Adaptive functional principal components analysis

Sunny Wang Guang Wei*   Valentin Patilea†   Nicolas Klutchnikoff‡

June 29, 2023

Abstract

Functional data analysis (FDA) almost always involves smoothing discrete observations into curves, because they are never observed in continuous time and rarely without error. Although smoothing parameters affect the subsequent inference, data-driven methods for selecting these parameters are not well-developed, frustrated by the difficulty of using all the information shared by curves while being computationally efficient. On the one hand, smoothing individual curves in an isolated, albeit sophisticated way, ignores useful signals present in other curves. On the other hand, bandwidth selection by automatic procedures such as cross-validation after pooling all the curves together quickly become computationally unfeasible due to the large number of data points. In this paper we propose a new data-driven, adaptive kernel smoothing, specifically tailored for functional principal components analysis (FPCA) through the derivation of sharp, explicit risk bounds for the eigen-elements. The minimization of these quadratic risk bounds provide refined, yet computationally efficient bandwidth rules for each eigen-element separately. Both common and independent design cases are allowed. Rates of convergence for the adaptive eigen-elements estimators are derived. An extensive simulation study, designed in a versatile manner to closely mimic characteristics of real data sets, support our methodological contribution, which is available for use in the R package FDAdapt.

Key words: Adaptive estimator; Functional Principal Components Analysis; Hölder exponent; Kernel smoothing

MSC2020: 62R10; 62G08; 62M99

1 Introduction

1.1 Motivation

The advent of modern data collection mechanisms, exemplified by the inclusion of sensors, has given rise to a fascinating array of intricate functional data sets. These data sets often exhibit rough trajectories, characterized by varying degrees of smoothness that challenge the conventional framework of functional data analysis (FDA). Traditional FDA methods have long focused on smooth sample paths, tacitly assuming knowledge of their regularity. However, the current big data landscape calls for adaptive methodologies that can accommodate the idiosyncrasies inherent in these complex functional data sets. While the estimation of mean, covariance, and regression functions has witnessed some advancements in adaptive techniques, the exploration of adaptive functional principal component analysis (FPCA) remains relatively unexplored. FPCA, with its capacity to extract salient modes of variation from curves, continues to hold a central position in the pantheon of dimension reduction techniques. Frequently employed as a preliminary step in downstream modeling endeavors such as regression and classification, FPCA has historically been executed through eigen-analysis of estimated covariance functions. In densely sampled scenarios, the 'smooth-first-then-estimate' approach dominates, while in sparsely sampled situations, 'weighting schemes' that pool all points together before estimation tend to prevail. These prevalent strategies assume knowns degree and patterns of smoothness, rendering them ill-suited for complex functional

---

*Ensai, CREST - UMR 9194, France; sunny.wang@ensai.fr

†Ensai, CREST - UMR 9194, France; valentin.patilea@ensai.fr

‡Univ Rennes, IRMAR - UMR 6625, France; Nicolas.klutchnikoff@univ-rennes2.fr

data sets characterized by heterogeneous degrees of smoothness and motivating the need for more flexible frameworks. Furthermore, in the context of FPCA, the notion of adaptivity goes beyond considering only the regularity of the underlying curves, as it also encompasses the adaptivity to each principal component and eigenvalue of the covariance.

Let us review a few of the most important contributions in FPCA. Bosq (2000) provided theoretical grounds in the ideal fully observed, noiseless case. See also Horváth and Kokoszka (2012). For the practical aspects, the book of Ramsay and Silverman (2005) is a landmark reference. The active research of the recent years provided advances in more realistic data settings where the curves are contaminated by noise and necessarily subjected to discretization errors. When the data curves admit derivatives up to some known order, Hall et al. (2006) established L^2 optimal rates of convergence for the principal components, using a covariance function estimated by pooling points together and applying a local linear smoother, while Li and Hsing (2010) obtained the uniform rates. Zhang and Wang (2016) further refined the convergence results for the pooling observations approach. Benko et al. (2009) propose alternative principal components estimation using a well-known duality relation between row and column spaces of a data matrix, and kernel smoothing. However, the existing body of methods is largely silent on the practical issue of bandwidth selection, and usually requires some mix of ad-hoc judgment and computationally intensive tools such as cross-validation. For example, in Zhang and Wang (2016), one recommendation is to use a mix of different cross-validation methods, and select the bandwidth that obtains the best fit or is the most interpretable. Let us end our brief review by mentioning Hall and Hosseini-Nasab (2006), who first proposed asymptotic expansions for eigen-elements associated to the empirical covariance estimator and their corresponding rates of convergence. Our new approach relies on this type of expansion.

This work aims to develop an FPCA algorithm that meets four important requirements: (i) the algorithm should be flexible enough to adapt to complex functional data sets with irregular sample paths of possibly varying smoothness, often occurring for instance in energy, meteorology, and medical applications; (ii) it should effectively utilize the replication structure of functional data by incorporating signals from other curves; (iii) be computationally simple and efficient to compute; and (iv) be tailored for each eigen-element of the covariance operator. The last property stems from risk bounds and error rates that we derive, and are expressed as explicit functions of the bandwidth. Our contribution is synthesized in the formulation of such an FPCA algorithm, made attainable through a refined plug-in bandwidth rule that minimizes explicit, sharp quadratic risk bounds that are easy to implement. Our risk bounds, more precisely the squared bias terms in the bounds, importantly feature the regularity of the process generating the true sample paths, which are usually unknown in practice, motivating the need for methods that automatically adapt to this regularity that typically governs the rate of convergence. Instead of following the direction of more traditional adaptive methods, such as block thresholding, we obtain adaptation by building upon the work of Golovkine et al. (2022), and explicitly estimating the local regularity of the process generating the data curves instead. The estimation of this regularity is made possible due to the replication nature of functional data. We show that the rates of convergence are governed through the integrals of the local regularity over its sampling domain, and in particular, is driven by the lowest regularity. Contrary to the conventional wisdom of a universal smoothing method for all FDA tasks, we propagate the findings by Carroll et al. (2013) and Golovkine et al. (2023), by showing that for the purposes of FPCA, the optimal bandwidth for smoothing curves is generally different from that of optimal curve recovery, or mean and covariance function estimation. Moreover, each principal component requires a different optimal bandwidth. The theory reveals that the difference between the optimal bandwidths for different principal components comes from the constants multiplying the terms in the quadratic risk bounds. However, like it was established several decades ago for the kernel smoothing for regression curves, the constants matters in applications. We finally want to emphasize that our adaptive approach based on the regularity of the process generating the data curves, requires the number of observation points on each curve to be sufficiently large. The estimation of the process regularity cannot be performed in the case of extremely sparse designs with very few observations per curve, as considered by Hall et al. (2006).

Our general methodology and the FPCA algorithm are presented in Section 2. It applies to discretely observed functional data, at random (independent design) or fixed (common design) points. Moreover, the

noise is allowed to be heteroscedastic. The justification of our risk bounds and their rates of convergence, as well as the rates of the data-driven bandwidths obtained by the algorithm, are provided in Section 3. The conclusions of an extensive empirical study are reported in Section 4. They are obtained with the newly built R package FDAdapt. For comparison purposes, we use simulated data, but the data are mimicking the features of a real power consumption dataset. We also build a general purpose simulator, which we consider of independent interest, as it can match arbitrary process mean, covariance functions and noise variance situations. This type of device is an effective tool for comparing different approaches in various realistic setups. The simulations illustrate the good performance of our method, and reasonable computation times. The Appendix gathers the assumptions, and the theoretical statements behind the local regularity estimation, a new result of independent interest. The Appendix also contains the sketch of the proofs of the results in Section 3. The technical details are relegated to the Supplementary Material, where we also present additional simulation results in several different setups.

2 Methodology

2.1 The data and curves reconstruction

Let $X = \{X_t\}_{t \in \mathcal{T}}$ be a stochastic process defined on a compact interval $\mathcal{T}$, say $[0, 1]$. The mean and covariance functions of the process X are given by

$$\mu(t) = \mathbb{E}(X_t) \quad \text{and} \quad \Gamma(s, t) = \gamma(s, t) - \mu(s)\mu(t) \quad \text{with} \quad \gamma(s, t) = \mathbb{E}(X_s X_t), \qquad s, t \in \mathcal{T}.$$

Let $\lambda_1 \geq \lambda_2 \geq \ldots \geq 0$ be the ordered eigenvalues of the covariance operator defined by Γ, and ψ_j be orthonormal eigenfunctions associated to λ_j, $j \geq 1$.

The data are collected from N independent realizations $X^{(1)}, \ldots X^{(i)}, \ldots X^{(N)}$ of X. The observed time points along each curve i are denoted $T_m^{(i)}$, they belong to $\mathcal{T}$ and are indexed by $1 \leq m \leq M_i$. For the sake of readability, we use the notation $X^{(i)}(T_m^{(i)})$ for $X_t^{(i)}$ with $t = T_m^{(i)}$. Each curve is contaminated by noise, due to independent, centered errors that are possibly heteroscedastic. Formally, for each $1 \leq i \leq N$, the observations associated to the sample path $X^{(i)}$ are the pairs $(Y_m^{(i)}, T_m^{(i)}) \in \mathbb{R} \times \mathcal{T}$ where

$$Y_m^{(i)} = X^{(i)}(T_m^{(i)}) + \sigma(T_m^{(i)})e_m^{(i)}, \qquad 1 \leq m \leq M_i. \tag{1}$$

The total number of observed pairs $(Y_m^{(i)}, T_m^{(i)})$ is $M_1 + \ldots + M_N$. The $e_m^{(i)}$'s are independently and identically distributed (iid) centered variables with unit variance. The errors' conditional variance function $\sigma^2(t)$ is bounded and unknown. In the independent design case we assume that the M_i are iid and generated from a positive integer variable M, with expectation $\mathbf{m}$. Moreover, we assume that the sampling points $T_m^{(i)}$ are iid and generated from $T \in \mathcal{T}$, and that X, M and T are mutually independent. In the common design case, $M_i = \mathbf{m}$ and the observed time points $T_1^{(i)}, \ldots, T_{\mathbf{m}}^{(i)}$ are the same for all i. Without loss of generality, whatever the design is, we assume that for each i, $T_1^{(i)}, \ldots, T_{M_i}^{(i)}$ are increasingly ordered.

In order to proceed with FPCA, we have to estimate $\Gamma(s, t)$ for any pair (s, t). In the ideal situation where the curves $X^{(i)}$ are observed at any $t \in \mathcal{T}$ without error, the natural estimates of λ_j and ψ_j are the empirical ones, that are obtained from the empirical covariance function

$$\widetilde{\Gamma}(s, t) = \frac{1}{N} \sum_{i=1}^{N} X_s^{(i)} X_t^{(i)} - \widetilde{\mu}(s)\widetilde{\mu}(t) \qquad \text{where} \qquad \widetilde{\mu}(t) = \frac{1}{N} \sum_{i=1}^{N} X_t^{(i)}, \quad s, t \in \mathcal{T}. \tag{2}$$

Let $\widetilde{\lambda}_j$ and $\widetilde{\psi}_j$ be the empirical estimators of the eigenvalue λ_j and eigenfunction ψ_j, respectively, that means

$$\int_{\mathcal{T}} \widetilde{\Gamma}(s, t)\widetilde{\psi}_j(s)ds = \widetilde{\lambda}_j \widetilde{\psi}_j(t), \quad \forall t \in \mathcal{T}, \qquad \int_{\mathcal{T}} \widetilde{\psi}_j(t)\widetilde{\psi}_{j'}(t)dt = \mathbf{1}\{j = j'\}, \quad \forall j, j' \geq 1.$$

Here, $\mathbf{1}\{.\}$ denote the indicator function. The properties of $\widetilde{\lambda}_j$ and $\widetilde{\psi}_j$ have been extensively studied, see for instance Horváth and Kokoszka (2012). Herein, for each estimated eigenfunction we adopt the standard convention and consider the version having positive inner product with the true eigenfunction.

In practice, the curves $X^{(i)}$ are not observed and the estimate of $\Gamma(s,t)$ has to be built using the data points $(Y_m^{(i)}, T_m^{(i)})$. A convenient way to proceed with real data is the so-called 'first smooth, then estimate' approach : first, using some nonparametric method, build an estimate $\widehat{X}^{(i)}$ for each curve $X^{(i)}$ separately; next, build a version $\widehat{\Gamma}$ of empirical covariance with the reconstructed curves $\widehat{X}^{(i)}$ replacing the true $X^{(i)}$'s. The formal definition of $\widehat{\Gamma}$ is provided in Section 2.3. We follow this simple idea and use a linear kernel smoother for each curve $X^{(i)}$, that is

$$\widehat{X}_t^{(i)} = \widehat{X}_t^{(i)}(h) = \sum_{m=1}^{M_i} W_m^{(i)}(t;h) Y_m^{(i)}, \qquad t \in \mathcal{T},\ 1 \leq i \leq N, \tag{3}$$

where the weights $W_m^{(i)}(t;h)$ are that of Nadaraya-Watson's or local polynomial smoothing. The weights depend on the observed time points $T_m^{(i)}$ and a bandwidth h. Let $\widehat{\lambda}_j$ and $\widehat{\psi}_j$, $j \geq 1$, denote the eigenelements obtained from $\widehat{\Gamma}$. They depend on the bandwidth h and the key issue is the choice of h.

2.2 Risk bounds and sample paths regularity

The goal is to construct adaptive estimates $\widehat{\lambda}_j$ and $\widehat{\psi}_j$ of the eigenvalues λ_j and eigenfunctions ψ_j, respectively. We consider the quadratic risks associated to the estimators $\widehat{\lambda}_j$ and $\widehat{\psi}_j$, that are

$$\mathcal{R}_N(\widehat{\lambda}_j; h) := \mathbb{E}\left[\left\{\widehat{\lambda}_j - \lambda_j\right\}^2\right] \quad \text{and} \quad \mathcal{R}_N(\widehat{\psi}_j; h) := \mathbb{E}\left\|\widehat{\psi}_j - \psi_j\right\|_2^2, \tag{4}$$

respectively. Here, $\|\cdot\|_2$ denotes the L^2- norm. The risks depend on the bandwidth and the challenge is to find a way to select the bandwidths which minimize them. By the inequality $(a+b)^2 \leq 2a^2 + 2b^2$,

$$\mathcal{R}_N(\widehat{\lambda}_j; h) \leq 2\mathbb{E}\left[\left\{\widehat{\lambda}_j - \widetilde{\lambda}_j\right\}^2\right] + 2\mathbb{E}\left[\left\{\widetilde{\lambda}_j - \lambda_j\right\}^2\right], \tag{5}$$

and

$$\mathcal{R}_N(\widehat{\psi}_j; h) \leq 2\mathbb{E}\left\|\widehat{\psi}_j - \widetilde{\psi}_j\right\|_2^2 + 2\mathbb{E}\left\|\widetilde{\psi}_j - \psi_j\right\|_2^2.$$

It is well-known that the infeasible empirical estimators $\widetilde{\lambda}_j$ and $\widetilde{\psi}_j$ converge at the parametric rate $O_{\mathbb{P}}(N^{-1/2})$; see Horváth and Kokoszka (2012). On the other hand, no estimator of the eigen-elements is expected to converge faster than the ones obtained in the ideal situation where the curves $X^{(i)}$ are observed at any time point without error. Therefore, minimizing the mean squared differences

$$\mathbb{E}\left[\left\{\widehat{\lambda}_j - \widetilde{\lambda}_j\right\}^2\right] \quad \text{and} \quad \mathbb{E}\left\|\widehat{\psi}_j - \widetilde{\psi}_j\right\|_2^2, \tag{6}$$

with respect to the bandwidth, would yield rate optimal estimators of λ_j and ψ_j, respectively. We show that the mean squared differences in (2.2) can be conveniently bounded when the covariance function is estimated using the 'first smooth, then estimate' approach. We are then able to define simple, data-driven bandwidth rules for the kernel smoothing of the curves, to be used next for building covariance estimates, from which the adaptive estimators $\widehat{\lambda}_j$ and $\widehat{\psi}_j$ are derived.

A key element for deriving workable risk bounds, is the local regularity of the process X. For the sake of simplicity, we here focus on the case where the sample paths of X are non-differentiable. There is now extensive evidence that many functional datasets can be reasonably considered as being generated by irregular sample paths $X^{(i)}$. See, for instance, Poß et al. (2020), Mohammadi and Panaretos (2021) and Mohammadi et al. (2022) for examples. Another example of a real data set, along with simulated curves
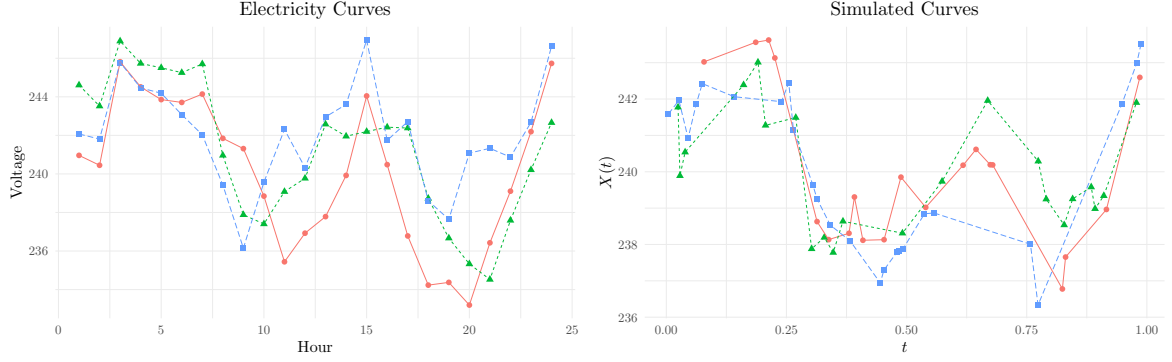
Figure 1: Sample of curves from electricity dataset (left) & simulated irregular curves mimicking the characteristics of real data (right)

based on learnt characteristics can be found in Figure 1. See Section 4 for more details on the dataset and the simulation procedure. In Section II.3 in the Supplementary Material we discuss the extension of our approach to FDA with smooth sample paths. In the case where X has non-differentiable (irregular) sample paths, we assume that functions $H : \mathcal{T} \to (0,1)$ and $L : \mathcal{T} \to (0,\infty)$ exist such that, for any $t \in \mathcal{T}$,

$$\mathbb{E}\left[(X_u - X_v)^2\right] \approx L_t^2 |u - v|^{2H_t}, \tag{7}$$

when $u \leq t \leq v$ lie in a small neighborhood of t. The formal definition behind (2.2) is provided in Section A.1. The function H provides the local Hölder exponent, while the function L gives the local Hölder constant. They are both allowed to depend on t in order to allow for curves with general patterns, in particular with varying regularity over $\mathcal{T}$. The function H is also connected to the rate of decay of eigenvalues. For illustration, consider $H \in (0,1)$ is constant. If, for some $1 < \nu < 3$, the eigenvalues λ_j have polynomial decrease rate $\lambda_j \sim j^{-\nu}$, $j \geq 1$, then, under mild conditions, H is constant and equal to $(\nu - 1)/2$. For instance, for the Brownian motion $H \equiv 0.5$ and $\nu = 2$. However, our framework is more general as we do not impose specific rates of decrease for the eigenvalues.

We show that if H and L are given, sharp bounds for the risk of kernel-based estimates of the eigenvalues and eigenfunctions can be derived. Replacing H and L by their nonparametric estimates in the risk bound of an eigen-element, yields an easy to minimize criterion in order to obtain a data-driven bandwidth adapted to the eigen-element. Building the covariance estimate $\widehat{\Gamma}$ as in (2.3) below with this bandwidth, yields the eigen-element optimal estimate.

Let us briefly describe the rationale behind the nonparametric estimators of H and L. Let

$$\theta(u, v) := \mathbb{E}\left[(X_u - X_v)^2\right], \quad u, v \in \mathcal{T},$$

and consider two points $t_1, t_3 \in \mathcal{T}$ such that t belongs to the interval defined by t_1 and t_3. Let $t_2 = (t_1 + t_3)/2$. Then, using (2.2), it easy to see that

$$\widetilde{H}_t = \frac{\log(\theta(t_1, t_3)) - \log(\theta(t_1, t_2))}{2\log(2)} \approx H_t \qquad \text{and} \qquad \widetilde{L}_t = \frac{\sqrt{\theta(t_1, t_3)}}{|t_1 - t_3|^{\widetilde{H}_t}} \approx L_t,$$

provided $|t_1 - t_3|$ is small. The estimators $\widehat{H}_t$ and $\widehat{L}_t$ of H_t and L_t are obtained by simply plugging nonparametric estimators of $\theta(\cdot, \cdot)$ into the expressions of the proxies $\widetilde{H}_t$ and $\widetilde{L}_t$. Details are provided in Section 2.6.

2.3 Covariance estimator

We now provide the formal definition of our covariance estimator $\widehat{\Gamma}$ for a given bandwidth h. The estimator is built using kernel estimators as in (2.1). With non-differentiable sample paths, the weights

5

are those of the Nadaraya-Watson estimator, that is, with the rule $0/0 = 0$,

$$W_m^{(i)}(t) = W_m^{(i)}(t; h) = K\left((T_m^{(i)} - t)/h\right)\left[\sum_{m=1}^{M_i} K\left((T_m^{(i)} - t)/h\right)\right]^{-1}.$$

The kernel K is a symmetric density that, for simplicity, we can consider to be supported on $[-1, 1]$. It is important to notice that some values $\widehat{X}_t^{(i)}$ can be degenerate, in the sense that, for some points t, $W_m^{(i)}(t) = 0$ for all $1 \leq m \leq M_i$. This likely happens for some curves in the sparse independent design case. In the common design case, either all or none of the smoothed curves are degenerated at some points t. When $\widehat{X}_s^{(i)}$ or $\widehat{X}_t^{(i)}$ is degenerate, the i-th curve will be dropped for covariance estimation at any point (s, t). This issue of curve selection is not specific to our approach. It implicitly occurs with other estimators, see for example Li and Hsing (2010), Zhang and Wang (2016) and Rubín and Panaretos (2020). An adaptive bandwidth rule should regularize for the number of curves not used in estimation. By construction, our approach includes a penalization scheme, which automatically adapts to both sparse and dense regimes in a data-driven way.

Let

$$w_i(t; h) = \begin{cases} 1 & \text{if} \quad \sum_{m=1}^{M_i} \mathbf{1}\{|T_m^{(i)} - t| \leq h\} \geq 1, \\ 0 & \text{otherwise}, \end{cases}$$

indicate if there is at least one point along the curve $X^{(i)}$ to be selected in the estimation of $X_t^{(i)}$. Let

$$\mathcal{W}_N(t; h) = \sum_{i=1}^{N} w_i(t; h) \quad \text{and} \quad \mathcal{W}_N(s, t; h) = \sum_{i=1}^{N} w_i(t; h) w_i(s; h).$$

By construction, $w_i(t; h) = 0$ if and only if $W_m^{(i)}(t; h) = 0$ for all $1 \leq m \leq M_i$. With at hand the smoothed curves $\widehat{X}^{(i)}(h)$ defined as in (2.1), a natural covariance estimator would be

$$\widehat{\Gamma}_N(s, t; h) = \frac{1}{\mathcal{W}_N(s, t; h)} \sum_{i=1}^{N} w_i(s; h) w_i(t; h) \left\{\widehat{X}_t^{(i)}(h) - \widehat{\mu}_N(t; h)\right\} \left\{\widehat{X}_s^{(i)}(h) - \widehat{\mu}_N(s; h)\right\}, \quad s, t \in \mathcal{T}, \tag{8}$$

with the mean estimator

$$\widehat{\mu}_N(t; h) = \frac{1}{\mathcal{W}_N(t; h)} \sum_{i=1}^{N} w_i(t; h) \widehat{X}_t^{(i)}(h).$$

The random positive integers $\mathcal{W}_N(t; h)$ and $\mathcal{W}_N(s, t; h)$ are thus the effective numbers of curves used in mean and covariance functions estimation, respectively.

It is well-known that a bias is usually induced on the diagonal set of the estimated covariance function. With a kernel K supported on $[-1, 1]$, the diagonal set is $|s - t| \leq 2h$. Since the eigen-elements are ultimately built by the eigen-decomposition of a covariance function, albeit with its own specific bandwidth, a correction is recommended to reduce the diagonal bias. The nonparametric covariance estimator we propose is

$$\widehat{\Gamma}(s, t; h) = \widehat{\Gamma}_N(s, t; h) - \widehat{d}_N(s, t; h), \tag{9}$$

with the diagonal correction

$$\widehat{d}_N(s, t; h) = \frac{\sigma(s)\sigma(t)}{\mathcal{W}_N(s, t; h)} \sum_{i=1}^{N} w_i(s; h) w_i(t; h) \sum_{m=1}^{M_i} W_m^{(i)}(s; h) W_m^{(i)}(t; h), \tag{10}$$

where $\sigma^2(t)$ is the conditional variance defined in (2.1). By construction, $\widehat{d}_N(s, t; h) = 0$ when $|s - t| > 2h$, and thus $\widehat{\Gamma} = \widehat{\Gamma}_N$ for (s, t) outside the diagonal set. The "first smooth, then estimate" approach induces a bias of rate $o_{\mathbb{P}}(1)$ on each curve. Our theoretical and empirical investigation show that the

diagonal correction (2.3) further improves the performance of the eigen-elements estimators. Our diagonal correction is different from that in Golovkine et al. (2023), since it corresponds to an integrated, instead of pointwise, risk.

The covariance estimator $\widehat{\Gamma}$ was formally defined in (2.3) for a generic bandwidth h. The bandwidth used to compute the adaptive estimator of an eigen-element, will be adapted for each eigen-element separately, and obtained by minimization of the corresponding risk bound presented in Section 2.4.

2.4 Risk bounds

Let $\widehat{\lambda}_j$ and $\widehat{\psi}_j$ be the $j-$th eigenvalue and the associated eigenfunction, respectively, obtained from our covariance estimator (2.3). Let

$$m_2(t) = \text{Var}(X_t) \qquad \text{and} \qquad c_2(s,t) = \text{Var}(X_s X_t), \quad s,t \in \mathcal{T}.$$

We prove in Section 3 that, up to terms which are negligible or do not depend on the bandwidth, the eigenvalue risk $\mathcal{R}_N(\widehat{\lambda}_j; h)$ can be bounded by twice

$$\begin{aligned}
\mathcal{B}_N(\widehat{\lambda}_j; h) = 4 &\int m_2(s)\psi_j^2(s)ds \int L_t^2 h^{2H_t}\psi_j^2(t)\int |u|^{2H_t}K(u)dudt \\
&+ 2\iint \left\{ \frac{\sigma^2(t)m_2(s)}{\mathcal{N}_\Gamma(s|t;h)} + \frac{\sigma^2(s)m_2(t)}{\mathcal{N}_\Gamma(t|s;h)} \right\} \psi_j^2(t)\psi_j^2(s)dtds \\
&+ \iint c_2(s,t) \left\{ \mathcal{W}_N^{-1}(s,t;h) - N^{-1} \right\} \psi_j^2(t)\psi_j^2(s)dtds \\
&=: \mathcal{B}_{1,N}(\widehat{\lambda}_j; h) + \mathcal{B}_{2,N}(\widehat{\lambda}_j; h) + \mathcal{B}_{3,N}(\widehat{\lambda}_j; h), \quad (11)
\end{aligned}$$

where

$$\mathcal{N}_\Gamma(t|s;h) = \left[\mathcal{W}_N^{-2}(s,t;h) \sum_{i=1}^{N} w_i(s;h)w_i(t;h) \max_{1 \leq m \leq M_i} |W_m^{(i)}(t;h)| \right]^{-1}.$$

Meanwhile, the part of the eigenfunction risk $\mathcal{R}_N(\widehat{\psi}_j; h)$ depending on h can be bounded using

$$\begin{aligned}
\mathcal{B}_N(\widehat{\psi}_j; h) = &\sum_{k\in\mathcal{K}:k\neq j} \frac{2}{(\lambda_j - \lambda_k)^2} \iint \psi_j^2(t)\psi_k^2(s) \\
&\times \left\{ m_2(t)L_s^2 h^{2H_s}\int |u|^{2H_s}K(u)du + m_2(s)L_t^2 h^{2H_t}\int |u|^{2H_t}K(u)du \right\} dtds \\
&+ \sum_{k\in\mathcal{K}:k\neq j} \frac{2}{(\lambda_j - \lambda_k)^2} \left[\iint \psi_k^2(s)\psi_j^2(t) \left\{ \frac{\sigma^2(s)m_2(t)}{\mathcal{N}_\Gamma(t|s;h)} + \frac{\sigma^2(t)m_2(s)}{\mathcal{N}_\Gamma(s|t;h)} \right\} dtds \right] \\
&+ \sum_{k\in\mathcal{K}:k\neq j} \frac{1}{(\lambda_j - \lambda_k)^2} \iint \psi_k^2(s)\psi_j^2(t)c_2(s,t) \left\{ \mathcal{W}_N^{-1}(s,t;h) - N^{-1} \right\} dtds \\
&=: \mathcal{B}_{1,N}(\widehat{\psi}_j; h) + \mathcal{B}_{2,N}(\widehat{\psi}_j; h) + \mathcal{B}_{3,N}(\widehat{\psi}_j; h). \quad (12)
\end{aligned}$$

The summation indices in the last three sums belong to a finite set $\mathcal{K}$, for instance the integers from 1 to some threshold K_0 to be set by the practitioner. The bounds in (2.4) and (2.4) are derived from a generalized version of the eigen-elements representations given by Hall and Hosseini-Nasab (2006) and Hall and Hosseini-Nasab (2009), using the local regularity property (2.2). Details are provided in the Appendix and the Supplementary Material.

The terms $\mathcal{B}_{1,N}(\widehat{\lambda}_j; h)$ and $\mathcal{B}_{1,N}(\widehat{\psi}_j; h)$ are the squared bias terms which depend on the local regularity, while $\mathcal{B}_{2,N}(\widehat{\lambda}_j; h)$ and $\mathcal{B}_{2,N}(\widehat{\psi}_j; h)$ are the variance terms. The terms $\mathcal{B}_{3,N}(\widehat{\lambda}_j; h)$ and $\mathcal{B}_{3,N}(\widehat{\psi}_j; h)$ are the penalty terms which regularize for the curves which are dropped in the covariance estimation procedure.

It is likely that all curves are selected with large h, so that $\mathcal{W}_N(s, t; h) = N$ and the third term vanishes, getting us back to the classical bias-variance trade off. On contrary, low h values lead to large $\mathcal{B}_{2,N}$ and $\mathcal{B}_{3,N}$ terms. In particular, in the common design case where all the M_i are equal and the set of $T_m^{(i)}$ is the same for all i, the bandwidth h cannot go below the smallest distance between two consecutive observed time points $T_m^{(i)}$. Our risk bounds realize an automatic balancing in the bandwidth selection process.

Both squared bias terms $\mathcal{B}_{1,N}(\widehat{\lambda}_j; h)$ and $\mathcal{B}_{1,N}(\widehat{\psi}_j; h)$ feature the integrated regularity of X represented by integrals of functions of t involving the factor h^{2H_t}. This governs the rate of convergence of the bias term, and a good estimator should adapt automatically to the unknown smoothness of the sample paths. We overcome this hurdle by explicitly estimating the regularity, and building estimators of eigen-elements that automatically take these smoothness estimates into account. It is important to notice that the rate of the integrals involving h^{2H_t} are driven by the smallest values of H_t when h tends to zero.

2.5 Estimation of auxiliary quantities in the risk bounds

In addition to the functions H and L which characterize the local regularity of X, the risk bounds (2.4) and (2.4) contain several other quantities which play the role of constants and can be easily estimated. We will see in Section 2.6 that estimators for the moment functions $m_2(t)$ and $c_2(s, t)$ can be obtained as byproducts of the procedure for the estimation of H and L. In order to obtain preliminary estimates of eigen-elements λ_j and ψ_j involved in the risk bounds, we simply perform a preliminary data-driven eigen-elements estimation step where we minimize the risk bound (2.4) with the ψ_j's replaced by 1. It is then easy to see from (2.4) and (2.4) that this simplification yields a same data-driven bandwidth for all eigenvalues and all eigenfunctions. We use it to compute preliminary estimates for λ_j and ψ_j.

Finally, we propose a simple estimator for the conditional variance of the error terms, from which an estimate of $\sigma^2(t)$ can be deduced. As presented in (2.1), we here allow heteroscedastic error terms

$$\varepsilon_m^{(i)} = \sigma(T_m^{(i)}) e_m^{(i)}, \qquad \forall 1 \le m \le M_i, 1 \le i \le N,$$

where the $e_m^{(i)}$'s are a random sample of a standardized variable e. Whenever the observed time points $T_m^{(i)}$ admit a density f_T which is bounded away from zero on $\mathcal{T}$, a simple estimator of $\sigma^2(t)$ is

$$\widehat{\sigma}^2(t; b) = \frac{1}{2\mathcal{I}_N(t; b)} \sum_{i=1}^N \left(Y_{m(t,i,1)}^{(i)} - Y_{m(t,i,2)}^{(i)} \right)^2 \mathcal{I}^{(i)}(t; b) \quad \text{with} \quad \mathcal{I}^{(i)}(t; b) = \mathbf{1}\left\{ \left| T_{m(t,i,2)}^{(i)} - t \right| \le b \right\}, \quad (13)$$

where $\mathcal{I}_N(t; b) = \sum_{i=1}^N \mathcal{I}^{(i)}(t; b)$, $b > 0$. For each i, $(Y_{m(t,i,1)}^{(i)}, T_{m(t,i,1)}^{(i)})$ and $(Y_{m(t,i,2)}^{(i)}, T_{m(t,i,2)}^{(i)})$ are the two pairs with the observed time points closest and the second closest to t, respectively. The estimator $\widehat{\sigma}^2(t; b)$ is inspired by Müller and Stadtmüller (1987) and adapted for FDA. Our estimator adjusts for possibly large spacings between the observed time points, which for instance is more relevant to the sparse regime. It suffices to average over the values i for which $\left| T_{m(t,i,1)}^{(i)} - T_{m(t,i,2)}^{(i)} \right|$ is sufficiently small, according to the value of b, which decreases to zero with the sample size. The rate of decrease of b has little impact when $\widehat{\sigma}^2(t; b)$ is plugged into the risk bounds (2.4) and (2.4), as only consistency is required at that level. However, the decrease of b should be suitably set when $\widehat{\sigma}^2(t; b)$ is used in the final diagonal bias correction (2.3). Details on the data-driven rules to choose b in both situations are provided in the next section.

2.6 The algorithm

We now summarize our adaptive FPCA methodology which is based on the minimization of the risk bounds (2.4) and (2.4) with respect to h in a bandwidth range $\mathcal{H}_N$. Our algorithm is fast and the tuning parameters are fixed in a data-driven way. Since the risk bounds depend on the unknown eigenvalues λ_j and eigenfunctions ψ_j, to make our procedure fully self-contained, we propose to run it twice, as explained at the end of this section.

Adaptive FPCA Algorithm

Input Proxies of λ_j and ψ_j; default values for the bandwidth b; $\zeta \in (0, 1/2)$; integer threshold K_0; bandwidth range $\mathcal{H}_N$.

Step 1: **Presmoothing.** Presmooth each curve i, for example using the Nadaraya-Watson estimator with a simple bandwidth rule. Denote by $\widetilde{X}_t^{(i)}$ the presmoothing estimator of $X^{(i)}$ at point t.

Step 2: **Estimation of H and L.** For any t, choose suitable $t_1, t_2, t_3 \in \mathcal{T}$, such that $|t_1 - t_3| = 2|t_1 - t_2|$ and t belongs to the interval defined by t_1 and t_3. Then build estimates

$$\widehat{H}_t = \frac{\log\big(\widehat{\theta}(t_1, t_3)\big) - \log\big(\widehat{\theta}(t_1, t_2)\big)}{2\log(2)}, \quad \widehat{L}_t^2 = \frac{\widehat{\theta}(t_1, t_3)}{|t_1 - t_3|^{2\widehat{H}_t}} \quad \text{with} \quad \widehat{\theta}(u, v) = \frac{1}{N}\sum_{i=1}^{N}\left\{\widetilde{X}_u^{(i)} - \widetilde{X}_v^{(i)}\right\}^2. \tag{14}$$

Step 3: **Estimation of the moment functions.** Build the estimates

$$\widehat{m}_2(t) = \frac{1}{N}\sum_{i=1}^{N}\left\{\widetilde{X}_t^{(i)}\right\}^2 - \left\{\frac{1}{N}\sum_{i=1}^{N}\widetilde{X}_t^{(i)}\right\}^2, \quad \widehat{c}_2(s,t) = \frac{1}{N}\sum_{i=1}^{N}\left\{\widetilde{X}_s^{(i)}\widetilde{X}_t^{(i)}\right\}^2 - \left\{\frac{1}{N}\sum_{i=1}^{N}\widetilde{X}_s^{(i)}\widetilde{X}_t^{(i)}\right\}^2,$$

of $m_2(t) = \mathrm{Var}(X_t)$ and $c_2(s,t) = \mathrm{Var}(X_s X_t)$, respectively.

Step 4: **Estimation of the conditional variance.** Apply (2.5) and build the estimate $\widehat{\sigma}^2(t; b)$ of the errors' conditional variance $\sigma^2(t)$.

Step 5: **Minimize the risk bounds with respect to h.** Let $\widehat{\mathcal{B}}_N(\widehat{\lambda}_j; h)$ and $\widehat{\mathcal{B}}_N(\widehat{\psi}_j; h)$ be the functions of h obtained by replacing in the expressions of $\mathcal{B}_N(\widehat{\lambda}_j; h)$ and $\mathcal{B}_N(\widehat{\psi}_j; h)$ the unknown quantities with the estimates from Step 3 and Step 4. Numerically approximate the integrals in (2.4) and (2.4) and compute $\widehat{\mathcal{B}}_N(\widehat{\lambda}_j; h)$ and $\widehat{\mathcal{B}}_N(\widehat{\psi}_j; h)$ for h on a grid in a range of bandwidths $\mathcal{H}_N$. For the computation of $\widehat{\mathcal{B}}_N(\widehat{\psi}_j; h)$, take $\mathcal{K} = \{1, \ldots, K_0\}$ in (2.4). Select the bandwidths $\widehat{h}^*(\lambda_j)$ and $\widehat{h}^*(\psi_j)$ that minimize $\widehat{\mathcal{B}}_N(\widehat{\lambda}_j; h)$ and $\widehat{\mathcal{B}}_N(\widehat{\psi}_j; h)$, respectively.

Step 6: **Adaptive covariance function estimation and eigen-decomposition.** For each $j \geq 1$, compute the covariance estimates $\widehat{\Gamma}(s, t; \widehat{h}^*(\lambda_j))$ and $\widehat{\Gamma}(s, t; \widehat{h}^*(\psi_j))$ using definition (2.3). For the diagonal corrections (2.3), use $\widehat{\sigma}^2(\cdot; b^*)$ from (2.5) instead of $\sigma^2(\cdot)$, with $b^*(\lambda_j) = \{\widehat{h}^*(\lambda_j)\}^{1-\zeta}$ and $b^*(\psi_j) = \{\widehat{h}^*(\psi_j)\}^{1-\zeta}$, respectively. Build the adaptive estimates $\widehat{\lambda}_j$ and $\widehat{\psi}_j$ with the covariance estimates $\widehat{\Gamma}(s, t; \widehat{h}^*(\lambda_j))$ and $\widehat{\Gamma}(s, t; \widehat{h}^*(\psi_j))$, respectively.

For presmoothing in Step 1 we propose a simple least-squares cross-validation (LS-CV) approach. We use LS-CV on a small subset of curves, for instance 20 curves, and take the median of the 20 bandwidths selected. With this bandwidth we smooth separately each curve i in the sample at the observed points $T_m^{(i)}$, and between the $T_m^{(i)}$'s we simply use linear interpolation. Details on the choice of t_1, t_3 in Step 2 are provided in Theorem 3 below. A rich toolkit for practitioners to perform eigenanalysis in Step 6 is available, see for instance Chapter 8.4 of Ramsay and Silverman (2005).

For a self-contained procedure, we first run the algorithm with a *ad-hoc* proxy for all ψ_j considered, that is the constant function equal to 1. The λ_j's then no longer matter in the optimization of the feasible bounds $\widehat{\mathcal{B}}_N(\psi_j; h)$. In this simplified first run, since b only serves to calculate constants, with independent design we propose a default choice $b = \mathrm{length}(\mathcal{T})/10$. In the common design case, we simply take b equal to the largest spacing between consecutive observed points $T_m^{(i)}$ and $T_{m+1}^{(i)}$. The simplified first run yields preliminary estimates for λ_j and ψ_j obtained in Step 6 from a single adaptive, data-driven bandwidth,

say, $\widehat{h}^*$. For the diagonal correction, we use the variance estimator with $b = \{\widehat{h}^*\}^{1-\zeta}$, for some $\zeta > 0$. We propose the default value $\zeta = 0.1$. Next, in both independent and common design cases, we run again the Steps 5 to 6 of the algorithm with the preliminary estimates of λ_j and ψ_j. We then have an optimal bandwidth $\widehat{h}^*$ and an associated b^* for each eigen-element separately. The rationale for the choice of b^* in Step 6 is provided by Lemmas A.6 and A.7 in the Appendix.

3   Theoretical properties

Our approach adapts to the regularity of the process X, a notion informally introduced in (2.2). The idea of local regularity estimators (2) was introduced by Golovkine et al. (2022). For the purpose of FPCA, we reconsider their construction and provide new exponential bounds for the uniform concentration of the regularity estimators $\widehat{H}$ and $\widehat{L}$, a result of independent interest. However, for the sake of readability and given our focus on the new approach for FPCA, we postpone the formal definition of the local regularity and the uniform concentration bounds of $\widehat{H}$ and $\widehat{L}$ to the Appendix. See Sections A.1 and A.2.

In the following, we state the formal results on the risk bounds for the risks defined in (2.2). Next, we provide the rate of the bandwidths minimizing the risk bounds, from which we derive the rates of convergence for our adaptive estimators of the eigen-elements. Finally, we show that using the feasible risk bounds $\widehat{\mathcal{B}}_N(\widehat{\lambda}_j; h)$ and $\widehat{\mathcal{B}}_N(\widehat{\psi}_j; h)$, where the unknown quantities are replaced by estimates, instead of $\mathcal{B}_N(\widehat{\lambda}_j; h)$ and $\mathcal{B}_N(\widehat{\psi}_j; h)$, does not alter the rates of convergence. The assumptions used for the theoretical results are provided in the Appendix. Recall that we consider λ_j decreasingly ordered.

3.1   Risk bounds and rates of convergence for eigen-elements estimators

Our first result shows that, modulo a constant and terms not depending on h, the quadratic risks $\mathcal{R}_N(\widehat{\lambda}_j; h)$ and $\mathcal{R}_N(\widehat{\psi}_j; h)$ can be bounded by $\mathcal{B}_N(\widehat{\lambda}_j; h)$ and $\mathcal{B}_N(\widehat{\psi}_j; h)$ defined in (2.4) and (2.4), respectively.

Theorem 1. *Let Assumption 1 hold true and let $\mathcal{H}_N$ be a bandwidth range as in Assumption 3. For $j \geq 2$, assume*

$$\min(\lambda_j - \lambda_{j+1}, \lambda_{j-1} - \lambda_j) > 0,$$

and $\lambda_1 - \lambda_2 > 0$ when $j = 1$. Then, uniformly over $\mathcal{H}_N$

$$\mathcal{R}_N(\widehat{\lambda}_j; h) \leq 2\mathcal{B}_N(\widehat{\lambda}_j; h)\{1 + o_{\mathbb{P}}(1)\} + terms\ not\ depending\ on\ h.$$

Moreover, given a constant $C > 0$, an integer K_0 depending on C exists such that, for $\mathcal{B}_N(\widehat{\psi}_j; h)$ defined as in (2.4) with $\mathcal{K} = \{1, \ldots, K_0\}$, we have

$$\mathcal{R}_N(\widehat{\psi}_j; h) \leq \{2 + C\}\mathcal{B}_N(\widehat{\psi}_j; h)\{1 + o_{\mathbb{P}}(1)\} + terms\ not\ depending\ on\ h.$$

We next study the rates of the bandwidths that minimize the risk bounds $\mathcal{B}_N(\widehat{\lambda}_j; h)$ and $\mathcal{B}_N(\widehat{\psi}_j; h)$. These drive the rates of convergence of the quadratic risks $\mathcal{R}_N(\widehat{\lambda}_j; h)$ and $\mathcal{R}_N(\widehat{\psi}_j; h)$, respectively. The rates of $\mathcal{B}_N(\widehat{\lambda}_j; h)$ and $\mathcal{B}_N(\widehat{\psi}_j; h)$, and the corresponding minimum h, are determined by

$$\underline{H} = \min_{t \in \mathcal{T}} H_t.$$

While in practice the numerical optimization will select the bandwidth without particular difficulty, in general the exact theoretical rate of the bandwidth can only be determined up to a logarithmic factor; see Lemma A.2 in the Appendix. The reason is the exact rate of an integral involving the factor h^{2H_t}, which depends on the behavior of H_t in the neighborhood of its minimum. If for instance, H is constant equal to $\underline{H}$ on a non-degenerate interval, the integrals have the rate given by $h^{2\underline{H}}$. If $\underline{H}$ is attained at isolated points, the rate of the integrals depend on the local behavior of H in the neighborhood of these

points. We therefore present a range of rates for the bandwidth and an upper bound for the rates of the risks, under the mild condition of Lipschitz continuity for the function H.

Let us first consider the case of independent design and define

$$\underline{r}(N, \mathfrak{m}) = \max \left\{ \left(\frac{1}{N\mathfrak{m}^2} \right)^{\frac{1}{2H+2}}, \left(\frac{1}{N\mathfrak{m}} \right)^{\frac{1}{2H+1}} \right\},$$

and

$$\overline{r}(N, \mathfrak{m}) = \max \left\{ \left(\frac{\log(N\mathfrak{m})}{N\mathfrak{m}^2} \right)^{\frac{1}{2H+2}}, \left(\frac{\log(N\mathfrak{m})}{N\mathfrak{m}} \right)^{\frac{1}{2H+1}} \right\}.$$

In the following, for a sequence of strictly positive random variables U_n and a sequence of numbers $b_n > 0$, the notation $U_n \lesssim_{\mathbb{P}} b_n$ means $U_n = O_{\mathbb{P}}(b_n)$, while $b_n \lesssim_{\mathbb{P}} U_n$ means $1/U_n = O_{\mathbb{P}}(1/b_n)$.

Corollary 1. *Assume the conditions of Theorem 1 hold true. Let $h^*(\widehat{\lambda}_j)$ and $h^*(\widehat{\psi}_j)$ be the bandwidths obtained by minimization over $\mathcal{H}_N$ of $\mathcal{B}_N(\widehat{\lambda}_j; h)$ and $\mathcal{B}_N(\widehat{\psi}_j; h)$, respectively. Then, in the case of independent design,*

$$\underline{r}(N, \mathfrak{m}) \lesssim_{\mathbb{P}} \min \left\{ h^*(\widehat{\lambda}_j), h^*(\widehat{\psi}_j) \right\} \leq \max \left\{ h^*(\widehat{\lambda}_j), h^*(\widehat{\psi}_j) \right\} \lesssim_{\mathbb{P}} \overline{r}(N, \mathfrak{m}).$$

Moreover, if $\widehat{\lambda}_j$ and $\widehat{\psi}_j$ are obtained with the bandwidths $h^*(\widehat{\lambda}_j)$ and $h^*(\widehat{\psi}_j)$, respectively, then

$$\mathbb{E}\left[\left\{ \widehat{\lambda}_j - \lambda_j \right\}^2 \right] + \mathbb{E}\left\| \widehat{\psi}_j - \psi_j \right\|_2^2 \lesssim_{\mathbb{P}} \overline{r}(N, \mathfrak{m})^{2H} + N^{-1}.$$

For simplicity, in the case of common design, we impose the following mild assumption on the spacings between consecutive observed time points: with $T_m^{(i)}$ increasingly ordered, a constant $C_d > 0$ exists such that

$$C_d^{-1}\mathfrak{m} \leq \left| T_{m+1}^{(i)} - T_m^{(i)} \right| \leq C_d\mathfrak{m}, \qquad \forall m \in \{1, \ldots, \mathfrak{m}\}. \tag{15}$$

Corollary 2. *Assume the conditions of Theorem 1 and (3.1) hold true. Let $h^*(\widehat{\lambda}_j)$, $h^*(\widehat{\psi}_j)$, $\widehat{\lambda}_j$ and $\widehat{\psi}_j$ be defined as in Corollary 1. Then, in the case of common design,*

$$\max \left\{ \mathfrak{m}^{-1}, (N\mathfrak{m})^{-\frac{1}{2H+1}} \right\} \lesssim_{\mathbb{P}} \min \left\{ h^*(\widehat{\lambda}_j), h^*(\widehat{\psi}_j) \right\}$$

$$\leq \max \left\{ h^*(\widehat{\lambda}_j), h^*(\widehat{\psi}_j) \right\} \lesssim_{\mathbb{P}} \max \left\{ \log(\mathfrak{m})/\mathfrak{m}, \{\log(N\mathfrak{m})/(N\mathfrak{m})\}^{\frac{1}{2H+1}} \right\}.$$

Moreover,

$$\mathbb{E}\left[\left\{ \widehat{\lambda}_j - \lambda_j \right\}^2 \right] + \mathbb{E}\left\| \widehat{\psi}_j - \psi_j \right\|_2^2 \lesssim_{\mathbb{P}} \{\log(\mathfrak{m})/\mathfrak{m}\}^{2H} + N^{-1}.$$

There are significant differences in the rates for the common design case. These are caused by the penalty terms $\mathcal{B}_{3,N}(\widehat{\lambda}_j; h)$ and $\mathcal{B}_{3,N}(\widehat{\psi}_j; h)$ in (2.4) and (2.4) which prevent the bandwidths to be smaller than half of the spacings between consecutive observed time points. This aspect is automatically taken into account by our risk bounds which can be used with both independent or common design.

Finally, it remains to show that, up to constants and uniformly with respect to the bandwidth, the difference between the estimated bounds $\widehat{\mathcal{B}}_N(\widehat{\lambda}_j; h)$ and $\widehat{\mathcal{B}}_N(\widehat{\psi}_j; h)$ from Step 5 of the Algorithm in Section 2.6, and the infeasible ones in (2.4) and (2.4), is negligible. The following result applies to both first and second runs of the Algorithm.

Theorem 2. *Let Assumptions 1, 2 and 3 in the Appendix hold true. Then, uniformly over $\mathcal{H}_N$,*

$$\widehat{\mathcal{B}}_N(\widehat{\lambda}_j; h) = \mathcal{B}_N(\widehat{\lambda}_j; h)\{1 + o_{\mathbb{P}}(1)\} \quad and \quad \widehat{\mathcal{B}}_N(\widehat{\psi}_j; h) = \mathcal{B}_N(\widehat{\psi}_j; h)\{1 + o_{\mathbb{P}}(1)\}.$$

Theorem 2 guarantees that Corollaries 1 and 2 remain true with $h^*(\widehat{\lambda}_j)$ and $h^*(\widehat{\psi}_j)$ replaced by the feasible versions $\widehat{h}^*(\widehat{\lambda}_j)$ and $\widehat{h}^*(\widehat{\psi}_j)$, and the eigen-elements $\widehat{\lambda}_j$ and $\widehat{\psi}_j$ obtained with these feasible bandwidths.

4 Numerical properties

We below provide the results of an extensive simulation study. The simulated data mimic the characteristics of a real data set of power consumption curves described in Section 4.3. The results are obtained using the package FDAdapt dedicated to our adaptive FPCA. The new, adaptive approach performs well when compared to existing FPCA approaches.

Simulated functional data are generated according to property (2.2). For this purpose, we build in Section 4.1 a wide class of Gaussian processes which satisfy (2.2) with given H and L. This class is interesting *per se*, as it offers an easy to implement and effective simulation setup that allows the data generating process (DGP) to inherit the characteristics of real data.

4.1 A general purpose simulator

The class of processes satisfying (2.2) is general. Examples include, but are not limited to stationary or stationary increment processes. See Golovkine et al. (2022) for examples. Here, we consider the example of multifractional Brownian motion (MfBm) processes. See, *e.g.*, Balança (2015) and the references therein for the formal definition. The MfBm, say $W = (W(t))_{t \geq 0}$, with Hurst index function, say $t \mapsto H_t \in (0, 1)$, is a centered Gaussian process with covariance function

$$C(s,t) = \mathbb{E}\left[W(s)W(t)\right] = D(H_s, H_t)\left[s^{H_s + H_t} + t^{H_s + H_t} - |t - s|^{H_s + H_t}\right], \qquad s, t \geq 0,$$

where

$$D(x, y) = \frac{\sqrt{\Gamma(2x+1)\Gamma(2y+1)\sin(\pi x)\sin(\pi y)}}{2\Gamma(x+y+1)\sin(\pi(x+y)/2)}, \qquad D(x, x) = 1/2, \qquad x, y > 0.$$

We show in Lemma SM.1 in the Supplementary Material that W satisfies (2.2) with the Hurst index as exponent and $L \equiv 1$. The fractional Brownian motion is an example of MfBm with constant Hurst index.

In order to accommodate more complex covariance functions that may be encountered in practice, in particular to allow more general functions L and variance functions, we consider a deterministic time deformation defined by a map $A : \mathbb{R}_+ \to \mathbb{R}_+$. This map is assumed to be continuously differentiable, with strictly positive derivative. Let A^{-1} be the inverse of A, and

$$H_{A,u} = H_{A^{-1}(u)}.$$

Given the function H and the time deformation A, the processes X we consider in our simulations are defined as

$$X(t) = \mu(t) + \tau(t)W_A(A(t)), \qquad t \geq 0, \tag{16}$$

where $(W_A(u))_{u \geq 0}$ is a MfBm with Hurst index function $H_{A,u}$, μ is a mean function, and τ is a scaling function. The covariance function of the Gaussian process X is

$$C_A(s,t) = \tau(s)\tau(t)D(H_s, H_t)\left[A(s)^{H_s + H_t} + A(t)^{H_s + H_t} - |A(t) - A(s)|^{H_s + H_t}\right], \tag{17}$$

and allows for flexible patterns matching real data characteristics. It suffices to suitably define the deformation A and the scaling τ. In our simulations, we consider two situations. First, we consider A

and τ such that X defined in (4.1) satisfies (2.2) with given functions H and L. In this case, we take

$$A(t) = A(0) + \int_0^t L_s^{1/H_s} ds, \quad A(0) \geq 0, \qquad \text{and} \quad \tau(t) = 1, \ t \in \mathcal{T},$$

and we have $\text{Var}(X(t)) = A(t)^{2H_t}$. In the second case, we want X to match functions H, L and a given variance function $m_2(t) = \text{Var}(X(t))$. For this, we consider

$$A(t) = A(0) \exp\left(\int_0^t \{m_2(s)^{-1} L_s\}^{1/H_s} ds\right), \quad A(0) > 0, \qquad \text{and} \quad \tau(t) = m_2(t) A(t)^{-H_t}, \ t \in \mathcal{T}. \quad (18)$$

Then X defined in (4.1) still satisfies (2.2), and $\text{Var}(X(t)) = m_2(t)$. The justification of these statements is provided in Section I of the Supplementary Material.

4.2 Data generating process

We use a data set containing household electric power consumption, which can be downloaded at `https://archive.ics.uci.edu/ml/datasets/individual+household+electric+power+consumption`. Different electrical quantities are measured, with 1 minute sampling rates over a period of 4 years, resulting in almost 2 million data points.

To restrict ourselves to an univariate analysis, we focus on the daily voltage variable, considering days without missing values. This data subset contains 1351 voltage curves observed on an equally spaced grid of 1440 sampling points which are re-scaled on $[0, 1]$. We estimate the local regularity parameters H_t and L_t according to Step 2 of the Algorithm, with simple linear interpolation presmoothing. See also the procedure described in the Section A.2. The variance function $m_2(t) = \text{Var}(X_t)$ is similarly estimated by interpolation according to Step 3. These three quantities were estimated on a grid containing 40 equally spaced evaluation points in the unit interval. We then smooth the estimated parameters with 9 Fourier basis functions. The constructed functions can be seen in Figure 2. Finally, we also define a mean function $\mu(t)$ as the empirical mean of the densely sampled electricity curves, shifted downward by 240, and smoothed by running a LASSO regression on linear combinations of sine and cosine functions using the `glmnet` package. A plot of the mean function can be seen in Figure 4, full details are provided in Section VI in the Supplement.

A time deformation $A(t)$ and a scaling function $\tau(t)$ are built using (4.1) and the three constructed functions H, L and m_2. The covariance function (4.1) and the mean function $\mu(t)$ are then used to simulate the sample paths. More precisely, for each $1 \leq i \leq N$, after drawing M_i and the $T_m^{(i)}$'s, we use that covariance function and the mean vector $(\mu(T_1^{(i)}), \ldots, \mu(T_{M_i}^{(i)}))$ to generate the Gaussian random vectors $(X^{(i)}(T_1^{(i)}), \ldots, X^{(i)}(T_{M_i}^{(i)}))$, and ultimately the values $Y_1^{(i)}, \ldots, Y_{M_i}^{(i)}$ using (2.1), by adding a heteroscedastic noise. Two levels of noise are considered. A plot of the square root of the variance function of the noise in the lower noise case can be seen in Figure 3. In the higher noise case, the signal-to-noise is divided by 4. We thus get random pairs $(Y_m^{(i)}, T_m^{(i)})$ corresponding to noisy measurements of sample paths from a process as in (4.1) and (4.1).

We consider 16 experimental setups, each consisting of 500 replications. They correspond to all combinations of the varying parameters $N \in \{50, 100, 200\}$, $\mathfrak{m} \in \{25, 50, 100\}$, and $\sigma_0 \in \{0.25, 1\}$, except for the two combinations with $(N = 100, \mathfrak{m} = 100)$, due to the excessive computational time required for the competing methods. The number of points M_i are generated from a Poisson distribution with mean parameter $\mathfrak{m}$. In all experiments, the sampling points $T_m^{(i)}$ were generated from a uniform distribution in $[0, 1]$, and the true eigen-elements used for computing the comparison measures were obtained by taking an accurate numerical eigen-decomposition of the true covariance function constructed. We focus on the first 9 eigen-elements because they accounted for more that 97.5% of explained variance; see Table 1 for more details on the eigenvalues and their cumulative explained variance. A plot of the covariance function along with its first 2 eigenfunctions can be seen in Figure 4.

Table 1: True Eigenvalues & their Cumulative Explained Variance (CEV)

Eigenvalues	λ_1	λ_2	λ_3	λ_4	λ_5	λ_6	λ_7	λ_8	λ_9
Value	7.354	0.296	0.152	0.074	0.039	0.035	0.023	0.0167	0.0161
CEV	89.7%	93.3%	95.2%	96.1%	96.6%	97%	97.3%	97.5%	97.7%
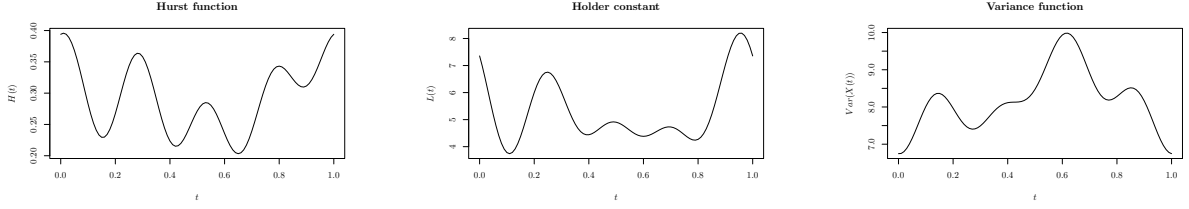
Figure 2: Local Regularity Parameters H and L and the Variance Function m_2 of X
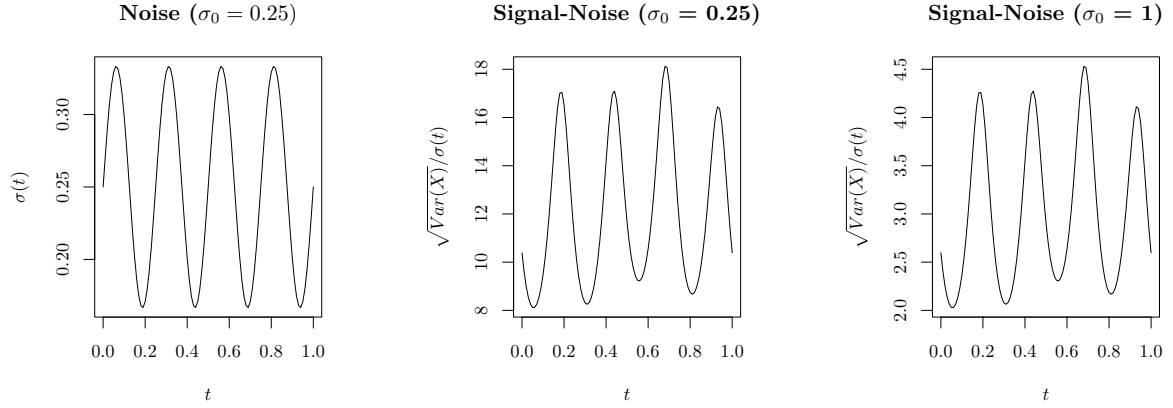

Figure 3: Noise (left plot) & Signal-to-Noise Ratios (middle and right plot)
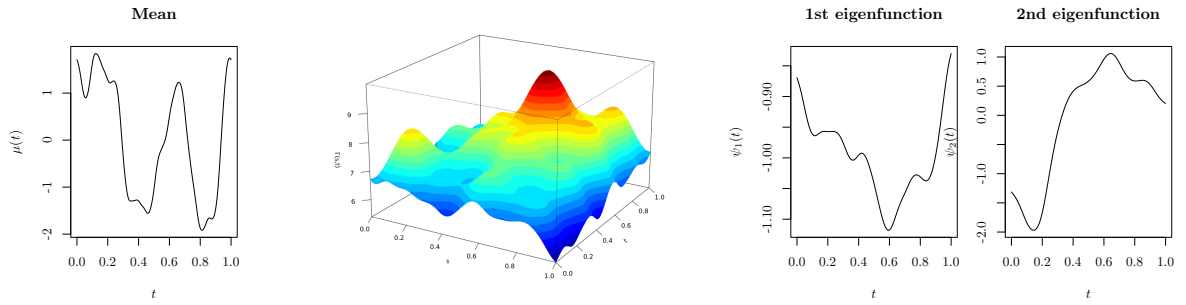

Figure 4: True Mean, Covariance & Eigenfunctions

14

4.3 Parameter settings, estimation and comparison measures

With a simulated sample of pairs $(Y_m^{(i)}, T_m^{(i)})$, we proceed as indicated in Section 2.6, running twice the algorithm for the sake of self-consistency. To get the $\widetilde{X}^{(i)}$'s, we perform local constant kernel presmoothing of our curves at the points $\{0, T_1^{(i)}, \ldots, T_{M_i}^{(i)}, 1\}$, and we linearly interpolate between these points. The bandwidth used for presmoothing is learnt by LS-CV on a randomly selected subset of 20 curves, where we selected the median of the 20 bandwidths. For the estimation of $\widehat{H}_t$ and $\widehat{L}_t$, as stated in Theorem 3 in the Appendix, a parameter $\gamma \in (0, 1)$ determining the spacings Δ_* has also to be fixed. This parameter, which defines the "local neighborhood" for estimating H_t and L_t, is set equal to 0.75. Slightly different other values we considered yielded very similar results. Then, $\widehat{H}_t$ and $\widehat{L}_t$ is first computed on a grid $\mathcal{T}_{param} \subset [0, 1]$. To ensure enough points in the local neighborhoods when estimating the local regularity parameters, we adapt the size of the parameter estimation grid such that $\#\mathcal{T}_{param} = [\mathfrak{m}/3]$. Here, $[a]$ denotes the integer part of a. In the applications, $\mathfrak{m}$ is simply estimated by $\overline{M}$, the average of the M_i's. We next complete the estimates $\widehat{H}_t$ and $\widehat{L}_t$ on a more refined grid $\mathcal{T}_{smooth}$ with 101 points, by applying a smoothing splines procedure on the unit interval, with a data-driven number of knots $[\mathfrak{m}/4] + 1$. In Steps 3 and 4 of the Algorithm, we start by computing $\widehat{m}_2$, $\widehat{\sigma}^2$ and $\widehat{c}_2$ on the grids $\mathcal{T}_{smooth}$ and $\mathcal{T}_{smooth} \times \mathcal{T}_{smooth}$, respectively. Finally, in Step 6 where the diagonal correction is performed, we set $\zeta = 0.1$.

Subsequently, the optimal bandwidth values $\widehat{h}^*(\lambda_j)$ and $\widehat{h}^*(\psi_j)$ were computed for each $1 \leq j \leq J = 9$, over a geometric grid $\mathcal{H}_N$ of 61 points, with

$$\min \mathcal{H}_N = \log(N)/(\overline{M}\sqrt{N}) \qquad \text{and} \qquad \max \mathcal{H}_N = 0.1.$$

The definition of the left end-point of $\mathcal{H}_N$ is in line with condition (2) on the bandwidth range. We learn from Lemma 2 in the Appendix, that numerical approximation of the squared bias terms $\mathcal{B}_{1,N}(\widehat{\lambda}_j; h)$ and $\mathcal{B}_{1,N}(\widehat{\psi}_j; h)$ could artificially increase the squared bias by a $\log(1/h)$ factor. Then the Algorithm would yield an optimal bandwidth diminished by a logarithmic factor. Based on these considerations and our extensive simulation study, we finally apply an upward correction of the values $\widehat{h}^*(\lambda_j)$ and $\widehat{h}^*(\psi_j)$ obtained by minimization of $\widehat{\mathcal{B}}_N(\widehat{\lambda}_j; h)$ and $\widehat{\mathcal{B}}_N(\widehat{\psi}_j; h)$, respectively. More precisely, we redefine the final bandwidths as

$$\widehat{h}^*_{WPK}(\lambda_j) = \log(1/\widehat{h}^*(\lambda_j)) \times \widehat{h}^*(\lambda_j) \quad \text{and} \quad \widehat{h}^*_{WPK}(\psi_j) = \log(1/\widehat{h}^*(\psi_j)) \times \widehat{h}^*(\psi_j),$$

and compute the our adaptive eigenvalues and eigenfunctions estimates, denoted $\widehat{\lambda}_{j,WPK}$ and $\widehat{\psi}_{j,WPK}$, using these bandwidths. The bi-dimensional grid for computing the final covariance estimates is $\mathcal{T}_{smooth} \times \mathcal{T}_{smooth}$.

Our estimates $\widehat{\lambda}_{j,WPK}$ and $\widehat{\psi}_{j,WPK}$ are then compared to the eigenvalue and eigenfunction estimates of Zhang and Wang (2016), obtained from the **R** package `fdapace`. See Zhou et al. (2022). The estimation of the eigen-elements of Zhang and Wang (2016) requires a bandwidth selection. Due to the exorbitant computational time of cross-validation, we performed comparisons against the default bandwidth setting $b_{ZW,2} = 0.1$ in the implementation Zhou et al. (2022), corresponding to 10% of the width of the sampling interval $\mathcal{T}$. We denote the corresponding eigen-element estimates as $(\widehat{\lambda}_{j,ZW,2}, \widehat{\psi}_{j,ZW,2})$. Additional comparisons against a different default bandwidth $b_{ZW,1} = 0.05$ can be found in the Supplementary Material. As a measure of comparison for the eigenvalues, we use the ratio of absolute error

$$\mathcal{R}_{ZW,2}(\lambda_j) = \left|\widehat{\lambda}_{j,WPK} - \lambda_j\right| \Big/ \left|\widehat{\lambda}_{j,ZW,2} - \lambda_j\right|.$$

For the eigenfunctions, we use the ratio of the L^2−norm errors, given by the square root of

$$\mathcal{R}^2_{ZW,2}(\psi_j) = \frac{\int \left(\widehat{\psi}^c_{j,WPK}(t) - \psi_j(t)\right)^2 dt}{\int \left(\widehat{\psi}^c_{j,ZW,2}(t) - \psi_j(t)\right)^2 dt}, \quad \text{where} \quad \widehat{\psi}^c_j = \text{sign}\left(\langle \widehat{\psi}_j, \psi_j \rangle\right) \times \widehat{\psi}_j.$$

4.4   Empirical results

The results are presented in Figures 5 to 8. For clearer visualization of the plots, in all our boxplots, we remove values below the 5th quantile and above the 95th quantile, since the number of extreme values on each ends were roughly the same. It can be seen that our estimates of the eigen-elements perform favorably relative to Zhang and Wang (2016), where for virtually all setups, our error ratios are below 1, the benchmark for equal performance. Our estimates obtain up to 4 times less error for the eigenvalues, and almost 2 times less error for the eigenfunctions, demonstrating a big improvement. More empirical results are available in the Supplementary Material, which include comparisons to other default settings of Zhou et al. (2022), lower signal-to-noise ratios, and a simpler simulation setup involving different constant Hurst functions for the fractional Brownian motion. We also consider a setup with common design. The conclusion is always the same : our method clearly outperforms the competitors.
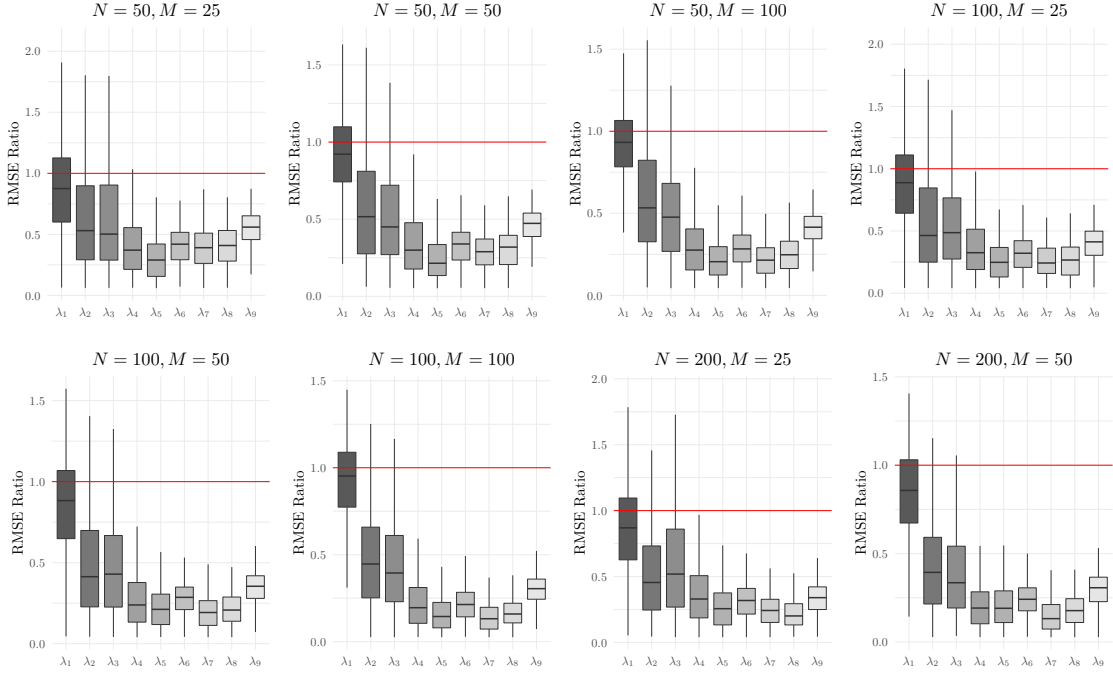

Figure 5: Eigenvalue boxplots for $\sigma_0 = 0.25$, $b_{ZW,2} = 0.1$

Due to the explicit nature of our risk bounds, our bandwidth rule is also computationally efficient, even accounting for the various quantities that need to be estimated in calculations. As seen in Figure 9, our computational times are comparable for small sample sizes, in the sense that it does not take more than 100 seconds to run. However, for larger sample sizes, say with 100 curves and 100 points on average along each curve, our approach performs significantly faster by around an order of magnitude. Furthermore, this comparison is with respect to the default setting, plug-in bandwidths in Zhou et al. (2022). With cross-validation instead of the default setting, the comparison of the computational times will be significantly more favorable for us.

A   Appendix

Our approach adapts to the regularity of the process, a notion introduced in (2.2) and formally defined in Section A.1. In Section A.2 we provide a new uniform concentration result for the regularity estimators.
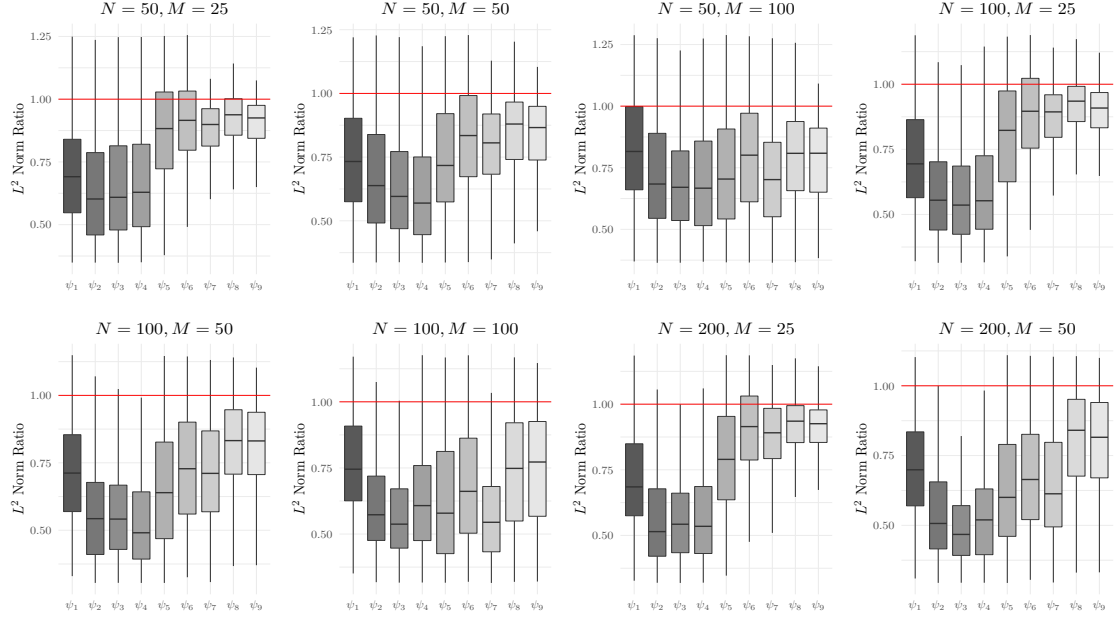
Figure 6: Eigenfunction boxplots for $\sigma_0 = 0.25$, $b_{ZW,2} = 0.1$
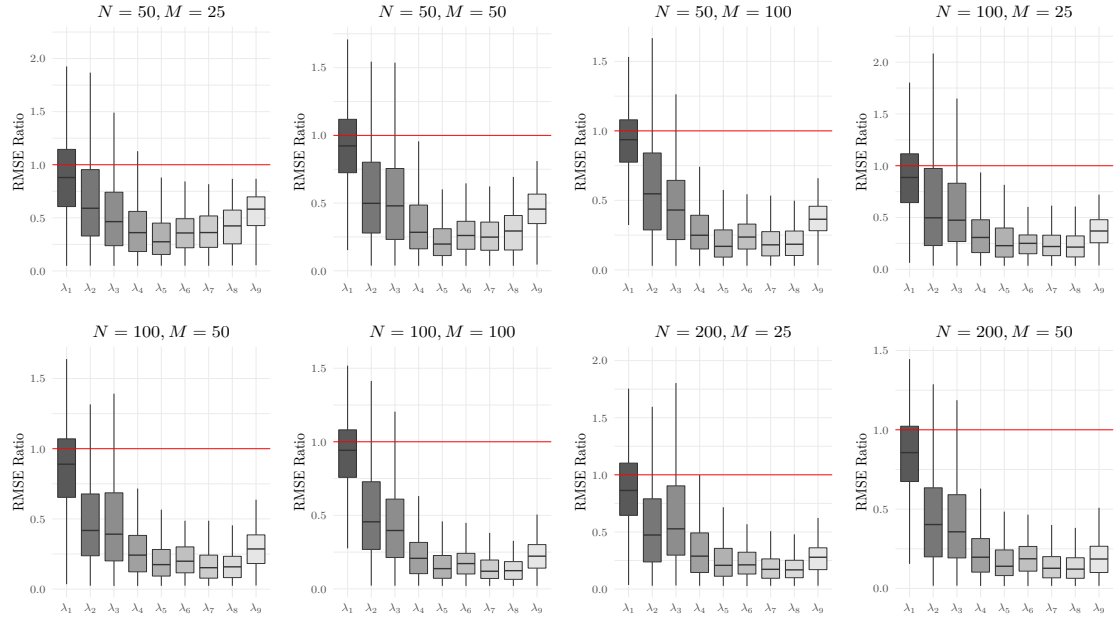

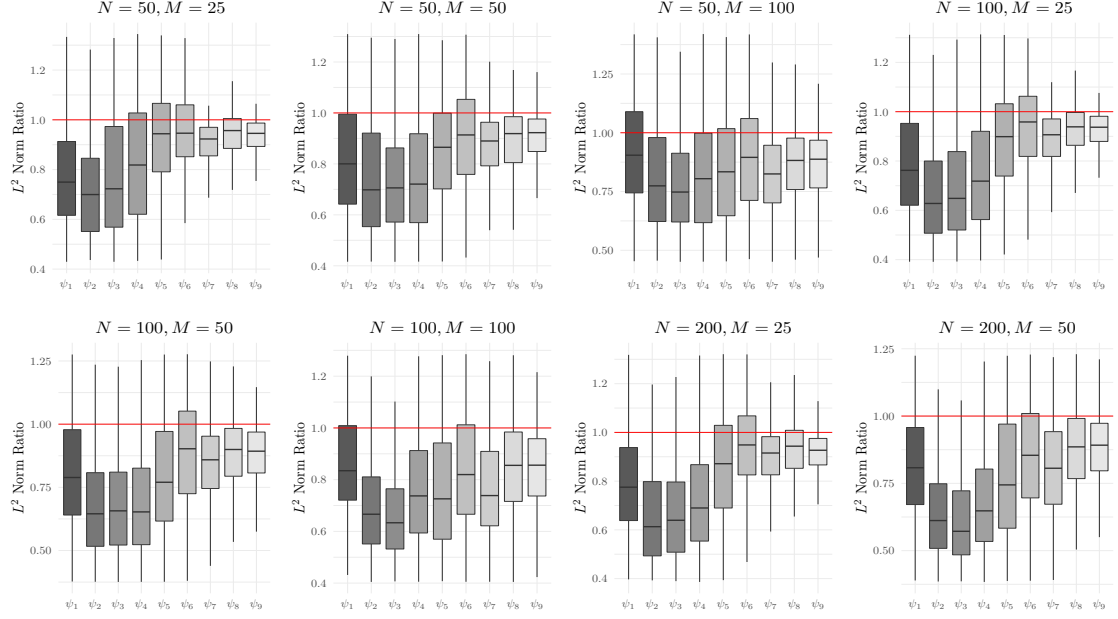Figure 7: Eigenvalue boxplots for $\sigma_0 = 1$, $b_{ZW,2} = 0.1$

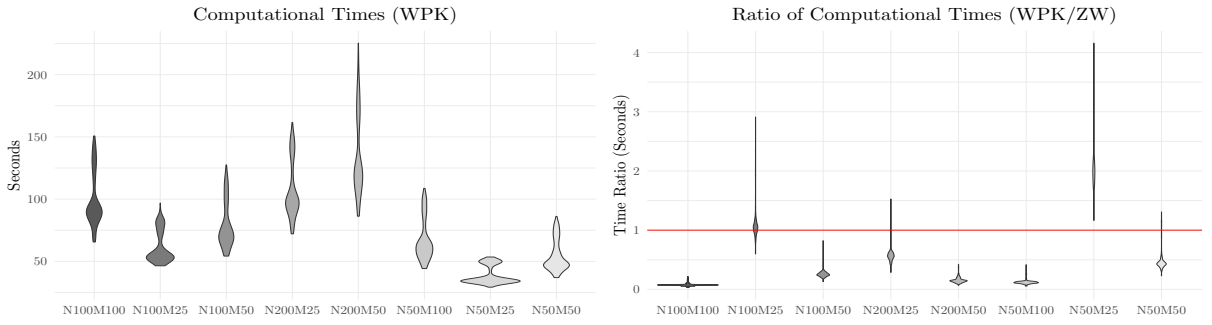Figure 8: Eigenfunction boxplots for $\sigma_0 = 1$, $b_{ZW,2} = 0.1$

Figure 9: Violin Plots of Computational Times & Ratios
for different N and M values

A.1   Local regularity in quadratic mean

In the following, $a \vee b$ and $a \wedge b$ denote the maximum and the minimum of a and b, respectively. Let H and L be two functions defined on a compact interval $\mathcal{T}$ such that

$$0 < \underline{H} := \min_{u \in \mathcal{T}} H_u \leq \max_{u \in \mathcal{T}} H_u =: \overline{H} < 1 \quad \text{and} \quad 0 < \underline{L} := \min_{u \in \mathcal{T}} L_u \leq \max_{u \in \mathcal{T}} L_u =: \overline{L} < \infty. \qquad (19)$$

Definition 1. *For given functions H and L satisfying* (A.1)*, the class $\mathcal{X}(H, L; \mathcal{T})$, or simply $\mathcal{X}(H, L)$, is the set of stochastic processes X satisfying the following conditions : a constant $\Delta_0 > 0$ exists such that :*

(H1) constants $\mathfrak{a} > 0$ and $\mathfrak{A} > 0$ exist such that, for any $p \geq 2$,

$$\sup_{u,v \in \mathcal{T}, |u-v| \leq \Delta_0} \mathbb{E}\left[|X_u - X_v|^{2p}\right] \leq \frac{p!}{2} \mathfrak{a} \mathfrak{A}^{p-2};$$

(H2) constants $S \geq 0$ and $\beta > 0$ exist such that, for any $u, v, t \in \mathcal{T}$ with $u \leq t \leq v$ and $|u - v| \leq \Delta_0$,

$$\left| \mathbb{E}\left[(X_u - X_v)^2\right] - L_t^2 |u-v|^{2H_t} \right| \leq S|u-v|^{2H_t + 2\beta}.$$

The function H defines the local regularity of the process, while L determines the local Hölder constant.

Definition 1 extends the definition of the local regularity proposed by Golovkine et al. (2023) to the whole domain $\mathcal{T}$. Condition (H1), which imposes sub-Gaussian increments, serves to derive the exponential bound for the concentration of the local regularity estimator using Bernstein's inequality for the squared increments.

The classes $\mathcal{X}(H, L)$ are general, examples are provided in Golovkine et al. (2023). More examples can be obtained by time deformation of processes in some $\mathcal{X}(H, L)$, see Section 4.1 above. The following result, for which the proof is given in the Supplementary Material, provides few additional examples.

Lemma A.1. *Let $X \in \mathcal{X}(H_X, L_X)$ and $Y \in \mathcal{X}(H_Y, L_Y)$, for some continuous functions (H_X, L_X) and (H_Y, L_Y).*

(1). Assume a constant $C > 0$ exists such that $\mathbb{E}\left[|X_u - X_v|^4\right] \leq C\mathbb{E}^2\left[|X_u - X_v|^2\right], \forall u, v \in \mathcal{T}$. Let $\phi : \mathbb{R} \to \mathbb{R}$ be a Lipschitz continuous map with bounded derivative ϕ'. Let $L_{\phi,t} = L_{X,t}\mathbb{E}^{1/2}[\phi'(X_t)^2]$, $t \in \mathcal{T}$. If $X \in \mathcal{X}(H_X, L_X)$, then $\phi(X) \in \mathcal{X}(H_X, L_\phi)$.

(2). Assume that $H_X \leq H_Y + \mathfrak{b}$, for some $\mathfrak{b} > 0$. Then $X + Y \in \mathcal{X}(H_X, L_X)$.

(3). Assume that X and Y are, zero-mean, independent and $H_X = H_Y$. Then $X + Y \in \mathcal{X}(H_X, L_X + L_Y)$.

(4). Assume that $H_X \leq H_Y + \mathfrak{b}$, for some $\mathfrak{b} > 0$, X and Y are independent, and Y is bounded. Then $XY \in \mathcal{X}(H_X, L)$, where $L_t = L_{X,t}\mathbb{E}[Y_t^2]$.

A.2   Regularity estimation

Let $X \in \mathcal{X}(H, L)$. The quality of $\widehat{H}_t$, $\widehat{L}_t$ defined in (2), depends on the quality of the presmoothing estimators $\widetilde{X}_u$ of X_u. To quantify their behavior, we consider the uniform L^p-risk

$$R_{\mathfrak{m}}(p) = \mathbb{E}\left(\left|\widetilde{X} - X\right|_\infty^p\right), \quad p \geq 1,$$

where $\mathfrak{m}$ is the expected number of observations on a curve. Here, and in the following, $|\cdot|_\infty$ is the uniform norm for the continuous functions defined on $\mathcal{T}$. The uniform risk above is averaged over the values of M, and therefore we let it depend on $\mathfrak{m}$. Any type of nonparametric estimator $\widetilde{X}$ (local polynomials, splines, *etc*) can be used, as soon as, its uniform L^p-risk is suitably bounded for $p = 2$ and $p > 2$, respectively.

To state our non-asymptotic uniform concentration result for the estimators $\widehat{H}_t$ and $\widehat{L}_t$, we need to define more precisely t_1, t_2 and t_3 used in (2), in terms of $t \in \mathcal{T}$ and some small $\Delta_* > 0$, namely when t is close to 0 or 1. Let us consider $\underline{t} = 0 \vee (t - \Delta_*)$, $\overline{t} = (t + \Delta_*) \wedge 1$, and define the pair (t_1, t_3) as follows: $(t_1, t_3) = (\underline{t}, \overline{t})$ if $t \leq 1/2$ or $(t_1, t_3) = (\overline{t}, \underline{t})$ otherwise. Then, $t \in [t_1 \wedge t_2, t_1 \vee t_2] \subset [t_1 \wedge t_3, t_1 \vee t_3]$. Finally, consider the following mild condition : a constant c exists such that

$$0 < c^{-1} \leq \frac{\log(N)}{\log(\mathfrak{m})} \leq c. \tag{20}$$

Theorem 3. *Assume that X belongs to $\mathcal{X}(H, L)$, and H and L are Lipschitz continuous. Moreover, (A.2) holds true, and positive constants $\mathfrak{c}$ and $\mathfrak{C}$ exist such that*

$$R_{\mathfrak{m}}(2p) \leq \frac{p!}{2}\mathfrak{c}\mathfrak{C}^{p-2}, \qquad \forall p, \mathfrak{m} \geq 1. \tag{21}$$

Assume also, constants $\tau > 0, B > 0$ exists such that

$$R_{\mathfrak{m}}(2) \leq B\mathfrak{m}^{-\tau}, \quad \forall \mathfrak{m} \geq 1. \tag{22}$$

Consider

$$\epsilon = \log^{-\varrho}(\mathfrak{m}) \quad and \quad \Delta_* = \exp(-\log^{\gamma}(\mathfrak{m})),$$

for some $\varrho > 1$ and $0 < \gamma < 1$. Then, for any $\mathfrak{m}$ larger than some constant $\mathfrak{m}_0$ depending on B, τ, γ, ϱ, H, β, L, and for some positive constants $\mathfrak{f}$ and $\mathfrak{g}$ we have

$$\mathbb{P}\left(\left|\widehat{H} - H\right|_{\infty} > \epsilon\right) \leq \exp\left(-\mathfrak{f}N\epsilon^2\Delta_*^{4\overline{H}}\right), \qquad \mathbb{P}\left(\left|\widehat{L}^2 - L^2\right|_{\infty} > \epsilon\right) \leq \exp\left(-\mathfrak{g}N\epsilon^2\frac{\Delta_*^{4\overline{H}}}{\log^2(\Delta_*)}\right). \tag{23}$$

Theorem 3 is proved in the Supplementary Material. A condition like (3) is satisfied by common estimators given the realization of M. See for instance Theorem 1 in Gaïffas (2007) for the case of local polynomials. Condition (3) is also a mild uniform convergence condition satisfied by usual nonparametric regression estimators, given the number of points on a curve. See for instance Tsybakov (2009) and Belloni et al. (2015). In particular, the required conditions for the uniform risk of $\widetilde{X}$ can be obtained under general forms of heteroscedasticity and mild conditons on the distribution of T. In order to guarantee that conditions (3) and (3) remain true when taking expectation with respect to M, it suffices to impose a mild condition like $C\mathfrak{m}^b \leq M$ for some constants $b \in (0, 1]$ and $C > 0$, which can reasonably be used for a wide panel of practical situations. Concerning the quantities $R_{\mathfrak{m}}(2)$, Δ_* and ϵ, they are required to be such that, for some suitable $a > 0$, $R_{\mathfrak{m}}(2)/\Delta_*^a + \Delta_*^{1/a}/\epsilon$ becomes negligible as $\mathfrak{m}$ increases. With our choices for ϵ and Δ_*, this holds true for any $a > 0$. On the other hand, for the purpose of adaptive kernel smoothing in the FPCA context, under the mild condition that $\log(N)/\log(\mathfrak{m})$ is bounded, as required in (A.2), the effect of estimating H is negligible as soon as ϵ is negligible compared to $\log^{-1}(\mathfrak{m})$. This explains our condition $\varrho > 1$. Finally, the condition $\gamma < 1$ combined with the lower bound in (A.2) make the bounds for the concentration of $\widehat{H}$ and $\widehat{L}$ to be exponentially small when N increases. In conclusion, the only practical choice we have to make is that of γ, which was set equal to 0.75 in simulations.

A.3   Assumptions

Assumption 1. *The data generating process satisfies the following conditions.*

1. *The curves $X^{(i)}$ are independent realizations of X which belongs to $\mathcal{X}(H, L, \mathcal{T})$, the set of stochastic processes introduced in Definition 1, where $\mathcal{T}$ is a compact interval, and H and L are Lipschitz continuous, bounded from below and from above, like in (A.1).*

2. *The observations are the pairs* $(Y_m^{(i)}, T_m^{(i)}) \in \mathbb{R} \times \mathcal{T}$, $1 \leq m \leq M_i$, $1 \leq i \leq N$, *obtained according to* (2.1). *The* M_i *are independent with mean* $\mathfrak{m}$, *and a constant* $C_\mathfrak{m} > 1$ *(independent of* $\mathfrak{m}$*) exists such that*

$$C_\mathfrak{m}^{-1} \mathfrak{m} \leq M_i \leq C_\mathfrak{m} \mathfrak{m}, \qquad \forall 1 \leq i \leq N. \tag{24}$$

The $X^{(i)}$*'s,* M_i*'s and* $T_m^{(i)}$*'s are mutually independent. The conditional variance function* $t \mapsto \sigma(t)$ *is Lipschitz continuous, and the* $e_m^{(i)}$*'s are iid centered, sub-Gaussian variables with unit variance, and independent of* $X^{(i)}$*'s,* M_i*'s and* $T_m^{(i)}$*'s.*

3. *In the independent design case, the independent* $T_m^{(i)}$*'s admit a Lipschitz continuous density* f_T *with Lipschitz constant* C_f, *and positive constants* $C_{f,L}, C_{f,U} > 0$ *exist such that*

$$C_{f,L} \leq f_T(t) \leq C_{f,U}, \qquad \forall t \in \mathcal{T}. \tag{25}$$

Condition (2) is a convenient technical condition which shortens the technical arguments required in the following, and has little practical relevance. It can relaxed by imposing probability bounds on the deviation of the M_i's from the mean. Without any technical difficulty, condition (3) can be relaxed to allow the density of the design points $T_m^{(i)}$ to be different for different curves.

Assumption 2. *1. The presmoothed curves* $\widetilde{X}^{(i)}$, $1 \leq i \leq N$, *satisfy conditions* (3) *and* (3).

Assumption 3. *The adaptive smoothing requires the following conditions.*

1. *The smoothing kernel* K *has the support* $[-1,1]$ *and is continuous and strictly positive on any non-degenerate subinterval of the support.*

2. *The bandwidth range* $\mathcal{H}_N$ *is a set of positive numbers and a constant* $\nu > 1$ *exists such that*

$$N \{\mathfrak{m} \inf \mathcal{H}_N\}^2 / \log^\nu(N) \to \infty \quad \text{and} \quad \sup \mathcal{H}_N \to 0. \tag{26}$$

A.4 Technical results

The results in this section are the technical building blocks for the proofs of the main results. Their proofs are provided in the Supplementary Material.

Lemma A.2. *Let* $H \in (0,1)$ *be a Lipschitz continuous function on* $\mathcal{T}$, *and*

$$I(h) = \int_0^1 D(t) h^{2H_t} dt,$$

where D *is some function such that* $0 < \underline{D} \leq D(t) \leq \overline{D} < \infty$. *Then a constant* $\underline{C}$ *exists such that*

$$0 < \underline{C} \frac{h^{2\overline{H}}}{\log(1/h)} \leq I(h) \leq \overline{D} h^{2\underline{H}}, \quad \forall h \in \mathcal{H}_N. \tag{27}$$

Moreover, if $\widehat{H}$ *is an estimator of* H *with uniform concentration like in* (3) *with* $\epsilon \log(\mathfrak{m}) \to 0$, *then*

$$\sup_{h \in \mathcal{H}_N} \left| \int_0^1 D(t) h^{2\widehat{H}_t} dt - I(h) \right| = o_\mathbb{P}(1). \tag{28}$$

The lower bound in (2) cannot be narrowed without further assumptions on H. Property (2) shows that the rates of convergence are not deteriorated as soon as the estimator of H uniformly concentrates faster than $\log^{-1}(\mathfrak{m})$, which is a very mild condition.

Let us next describe the behavior of the errors' conditional variance estimator defined in (2.5).

Lemma A.3. *Assumption 1.2 and 1.3, and Assumption 3 hold true. Then :*

1. *Constants $\underline{b}_W, \overline{b}_W \in (0, 1]$ then exist such that*

$$\underline{b}_W N \leq \frac{\mathbb{E}_M \left[\mathcal{W}_N(t; h)\right]}{\min\{1, \mathfrak{m}h\}} \leq \overline{b}_W N, \qquad \forall h \in \mathcal{H}_N, \forall t \in \mathcal{T}.$$

2. *Constants $\underline{c}_W, \overline{c}_W \in (0, 1]$ then exist such that*

$$\underline{c}_W N \leq \frac{\mathbb{E}_M \left[\mathcal{W}_N(s, t; h)\right]}{\min\{1, (\mathfrak{m}h)^2\}} \leq \overline{c}_W N, \qquad \forall h \in \mathcal{H}_N, |s - t| > 2h.$$

3. *Constants $\underline{d}_W, \overline{d}_W \in (0, 1]$ then exist such that*

$$\underline{d}_W N \leq \frac{\mathbb{E}_M \left[\mathcal{W}_N(s, t; h)\right]}{\min\{1, \mathfrak{m}^2 h |t - s|/2\} + \min\{1, \mathfrak{m}(h - |t - s|/2)\}} \leq \overline{d}_W N, \quad \forall h \in \mathcal{H}_N, \ \forall |s - t| \leq 2h.$$

Lemma A.4. *Assume that the conditions of Lemma A.3 hold true. Then*

$$\sup_{h \in \mathcal{H}_N} \sup_{s, t \in \mathcal{T}} \left\{ \left| \frac{\mathcal{W}_N(s, t; h)}{\mathbb{E}_M \left[\mathcal{W}_N(s, t; h)\right]} - 1 \right| + \left| \frac{\mathbb{E}_M \left[\mathcal{W}_N(s, t; h)\right]}{\mathcal{W}_N(s, t; h)} - 1 \right| \right\} = o_{\mathbb{P}}(1).$$

Lemma A.5. *Assume that the conditions of Lemma A.3 hold true. Then :*

1. *Constants $\underline{c}_{\mathcal{N}}, \overline{c}_{\mathcal{N}} \in (0, 1]$ then exist such that*

$$\frac{\{1 + o_{\mathbb{P}}(1)\} \underline{c}_{\mathcal{N}}}{N \min\{\mathfrak{m}h, (\mathfrak{m}h)^2\}} \leq \frac{1}{\mathcal{N}_\Gamma(s|t; h)} \leq \frac{\overline{c}_{\mathcal{N}} \{1 + o_{\mathbb{P}}(1)\}}{\max\{1, \mathfrak{m}h\} \times \mathbb{E}[\mathcal{W}_N(s, t; h)]}, \qquad \forall |s - t| > 2h;$$

with the $o_{\mathbb{P}}(1)$ uniform with respect to $h \in \mathcal{H}_N$;

2. *Constants $\underline{d}_{\mathcal{N}}, \overline{d}_{\mathcal{N}} \in (0, 1]$ then exist such that*

$$\frac{\underline{d}_{\mathcal{N}} \{1 + o_{\mathbb{P}}(1)\}}{\underline{r}(s, t; h)} \leq \frac{1}{\mathcal{N}_\Gamma(s|t; h)} \leq \frac{\overline{d}_{\mathcal{N}} \{1 + o_{\mathbb{P}}(1)\}}{\max\{1, \mathfrak{m}h\} \times \mathbb{E}[\mathcal{W}_N(s, t; h)]}, \quad \forall |s - t| \leq 2h,$$

with the $o_{\mathbb{P}}(1)$ uniform with respect to $h \in \mathcal{H}_N$, where

$$\underline{r}(s, t; h) = \min\{\mathfrak{m}h, (\mathfrak{m}h)^2, \mathfrak{m}^2 h |t - s|/2\} + \mathfrak{m}\{h - |t - s|/2\}.$$

Lemma A.6. *Assumption 1 holds true and $N\mathfrak{m}b^2 / \log(N) \to \infty$. Then*

$$\sup_{t \in \mathcal{T}} |\widehat{\sigma}(t; b) - \sigma(t)| = O_{\mathbb{P}} \left(b + b^{2\underline{H}} + \{\log(N)/[N \min(1, \mathfrak{m}b^2)]\}^{1/2} \right),$$

with defined $\widehat{\sigma}^2(t; b)$ in (2.5). In the common design case, for any $b > C_d^{-1}\mathfrak{m}$, with C_d in (3.1), we have

$$\sup_{t \in \mathcal{T}} |\widehat{\sigma}(t; b) - \sigma(t)| = O_{\mathbb{P}} \left(b + b^{2\underline{H}} + \{\log(N)/N\}^{1/2} \right).$$

The next result concerns the diagonal correction (2.3). Let

$$\widehat{\widehat{d}}_N(s, t; h) = \frac{\widehat{\sigma}(s; b)\widehat{\sigma}(t; b)}{\mathcal{W}_N(s, t; h)} \sum_{i=1}^{N} w_i(s; h) w_i(t; h) \sum_{m=1}^{M_i} W_m^{(i)}(s; h) W_m^{(i)}(t; h), \tag{29}$$

with $\widehat{\sigma}^2(t; b)$ defined in (2.5). Note that, by construction, $\widehat{\widehat{d}}_N(s, t; h) = \widehat{d}_N(s, t; h) = 0$ when $|s - t| > 2h$.

Lemma A.7. *Assume the conditions of Lemma A.6 hold true. Moreover, $\sigma(s)\sigma(t)$ is estimated by $\widehat{\sigma}(s;b)\widehat{\sigma}(t;b)$ with $b = h^a$ for some $\max(1/2, 2\underline{H} - 1) < a < 1$. Let $\widetilde{\Gamma}$ and $\widehat{\Gamma}$ be defined as in (2.1) and (2.3), respectively. Then*

$$\iint_{\mathcal{T}\times\mathcal{T}} \left\{\widehat{d}_N(s,t;h) - \widehat{\tilde{d}}_N(s,t;h)\right\}^2 dsdt \le o_{\mathbb{P}}(1) \times \iint_{\mathcal{T}\times\mathcal{T}} \left\{\widehat{\Gamma}(s,t;h) - \widetilde{\Gamma}(s,t)\right\}^2 dsdt,$$

and the $o_{\mathbb{P}}(1)$ term does not depend on $h \in \mathcal{H}_N$. Moreover,

$$\iint_{\mathcal{T}\times\mathcal{T}} \widehat{d}_N^2(s,t;h)dsdt \le 4h^2 \operatorname{length}(\mathcal{T})^2 \sup_{t\in\mathcal{T}} \sigma^2(t).$$

Theorems A.1 and A.2 below are the cornerstone for justifying Theorem 1. They provide Taylor expansion (representation) for each eigen-element. Theorems A.1 and A.2 are versions of independent interest of (Hall and Hosseini-Nasab, 2006, equations (2.8) and (2.9)), Hall and Hosseini-Nasab (2009); see also Jirak and Wahl (2023). They provide a representation between the eigen-elements of Γ and those of the (infeasible) empirical estimator $\widetilde{\Gamma}$. Here, we provide an extended version of the results of Hall and Hosseini-Nasab (2009), such that we can derive the Taylor expansions with more general estimators of the covariance function.

Consider the following notation: for M a square-integrable functions of two variables on $\mathcal{T} \times \mathcal{T}$, write

$$\||M\||^2 = \iint_{\mathcal{T}\times\mathcal{T}} M^2(s,t)dsdt = \iint M^2.$$

For the covariance function Γ, assumed to be continuous on $\mathcal{T} \times \mathcal{T}$, we can write the spectral decomposition

$$\Gamma(u,v) = \sum_{j=1}^{\infty} \lambda_j \psi_j(u)\psi_j(v), \tag{30}$$

where $\psi_1, \psi_2, \ldots$ denotes a complete orthonormal sequence of continuous eigenfunctions in $\mathcal{C}$, corresponding to the respective eigenvalues $\lambda_1 \ge \lambda_2 \ge \ldots \ge 0$. Let

$$\rho_1 = \lambda_1 - \lambda_2 > 0 \quad \text{and} \quad \rho_j = \min(\lambda_j - \lambda_{j+1}, \lambda_{j-1} - \lambda_j) > 0, \quad \forall j \ge 2,$$

and

$$\xi_j = \inf\{1 - (\lambda_k/\lambda_j) : \lambda_k < \lambda_j\} \quad \text{and} \quad \eta_j = \inf\{(\lambda_k/\lambda_j) - 1 : \lambda_k > \lambda_j\}.$$

Let L denote a second symmetric kernel with associated non negative eigenvalues sequence $\theta_1 \ge \theta_2 \ge \ldots$ and the complete orthonormal sequence of eigenfunctions $\phi_1, \phi_2, \ldots$, such that we have

$$L(u,v) = \sum_{j=1}^{\infty} \theta_j \phi_j(u)\phi_j(v). \tag{31}$$

Theorem A.1. *Let Γ and L be two operators as in (A.4) and (A.4). Then, for each $j \ge 1$ for which*

$$\rho_j > 0 \qquad \text{and} \qquad \||L - \Gamma\|| \le \frac{1}{2}\lambda_j \min(\xi_j, \eta_j),$$

we have

$$\theta_j - \lambda_j = \iint_{\mathcal{T}\times\mathcal{T}} (L - \Gamma)(s,t)\psi_j(s)\psi_j(t)dsdt + \mathfrak{L}_j, \qquad \text{with} \quad |\mathfrak{L}_j| \le C_\lambda \times \||L - \Gamma\||^2,$$

for some constant $C_\lambda > 0$ depending on the kernel Γ, λ_j, ρ_j, ξ_j and η_j, but not on L.

Theorem A.2. *Assume that the conditions of Theorem A.1 hold true. Then*

$$\phi_j(t) - \psi_j(t) = \sum_{k:k\neq j} (\lambda_j - \lambda_k)^{-1} \psi_k(t) \iint (L - \Gamma)(u,v)\psi_j(u)\psi_k(v)dudv + \mathfrak{S}_j(t), \qquad t \in \mathcal{T},$$

and

$$\|\mathfrak{S}_j\|_2 \leq C_\psi \times \||L - \Gamma\||^2$$

for some constant $C_\psi > 0$ depending on the kernel Γ, λ_j, ρ_j, ξ_j and η_j, but not on L.

A.5 Main results: outline of the proofs

In the following, $C, c, \ldots$ denotes constants with possibly different values at different occurrences. Moreover, the symbol $\lesssim$ means that the left side is bounded by a constant times the right side, while $\asymp$ means left side bounded above and below by constants times right side.

Proof of Theorem 1. For this result, all the unknown quantities appearing in the risk bounds (2.4) and (2.4), as well in the diagonal correction $\widehat{d}_N(s,t;h)$, defined in (2.3) and used to build $\widehat{\Gamma}(s,t;h)$, are supposed given. Let us consider another infeasible covariance function estimator, that is

$$\widetilde{\Gamma}_W(s,t;h) = \frac{1}{\mathcal{W}_N(s,t;h)} \sum_{i=1}^{N} w_i(s;h)w_i(t;h) \left\{ X_s^{(i)} - \mu_W(s;h) \right\} \left\{ X_t^{(i)} - \mu_W(t;h) \right\}, \quad s,t \in \mathcal{T},$$

with

$$\mu_W(t;h) = \frac{1}{\mathcal{W}_N(t;h)} \sum_{i=1}^{N} w_i(t;h) X_t^{(i)}.$$

It is the empirical covariance function computed with the curves as selected for building $\widehat{\Gamma}(s,t;h)$. Let $\widetilde{\lambda}_{j,W}$ and $\widetilde{\psi}_{j,W}$ denote the $j-$th eigenvalue and the corresponding eigenfunction of the covariance operator defined by $\widetilde{\Gamma}_W(s,t;h)$. Let $\widehat{\lambda}_{N,j}$ and $\widehat{\psi}_{N,j}$ denote the $j-$th eigenvalue and eigenfunction obtained with $\widehat{\Gamma}_N(s,t;h)$ defined in (2.3), the uncorrected for diagonal bias estimator of $\Gamma(s,t)$. Weyl's inequality and the last part of Lemma A.7 guarantees that

$$\mathbb{E}_{M,T}\left[\left\{ \widehat{\lambda}_{N,j} - \widehat{\lambda}_j \right\}^2 \right] \leq Ch^2,$$

for some constant C depending only on the function $\sigma^2(\cdot)$ and the length of $\mathcal{T}$. We then decompose

$$\widehat{\lambda}_j - \widetilde{\lambda}_j = \{\widehat{\lambda}_j - \widetilde{\lambda}_{W,j}\} + \{\widetilde{\lambda}_{W,j} - \widetilde{\lambda}_j\} \quad \text{and} \quad \widehat{\lambda}_j - \widetilde{\lambda}_{W,j} = \{\widehat{\lambda}_j - \widehat{\lambda}_{N,j}\} + \{\widehat{\lambda}_{N,j} - \widetilde{\lambda}_{W,j}\},$$

and, using Theorem A.1 and Lemma A.7, we derive the representations

$$\widehat{\lambda}_j - \widetilde{\lambda}_{W,j} = \iint (\widehat{\Gamma} - \widetilde{\Gamma}_W)\psi_j\psi_j + R_{\lambda,1} \quad \text{and} \quad \widetilde{\lambda}_{W,j} - \widetilde{\lambda}_j = \iint (\widetilde{\Gamma}_W - \widetilde{\Gamma})\psi_j\psi_j + R_{\lambda,2}.$$

The remainder terms $R_{\lambda,1}$ and $R_{\lambda,2}$ are negligible compared to the double integrals. Next, since $(a+b)^2 \leq 2a^2 + 2b^2$, we show that, up to negligible terms,

$$\mathbb{E}_{M,T}\left[\left\{ \widehat{\lambda}_j - \widetilde{\lambda}_{j,W} \right\}^2 \right] \quad \text{and} \quad \mathbb{E}_{M,T}\left[\left\{ \widetilde{\lambda}_{j,W} - \widetilde{\lambda}_j \right\}^2 \right],$$

are bounded by $\mathcal{B}_{1,N}(\widehat{\lambda}_j; h) + \mathcal{B}_{2,N}(\widehat{\lambda}_j; h)$ (squared bias + variance terms) and $\mathcal{B}_{3,N}(\widehat{\lambda}_j; h)$ (penalty term), respectively. To derive the expression of $\mathcal{B}_{1,N}(\widehat{\lambda}_j; h) + \mathcal{B}_{2,N}(\widehat{\lambda}_j; h)$, we write $\widehat{X}_t^{(i)}(h) = X_t^{(i)} + B_t^{(i)}(h) + V_t^{(i)}(h)$, with

$$B_t^{(i)}(h) = \mathbb{E}_{M,T}[\widehat{X}_t^{(i)}(h)] - X_t^{(i)} \quad \text{and} \quad V_t^{(i)}(h) = \widehat{X}_t^{(i)}(h) - \mathbb{E}_{M,T}[\widehat{X}_t^{(i)}(h)], \quad t \in \mathcal{T},$$

which are the bias and the stochastic part of $\widehat{X}_t^{(i)}(h)$, respectively. The bias term depends on the $X^{(i)}$, M_i and the $T_m^{(i)}$'s, while the stochastic term depends on $e_m^{(i)}$, M_i and the $T_m^{(i)}$'s. For deriving the expression of $\mathcal{B}_{1,N}(\widehat{\lambda}_j; h)$ we need to bound the expectation of the square of $B_t^{(i)}(h)$, which depends on functions H and L. The details are provided in Section V in the Supplementary Material.

For the eigenfunctions, using Theorem A.2, we derive the expansions

$$\widehat{\psi}_j(t) - \widetilde{\psi}_{W,j}(t) = \sum_{k:k\neq j} (\lambda_j - \lambda_k)^{-1}\psi_k(t) \iint (\widehat{\Gamma} - \widetilde{\Gamma}_W)\psi_j\psi_k + R_{\psi,1}(t), \tag{32}$$

and

$$\widetilde{\psi}_{W,j}(t) - \widetilde{\psi}_j(t) = \sum_{k:k\neq j} (\lambda_j - \lambda_k)^{-1}\psi_k(t) \iint (\widetilde{\Gamma}_W - \widetilde{\Gamma})\psi_j\psi_k + R_{\psi,2}(t). \tag{33}$$

The L^2-norms of the remainders $R_{\psi,1}$ and $R_{\psi,2}$ are easily shown to be negligible compared to the L^2-norms of the sums of double integrals in (A.5) and (A.5), respectively. We then follow the same lines like for the eigenvalues, and derive the three terms of the risk bound, that are squared bias, variance terms and penalty term, respectively. For practical purposes, the calculations are done for versions of the representations (A.5) and (A.5) with truncated sums. Since the truncation error can be arbitrarily small, the rates of convergence of the estimates are not altered. More precisely, for any $\mathcal{K} = \{1,\ldots,K_0\}$, we can write

$$\sum_{k:k\neq j} = \sum_{k\in\mathcal{K}:k\neq j} + \sum_{k\notin\mathcal{K}:k\neq j} = \sum_{k\in\mathcal{K}:k\neq j} + \text{ truncation error.}$$

Since the $\sum_{k:k\neq j}$ is convergent, for any $C > 0$, an integer K_0 exists such that the norm of the truncation error is smaller than C times the norm of the complete sum $\sum_{k:k\neq j}$. The detailed derivations are provided in Section IV in the Supplementary Material. It also follows from the arguments provided there that replacing ψ_j and ψ_k with the constant function equal to 1 when computing $\mathcal{B}_N(\widehat{\lambda}_j; h)$ and $\mathcal{B}_N(\widehat{\psi}_j; h)$, may change the constants but does not change the rates of convergence of the bounds. This, combined with Corollary 1 or 2, guarantee that running the Algorithm in Section 2.6 with the *ad-hoc* proxy of ψ_j equal to the constant function equal to 1, leads to estimates $\widehat{\lambda}_j$ and $\widehat{\psi}_j$ with the same rates of convergence. $\square$

Proof of Corollary 1. Lemmas A.3, A.4 and A.5 imply that the double integrals in the expressions of $\mathcal{B}_{2,N}(\widehat{\lambda}_j; h)$, $\mathcal{B}_{3,N}(\widehat{\lambda}_j; h)$, $\mathcal{B}_{2,N}(\widehat{\lambda}_j; h)$ and $\mathcal{B}_{3,N}(\widehat{\lambda}_j; h)$ can be restricted to the domain $\{|s - t| > 2h\}$ only, because the integration over the diagonal set $\{|s - t| \leq 2h\}$ is negligible. Moreover, a constant $C > 0$ exists such that

$$\frac{C^{-1}\{1 + o_{\mathbb{P}}(1)\}}{N \min\{\mathfrak{m}h, (\mathfrak{m}h)^2\}} \leq \mathcal{B}_{2,N}(\widehat{\lambda}_j; h) + \mathcal{B}_{3,N}(\widehat{\lambda}_j; h) \leq \frac{C\{1 + o_{\mathbb{P}}(1)\}}{N \min\{\mathfrak{m}h, (\mathfrak{m}h)^2\}},$$

and similar bounds can be derived for $\mathcal{B}_{2,N}(\widehat{\psi}_j; h) + \mathcal{B}_{3,N}(\widehat{\psi}_j; h)$. On the other hand, the squared bias terms $\mathcal{B}_{1,N}(\widehat{\lambda}_j; h)$ and $\mathcal{B}_{1,N}(\widehat{\psi}_j; h)$ can rewritten under the form

$$\int B(t)h^{2H_t}dt, \tag{34}$$

with some specific function B determined by L, K, σ^2, m_2, c_2 and the true eigen-elements. Taking the derivative with respect to h of a function like in (A.5), and searching the roots of the derivative, we have to find h such that

$$h^{2\underline{H}-1} \int B(t)h^{2H_t-2\underline{H}}\,dt \asymp \frac{1}{N \min\{\mathfrak{m}h^2, \mathfrak{m}^2h^3\}}.$$

Let h^* be such solution, and note that by construction $\log(1/h^*) \lesssim \log(N\mathfrak{m})$. Lemmas A.2 then indicates that h^* either satisfies

$$(N\mathfrak{m})^{-\frac{1}{2\underline{H}+1}} \lesssim h^* \lesssim (\log(N\mathfrak{m})/N\mathfrak{m})^{\frac{1}{2\underline{H}+1}} \qquad \text{or} \qquad (N\mathfrak{m}^2)^{-\frac{1}{2\underline{H}+2}} \lesssim h^* \lesssim (\log(N\mathfrak{m})/N\mathfrak{m}^2)^{\frac{1}{2\underline{H}+2}},$$

25

depending on the range of h, that is either we have $\min\left\{\mathfrak{m}h^2, \mathfrak{m}^2 h^3\right\} = \mathfrak{m}^2 h^3$ or $\min\left\{\mathfrak{m}h^2, \mathfrak{m}^2 h^3\right\} = \mathfrak{m}^2 h^3$, respectively. Gathering facts, we have that

$$\max\left[(N\mathfrak{m})^{-\frac{1}{2\underline{H}+1}}, \left(N\mathfrak{m}^2\right)^{-\frac{1}{2\underline{H}+2}}\right] \lesssim h^* \lesssim \max\left[(\log(N\mathfrak{m})/N\mathfrak{m})^{\frac{1}{2\underline{H}+1}}, \left(\log(N\mathfrak{m})/N\mathfrak{m}^2\right)^{\frac{1}{2\underline{H}+2}}\right].$$

Substituting h^* back into the square bias terms, and using (2.2), we obtain the upper bound for the rate of convergence of $\mathcal{R}_N(\widehat{\lambda}_j; h)$ and $\mathcal{R}_N(\widehat{\psi}_j; h)$. $\qquad\square$

Proof of Corollary 2. In the common design case, when $\mathfrak{m}^{2\underline{H}}/N \to \infty$, we have $(N\mathfrak{m})^{2\underline{H}/(2\underline{H}+1)}/N \to \infty$. The risks of our estimators are then bounded by the risks on the infeasible, empirical estimators. When $\mathfrak{m}^{2\underline{H}}/N \to 0$, the squared bias term remains larger than two others because the bandwidths cannot decrease faster than m^{-1}. In that case the risks' rate is bounded by $\{\log(\mathfrak{m})/\mathfrak{m}\}^{2\underline{H}}$. $\qquad\square$

Proof of Theorem 2. Let us first notice that that the second part of Lemma A.2 guarantees that the estimation of the function H will not change the rates of convergence in probability. Consider next the feasible versions of $\widehat{\lambda}_j$ and $\widehat{\psi}_j$ computed with all the unknown quantities replaced by estimates, as described in the Algorithm in Section 2.6. In particular, $\widehat{\Gamma}(s,t; h)$ is obtained with the diagonal correction $\hat{\hat{d}}_N(s,t; h)$ in (A.4). Theorem 2 is then a direct consequence of several facts. On the one hand, the uniform convergence of $\widehat{m}_2(t)$ and $\widehat{c}_2(s,t)$ from Step 3 of the Algorithm in Section 2.6. This uniform convergence follows from the uniform convergence of the presmoothing estimator $\widetilde{X}^{(i)}$. On the other hand, from the proof of Theorem 1, we have

$$\widehat{\lambda}_j - \lambda_j = o_{\mathbb{P}}(1) \qquad \text{and} \qquad \|\widehat{\psi}_j - \psi_j\|_2 = o_{\mathbb{P}}(1),$$

if $\widehat{\lambda}_j$ and $\widehat{\psi}_j$ are obtained from the Algorithm with a constant function as input proxy of ψ_j. $\qquad\square$

Acknowledgements

Supplementary Material

In the Supplement, we provide detailed justification for the theoretical statements above. We also provide additional results from additional extensive simulation experiments with several different data generating processes. The content of the Supplement is the following. In Section I we provide details on the properties of the class of multifractional Brownian motion introduced in Section 4.1. Section II is dedicated to the proofs of the properties of the local regularity and the proof of uniform convergence of our local regularity estimators. Section III contain proofs of the technical results stated in the Appendix. In Section IV we prove our first-order Taylor expansions for the eigen-elements, as previewed in Theorems A.1 and A.2 in the Appendix. In Section V we provide the details for the proof of Theorem 1. Finally, Section VI contains additional simulation results.

References

Balança, P. (2015). Some sample path properties of multifractional Brownian motion. *Stochastic Processes Appl.*, 125(10):3823–3850.

Belloni, A., Chernozhukov, V., Chetverikov, D., and Kato, K. (2015). Some new asymptotic theory for least squares series: pointwise and uniform results. *J. Econometrics*, 186(2):345–366.

Benko, M., Härdle, W., and Kneip, A. (2009). Common functional principal components. *Ann. Statist.*, 37(1):1–34.

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

Carroll, R. J., Delaigle, A., and Hall, P. (2013). Unexpected properties of bandwidth choice when smoothing discrete data for constructing a functional data classifier. *Ann. Statist.*, 41(6):2739–2767.

Gaïffas, S. (2007). On pointwise adaptive curve estimation based on inhomogeneous data. *ESAIM: Probability and Statistics*, 11:344–364.

Golovkine, S., Klutchnikoff, N., and Patilea, V. (2022). Learning the smoothness of noisy curves with application to online curve estimation. *Electronic Journal of Statistics*, 16(1):1485–1560.

Golovkine, S., Klutchnikoff, N., and Patilea, V. (2023). Adaptive optimal estimation of irregular mean and covariance functions. arxiv:2108.06507v2.

Hall, P. and Hosseini-Nasab, M. (2006). On properties of functional principal components analysis. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, 68(1):109–126.

Hall, P. and Hosseini-Nasab, M. (2009). Theory for high-order bounds in functional principal components analysis. *Math. Proc. Cambridge Philos. Soc.*, 146(1):225–256.

Hall, P., Müller, H. G., and Wang, J. L. (2006). Properties of principal component methods for functional and longitudinal data analysis. *Annals of Statistics*, 34(3):1493–1517.

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

Jirak, M. and Wahl, M. (2023). Relative perturbation bounds with applications to empirical covariance operators. *Advances in Mathematics*, 412:108808.

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

Mohammadi, N. and Panaretos, V. M. (2021). Functional data analysis with rough sample paths? arxiv:2105.12035.

Mohammadi, N., Santoro, L., and Panaretos, V. M. (2022). Nonparametric estimation for sde with sparsely sampled paths: an fda perspective. arxiv:2110.14433.

Müller, H.-G. and Stadtmüller, U. (1987). Estimation of Heteroscedasticity in Regression Analysis. *The Annals of Statistics*, 15(2):610 – 625.

Poß , D., Liebl, D., Kneip, A., Eisenbarth, H., Wager, T. D., and Barrett, L. F. (2020). Superconsistent Estimation of Points of Impact in Non-Parametric Regression with Functional Predictors. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 82(4):1115–1140.

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

Rubín, T. and Panaretos, V. M. (2020). Sparsely observed functional time series: estimation and prediction. *Electronic Journal of Statistics*, 14(1):1137 – 1210.

Tsybakov, A. B. (2009). *Introduction to nonparametric estimation.* Springer Series in Statistics. Springer, New York.

Zhang, X. and Wang, J. L. (2016). From sparse to dense functional data and beyond. *Annals of Statistics*, 44(5):2281–2321.

Zhou, Y., Bhattacharjee, S., Carroll, C., Chen, Y., Dai, X., Fan, J., Gajardo, A., Hadjipantelis, P. Z., Han, K., Ji, H., Zhu, C., Müller, H.-G., and Wang, J.-L. (2022). *fdapace: Functional Data Analysis and Empirical Dynamics.* R package version 0.5.9.