proof shows that smooth, regular processes (with rapid decay of λ_j) have good approximation properties but ill-conditioned $\Gamma_n^\dagger$ (i.e., with rapidly increasing norm) damaging the rate of convergence of $\widehat{S}_n$ which depends on it. But we readily see that irregular processes (with slowly decreasing λ_j), despite their poor approximation properties, lead to a slowly increasing $\Gamma_n^\dagger$ and to solving an easier inverse problem.

2.2. Weak convergence

The next and last result deals with weak convergence. We start with a negative result which shows that due to the underlying inverse problem, the issue of weak convergence cannot be addressed under too strong topologies.

Theorem 8. *It is impossible for S_n to converge in distribution for the Hilbert–Schmidt norm.*

Once again turning to the predictor, hence smoothing the estimated operator, will produce a positive result. We improve twofold the results by Cardot, Mas and Sarda [7] since, first, the model is more general and, second, we remove the bias term. Weak convergence (convergence in distribution) is denoted $\overset{w}{\to}$. The reader should pay attention to the fact that the following theorem holds in space of functions (here H). Within this theorem, two results are proved. The first assesses weak convergence for the predictor with a bias term. The second removes this bias at the expense of a more specific assumption on the sequence k_n.

Theorem 9. *If the condition* $(k\log k)^2/n \to 0$ *holds, then*

$$\sqrt{\frac{n}{k}}\big[\widehat{S}_n(X_{n+1}) - S\Pi_k(X_{n+1})\big] \overset{w}{\to} \mathcal{G}_\varepsilon,$$

where $\mathcal{G}_\varepsilon$ is a centered Gaussian random element with values in H and covariance operator Γ_ε. Besides, denoting $\gamma_k = \sup_{j\geq k}\{j\log j\|S(e_j)\|\sqrt{\lambda_j}\}$ (it is plain that $\gamma_k \to 0$) and choosing k such that $n \leq (k\log k)^2/\gamma_k$ (which means that $(k\log k)^2/n$ should not decay too quickly to zero), the bias term can be removed and we obtain

$$\sqrt{\frac{n}{k}}\big[\widehat{S}_n(X_{n+1}) - S(X_{n+1})\big] \overset{w}{\to} \mathcal{G}_\varepsilon.$$

From Theorem 9, we deduce general prediction intervals for the predictor: let $\mathcal{K}$ be a continuous set for the measure induced by $\mathcal{G}_\varepsilon$, that is, $\mathbb{P}(\mathcal{G}_\varepsilon \in \partial\mathcal{K}) = 0$ where $\partial\mathcal{K} = \overline{\mathcal{K}} \setminus \operatorname{int}(\mathcal{K})$ is the frontier of $\mathcal{K}$ then $\mathbb{P}(\widehat{S}_n(X_{n+1}) \in S(X_{n+1}) + \sqrt{\frac{k}{n}}\mathcal{K}) \to \mathbb{P}(\mathcal{G}_\varepsilon \in \mathcal{K})$ when $n \to +\infty$. As an application, we propose the two following corollaries of Theorem 9. The notation Y_{n+1}^* stands for $S(X_{n+1}) = \mathbb{E}(Y_{n+1}|X_{n+1})$. The first corollary deals with asymptotic prediction intervals for general functionals of the theoretical predictor such as weighted integrals.

Corollary 10. *Let m be a fixed function in the space $H = L^2([0,1])$. We have the following asymptotic confidence interval for $\int Y_{n+1}^*(t)m(t)\,\mathrm{d}t$ at level $1-\alpha$:*

$$\mathbb{P}\left(\int_0^1 Y_{n+1}^*(t)m(t)\,\mathrm{d}t \in \left[\int_0^1 \widehat{Y}_{n+1}(t)m(t)\,\mathrm{d}t \pm \sqrt{\frac{k}{n}}\sigma_m q_{1-\alpha/2}\right]\right) = 1-\alpha,$$

where $\sigma_m^2 = \langle m, \Gamma_\varepsilon m\rangle = \iint \Gamma_\varepsilon(s,t)m(t)m(s)\,\mathrm{d}t\,\mathrm{d}s$ rewritten in 'kernel' form and $q_{1-\alpha/2}$ is the quantile of order $1-\alpha/2$ of the $\mathcal{N}(0,1)$ distribution.

Theorem 9 holds for the Hilbert norm. In order to derive a prediction interval for $Y_{n+1}^*(t_0)$ (where t_0 is fixed in $[0,1]$), we have to make sure that the evaluation (linear) functional $f \in H \longmapsto f(t_0)$ is continuous for the norm $\|\cdot\|$. This functional is always continuous in the space $(C([0,1]), |\cdot|_\infty)$ but is not in the space $L^2([0,1])$. A slight change in H will yield the desired result, stated in the next corollary.

Corollary 11. *When $H = W_0^{2,1}([0,1]) = \{f \in L^2([0,1]) : f(0) = 0, f' \in L^2([0,1])\}$ endowed with the inner product $\langle u, v\rangle = \int_0^1 u'v'$, the evaluation functional is continuous with respect to the norm of H and we can derive from Theorem 9:*

$$\mathbb{P}\left(Y_{n+1}^*(t_0) \in \left[\widehat{Y}_{n+1}(t_0) \pm \sqrt{\frac{k}{n}}\sigma_{t_0} q_{1-\alpha/2}\right]\right) = 1-\alpha,$$

where $\sigma_{t_0}^2 = \Gamma_\varepsilon(t_0, t_0)$.

Note that data $(Y_i)_{1\le i\le n}$ reconstructed by cubic splines and correctly rescaled to match the condition $[f(0) = 0]$ belong to the space $W_0^{2,1}([0,1])$ mentioned in the corollary.

2.3. Comparison with existing results – Conclusion

The literature on linear models for functional data gave birth to impressive and brilliant recent works. We discuss briefly here our contribution with respect to some articles, close in spirit to this present paper.

We consider exactly the same model (with functional outputs) as Yao, Müller and Wang [23] and our estimate is particularly close to the one they propose. In their work, the case of longitudinal data was studied with care with possibly sparse and irregular data. They introduce a very interesting functional version of the R^2 and prove convergence in probability of their estimates in Hilbert–Schmidt norm. We complete their work by providing the rates and optimality for convergence in mean square.

Our initial philosophy is close to the article by Crambes, Kneip and Sarda [9]. Like these authors we consider the prediction with random design. We think that this way seems to be justified from a statistical point of view. The case of a fixed design gives birth to several situations and different rates (with possible oversmoothing which entails parametric rates of convergence which are odd in this truly nonparametric model) and does not necessarily correspond to the statistical reality. The main differences rely in the fact that our results hold in mean square norm rather than in probability for a larger class of data and parameter at the expense of more restricted moment assumptions.

Our methodology is closer to the articles by Hall and Horowitz [14]. They studied the prediction risk at a fixed design in the model with real outputs (2) but with specified eigenvalues namely $\lambda_j \sim Cj^{-1-\alpha}$ and parameter spectral decomposition $\langle \beta, e_j \rangle \sim Cj^{-1-\gamma}$ with $\alpha, \gamma > 0$. The comparisons may be simpler with these works since we share the approach through spectral decomposition of operators or Karhunen–Loeve development for the design X.

The problem of weak convergence is considered only in Yao, Müller and Wang [23]: they provide very useful and practical pointwise confidence sets which imply estimation of the covariance of the noise. Our result may allow to consider a larger class of testing issues through delta-methods (we have in mind testing of hypotheses like $S = S_0$ versus $S_{(n)} = S_0 + \eta_n v$ where $\eta_n \to 0$ and v belongs to a well-chosen set in H).

The contribution of this article essentially deals with a linear regression model – the concerns related to the functional outputs concentrate on lower bounds in optimality results and in proving weak convergence with specific techniques adapted to functional data. We hope that our methods will demonstrate that optimal results are possible in a general framework and that regularity assumptions can often be relaxed thanks to the compensation (or regularity/inverse problem trade-off) phenomenon mentioned within Remark 7. The Hilbert space framework is necessary at least in the section devoted to weak convergence. Generalizations to Banach spaces of functions could be investigated, for instance, in $C([0,1])$, Hölder or Besov spaces.

3. Practical implementation

In order to illustrate our theoretical results, we made a short simulation study. In practice, it is worth noticing that all the curves are observed on a grid of points of the interval $[0,1]$, while our theoretical results do not take the discretization into account. From now on, we consider that we observe the X_i's at the values $s_1 < \cdots < s_p$ of $[0,1]$ and the Y_i's at the values $t_1 < \cdots < t_q$. Moreover, we implicitly assume that the X_i's are observed without error. There are some possible approaches to deal with covariates contaminated with measurement errors (see, e.g., in Cardot *et al.*, [5]), but it is not in the scope of this paper, and would damage the clarity of the work. The number of discretization points in our simulations are $p = 100$ for the X_i's and $q = 100$ for the Y_i's. The objective of this simulation is simply to see how our theoretical results behave in practice, especially the smoothing parameter k (number of principal components).

We simulate the X_i's using the Karhunen–Loeve decomposition, with $k_{\text{real}} = 8$ principal components. In other words, we write $X_i(t) = \sum_{j=1}^{k_{\text{real}}} \lambda_j \xi_{ij} e_j(t)$, where $\lambda_j = \frac{1}{\pi^2 (j-0.5)^2}$, ξ_{ij} are standard Gaussian random variables and $e_j(t) = \sqrt{2}\sin((j-0.5)\pi t)$. The functional parameter is the function $\mathcal{S}(s,t)=s^2+t^2$. The noise of the model is simulated as a Brownian motion with $\sigma_\varepsilon^2=0.127$.

3.1. Estimation of the variance of the noise

In order to see the behavior of our theoretical results, especially the result of Theorem 9, we first have to estimate the variance of the noise σ_ε^2. This problem is not trivial. Indeed, given a new curve X_{n+1}, an estimator of ε_{n+1} is $\widehat{\varepsilon}_{n+1}=Y_{n+1}-\widehat{Y}_{n+1}$. However, we can see that

$$\begin{aligned}\widehat{\varepsilon}_{n+1}\otimes\widehat{\varepsilon}_{n+1}&=\big(\widehat{S}(X_{n+1})-S(X_{n+1})\big)\otimes\big(\widehat{S}(X_{n+1})-S(X_{n+1})\big)\\&\quad-2\varepsilon_{n+1}\otimes\big(\widehat{S}(X_{n+1})-S(X_{n+1})\big)\\&\quad+\varepsilon_{n+1}\otimes\varepsilon_{n+1}.\end{aligned}$$

This proves that the naive estimator of σ_ε^2 given by $\widetilde{\sigma}_\varepsilon^2=\operatorname{tr}\Gamma_{\widehat{\varepsilon}}$ is biased: it overestimates the true value σ_ε^2. In order to remove bias, we adopt the following procedure. We split our sample into three parts. With the first part of the sample, we compute an estimator $\widehat{S}^{[1]}$ of S. With the second part of the sample, we compute another estimator $\widehat{S}^{[2]}$ of S and with the last part of the sample, we approximate the term $(\widehat{S}(X_{n+1})-S(X_{n+1}))\otimes(\widehat{S}(X_{n+1})-S(X_{n+1}))$, with $\frac{1}{2}(\widehat{S}^{[1]}(X_{n+1})-\widehat{S}^{[2]}(X_{n+1}))\otimes(\widehat{S}^{[1]}(X_{n+1})-\widehat{S}^{[2]}(X_{n+1}))$ which allows to build an unbiased estimator of σ_ε^2, given by $\widehat{\sigma}_\varepsilon^2=\operatorname{tr}\Gamma_{\widehat{\varepsilon}}-\frac{1}{2}\operatorname{tr}\Gamma_{\widehat{S}^{[1]}(X_{n+1})-\widehat{S}^{[2]}(X_{n+1})}$, where $\widehat{\varepsilon}$ is computed on the third part of the sample. This procedure has been tested on an initial sample with length $n=600$, divided into three samples with sizes 200 each. The results are synthesized in Table 1 with the means and the standard deviations of $\widehat{\sigma}_\varepsilon^2$ and $\widetilde{\sigma}_\varepsilon^2$, computed on $N=500$ repeated simulations. We can see that the unbiasing procedure works quite well on our simulations, since the objective value $\sigma_\varepsilon^2=0.127$ is correctly approximated by $\widehat{\sigma}_\varepsilon^2$. Simulations also give a confirmation that the estimator $\widetilde{\sigma}_\varepsilon^2$ is biaised.

3.2. Empirical illustration of a theoretical result

We analyse in this subsection how our theoretical result given in Theorem 9 behaves in practice. In order to do that, we compare an empirical estimation of the term $\mathbb{E}\|\widehat{S}_n(X_{n+1})-S(X_{n+1})\|^2$

Table 1. Means and standard deviations of $\widehat{\sigma}_\varepsilon^2$ and $\widetilde{\sigma}_\varepsilon^2$, computed on $N=500$ repeated simulations, for a sample size $n=600$

	$\widehat{\sigma}_\varepsilon^2$	$\widetilde{\sigma}_\varepsilon^2$
Mean	0.127133	0.131461
Standard deviation	0.012293	0.012171

Table 2. Means and standard deviations of k_{emp}, k_{th} and $k_{\text{th,est}}$ (optimal values of k), computed on $N = 500$ repeated simulations

	k_{emp}	k_{th}	$k_{\text{th,est}}$
Mean	7.725	7.655	7.691
Standard deviation	1.077	1.169	1.193

with the theoretical value $\text{MSE}_{\text{th}} := \sigma_\varepsilon^2 \frac{k}{n} + \sum_{j=k+1}^{+\infty} \lambda_j \|S(e_j)\|^2$, and with the estimated theoretical value $\text{MSE}_{\text{th,est}} := \widehat{\sigma}_\varepsilon^2 \frac{k}{n} + \sum_{j=k+1}^{+\infty} \widehat{\lambda}_j \|\widehat{S}(\widehat{e}_j)\|^2$. For that aim, we approximate the value of $\mathbb{E}\|\widehat{S}_n(X_{n+1}) - S(X_{n+1})\|^2$ with an empirical version of it, which consists in computing the mean of the error $\|\widehat{S}_n(X_{n+1}) - S(X_{n+1})\|^2$ on $M = 1000$ simulated curves X_{n+1}. We denote it MSE_{emp}. The second term, MSE_{th}, is computed using the theoretical values (unknown in real applications, but known in the simulation study). The third term, $\text{MSE}_{\text{th,est}}$, is computed using the unbiased estimation of σ_ε^2, presented in the previous subsection. These risks MSE_{emp}, MSE_{th} and $\text{MSE}_{\text{th,est}}$ have been computed for several values of k, on the basis of a sample with length $n = 600$, divided into three samples with sizes 200 each. The two first samples are used for the procedure to compute the unbiased estimator $\widehat{\sigma}_\varepsilon^2$. The third sample is used to compute the values of MSE_{emp}, MSE_{th} and $\text{MSE}_{\text{th,est}}$, as explained in Section 3.1.

We give on Table 2 the mean values and standard deviations for the optimal chosen values of k with respect to the risks MSE_{emp}, MSE_{th} and $\text{MSE}_{\text{th,est}}$. We denote these optimal values by k_{emp}, k_{th} and $k_{\text{th,est}}$. We can see that the optimal values chosen by the three risks are close. All these values remain close to the real value $k_{\text{real}} = 8$, even if they seem to slightly underestimate it. Hence, the theoretical criterion given in Theorem 9 can bring a possible solution in practice to select the number of eigenvalues of the covariance operator. We have collected on Table 3 the mean values of MSE_{emp}, MSE_{th} and $\text{MSE}_{\text{th,emp}}$ for several values of k. The values of MSE_{th} and $\text{MSE}_{\text{th,emp}}$ are decomposed into square bias and variance. We can see that the values are close. The three MSE criteria are convex, and we notice moreover that the square bias decreases and the variance increases as k increases, as expected.

4. Mathematical derivations

In the sequel, the generic notation C stands for a constant which does not depend on k, n or S. All our results are related to the decomposition given below:

$$\widehat{S}_n = S\Gamma_n\Gamma_n^\dagger + U_n\Gamma_n^\dagger = S\widehat{\Pi}_k + \frac{1}{n}\sum_{i=1}^{n} \varepsilon_i \otimes \Gamma_n^\dagger X_i. \tag{11}$$

It is plain that a bias–variance decomposition is exhibited just above. The random projection $\widehat{\Pi}_k$ is not a satisfactory term and we intend to remove it and to replace it with its non-random

Table 3. Means and standard deviations of $\mathrm{MSE}_{\mathrm{emp}}$, $\mathrm{MSE}_{\mathrm{th}}$ and $\mathrm{MSE}_{\mathrm{th,est}}$, computed on $N = 500$ repeated simulations. The values of $\mathrm{MSE}_{\mathrm{th}}$ and $\mathrm{MSE}_{\mathrm{th,est}}$ are decomposed into square bias and variance

Value of k	3	4	5	6	7	**8**	9	10
Mean of $\mathrm{MSE}_{\mathrm{emp}} \times 10^3$	15.849	12.779	11.135	10.365	9.977	**9.873**	9.929	10.110
	(4.041)	(3.637)	(3.233)	(2.829)	(2.424)	(**1.213**)	(1.616)	(2.021)
Mean of $\mathrm{MSE}_{\mathrm{th}} \times 10^3$	15.707	12.665	11.035	10.272	9.887	**9.785**	9.840	10.019
	(3.985)	(3.587)	(3.188)	(2.790)	(2.391)	(**1.196**)	(1.594)	(1.993)
Square bias for $\mathrm{MSE}_{\mathrm{th}} \times 10^3$	13.802	10.125	7.860	6.462	5.442	4.705	4.125	3.669
Variance for $\mathrm{MSE}_{\mathrm{th}} \times 10^3$	1.905	2.540	3.175	3.810	4.445	5.080	5.715	6.350
Mean of $\mathrm{MSE}_{\mathrm{th,est}} \times 10^3$	15.723	12.678	11.047	10.283	9.898	**9.795**	9.850	10.030
	(4.017)	(3.616)	(3.214)	(2.812)	(2.410)	(**1.206**)	(1.607)	(2.009)
Square bias for $\mathrm{MSE}_{\mathrm{th,est}} \times 10^3$	13.816	10.135	7.869	6.469	5.448	4.710	4.129	3.673
Variance for $\mathrm{MSE}_{\mathrm{th,est}} \times 10^3$	1.907	2.543	3.178	3.814	4.450	5.085	5.721	6.357

counterpart. When turning to the predictor, (11) may be enhanced:

$$\begin{aligned}\widehat{S}_n(X_{n+1}) - S(X_{n+1}) &= S(\Pi_k - I)(X_{n+1}) + S[\widehat{\Pi}_k - \Pi_k](X_{n+1}) \\ &\quad + \frac{1}{n}\sum_{i=1}^{n} \varepsilon_i \langle \Gamma_n^{\dagger} X_i, X_{n+1}\rangle,\end{aligned} \tag{12}$$

where Π_k is defined in the same way as we defined $\widehat{\Pi}_k$ previously, that is, the projection on the k first eigenvectors of Γ.

In terms of mean square error, the following easily stems from $\mathbb{E}(\varepsilon_i | X) = 0$:

$$\mathbb{E}\bigl\|\widehat{S}_n(X_{n+1}) - S(X_{n+1})\bigr\|^2 = \mathbb{E}\bigl\|S\widehat{\Pi}_k(X_{n+1}) - S(X_{n+1})\bigr\|^2 + \mathbb{E}\left\|\frac{1}{n}\sum_{i=1}^{n} \varepsilon_i \langle \Gamma_n^{\dagger} X_i, X_{n+1}\rangle\right\|^2 .$$

We prove below that:

$$\mathbb{E}\bigl\|S[\widehat{\Pi}_k - \Pi_k](X_{n+1})\bigr\|^2 = \mathrm{o}\left(\mathbb{E}\left\|\frac{1}{n}\sum_{i=1}^{n} \varepsilon_i \langle \Gamma_n^{\dagger} X_i, X_{n+1}\rangle\right\|^2\right), \tag{13}$$

and that the two terms that actually influence the mean square error are the first and the third in line (12). The first term $S(\Pi_k - I)(X_{n+1})$ is the bias term and the third a variance term (see line (9)).

The proofs are split into two parts. In the first, part we provide some technical lemmas which are collected there to enhance the reading of the second part devoted to the proof of the main results. In all the sequel, the sequence $k = k_n$ depends on n even if this index is dropped. We assume that all the assumptions mentioned earlier in the paper hold; they will be however recalled when addressing crucial steps. We assume once and for all that $(k \log k)^2/n \to 0$. The rate of convergence to 0 of $(k \log k)^2/n$ will be tuned when dealing with weak convergence.

4.1. Preliminary material

All along the proofs, we will make an intensive use of perturbation theory for bounded operators. It may be useful to have basic notions about spectral representation of bounded operators and perturbation theory. We refer to Kato [15] and Dunford and Schwartz ([11], Chapter VII.3) for an introduction to functional calculus for operators related with Riesz integrals. Roughly speaking, several results mentioned below and throughout the article may be easily understood by considering the formula of residues for analytic functions on the complex plane (see Rudin [20]) and extending it to functions still defined on the complex plane but with values in the space of operators.

Let us denote by $\mathcal{B}_j$ the oriented circle of the complex plane with center λ_j and radius $\delta_j/2$ where $\delta_j = \min\{\lambda_j - \lambda_{j+1}, \lambda_{j-1} - \lambda_j\} = \lambda_j - \lambda_{j+1}$, the last equality coming from the convexity associated to the λ_j's. Let us define $\mathcal{C}_k = \bigcup_{j=1}^{k} \mathcal{B}_j$. The open domain whose boundary is $\mathcal{C}_k$ is not connected but we can apply the functional calculus for bounded operators (see Dunford–Schwartz, Section VII.3, Definitions 8 and 9). With this formalism at hand, it is easy to prove the following formulas:

$$\Pi_{k_n} = \frac{1}{2\pi\iota}\int_{\mathcal{C}_k} (zI - \Gamma)^{-1}\,\mathrm{d}z, \qquad \Gamma^{\dagger} = \frac{1}{2\pi\iota}\int_{\mathcal{C}_k} \frac{1}{z}(zI - \Gamma)^{-1}\,\mathrm{d}z. \tag{14}$$

The same is true with the random Γ_n, but the contour $\mathcal{C}_k$ must be replaced by its random counterpart $\widehat{\mathcal{C}}_k = \bigcup_{j=1}^{k_n} \widehat{\mathcal{B}}_j$ where each $\widehat{\mathcal{B}}_j$ is a random ball of the complex plane with center $\widehat{\lambda}_j$ and, for instance, a radius $\widehat{\delta}_j/2$ with plain notations. Then $\widehat{\Pi}_{k_n} = \frac{1}{2\pi\iota}\int_{\widehat{\mathcal{C}}_k} (zI - \Gamma_n)^{-1}\,\mathrm{d}z$ and $\Gamma_n^{\dagger} = \frac{1}{2\pi\iota}\int_{\widehat{\mathcal{C}}_k} \frac{1}{z}(zI - \Gamma_n)^{-1}\,\mathrm{d}z$. This first lemma is based on convex inequalities. In the sequel, much depends on the bounds derived in this lemma.

Lemma 12. *Consider two large enough positive integers j and k such that $k > j$. Then*

$$j\lambda_j \geq k\lambda_k, \qquad \lambda_j - \lambda_k \geq \left(1 - \frac{j}{k}\right)\lambda_j, \qquad \sum_{j \geq k} \lambda_j \leq (k+1)\lambda_k,$$
$$\sum_{j \geq 1, j \neq k} \lambda_j/|\lambda_k - \lambda_j| \leq Ck\log k. \tag{15}$$

Besides, $\mathbb{E}\sup_{z \in \mathcal{B}_j} \|(zI - \Gamma)^{-1/2}(\Gamma - \Gamma_n)(zI - \Gamma)^{-1/2}\|_2^2 \leq C(j\log j)^2/n$.

The proof of this lemma will be found in Cardot, Mas and Sarda [7], pages 339–342. We introduce the event $\mathcal{A}_n = \{\forall j \in \{1,\ldots,k_n\}, |\widehat{\lambda}_j - \lambda_j|/\delta_j < 1/2\}$ which describes the way the estimated eigenvalues concentrate around the population ones: the higher the index j the closer are the $\widehat{\lambda}_j$'s to the λ_j's.

Proposition 13. *If* $(k \log k)^2/n \to 0$, $\mathbb{P}(\limsup \overline{\mathcal{A}}_n) = 0$.

Proof. We just check that the Borel–Cantelli lemma holds $\sum_{n=1}^{+\infty} \mathbb{P}(\overline{\mathcal{A}}_n) < +\infty$ where

$$\mathbb{P}(\overline{\mathcal{A}}_n) \leq \sum_{j=1}^{k} \mathbb{P}\big(|\widehat{\lambda}_j - \lambda_j|/\lambda_j > \delta_j/(2\lambda_j)\big) \leq \sum_{j=1}^{k} \mathbb{P}\big(|\widehat{\lambda}_j - \lambda_j|/\lambda_j > 1/2(j+1)\big).$$

Now, applying the results proved in Bosq [4] at pages 122–124, we see that the asymptotic behaviour of $\mathbb{P}(|\widehat{\lambda}_j - \lambda_j|/\lambda_j > \frac{1}{2j})$ is the same as $\mathbb{P}(|\frac{1}{n}\sum_{i=1}^{n} \langle \Gamma_n(e_j), e_j\rangle - \lambda_j| > \frac{\lambda_j}{2(j+1)})$. We apply Bernstein's exponential inequality – which is possible due to assumption (6) – to the latter, and we obtain (for the sake of brevity, $j+1$ was replaced by j in the right-hand side of the probability but this does not change the final result):

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n} \langle X_i, e_j\rangle^2 - \lambda_j\right| > \frac{\lambda_j}{2j}\right) \leq 2\exp\left(-\frac{n}{j^2}\frac{1}{8c + 1/(6j)}\right) \leq 2\exp\left(-C\frac{n}{j^2}\right)$$

and then $\sum_{j=1}^{k} \mathbb{P}(|\widehat{\lambda}_j - \lambda_j| > \lambda_j/2j) \leq 2k\exp(-Cn/k^2)$. Now it is plain from $(k\log k)^2/n \to 0$ that $k\exp(-C\frac{n}{k^2}) \leq 1/n^{1+\varepsilon}$ for some $\varepsilon > 0$ which leads to checking that $\sum_n k_n \exp(-C\frac{n}{k_n^2}) < +\infty$, and to the statement of Proposition 13 through Borel–Cantelli's lemma. □

Corollary 14. *We may write*

$$\widehat{\Pi}_{k_n} = \frac{1}{2\pi\iota}\int_{\mathcal{C}_k} (zI - \Gamma_n)^{-1}\,\mathrm{d}z, \qquad \Gamma_n^{\dagger} = \frac{1}{2\pi\iota}\int_{\mathcal{C}_k} \frac{1}{z}(zI - \Gamma_n)^{-1}\,\mathrm{d}z \qquad \textit{a.s.},$$

where this time the contour is $\mathcal{C}_k$ *hence no more random.*

Proof. The formulae above easily stem from Proposition 13 and perturbation theory (see Kato [15], Dunford and Schwartz [11], e.g.). □

4.2. Proofs of the main results

We denote $(zI - \Gamma)^{-1} = \Delta(z)$. We start with proving (13) as announced in the foreword of this section. What we give here is nothing but the term A_n in Theorem 2.

Proposition 15. *The following bound holds* $\mathbb{E}\|S(\widehat{\Pi}_k - \Pi_k)(X_{n+1})\|^2 \leq Ck^2\lambda_k\|S\|_2/n$.

Proof. We start with noting that

$$\mathbb{E}\big\|S(\widehat{\Pi}_k - \Pi_k)(X_{n+1})\big\|^2 = \sum_{j=1}^{+\infty}\sum_{\ell=1}^{+\infty} \mathbb{E}\big\langle S(\widehat{\Pi}_k - \Pi_k)\Gamma^{1/2}(e_j), e_\ell\big\rangle^2.$$

By Corollary 14, we have

$$\widehat{\Pi}_k - \Pi_k = \frac{1}{2\pi\iota} \sum_{m=1}^{k} \int_{\mathcal{B}_m} \left\{ (zI - \Gamma_n)^{-1} - (zI - \Gamma)^{-1} \right\} \mathrm{d}z = \sum_{m=1}^{k} T_{m,n}, \tag{16}$$

where $T_{m,n} = (1/2\pi\iota) \int_{\mathcal{B}_m} (zI - \Gamma_n)^{-1} (\Gamma - \Gamma_n)(zI - \Gamma)^{-1} \,\mathrm{d}z$. To go ahead now, we ask the reader to accept momentaneously that for all $m \le k$, the asymptotic behaviour of $T_{m,n}$ is the same as $T^*_{m,n} = (1/2\pi\iota) \int_{\mathcal{B}_m} \Delta(z)(\Gamma - \Gamma_n)\Delta(z) \,\mathrm{d}z$, where the random $(zI - \Gamma_n)^{-1}$ was replaced by the non-random $(zI - \Gamma)^{-1}$ and that studying $\widehat{\Pi}_k - \Pi_k$ comes down to studying $(1/2\pi\iota) \sum_{m=1}^{k} \int_{\mathcal{B}_m} \Delta(z)(\Gamma - \Gamma_n)\Delta(z) \,\mathrm{d}z$. The proof that this switch is allowed is postponed to Lemma 16. We go on with

$$\left\langle S(\widehat{\Pi}_k - \Pi_k)\Gamma^{1/2}(e_j), e_\ell \right\rangle = \frac{\sqrt{\lambda_j}}{2\pi\iota} \sum_{m=1}^{k} \int_{\mathcal{B}_m} \left\langle \Delta(z)(\Gamma - \Gamma_n)(e_j), S^* e_\ell \right\rangle \frac{\mathrm{d}z}{z - \lambda_j},$$

where S^* is the adjoint operator of S. We obtain

$$\int_{\mathcal{B}_m} \left\langle \Delta(z)(\Gamma - \Gamma_n)(e_j), S^* e_\ell \right\rangle \frac{\mathrm{d}z}{z - \lambda_j} = \int_{\mathcal{B}_m} \sum_{j'=1}^{+\infty} \frac{\langle (\Gamma - \Gamma_n)(e_j), e_{j'} \rangle \langle S^* e_\ell, e_{j'} \rangle}{(z - \lambda_j)(z - \lambda_{j'})} \,\mathrm{d}z.$$

We deduce that

$$\left\langle S(\widehat{\Pi}_k - \Pi_k)\Gamma^{1/2}(e_j), e_\ell \right\rangle = \frac{\sqrt{\lambda_j}}{2\pi\iota} \sum_{j'=1}^{+\infty} \left\langle (\Gamma - \Gamma_n)(e_j), e_{j'} \right\rangle \left\langle S^* e_\ell, e_{j'} \right\rangle \mathcal{I}_{k,j,j'}$$

with

$$\mathcal{I}_{k,j,j'} = \sum_{m=1}^{k} \int_{\mathcal{B}_m} \frac{\mathrm{d}z}{(z - \lambda_j)(z - \lambda_{j'})} = \begin{cases} 0, & \text{if } j, j' > m, \text{ and if } j, j' \le m, \\ (\lambda_j - \lambda_{j'})^{-1}, & \text{if } j' > m, j \le m, \\ (\lambda_{j'} - \lambda_j)^{-1}, & \text{if } j' \le m, j > m. \end{cases}$$

Then $\sum_{j=1}^{+\infty} \langle S(\widehat{\Pi}_k - \Pi_k)\Gamma^{1/2}(e_j), e_\ell \rangle^2 = A + B$ where

$$A = \frac{1}{4\pi^2} \sum_{j=1}^{k} \lambda_j \left[\sum_{j'=k+1}^{+\infty} \frac{\langle (\Gamma - \Gamma_n)(e_j), e_{j'} \rangle}{(\lambda_j - \lambda_{j'})} \left\langle S^* e_\ell, e_{j'} \right\rangle \right]^2,$$

$$B = \frac{1}{4\pi^2} \sum_{j=k+1}^{+\infty} \lambda_j \left[\sum_{j'=1}^{k} \frac{\langle (\Gamma - \Gamma_n)(e_j), e_{j'} \rangle}{(\lambda_{j'} - \lambda_j)} \left\langle S^* e_\ell, e_{j'} \right\rangle \right]^2.$$

We first compute $\mathbb{E}A$. We develop the series under the square and take expectations to obtain

$$\mathbb{E}\left[\sum_{j'=k+1}^{+\infty} \frac{\langle (\Gamma - \Gamma_n)(e_j), e_{j'} \rangle}{(\lambda_j - \lambda_{j'})} \left\langle S^* e_\ell, e_{j'} \right\rangle \right]^2 \le C \frac{\lambda_j}{n} \left(\sum_{j'=k+1}^{+\infty} \frac{\sqrt{\lambda_{j'}}}{(\lambda_j - \lambda_{j'})} \left\langle S^* e_\ell, e_{j'} \right\rangle \right)^2.$$

We split the series $\sum_{j'=k+1}^{+\infty}\frac{\sqrt{\lambda_{j'}}}{(\lambda_j-\lambda_{j'})}\langle S^*e_\ell, e_{j'}\rangle$ into two terms and get

$$\mathbb{E}A \le \frac{C}{n}\sum_{j=1}^{k}\frac{\lambda_j^2\lambda_{k+1}}{(\lambda_j-\lambda_{k+1})^2}\left(\sum_{j'=k+1}^{2k}\left|\langle S^*e_\ell, e_{j'}\rangle\right|\right)^2 + \frac{Ck}{n}\left(\sum_{j'=2k+1}^{+\infty}\sqrt{\lambda_{j'}}\left|\langle S^*e_\ell, e_{j'}\rangle\right|\right)^2. \tag{17}$$

The second term above is bounded by $C(k^2/n)\lambda_k\sum_{j'=2k+1}^{+\infty}|\langle S^*e_\ell, e_{j'}\rangle|^2$ because, by Lemma 12, we have $\sum_{j'=2k+1}^{+\infty}\lambda_{j'} \le (2k+1)\lambda_{2k+1} \le k\lambda_k$. We focus on the other term on line (17) and get

$$\begin{aligned}
&\sum_{j=1}^{k}\frac{\lambda_j^2\lambda_{k+1}}{(\lambda_j-\lambda_{k+1})^2}\left(\sum_{j'=k+1}^{2k}\left|\langle S^*e_\ell, e_{j'}\rangle\right|\right)^2\\
&\quad\le \lambda_{k+1}\sum_{j=1}^{k}\left[\left(\frac{k+1}{k+1-j}\right)^2\left(\sum_{j'=k+1}^{2k}\left|\langle S^*e_\ell, e_{j'}\rangle\right|\right)^2\right]\\
&\quad\le \left(\sum_{j'=k+1}^{2k}\left|\langle S^*e_\ell, e_{j'}\rangle\right|^2\right)(k+1)^2\lambda_{k+1}\sum_{j=1}^{k}\frac{1}{j^2}\\
&\quad\le C\left(\sum_{j'=k+1}^{2k}\left|\langle S^*e_\ell, e_{j'}\rangle\right|^2\right)k^2\lambda_{k+1},
\end{aligned}$$

hence $\mathbb{E}A \le \frac{C}{n}(\sum_{j'=k+1}^{+\infty}|\langle S^*e_\ell, e_{j'}\rangle|^2)k^2\lambda_k$. A similar bound may be proven for B. The method is given because it is significantly distinct but sketched. Denote $\lfloor x\rfloor$ the largest integer smaller than x and $\langle S^*e_\ell, e_{j'}\rangle = s_{\ell,j'}$. Here we bound $(\lambda_j/n)\sum_{j'=1}^{k}\sqrt{\lambda_{j'}}\langle S^*e_\ell, e_{j'}\rangle/(\lambda_{j'}-\lambda_j)$, hence

$$\begin{aligned}
\frac{\lambda_j}{n}\left(\sum_{j'=1}^{k}\frac{\sqrt{\lambda_{j'}}}{(\lambda_{j'}-\lambda_j)}s_{\ell,j'}\right) &\le \frac{\lambda_j}{n}\left[\left(\sum_{j'=1}^{\lfloor k/2\rfloor}\frac{\sqrt{\lambda_{j'}}}{\lambda_{j'}-\lambda_j}|s_{\ell,j'}|\right)^2 + \left(\sum_{j'=\lfloor k/2\rfloor}^{k}\frac{\sqrt{\lambda_{j'}}}{\lambda_{j'}-\lambda_j}|s_{\ell,j'}|\right)^2\right]\\
&\le C\frac{k}{n}\sum_{j'=1}^{k}s_{\ell,j'}^2 + \frac{k}{n}\left(\frac{j}{j-k}\right)^2\sum_{j'=\lfloor k/2\rfloor}^{k}s_{\ell,j'}^2.
\end{aligned}$$

From the definition of B, we get finally

$$\mathbb{E}B \le C\frac{k}{n}\left(\sum_{j'=1}^{k}s_{\ell,j'}^2\right)\sum_{j=k+1}^{+\infty}\lambda_j + \left(\sum_{j'=\lfloor k/2\rfloor}^{k}s_{\ell,j'}^2\right)\frac{k}{n}\sum_{j=k+1}^{+\infty}\lambda_j\left(\frac{j}{j-k}\right)^2.$$

It is plain that, for sufficiently large k, $\sum_{j'=\lfloor k/2\rfloor}^{k}\langle S^*e_\ell, e_{j'}\rangle^2 \le C/k$ (otherwise $\sum_{j'}\langle S^*e_\ell, e_{j'}\rangle^2$ cannot converge), whence

$$\left(\sum_{j'=\lfloor k/2\rfloor}^{k}\langle S^*e_\ell, e_{j'}\rangle^2\right)\frac{k}{n}\sum_{j=k+1}^{+\infty}\lambda_j\left(\frac{j}{j-k}\right)^2 \le \frac{C}{n}\left[\sum_{j=k+1}^{2k}\lambda_j\left(\frac{j}{j-k}\right)^2 + 4\sum_{j=2k}^{+\infty}\lambda_j\right].$$

Denoting $\varkappa_k = \sup_{k+1\le j\le 2k}(j\log j\lambda_j)$, we get at last $\sum_{j=k+1}^{2k}\lambda_j\frac{j^2}{(j-k)^2} \le Ck\varkappa_k$, and consequently $\mathbb{E}B \le C\frac{k}{n}\varkappa_k(\sum_{j'=1}^{k}|\langle S^*e_\ell, e_{j'}\rangle|^2)$, with $\varkappa_k \to 0$. Finally:

$$\sum_{j=1}^{+\infty}\sum_{\ell=1}^{+\infty}\langle S(\widehat{\Pi}_k - \Pi_k)\Gamma^{1/2}(e_j), e_\ell\rangle^2 \le C\frac{k}{n}\varkappa_k\sum_{j=1}^{+\infty}\sum_{\ell=1}^{+\infty}|\langle S^*e_\ell, e_j\rangle|^2.$$

This last bound almost concludes the proof of Proposition 15. It remains to ensure that switching $T_{m,n}^*$ and $T_{m,n}$ as announced just below line (16) is possible. □

Lemma 16. *We have*

$$\mathbb{E}\sum_{j=1}^{+\infty}\sum_{\ell=1}^{+\infty}\langle S(\widehat{\Pi}_k - \Pi_k)\Gamma^{1/2}(e_j), e_\ell\rangle^2 \sim \mathbb{E}\sum_{j=1}^{+\infty}\sum_{\ell=1}^{+\infty}\sum_{m=1}^{k}\langle ST_{m,n}^*\Gamma^{1/2}(e_j), e_\ell\rangle^2.$$

In other words, switching $T_{m,n}^*$ and $T_{m,n}$ is possible in line (16).

Proof. The proof of this lemma is close to the control of second order term at pages 351 and 352 of Cardot, Mas and Sarda [7] and we will give a sketch of it. We start from:

$$T_{m,n} = \frac{1}{2\pi\iota}\int_{\mathcal{B}_m}(zI - \Gamma)^{-1/2}R_n(z)(zI - \Gamma)^{-1/2}(\Gamma - \Gamma_n)\Delta(z)\,\mathrm{d}z$$

with $R_n(z) = (zI - \Gamma)^{1/2}(zI - \Gamma_n)^{-1}(zI - \Gamma)^{1/2}$. Besides, as can be seen from Lemma 4 in Cardot, Mas and Sarda [7] $[I + (zI - \Gamma)^{-1/2}(\Gamma - \Gamma_n)(zI - \Gamma)^{-1/2}]R_n(z) = I$. Denoting $S_n(z) = (zI - \Gamma)^{-1/2}(\Gamma - \Gamma_n)(zI - \Gamma)^{-1/2}$, it is plain that when $\|S_n(z)\| \le 1$ for all $z \in \mathcal{C}_k$, we have $R_n(z) = [I + S_n(z)]^{-1} := I + R_n^0(z)$ with $\|R_n^0(z)\|_\infty \le C\|S_n(z)\|_\infty$ for all $z \in \mathcal{C}_k$. Turning back to our initial equation, we then get:

$$T_{m,n} - T_{m,n}^* = \frac{1}{2\pi\iota}\int_{\mathcal{B}_m}(zI - \Gamma)^{-1/2}R_n^0(z)(zI - \Gamma)^{-1/2}(\Gamma - \Gamma_n)(zI - \Gamma)^{-1}\,\mathrm{d}z,$$

and we confine to considering only the first term in the development of $R_n^0(z)$ which writes

$$(2\pi\iota)^{-1}\int_{\mathcal{B}_m}(zI - \Gamma)^{-1/2}S_n^2(z)(zI - \Gamma)^{-1/2}\,\mathrm{d}z.$$

Now, if we denote $\mathcal{J} = \{\sup_{z\in\mathcal{C}_k} \|S_n(z)\|_2^2 < \tau_n k_n/n\}$ where τ_n will be tuned later, then we can write $S(\widehat{\Pi}_k - \Pi_k)\Gamma^{1/2} = S(\widehat{\Pi}_k - \Pi_k)\Gamma^{1/2}\mathbb{1}_{\mathcal{J}} + S(\widehat{\Pi}_k - \Pi_k)\Gamma^{1/2}\mathbb{1}_{\overline{\mathcal{J}}}$. We have:

$$\mathbb{E}\big\|S(\widehat{\Pi}_k - \Pi_k)\Gamma^{1/2}\mathbb{1}_{\overline{\mathcal{J}}}\big\|_2^2 \le 4\big\|S\Gamma^{1/2}\big\|_2^2\mathbb{P}(\overline{\mathcal{J}}) \tag{18}$$

and recalling the notation $(zI - \Gamma)^{-1} = \Delta(z)$

$$\begin{aligned}\left\|S\left[\widehat{\Pi}_k - \Pi_k - \sum_{m=1}^{k} T_{m,n}^*\right]\Gamma^{1/2}\mathbb{1}_{\mathcal{J}}\right\|_2 &\le \frac{1}{2\pi}\left\|\sum_{m=1}^{k}\int_{\mathcal{B}_m} S\Delta^{1/2}(z)S_n^2(z)\Delta^{1/2}(z)\Gamma^{1/2}\,\mathrm{d}z\,\mathbb{1}_{\mathcal{J}}\right\|_2 \\ &\le \frac{\tau_n^2 k_n^2}{2\pi n^2}\sum_{m=1}^{k}\delta_m \sup_{z\in\mathcal{B}_m}\big\{\big\|\Delta^{1/2}(z)\Gamma^{1/2}\big\|_\infty\big\|S\Delta^{1/2}(z)\big\|_\infty\big\} \\ &\le \|S\|_\infty\frac{\tau_n^2 k_n^2}{2\pi n^2}\sum_{m=1}^{k}\sqrt{\delta_m m}.\end{aligned}$$

Now from $\sum_{m=1}^{+\infty} m\delta_m < +\infty$ we get $\sqrt{\delta_m m} \le c/\sqrt{m\log m}$ hence $\frac{\tau_n^2 k_n^2}{n^2}\sum_{m=1}^{k}\sqrt{\delta_m m} = \mathrm{o}(\sqrt{k_n/n})$ whenever $k_n^4\tau_n^4/n^3 \to 0$.

The last step consists in controlling the right-hand side of (18). In Cardot, Mas and Sarda [7] this is done by classical Markov moment assumptions under the condition that $k_n^5\log^4 n/n$ tends to zero. Here, Bernstein's exponential inequality yields a tighter bound and ensures that $\mathbb{P}(\overline{\mathcal{J}}) = \mathrm{o}(k_n/n)$ when $k_n^2\log^2 k_n/n$ tends to zero. The method of proof is close in spirit though slightly more intricate than Proposition 13. □

Proposition 17. *Let* $T_n = \frac{1}{n}\sum_{i=1}^{n}\varepsilon_i\langle\Gamma_n^\dagger X_i, X_{n+1}\rangle$, *then* $\mathbb{E}\|T_n\|^2 = \frac{\sigma_\varepsilon^2}{n}k + \mathrm{tr}[\Gamma\mathbb{E}(\Gamma_n^\dagger - \Gamma^\dagger)]/n$.

Remark 18. We see that the right-hand side in the line above matches the decomposition in (9) and $\mathrm{tr}[\Gamma\mathbb{E}(\Gamma_n^\dagger - \Gamma^\dagger)]/n$ is precisely B_n in Theorem 2.

Proof of Proposition 17. We have

$$\|T_n\|^2 = \frac{1}{n^2}\sum_{i=1}^{n}\|\varepsilon_i\|^2\big\langle\Gamma_n^\dagger X_i, X_{n+1}\big\rangle^2 + \frac{1}{n^2}\sum_{i\neq i'}^{n}\langle\varepsilon_i, \varepsilon_{i'}\rangle\big\langle\Gamma_n^\dagger X_i, X_{n+1}\big\rangle\big\langle\Gamma_n^\dagger X_{i'}, X_{n+1}\big\rangle.$$

We take expectations in the line above and we note that the distribution of each member of the first series on the right-hand side does not depend on n or i and, due to linearity of expectation and $\mathbb{E}(\varepsilon_i|X_i) = 0$, the expectation of the second series is null, hence $\mathbb{E}\|T_n\|^2 = \frac{1}{n}\mathbb{E}[\|\varepsilon_1\|^2|X_1]\mathbb{E}\langle\Gamma_n^\dagger\Gamma\Gamma_n^\dagger X_1, X_1\rangle$. We focus on $\mathbb{E}\langle\Gamma_n^\dagger\Gamma\Gamma_n^\dagger X_1, X_1\rangle = \mathbb{E}[\mathrm{tr}\,\Gamma_n^\dagger\Gamma\Gamma_n^\dagger\cdot(X_i\otimes X_i)] = \mathrm{tr}[\Gamma\mathbb{E}\Gamma_n^\dagger]$. At last, we get $\mathbb{E}\langle\Gamma_n^\dagger\Gamma\Gamma_n^\dagger X_1, X_1\rangle = \mathrm{tr}[\Gamma\Gamma^\dagger] + \mathrm{tr}[\Gamma\mathbb{E}(\Gamma_n^\dagger - \Gamma^\dagger)] = k + \mathrm{tr}[\Gamma\mathbb{E}(\Gamma_n^\dagger - \Gamma^\dagger)]$. From Lemma 19 just below, we deduce that $\mathrm{tr}[\Gamma\mathbb{E}(\Gamma_n^\dagger - \Gamma^\dagger)] = \mathrm{o}(k)$, which finishes the proof of Proposition 17. □

Lemma 19. *We have* $\operatorname{tr}[\Gamma\mathbb{E}(\Gamma_n^\dagger - \Gamma^\dagger)] \le Ck^2(\log k)^2/n$, *where* C *does not depend on* S, n *or* k. *The preceding bound is an* $\mathrm{o}(k)$ *when* $k(\log k)^2/n \to 0$.

Proof. Here we slightly change the contour C_n and take the rectangle with left vertice $z = \lambda_k - \delta_k + \mathrm{i}x$ $x \in [-2\lambda_1, 2\lambda_1]$. We focus on

$$\left(\Gamma_n^\dagger - \Gamma^\dagger\right) = -\int_{C_n} \frac{1}{z}\Delta(z)(\Gamma_n - \Gamma)\Delta(z)\,\mathrm{d}z - \int_{C_n} \frac{1}{z}(zI - \Gamma_n)^{-1}(\Gamma_n - \Gamma)\Delta(z)(\Gamma_n - \Gamma)\Delta(z)\,\mathrm{d}z.$$

But $\mathbb{E}\int_{C_n} \frac{1}{z}\Delta(z)(\Gamma_n - \Gamma)\Delta(z)\,\mathrm{d}z = \int_{C_n} \frac{1}{z}\Delta(z)\mathbb{E}(\Gamma_n - \Gamma)\Delta(z)\,\mathrm{d}z = 0$ so we consider the second term above $R_n = \int_{C_n} \frac{1}{z}(zI - \Gamma)^{-1/2}T_n(z)A_n(z)A_n(z)(zI - \Gamma)^{-1/2}\,\mathrm{d}z$ where

$$T_n(z) = (zI - \Gamma)^{1/2}(zI - \Gamma_n)^{-1}(zI - \Gamma)^{1/2},$$
$$A_n(z) = (zI - \Gamma)^{-1/2}(\Gamma_n - \Gamma)(zI - \Gamma)^{-1/2},$$

whence $|\operatorname{tr}[\Gamma R_n]| = |\int_{C_n} z^{-1}\operatorname{tr}[\Delta^{-1}(z)\Gamma T_n(z)A_n^2(z)]\,\mathrm{d}z|$ and

$$\begin{aligned}\left|\operatorname{tr}[\Gamma R_n]\right| &\le \sup_{z\in C_n}\left\|T_n(z)\right\|_\infty \sup_{z\in C_n}\left\|A_n(z)\right\|_2^2 \int_{C_n} |z|^{-1}\left\|\Delta^{-1}(z)\Gamma\right\|_\infty \mathrm{d}z\\ &\le c\sup_{z\in C_n}\left\|A_n(z)\right\|_2^2,\end{aligned}$$

because

$$\int_{C_n} |z|^{-1}\left\|\Delta^{-1}(z)\Gamma\right\|_\infty \mathrm{d}z = \int_0^{2\lambda_1} |\lambda_k - \delta_k + \mathrm{i}x|^{-1}\lambda_k/|\delta_k + \mathrm{i}x|\,\mathrm{d}x \le C$$

and $\sup_{z\in C_n}\|T_n(z)\|_\infty$ is almost surely bounded. Now, by Lemma 12, we can write $\mathbb{E}\sup_{z\in C_n}\|A_n(z)\|_2^2 \le C(k\log k)^2/n$, and consequently $\mathbb{E}|\operatorname{tr}[\Gamma R_n]| \le C(k\log k)^2/n$. Finally, $|\operatorname{tr}[\Gamma\mathbb{E}(\Gamma_n^\dagger - \Gamma^\dagger)]| \le Ck^2(\log^2 k)/n$ and we proved Lemma 19 because $(k\log^2 k)/n \to 0$. We turn to Theorem 2. □

Proof of Theorem 2. From equation (12), we obtain

$$\mathbb{E}\left\|S_n(X_{n+1}) - S(X_{n+1})\right\|^2 = \mathbb{E}\left\|S\widehat{\Pi}_k(X_{n+1}) - S(X_{n+1})\right\|^2 + \mathbb{E}\left\|(1/n)\sum_{i=1}^n \varepsilon_i\left\langle\Gamma_n^\dagger X_i, X_{n+1}\right\rangle\right\|^2.$$

From Proposition 17 followed by Lemma 19, the second term is $\sigma_\varepsilon^2 k/n + B_n$. Proposition 15 and basic calculus yield $\mathbb{E}\|S\widehat{\Pi}_k(X_{n+1}) - S(X_{n+1})\|^2 = \mathbb{E}\|S(\Pi_k - I)(X_{n+1})\|^2 + A_n$ where A_n matches the bound of the theorem. Lastly $\mathbb{E}\|S(\Pi_k - I)(X_{n+1})\|^2 = \sum_{j\ge k+1}\lambda_j\|Se_j\|^2$ which finishes the proof. □

Proof of Theorem 6. Our proof follows the lines of Cardot and Johannes [6] through a modified version of Assouad's lemma. To simplify notations, we set $k_n^* = k_n$. Take $S^\theta =$

$\sum_{j=1}^{k_n} \eta_i \omega_i e_i \otimes e_1$ where $\omega_i \in \{-1, 1\}$ and $\theta = [\omega_1, \dots, \omega_k]$ and $\eta_i \in \mathbb{R}^+$ will be fixed later such that $S^\theta \in \mathcal{L}_2(\varphi, C)$ for all θ. Denote $\theta_{-i} = [\omega_1, \dots, -\omega_i, \dots, \omega_k]$ and $\mathbb{P}_\theta := \mathbb{P}_\theta[(Y_1, X_1), \dots, (Y_n, X_n)]$ denote the distribution of the data when $S = S^\theta$. Let ρ stand for Hellinger's affinity, $\rho(\mathbb{P}_0, \mathbb{P}_1) = \int \sqrt{\mathrm{d}\mathbb{P}_0 \, \mathrm{d}\mathbb{P}_1}$ and $\mathbf{KL}(\mathbb{P}_0, \mathbb{P}_1)$ for Küllback–Leibler divergence then $\rho(\mathbb{P}_0, \mathbb{P}_1) \geq (1 - \frac{1}{2}\mathbf{KL}(\mathbb{P}_0, \mathbb{P}_1))$.

Note that considering models based on S^θ above comes down to projecting the model on a one-dimensional space. We are then faced with a linear model with real output and finally confine ourselves to proving that the optimal rate is unchanged (see Hall and Horowitz [14]):

$$\begin{aligned}\mathcal{R}_n(T_n) &= \sup_{S \in \mathcal{L}_2(\varphi, C)} \mathbb{E}\left\|(T_n - S)\Gamma^{1/2}\right\|_2^2 \geq \frac{1}{2^k} \sum_{\omega \in \{-1,1\}^k} \sum_{i=1}^{k_n} \lambda_i \mathbb{E}_\theta \left\langle (T_n - S^\theta) e_i, e_1 \right\rangle^2 \\ &= \frac{1}{2^k} \sum_{\omega \in \{-1,1\}^k} \frac{1}{2} \sum_{i=1}^{k_n} \lambda_i \left[\mathbb{E}_\theta \left\langle (T_n - S^\theta) e_i, e_1 \right\rangle^2 + \mathbb{E}_{\theta_{-i}} \left\langle (T_n - S^{\theta_{-i}}) e_i, e_1 \right\rangle^2\right] \\ &\geq \frac{1}{2^k} \sum_{\omega \in \{-1,1\}^k} \sum_{i=1}^{k_n} \lambda_i \eta_i^2 \rho^2(\mathbb{P}_\theta, \mathbb{P}_{\theta_{-i}}).\end{aligned}$$

The last line was obtained by a slight variant of the bound (A.9) in Cardot and Johannes [6], page 405, detailed below:

$$\begin{aligned}\rho(\mathbb{P}_\theta, \mathbb{P}_{\theta_{-i}}) &\leq \int \frac{\langle (T_n - S^\theta) e_i, e_1 \rangle}{|\langle (S^{\theta_{-i}} - S^\theta) e_i, e_1 \rangle|} \sqrt{\mathrm{d}\mathbb{P}_0 \, \mathrm{d}\mathbb{P}_1} + \int \frac{\langle (T_n - S^{\theta_{-i}}) e_i, e_1 \rangle}{|\langle (S^{\theta_{-i}} - S^\theta) e_i, e_1 \rangle|} \sqrt{\mathrm{d}\mathbb{P}_0 \, \mathrm{d}\mathbb{P}_1} \\ &\leq \frac{1}{2\eta_i} \left(\int \left\langle (T_n - S^\theta) e_i, e_1 \right\rangle^2 \mathrm{d}\mathbb{P}_\theta \right)^{1/2} + \left(\int \left\langle (T_n - S^{\theta_{-i}}) e_i, e_1 \right\rangle \mathbb{P}_{\theta_{-i}} \right)^{1/2}\end{aligned}$$

by Cauchy–Schwarz inequality and since $|\langle (S^{\theta_{-i}} - S^\theta) e_i, e_1 \rangle| = 2\eta_i$. Then

$$2\eta_i^2 \rho^2(\mathbb{P}_\theta, \mathbb{P}_{\theta_{-i}}) \leq \mathbb{E}_\theta \left\langle (T_n - S^\theta) e_i, e_1 \right\rangle^2 + \mathbb{E}_{\theta_{-i}} \left\langle (T_n - S^{\theta_{-i}}) e_i, e_1 \right\rangle^2$$

yields $\mathcal{R}_n(T_n) \geq \inf_{\omega \in \{-1,1\}^k} \inf_i \rho(\mathbb{P}_\theta, \mathbb{P}_{\theta_{-i}}) \sum_i \lambda_i \eta_i^2$. We show below that $\mathbf{KL}(\mathbb{P}_\theta, \mathbb{P}_{\theta_{-i}}) \leq 4n\lambda_i \eta_i^2 / \sigma_1^2$. Choosing $\eta_i = \sigma_1 / 2\sqrt{n\lambda_i}$ for $1 \leq i \leq k_n$ gives $S^\theta \in \mathcal{L}_2(\varphi, 1)$ and $\sup_{\omega,i} \mathbf{KL}(\mathbb{P}_\theta, \mathbb{P}_{\theta_{-i}}) \leq 1$, $\inf_{\omega,i} \rho(\mathbb{P}_\theta, \mathbb{P}_{\theta_{-i}}) \leq 1/2$ and $\mathcal{R}_n(T_n) \geq \sum_{i=1}^{k_n} \lambda_i \eta_i^2 / 2 = k_n / 2n$ whatever the choice of the estimate T_n. This proves the lower bound:

$$\lim \sup_{n \to +\infty} \varphi_n^{-1} \inf_{T_n} \sup_{S \in \mathcal{L}_2(\varphi, C)} E\left\|(T_n - S)\Gamma^{1/2}\right\|^2 > \frac{1}{2},$$

and the theorem stems from this last line.

We finish by proving that $\mathbf{KL}(\mathbb{P}_\theta, \mathbb{P}_{\theta_{-i}}) \leq 4n\lambda_i \eta_i^2 / \sigma_1^2$. It suffices to notice that $\mathbf{KL}(\mathbb{P}_\theta, \mathbb{P}_{\theta_{-i}}) = \int \log(\mathrm{d}\mathbb{P}_{\theta|X} / \mathrm{d}\mathbb{P}_{\theta_{-i}|X}) \, \mathrm{d}\mathbb{P}_\theta$ where $\mathbb{P}_{\theta|X}$ stand for the likelihood of Y conditionally to X. In this

Hilbert setting we must clarify the existence of this likelihood ratio. It suffices to prove that $\mathbb{P}_{\theta|X}(Y) \ll \mathbb{P}_{0|X}(Y)$ which in turn is true when $S^{\theta}X$ belongs to the RKHS associated to ε (see Lifshits [16]). With other words, we need that almost surely $\Gamma_{\varepsilon}^{-1/2}S^{\theta}X$ is finite where Γ_{ε} is the covariance operator of the noise. But $\Gamma_{\varepsilon}^{-1/2}S^{\theta} = S^{\theta}/\sigma_1$. Set $\omega_l' = \omega_l$ if $l \neq i$ with $\omega_i' = -\omega_i$:

$$\log \frac{\mathrm{d}\mathbb{P}_{\theta|X}(Y)}{\mathrm{d}\mathbb{P}_{\theta_{-i}|X}(Y)} = -2\omega_i \eta_i \frac{\langle X, e_i\rangle}{\sigma_1^2}\big(2\langle \varepsilon, e_1\rangle + 2\omega_i \eta_i \langle X, e_i\rangle\big)$$

and $\mathbb{E}_{\theta}[\log \mathrm{d}\mathbb{P}_{\theta|X}(Y)/\mathrm{d}\mathbb{P}_{\theta_{-i}|X}(Y)] = 4\eta_i^2\mathbb{E}_{\theta}\langle X, e_i\rangle^2/\sigma_1^2 = 4\eta_i^2\lambda_i/\sigma_1^2$. □

Now we focus on the problem of weak convergence.

Proof of Theorem 8. Consider (11). We claim that weak convergence of S_n will depend on the series $(1/n)\sum_{i=1}^n \varepsilon_i \otimes \Gamma_n^{\dagger} X_i$. This fact can be checked by inspecting the proof of Theorem 2. We are going to prove that $(1/n)\sum_{i=1}^n \varepsilon_i \otimes \Gamma^{\dagger} X_i$ cannot converge for the classical (supremum) operator norm. We replace the random $\Gamma_n^{\dagger}$ by the non-random $\Gamma^{\dagger}$. It is plain that non-convergence of the second series implies non-convergence of the first. Suppose that for some sequence $\alpha_n \uparrow +\infty$ the centered series $(\alpha_n/n)\sum_{i=1}^n \varepsilon_i \otimes \Gamma^{\dagger} X_i \xrightarrow{w} Z$, in operator norm, where Z is a fixed random operator (not necessarily Gaussian). Then for all fixed x and y in H, $\frac{\alpha_n}{n}\sum_{i=1}^n \langle \varepsilon_i, y\rangle\langle \Gamma^{\dagger}X_i, x\rangle \xrightarrow{w} \langle Zx, y\rangle$, as real random variables. First, take x in the domain of Γ^{-1}. From $\|\Gamma^{-1}x\| < +\infty$, we see that $\mathbb{E}\langle \varepsilon_i, y\rangle^2\langle \Gamma^{\dagger}X_i, x\rangle^2 < +\infty$ implies that $\alpha_n = \sqrt{n}$ (and Z is Gaussian since we apply the central limit theorem for independent random variables). Now take a x such that $\|\Gamma^{-1}x\| = +\infty$, then $\mathbb{E}\langle \varepsilon_1, y\rangle^2\langle \Gamma^{\dagger}X_1, x\rangle^2 = \mathbb{E}\langle \varepsilon_1, y\rangle^2\mathbb{E}\langle \Gamma^{\dagger}x, x\rangle$, and it is easily seen from the definition of $\Gamma^{\dagger}$ that $\mathbb{E}\langle \Gamma^{\dagger}x, x\rangle$ – which is positive and implicitly depends on n through k – tends to infinity. Consequently, $(1/\sqrt{n})\sum_{i=1}^n \varepsilon_i \otimes \Gamma^{\dagger}X_i$ cannot converge weakly anymore since the margins related to the x's do not converge in distribution. This proves the theorem. □

We recall that $T_n = \frac{1}{n}\sum_{i=1}^n \varepsilon_i \langle \Gamma_n^{\dagger}X_i, X_{n+1}\rangle$ and this series is the crucial term that determines weak convergence. The following lemma is close to Lemma 8, page 355, in Cardot, Mas and Sarda [7] and will not be proved.

Lemma 20. *The random sequence* $\sqrt{\frac{k_n}{n}}T_n$ *is flatly concentrated and uniformly tight. In fact, if* $\mathcal{P}_m$ *is the projection operator on the m first eigenvectors of* Γ_{ε} *and* $\eta > 0$ *is a real number*

$$\limsup_{m\to+\infty}\sup_{n} \mathbb{P}\big(\big\|\sqrt{n/k_n}(I - \mathcal{P}_m)T_n\big\| > \eta\big) = 0.$$

Besides for all fixed x in H, $\sqrt{n/k_n}\langle T_n, x\rangle \xrightarrow{w} \mathcal{N}(0, \sigma_{\varepsilon,x}^2)$, *where* $\sigma_{\varepsilon,x}^2 = \mathbb{E}\langle \varepsilon_k, x\rangle^2$.

Proof of Theorem 9. We only prove the second part of the theorem: weak convergence with no bias. The first part follows immediately. We start again from the decomposition (12). As announced just above, the two first terms vanish with respect to convergence in distribution. For $S[\widehat{\Pi}_k - \Pi_k](X_{n+1})$, we invoke Proposition 15 to claim that, whenever $k^2 \log^2 k/n \to 0$, $(n/k)\mathbb{E}\|S[\widehat{\Pi}_k - \Pi_k](X_{n+1})\|^2 \to 0$ and we just have to deal with the first term, related to bias: $S(\Pi_k - I)(X_{n+1})$. Assume first that the mean square of the latter reminder, $(n/k)\sum_{j=k+1}^{+\infty}\lambda_j\|S(e_j)\|^2$, decays to zero. Then the proof of the theorem is immediate from Lemma 20. The sequence $\sqrt{n/k_n}T_n$ is uniformly tight and its finite-dimensional distributions (in the sense of "all finite-dimensional projections of $\sqrt{n/k_n}T_n$") converge weakly to $\mathcal{N}(0, \sigma^2_{\varepsilon,x})$. This is enough to claim that Theorem 9 holds. We refer, for instance, to Araujo and Giné [3] for checking the validity of this conclusion.

It remains to prove $\lim_{n\to+\infty}(n/k)\sum_{j=k+1}^{+\infty}\lambda_j\|S(e_j)\|^2 = 0$ when tightening conditions on the sequence k_n. First, we know by previous remarks (since λ_j and $\|S(e_j)\|^2$ are convergent series) that $\lambda_j\|S(e_j)\|^2 = \tau_j(j^2\log^2 j)$, where τ_j tends to zero. Taking as in the first part of the theorem $n = k^2\log^2 k/\sqrt{\gamma_k}$, we can focus on $\lim_{k+\infty}\frac{k\log^2 k}{\sqrt{\gamma_k}}\sum_{j=k+1}^{+\infty}\tau_j/(j^2\log^2 j)$. We know that for a sufficiently large k and for all $j \geq k$, $0 \leq \tau_j \leq \epsilon$ where $\epsilon > 0$ is fixed. Then

$$\frac{1}{\sqrt{\gamma_k}}\sum_{j=k+1}^{+\infty}\tau_j\frac{k\log^2 k}{j^2\log^2 j} = \frac{1}{\sqrt{\gamma_k}}\sum_{m=1}^{+\infty}\sum_{j=km+1}^{km+k}\tau_j\frac{k\log^2 k}{j^2\log^2 j} \leq \frac{1}{\sqrt{\gamma_k}}\Big(\sup_{k\leq j}\tau_j\Big)\sum_{m=1}^{+\infty}\frac{1}{m^2} = C\sqrt{\gamma_k} \to 0,$$

which removes the bias term and is the desired result. □

Acknowledgements

We are grateful to the Associate Editor and two anonymous referees for valuable comments that helped us to improve the paper.

References

[1] Aguilera, A., Ocaña, F. and Valderrama, M. (2008). Estimation of functional regression models for functional responses by wavelet approximation. In *Functional and Operatorial Statistics* (S. Dabo-Niang and F. Ferraty, eds.). *Contrib. Statist.* 15–21. Heidelberg: Springer. MR2478482

[2] Antoch, J., Prchal, L., De Rosa, M. and Sarda, P. (2010). Electricity consumption prediction with functional linear regression using spline estimators. *J. Appl. Stat.* **37** 2027–2041. MR2740138

[3] Araujo, A. and Giné, E. (1980). *The Central Limit Theorem for Real and Banach Valued Random Variables*. New York: Wiley. MR0576407

[4] Bosq, D. (2000). *Linear Processes in Function Spaces*: *Theory and Applications*. *Lecture Notes in Statistics* **149**. New York: Springer. MR1783138

[5] Cardot, H., Crambes, C., Kneip, A. and Sarda, P. (2007). Smoothing splines estimators in functional linear regression with errors-in-variables. *Comput. Statist. Data Anal.* **51** 4832–4848. MR2364543

[6] Cardot, H. and Johannes, J. (2010). Thresholding projection estimators in functional linear models. *J. Multivariate Anal.* **101** 395–408. MR2564349

[7] Cardot, H., Mas, A. and Sarda, P. (2007). CLT in functional linear regression models. *Probab. Theory Related Fields* **138** 325–361. MR2299711

[8] Chiou, J.M., Müller, H.G. and Wang, J.L. (2004). Functional response models. *Statist. Sinica* **14** 675–693. MR2087968

[9] Crambes, C., Kneip, A. and Sarda, P. (2009). Smoothing splines estimators for functional linear regression. *Ann. Statist.* **37** 35–72. MR2488344

[10] Cuevas, A., Febrero, M. and Fraiman, R. (2002). Linear functional regression: The case of fixed design and functional response. *Canad. J. Statist.* **30** 285–300. MR1926066

[11] Dunford, N. and Schwartz, J.T. (1988). *Linear Operators*, *Vols. I & II*. New York: Wiley.

[12] Engl, H.W., Hanke, M. and Neubauer, A. (1996). *Regularization of Inverse Problems*. *Mathematics and Its Applications* **375**. Dordrecht: Kluwer Academic. MR1408680

[13] Ferraty, F. and Vieu, P. (2006). *Nonparametric Functional Data Analysis*: *Theory and Practice*. *Springer Series in Statistics*. New York: Springer. MR2229687

[14] Hall, P. and Horowitz, J.L. (2007). Methodology and convergence rates for functional linear regression. *Ann. Statist.* **35** 70–91. MR2332269

[15] Kato, T. (1976). *Perturbation Theory for Linear Operators*, 2nd ed. *Grundlehren der Mathematischen Wissenschaften* **132**. Berlin: Springer. MR0407617

[16] Lifshits, M.A. (1995). *Gaussian Random Functions*. *Mathematics and Its Applications* **322**. Dordrecht: Kluwer Academic. MR1472736

[17] Malfait, N. and Ramsay, J.O. (2003). The historical functional linear model. *Canad. J. Statist.* **31** 115–128. MR2016223

[18] Ramsay, J.O. and Dalzell, C.J. (1991). Some tools for functional data analysis. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **53** 539–572. With discussion and a reply by the authors. MR1125714

[19] Ramsay, J.O. and Silverman, B.W. (2005). *Functional Data Analysis*, 2nd ed. *Springer Series in Statistics*. New York: Springer. MR2168993

[20] Rudin, W. (1987). *Real and Complex Analysis*, 3rd ed. New York: McGraw-Hill. MR0924157

[21] Stone, C.J. (1982). Optimal global rates of convergence for nonparametric regression. *Ann. Statist.* **10** 1040–1053. MR0673642

[22] Tikhonov, A.N. and Arsenin, V. Y. (1977). *Solutions of Ill-Posed Problems*. New York: Wiley. MR0455365

[23] Yao, F., Müller, H.G. and Wang, J.L. (2005). Functional linear regression analysis for longitudinal data. *Ann. Statist.* **33** 2873–2903. MR2253106

Received February 2011 and revised June 2012