Adaptive estimation of irregular mean and covariance functions

Steven Golovkine
MACSI, Department of Mathematics and Statistics, University of Limerick, Ireland
E-mail: steven.golovkine@ul.ie
Nicolas Klutchnikoff
Univ. Rennes, CNRS, IRMAR - UMR 6625, F-35000 Rennes, France
E-mail: nicolas.klutchnikoff@univ-rennes2.fr
Valentin Patilea
Univ. Rennes, Ensai, CNRS, CREST - UMR 9194, F-35000 Rennes, France
E-mail: valentin.patilea@ensai.fr

Summary. We propose nonparametric estimators for the mean and the covariance functions of functional data. Our setup covers a wide range of practical situations. The random trajectories are, not necessarily differentiable, have unknown regularity, and are measured with error at discrete design points. The measurement error could be heteroscedastic. The design points could be either randomly drawn or common for all curves. The definition of our estimators depends on the local regularity of the stochastic process generating the functional data. We consider a simple estimator of this local regularity which takes strength from the replication and regularization features of functional data. Next, we use the "smoothing first, then estimate" approach for the mean and the covariance functions. They can be applied with both sparsely or densely sampled curves, are easy to calculate and to update, and perform well in simulations. Simulations built upon a real data example on household power consumption illustrate the effectiveness of the new approach.

Keywords: Functional data analysis; Kernel smoothing; Hölder exponent; Minimax optimality

1. Introduction

Motivated by a large number of applications, there is a great interest in models for observation entities in the form of a sequence of measurements recorded intermittently at several discrete points in time. Functional data analysis (FDA) considers such data as being values on the trajectories of a stochastic process, recorded with some error, at discrete random times. The mean and the covariances functions play a critical role in FDA.

To formalize the framework, let $\mathcal{T}$ be a compact interval, typically $[0,1]$. Data consist of random realizations of sample paths from a second-order stochastic process $X = (X_t : t \in \mathcal{T})$ with continuous trajectories. The mean and covariance functions are $\mu(t) = \mathbb{E}(X_t)$ and

$$\Gamma(s,t) = \mathbb{E}\left\{[X_s - \mu(s)][X_t - \mu(t)]\right\} = \mathbb{E}\left(X_s X_t\right) - \mu(s)\mu(t), \qquad s,t \in \mathcal{T},$$

respectively. If the independent realizations $X^{(1)}, \ldots, X^{(i)}, \ldots, X^{(N)}$ of X were observed, the ideal estimators would be

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

In real applications, the curves are rarely observed without error and never at each value $t \in \mathcal{T}$. This is why we consider the following common and more realistic setup. For each $1 \leq i \leq N$, and given a positive integer M_i, let $T_m^{(i)} \in \mathcal{T}$, $1 \leq m \leq M_i$, be the observation times for the curve $X^{(i)}$. The observations associated with a curve, or trajectory, $X^{(i)}$ consist of the pairs $(Y_m^{(i)}, T_m^{(i)}) \in \mathbb{R} \times \mathcal{T}$ where $Y_m^{(i)}$ is defined as

$$Y_m^{(i)} = X^{(i)}(T_m^{(i)}) + \varepsilon_m^{(i)}, \qquad 1 \leq m \leq M_i, \;\; 1 \leq i \leq N, \tag{1}$$

and $\varepsilon_m^{(i)}$ is an independent (centered) error variable. Here, and in the following, we use the notation $X_t^{(i)}$ for the value at a generic point $t \in \mathcal{T}$ of the realization $X^{(i)}$ of X, while $X^{(i)}(T_m^{(i)})$ denotes the measurement at $T_m^{(i)}$ of this realization.

A commonly used idea is to build feasible versions of $\widetilde{\mu}_N(\cdot)$ and $\widetilde{\Gamma}_N(\cdot, \cdot)$ using nonparametric estimates of $X_t^{(i)}$ and $X_s^{(i)} X_t^{(i)}$, such as obtained by smoothing splines or local polynomials. This approach, usually called "smoothing first, then estimate" or "two-stage procedure", have been considered, amongst others, by Hall et al. (2006) and Zhang and Chen (2007). In general, the sample trajectories are required to admit at least second-order derivatives over $\mathcal{T}$. Li and Hsing (2010), Zhang and Wang (2016) and Zhang and Wang (2018) propose an alternative local linear smoothing approach where the estimators are determined by suitably weighting schemes which involve the whole sample of curves. This idea exploits the so-called replication and regularization features of functional data (see Ramsay and Silverman, 2005, ch. 22). In this alternative approach, the regularity assumptions are imposed on the mean and covariance functions, which are required to admit second, or higher, order derivatives over the whole domain. Since, in general, the mean and covariance functions are more regular than the sample trajectories, the approach based on weighting schemes using all the sample curves might be preferable. However, in some cases, for instance in energy, chemistry and physics, astronomy and medical applications, the mean and covariance functions could be quite irregular, of unknown irregularity.

Cai and Yuan (2011) and Cai and Yuan (2010) derive the optimal rates of convergence, in the minimax sense, for the mean and covariance functions, respectively, and propose optimal estimators. The estimator of the mean function proposed by Cai and Yuan (2011) is a smoothing spline estimator which could be built only if the regularity of the sample paths is known. Cai and Yuan (2010) used the representation of the covariance function in a tensor product reproducing kernel Hilbert space (RKHS) space. Next, under some assumptions, they derived estimators for $\Gamma(s,t)$ using a low dimension version of this representation obtained by a regularization procedure, provided the values M_i are not very different. This procedure involves numerical optimization. See also Wong and Zhang (2019). The optimal rates for the mean and covariance functions are defined by the sum of two types of terms. One corresponds to the rate of convergence of the $\widetilde{\mu}_N(\cdot)$ and $\widetilde{\Gamma}_N(\cdot, \cdot)$, which is the standard rate of convergence for empirical means and covariances. The other contribution to the optimal rates is given by the differences between $\widetilde{\mu}_N(\cdot)$ and $\widetilde{\Gamma}_N(\cdot, \cdot)$ and their feasible versions. The optimal rates of the differences depend on the sample trajectories regularity, because the minimax lower bounds should also take into account the case where the functions to be estimated has the same regularity as the trajectories.

The estimation of the mean and covariance functions presents another specific feature. The optimal rates of convergence depend on the nature of the measurement times $T_m^{(i)}$. For now, two situations were investigated in the literature. On the one hand, the so-called *independent design* case where, given the M_i's, the $T_m^{(i)}$ are obtained as a random sample of size $M_1 + \cdots + M_N$ from the same continuous distribution. On the other hand, the so-called *common design* case where the M_i are all equal to some integer value $\mathfrak{m}$, and the $T_m^{(i)}$, $1 \leq m \leq \mathfrak{m}$, are the same

across the curves $X^{(i)}$. In both cases, the best rates for the nonparametric estimators depend on the regularity of the sample trajectories. These rates also depend on the number of different observation times $T_m^{(i)}$, that is equal to $M_1 + \cdots + M_N$ with independent design, and equal to $\mathfrak{m}$ with common design. In other words, the replication feature of functional data is less impactful with common design. See Cai and Yuan (2011) for the case of the mean function, and Cai and Yuan (2010) and Cai and Yuan (2016) for the covariance function case.

In this paper, we propose data-driven "smoothing first, then estimate" type methods, based on $1-$dimensional smoothing. The process is allowed to have a piecewise constant, unknown regularity. Our method does not require complex numerical optimization. It applies in the same way with common and independent design situations, and allows for general heteroscedastic measurement errors $\varepsilon_m^{(i)}$. Moreover, our approach is suitable with both sparsely or densely sampled curves. The definition of sparse and dense regimes is recalled in Section 2.

Let $\widehat{X}^{(i)}$ be a suitable nonparametric estimator of $X^{(i)}$ applied to the M_i pairs $(Y_m^{(i)}, T_m^{(i)})$, for instance a kernel estimator. What will make this estimator suitable is that it takes into account the regularity of the process X and the final estimation purpose, that is the mean or the covariance function. These features can be achieved in an easy, data-driven way, as will be explained below. With at hand the $\widehat{X}^{(i)}$'s tuned for the mean function estimation, we define

$$\widehat{\mu}_N(t) = \frac{1}{N}\sum_{i=1}^{N} \widehat{X}_t^{(i)}, \qquad t \in \mathcal{T}. \tag{2}$$

For the covariance function, we distinguish the diagonal from the non-diagonal points. With at hand the $\widehat{X}^{(i)}$'s tuned for the covariance function estimation, and for some diagonal set $\mathcal{D} \subset \mathcal{T}^2 := \mathcal{T} \times \mathcal{T}$ that we will determine using the data, let us define

$$\widehat{\Gamma}_N(s,t) = \frac{1}{N}\sum_{i=1}^{N} \widehat{X}_s^{(i)}\widehat{X}_t^{(i)} - \widehat{\mu}_N(s)\widehat{\mu}_N(t), \qquad (s,t) \in \mathcal{T}^2 \setminus \mathcal{D}. \tag{3}$$

It is well known that the variance function $\Gamma(s,s)$ induces a singularity when estimating the covariance function $\Gamma(\cdot,\cdot)$. See, for instance, Zhang and Wang (2016), Remark 4. We propose a simple way to build the diagonal set $\mathcal{D}$, which shrinks to the diagonal segment according to a data-driven rule that we provide in the following. Given $\mathcal{D}$, the estimates of $\Gamma(\cdot,\cdot)$ on $\mathcal{D}$ are directly obtained from the estimates $\widehat{\Gamma}_N(s,t)$ for the closest (s,t) on the boundary of $\mathcal{D}$.

Although the methodology we propose is general and can be used with different types of smoothers, we focus on the case where the $\widehat{X}_t^{(i)}$ are obtained by kernel smoothing. In this case, tuning the $\widehat{X}^{(i)}$'s means to suitably determine the rate of decrease and the constant defining the bandwidth. In our case, this is done completely data-driven by a one variable minimization of a new, suitable risk function.

To the best of our knowledge, there is no contribution which considers estimators of the curves $X^{(i)}$ adapted to their regularity and to the purpose of estimating mean or covariance functions. It is clear that trajectory-by-trajectory adaptive optimal smoothing, for instance using the Goldenshluger and Lepski (2011) method, in general yields sub-optimal rates of convergence for $\widehat{\mu}_N(t)$ and $\widehat{\Gamma}_N(s,t)$. The reason is that trajectory-by-trajectory smoothing ignores the information contained in the other $N-1$ curves in the sample generated according to the same stochastic process X. See Cai and Yuan (2011) for a discussion on the differences with the usual nonparametric rates. One can also use cross-validation for choosing the bandwidth with the suitably weighting schemes, such as proposed by Li and Hsing (2010) or Zhang and Wang (2016). However, this

would require significant computational effort, and, to the best of our knowledge, the idea has not yet received a theoretical justification. Using the replication and regularization features of functional data, we consider an effective estimator for the local regularity of the process X, a probabilistic concept which determines the analytic regularity of the trajectories of X. The local regularity estimator, a version of the one introduced by Golovkine et al. (2022), combines information both across and within curves. Moreover, it allows for general heteroscedastic measurement errors, does not involve any optimization and is obtained after a fast, possibly parallel, computation. With at hand the local regularity estimator, we derive the suitable estimators $\widehat{X}_t^{(i)}$, and finally our optimal mean and covariance functions estimators. The smoothing parameter used to build the $\widehat{X}_t^{(i)}$ depends on M_i and N, but can be easily computed given the estimate of the local regularity of X. We assert that the replication feature of the functional data makes the concept of local regularity of the process a more meaningful parameter than the usual curve regularity, which is an analytic concept designed for a single function.

In Section 2, we provide insight on why the local regularity of the process X is a natural feature to be considered. Moreover, we explain why the "smoothing first, then estimate" approach could achieve optimal rates when the regularity of X is known. In Section 3, we formally define the local regularity of the process X. Moreover, we introduce the estimator for this regularity and present exponential bounds for the concentration under mild conditions. In particular, both independent and common designs are allowed, and the process regularity is allowed to vary with t. Section 3 ends with a discussion on the relationship between the process regularity and the trajectories' analytical regularity. In Section 4, we use the regularity estimate to build sharp bounds of the pointwise quadratic risk function between our estimators and the unfeasible estimators $\widetilde{\mu}_N$ and $\widetilde{\Gamma}_N$, respectively. The bounds depend on quantities which could be estimated by sample averages. Minimizing the risk bounds with respect to the bandwidth, we derive the optimal bandwidth for the kernel estimates of the trajectories. These estimates are further used to estimate the mean and covariance functions. Our mean and covariance estimators, and the local regularity estimator, are computed on the same sample of curves. In other words, no data splitting is necessary with our approach. The finite sample performance of the new estimators is illustrated in Section 5 using simulated samples generated according to the setup of a real data set on the power consumption of households. The simulation method which we introduce in Section 5 is a simple device allowing to generate functional data with regularity features similar to those observed in real applications. Some conclusions and discussions are gathered in Section 6. Few proofs are relegated to the Appendix. A Supplementary Material contains more technical arguments and simulation results.

2. From unfeasible to feasible optimal estimators

The novelty of our approach is based on the local regularity of X, a mild condition on the second-order moments of the local increments of the process X. Before proceeding to formal definitions, let us first provide insight into the reason why the local regularity of the process generating the curves is a meaningful concept, and why our approach can achieve good performance. For this purpose, we analyze the difference $\widehat{\mu}_N(t) - \widetilde{\mu}_N(t)$, $s, t \in \mathcal{T}$, but similar apply to the covariance function estimation.

The data $(Y_m^{(i)}, T_m^{(i)}) \in \mathbb{R} \times \mathcal{T}$ are generated according to model (1) with

$$\varepsilon_m^{(i)} = \sigma(T_m^{(i)}, X^{(i)}(T_m^{(i)}))e_m^{(i)}, \quad 1 \leq m \leq M_i,\ 1 \leq i \leq N, \tag{4}$$

where the $X^{(i)}$ are independent trajectories of X, $e_m^{(i)}$ are independent copies of a centered variable e with unit variance, and $\sigma(t, x)$ is some unknown bounded function which account for possibly

heteroscedastic measurement errors. The integers $M_1, \ldots, M_N$ represent an independent sample of an integer-valued random variable M with expectation $\mathfrak{m}$ which increases with N. Thus, $M_1, \ldots, M_N$ is the Nth line of a triangular array of integer numbers. In the independent design case, for each $1 \leq i \leq N$, the observation times $T_m^{(i)}$ are random realizations of a variable $T \in \mathcal{T}$. We assume that the realizations of X, e, M and T are mutually independent. Let $\mathcal{T}_{obs}^{(i)}$ denote the set of observation times $T_m^{(i)}$, $1 \leq m \leq M_i$, over the trajectory $X^{(i)}$. In the common design case, $M \equiv \mathfrak{m}$, and the $\mathcal{T}_{obs}^{(i)}$ are the same for all i. Thus, if not stated differently, the issues discussed in this section apply to both independent design and common design cases.

Let

$$\mathbb{E}_i(\cdot) = \mathbb{E}(\cdot \mid M_i, \mathcal{T}_{obs}^{(i)}, X^{(i)}) \quad \text{and} \quad \mathbb{E}_{M,T}(\cdot) = \mathbb{E}\left(\cdot \mid M_i, \mathcal{T}_{obs}^{(i)}, 1 \leq i \leq N\right).$$

For any $t \in \mathcal{T}$, we consider a generic linear nonparametric estimator

$$\widehat{X}_t^{(i)} = \sum_{m=1}^{M_i} Y_m^{(i)} W_m^{(i)}(t), \quad 1 \leq i \leq N. \tag{5}$$

The weights $W_m^{(i)}(t)$ are defined as functions of the elements in $\mathcal{T}_{obs}^{(i)}$. The example we keep in mind is that of kernel smoothing which we investigate in detail in Section 4. Let

$$\widehat{X}_t^{(i)} - X_t^{(i)} = B_t^{(i)} + V_t^{(i)}, \quad t \in \mathcal{T}, \tag{6}$$

where

$$B_t^{(i)} := \mathbb{E}_i\left[\widehat{X}_t^{(i)}\right] - X_t^{(i)} \quad \text{and} \quad V_t^{(i)} := \widehat{X}_t^{(i)} - \mathbb{E}_i\left[\widehat{X}_t^{(i)}\right] = \sum_{m=1}^{M_i} \varepsilon_m^{(i)} W_m^{(i)}(t).$$

The pairs of random variables $(B_t^{(i)}, V_t^{(i)})$, $1 \leq i \leq N$, are independent and we could reasonably assume that they are squared integrable for all t. Then, for the mean, we can write

$$\widehat{\mu}_N(t) - \widetilde{\mu}_N(t) = \frac{1}{N}\sum_{i=1}^{N} B_t^{(i)} + \frac{1}{N}\sum_{i=1}^{N} V_t^{(i)}.$$

All the variables $\varepsilon_m^{(i)}$ are centered and conditionally independent, with bounded conditional variance, given all M_i, $\mathcal{T}_{obs}^{(i)}$ and $X^{(i)}$. Thus,

$$\mathbb{E}_{M,T}\left[\left\{N^{-1}\sum_{i=1}^{N} V_t^{(i)}\right\}^2\right] \leq N^{-1}\sup_x \sigma^2(t,x) \times N^{-1}\sum_{i=1}^{N}\left\{\max_m \left|W_m^{(i)}(t)\right| \times \sum_{m=1}^{M_i}\left|W_m^{(i)}(t)\right|\right\}. \tag{7}$$

For local polynomials with bandwidth h, under some mild conditions, the rate of decrease of the right-hand side in the last display, given the design, is $O_{\mathbb{P}}\left((N\mathfrak{m}h)^{-1}\right)$.

For simplicity, we suppose the trajectories are not differentiable. The case of smooth paths is discussed in the Supplement. On the bias part, by Cauchy-Schwarz inequality, we then have

$$\begin{aligned}\mathbb{E}_{M,T}&\left[\left\{N^{-1}\sum_{i=1}^{N} B_t^{(i)}\right\}^2\right] \\ &\leq N^{-1}\sum_{i=1}^{N}\left\{\sum_{m=1}^{M_i}\left|W_m^{(i)}(t)\right| \times \sum_{m=1}^{M_i}\mathbb{E}_{M,T}\left(\left\{X^{(i)}(T_m^{(i)}) - X_t^{(i)}\right\}^2 \mid \mathcal{T}_{obs}^{(i)}\right)\left|W_m^{(i)}(t)\right|\right\}.\end{aligned} \tag{8}$$

It now becomes clear that the rate of the square of the bias term in $\widehat{\mu}_N(t) - \widetilde{\mu}_N(t)$ is determined by the second-order moment of the increments $X^{(i)}(T_m^{(i)}) - X_t^{(i)}$. If, for $u, v \in \mathcal{T}$ close to t,

$$\mathbb{E}\left(\{X_u - X_v\}^2\right) \approx L_t^2 |u - v|^{2H_t}, \tag{9}$$

with some $0 < H_t \leq 1$ and $L_t > 0$, then the rate of the right-hand side in (8) is bounded by

$$N^{-1} \sum_{i=1}^{N} \left\{ \sum_{m=1}^{M_i} \left| W_m^{(i)}(t) \right| \times \sum_{m=1}^{M_i} L_t^2 \left| T_m^{(i)} - t \right|^{2H_t} |W_m^{(i)}(t)| \right\}. \tag{10}$$

For the Nadaraya-Watson estimator with bandwidth h, this has the rate $O_{\mathbb{P}}\left(h^{2H_t}\right)$.

Gathering facts, we deduce that, in the case of non differentiable trajectories, with the Nadaraya-Watson estimator and

$$h \sim (N\mathfrak{m})^{-1/(1+2H_t)}, \tag{11}$$

one can expect

$$\mathbb{E}_{M,T}\left[\{\widehat{\mu}_N(t) - \widetilde{\mu}_N(t)\}^2\right] = O_{\mathbb{P}}\left((N\mathfrak{m})^{-\frac{2H_t}{1+2H_t}}\right).$$

Thus, given the local regularity H_t, the estimator $\widehat{\mu}_N(t)$ can achieve the minimax optimal rate for the estimation of the mean function $\mu(t)$. See Cai and Yuan (2011).

Let us note that in some cases, in particular with kernel smoothing, the estimator defined in (5) could be degenerate, *i.e.*, the weights $W_m^{(i)}(t)$ are not well defined because h is too small. The trajectories for which this happens could change with t. Then, $\widehat{\mu}_N(t)$ is defined as an average over the trajectories for which the estimator (5) is not degenerate. This can more likely happen in the so-called *sparse* regime, where $\mathfrak{m}^{2H_t} \ll N$. A similar phenomenon occurs with estimators determined by suitably weighting schemes, see for instance equation (2.1) in Li and Hsing (2010), or equation (2.3) in Zhang and Wang (2016). However, in the independent case, one could benefit from the replication feature of functional data, because only a fraction of trajectories will yield non degenerate estimators $\widehat{X}_t^{(i)}$. The size of this fraction plays a central role in the sparse regime. This crucial aspect is taken into account in Sections 4.1 and 4.2, where we choose the bandwidths while penalizing the number of trajectories which yield degenerate estimators.

The case of common design requires some special attention. For simplicity, let us assume the common design points are equidistant and consider the kernel smoothing uses a kernel supported on $[-1, 1]$. In this case, the bandwidth cannot have a rate smaller than $\mathfrak{m}^{-1}$, otherwise the weights $W_m^{(i)}(t)$ could all be equal to zero. This means that with a common design, the optimal bandwidth is given by the minimization of $h^{2H_t} + (N\mathfrak{m}h)^{-1}$ under the constraint that $\mathfrak{m}h$ stays away from zero. Without loss of generality, we could set $h = k/\mathfrak{m}$ with k a positive integer and search k which minimizes $h^{2H_t} + (N\mathfrak{m}h)^{-1}$. Balancing the two terms, one expects the optimal $k/\mathfrak{m}$ to have the rate $(N\mathfrak{m})^{-1/\{1+2H_t\}}$. If $\mathfrak{m}^{2H_t}$ is larger than N, *i.e.* in the so-called *dense* regime, the optimal k is well defined and $k \sim (\mathfrak{m}^{2H_t}/N)^{1/\{1+2H_t\}}$ and, with this optimal choice, $\mathbb{E}_{M,T}\left[\{\widehat{\mu}_N(t) - \widetilde{\mu}_N(t)\}^2\right] = o_{\mathbb{P}}\left(N^{-1}\right)$. If $\mathfrak{m}^{2H_t} \ll N$, then the constraint that $k \geq 1$ becomes binding, and it is no longer possible to balance the squared bias term and the variance term. The rate of h^{2H_t} dominates the rate $(N\mathfrak{m}h)^{-1}$. The minimal rate for $\mathbb{E}_{M,T}\left[\{\widehat{\mu}_N(t) - \widetilde{\mu}_N(t)\}^2\right]$ then corresponds to $k = 1$, and is $O_{\mathbb{P}}\left(\mathfrak{m}^{-2H_t}\right)$. Gathering facts, we recover the optimal rate for mean estimation with common design, that is $O_{\mathbb{P}}\left(\mathfrak{m}^{-H_t} + N^{-1/2}\right)$, see Cai and Yuan (2011). Finally, let us recall the somehow surprising message from Cai and Yuan (2011), p. 2332: the interpolation is rate optimal when $\mathfrak{m}^{2H_t} \gg N$ in the case of common design; smoothing does not improve convergence rates. Our contribution on this aspect is a data-driven rule for the practitioner

which completes this theoretical fact about the interpolation. The adaptive bandwidth rule proposed in Section 4 automatically chooses between smoothing and interpolation.

We learn from the above that the "smoothing first, then estimate" approach can lead to optimal rates of convergence for estimating the mean function with independent and common design, as derived by Cai and Yuan (2011), provided the regularity parameter H_t in (9) is known. In the next section, we introduce a simple estimator of this parameter. Under some mild conditions, this estimator concentrates around the true local regularity faster than a suitable negative power of $\log(\mathfrak{m})$. This suffices to guarantee that our mean and covariance functions estimators achieve the same rates as when the local regularity is known.

Let us end this section with a discussion of the differences with the weighting schemes approach, as for instance considered by Li and Hsing (2010) and Zhang and Wang (2016). If the regularity of $\mu(\cdot)$ is known, then one could define $B_t^{(i)}$ and $V_t^{(i)}$ in (6) centering by the mean function instead of the trajectory $X_t^{(i)}$. Then, one could derive the rate of $\mathbb{E}\left[\{\widehat{\mu}_N(t) - \mu(t)\}^2\right]$ and find the bandwidth which minimizes this rate, exactly as done in Li and Hsing (2010) and Zhang and Wang (2016) where $\mu(\cdot)$ is assumed to be twice differentiable. However, the estimation of the regularity of $\mu(\cdot)$ remains an open problem.

3. Local regularity estimator

Our approach is based on the general regularity condition (9), which is a local property that we formally define in the following. In Subsection 3.2, we propose an estimator of H_t and in Subsection 3.3, we provide theoretical guarantees. Given this type of regularity, the Kolmogorov Continuity Theorem allows to determine the analytic regularity of the trajectories of X. Details are provided in Section 3.4. Hereafter, $t \in \mathcal{T}$ is an arbitrarily fixed point.

3.1. Local regularity in quadratic mean

Let $H : u \mapsto H_u \in (0,1)$ and $L : u \mapsto L_u > 0$ be Lipschitz functions defined on $\mathcal{T}$. Let $\Delta_* > 0$ and $\mathcal{O}_*(t) = [t - \Delta_*/2, t - \Delta_*/2] \cap \mathcal{T}$.

DEFINITION 1. *The class $\mathcal{X}(H, L; \mathcal{O}_*(t))$ is the set of stochastic processes X satisfying the following the conditions.*

(H1) Constants $\mathfrak{a} > 0$ and $\mathfrak{A} > 0$ exist such that, for any $p \geq 1$

$$0 < \inf_{u \in \mathcal{O}_*(t)} \mathbb{E}\left[|X_u|^2\right] \qquad \text{and} \qquad \sup_{u \in \mathcal{O}_*(t)} \mathbb{E}\left[|X_u - X_t|^{2p}\right] \leq \frac{p!}{2}\mathfrak{a}\mathfrak{A}^{p-2}.$$

(H2) Positive constants S and β exist such that

$$\left|\mathbb{E}\left[(X_u - X_v)^2\right] - L_t^2|u - v|^{2H_t}\right| \leq S^2|u - v|^{2H_t}\Delta_*^{2\beta}, \qquad u, v \in \mathcal{O}_*(t).$$

The quantity H_t is the local regularity of the process on $\mathcal{O}_*(t)$, while L_t is the Hölder constant of the trajectories.

In Section (5.1), we introduce a general class of processes satisfying Definition 1. See also Blanke and Vial (2014) and Golovkine et al. (2022) for more examples and references on processes satisfying the mild condition in (H2). Examples include, but are not limited to stationary or stationary increment processes. For some common processes with the ordered eigenvalues of the

covariance operator such that, for some $1 < \nu < 3$, $\lambda_j \sim j^{-\nu}$, $j \geq 1$, one has $H \equiv (\nu - 1)/2$. Golovkine et al. (2022) also considers the case of differentiable trajectories, in which case the local regularity H_t refers to the highest order derivative of the sample path in a neighborhood of t. The second part of condition in (H1) serves to derive the exponential bound for the concentration of the local regularity estimator, while the first part excludes the case of constant sample paths, a case where H_t and L_t are not well defined.

3.2. The local regularity estimation method

Assume that X belongs to $\mathcal{X}(H, L; \mathcal{O}_*(t))$. Our first goal is to construct an estimator of H_t. For simplicity, for $u, v \in \mathcal{O}_*(t)$, let us denote

$$\theta(u, v) = \mathbb{E}\left[(X_u - X_v)^2\right] \approx L_t^2 |u - v|^{2H_t} \quad \text{if } \Delta_* \text{ is small.}$$

Now, let t_1 and t_3 be such that $[t_1, t_3] \subset \mathcal{O}_*(t)$ and $t_3 - t_1 = \Delta_*/2$. Let $t_2 = (t_1 + t_3)/2$ and define

$$\widetilde{H}_t = \frac{\log(\theta(t_1, t_3)) - \log(\theta(t_1, t_2))}{2\log(2)} \quad \text{if } \Delta_* \text{ is small.}$$

When t is away from the left and right endpoints of $\mathcal{T}$ by more than $\Delta_*/2$, we set $t_2 = t$. Otherwise, we set $t_1 = \min \mathcal{T}$ or $t_3 = \max \mathcal{T}$, respectively. Since H is Lipschitz continuous and, by construction, $|t_2 - t| \leq \Delta_*/2$, the quantity $\widetilde{H}_t$ is a proxy of H_t.

Given nonparametric estimators $\widetilde{X}_u^{(i)}$ of $X_u^{(i)}$, we define a natural estimator of $\widetilde{H}_t$, and thus of H_t, as

$$\widehat{H}_t = \frac{\log\left(\widehat{\theta}(t_1, t_3)\right) - \log\left(\widehat{\theta}(t_1, t_2)\right)}{2\log(2)}, \tag{12}$$

where

$$\widehat{\theta}(u, v) = \frac{1}{N} \sum_{i=1}^{N} \left(\widetilde{X}_u^{(i)} - \widetilde{X}_v^{(i)}\right)^2, \quad u, v \in \mathcal{O}_*(t).$$

The estimate of L_t is readily obtained given $\widehat{H}_t$, the details are provided in Section 4.4.

3.3. Concentration properties of the local regularity estimator

The local regularity estimator (12) was studied by Golovkine et al. (2022) in the case of constant functions H and L in a neighborhood of t. The quality of the estimator $\widehat{H}_t$ depends on the quality of the generic nonparametric estimators $\widetilde{X}_u$ of X_u. To quantify their behavior, we consider the local $\mathbb{L}^p$-risk

$$R_N(t; p) = \sup_{u \in \mathcal{O}_*(t)} \mathbb{E}\left(\left|\widetilde{X}_u - X_u\right|^p\right), \quad p \geq 1.$$

The risks $R_N(t, p)$ depends on N because the distribution of the number of points on each curve is allowed to depend on N. Our methodology applies with any type of nonparametric estimator $\widetilde{X}$ (local polynomials, splines, ...) as soon as, for any $p \in \mathbb{N}$, its $\mathbb{L}^p$-risk is suitably bounded. The following mild condition is satisfied by common estimators, see for instance Theorem 1 in Gaïffas (2007) for the case of local polynomials.

Assumptions.

(LP1) There exist two positive constants $\mathfrak{c}$ and $\mathfrak{C}$ such that, for any $p \geq 1$,

$$R_N(t; 2p) \leq \frac{p!}{2} \mathfrak{c} \mathfrak{C}^{p-2}, \qquad \forall N \geq 1.$$

We can now state a non-asymptotic concentration result for the estimator $\widehat{H}_t$.

THEOREM 1. *Assume that X belongs to $\mathcal{X}(H, L; \mathcal{O}_*(t))$, and that (LP1) holds true. Assume also that there exists $\tau > 0$ and $B > 0$ such that $R_N(t; 2) \leq B\mathfrak{m}^{-\tau}$. Let $1 < \varrho < \gamma$, and consider*

$$\varphi(\mathfrak{m}) = \log^{-\varrho}(\mathfrak{m}) \quad \textit{and} \quad \Delta_* = \log^{-\gamma}(\mathfrak{m}).$$

Then, for any $\mathfrak{m}$ larger than some constant $\mathfrak{m}_0$ depending on B, τ, γ, Γ, H, β and for some constant $\mathfrak{f}$, we have

$$\mathbb{P}\left(\left|\widehat{H}_t - H_t\right| > \varphi(\mathfrak{m}),\, \widehat{H}_t > 0\right) \leq \exp\left(-\mathfrak{f} N \varphi^2(\mathfrak{m}) \Delta_*^{4H_t}\right).$$

The proof of Theorem 1 follows the lines of the proof of Theorem 5 in Golovkine et al. (2022) and is thus omitted. The three quantities $R_N(t; 2)$, Δ_* and $\varphi(\mathfrak{m})$ are required to decrease to zero, as $\mathfrak{m}$ tends to infinity with N, in such a way that $R_N(t; 2)/\Delta_* + \Delta_*/\varphi(\mathfrak{m}) \to 0$. It will be shown below that, in order to achieve optimal rates of convergence for the mean and covariance estimators, the local regularity has to be estimated with suitable rate $\varphi(\mathfrak{m})$. More precisely, to select the bandwidth like in (11), we replace H_t by $\widehat{H}_t$. Imposing the mild condition that $\log(N)/\log(\mathfrak{m})$ is bounded, since $\mathfrak{m}^{1/\log(\mathfrak{m})} = e$, the effect of this replacement is negligible as soon as $\varphi(\mathfrak{m}) \ll \log^{-1}(\mathfrak{m})$. Finally, let us point out that $\widehat{H}_t$ does not depend on τ. Indeed, since $\tau > 0$ could be arbitrarily small, the rate imposed on the nonparametric estimators $\widetilde{X}$ of X is a very mild consistency requirement which is achieved by the common estimators, with random or fixed design, under mild conditions, in particular on the distribution of the M_i. See, for instance, Tsybakov (2009) and Belloni et al. (2015). In particular, the required rate for the $\widetilde{X}$ can be obtained under general forms of heteroscedasticity. In conclusion, the only practical choice is that of γ, and we set $\gamma = 1.1$ in our empirical study.

3.4. *From the regularity of the process to the regularity of the trajectories*

Let us now connect the probabilistic concept of local regularity with the regularity of the sample paths considered as functions. For simplicity, assume that H is constant in a neighborhood of t. For the more general cases with non constant H, see Balança (2015) and the references therein.

By Assumption (H2), using the refined version of Kolmogorov's criterion stated in Revuz and Yor (2013), it can be proven that almost all sample paths of X belong to any Hölder space of functions defined on the neighborhood of t, with the Hölder exponent less than H. As an example, the Brownian motion has a constant local regularity equal to $1/2$. Moreover, almost surely, the sample paths of the Brownian motion belong to any Hölder space of local regularity less than $1/2$, but cannot be Hölder continuous with exponent larger than or equal to $1/2$.

Hence, the probability theory indicates that imposing assumptions on the regularity of the sample paths could be a delicate issue. Indeed, even for some widely used examples, this regularity is not well defined in the sense required by the nonparametric statistics theory. Since the sample paths have a regularity which can be arbitrarily close to the local regularity of the process X as defined above, the probabilistic concept of local regularity seems more appropriate for establishing the rates of convergence for the mean and covariance estimators.

4. Adaptive mean and covariance function estimators

We now explain how to select data-driven bandwidths for kernel smoothing of the trajectories and build adaptive mean and covariance function estimates. Hereafter, $\widehat{H}_t$ will be the estimator

of H_t defined in (12), considered on the event $\{\widehat{H}_t > 0\}$. Let $\widehat{\mathfrak{m}} = N^{-1}\sum_{i=1}^{N} M_i$. Let us consider a class of linear smoothers of the sample paths. For each $1 \leq i \leq N$, using the measurements $(Y_m^{(i)}, T_m^{(i)})$, $1 \leq m \leq M_i$, of the trajectory $X^{(i)}$, we define

$$\widehat{X}_t^{(i)} = \sum_{m=1}^{M_i} Y_m^{(i)} W_m^{(i)}(t), \qquad 1 \leq i \leq N,$$

where $W_m^{(i)}(t)$ are the weights depending on the $T_m^{(i)}$'s only, and on some smoothing parameter.

In the following, we focus on the case of Nadaraya-Watson (NW) estimators, but also indicate how to adapt the construction for local linear smoothing. Given the bandwidth h, with the convention $0/0 = 0$, the weights of the NW estimator of $X^{(i)}$ are

$$W_m^{(i)}(t) = W_m^{(i)}(t; h) = K\left(\frac{T_m^{(i)} - t}{h}\right)\left[\sum_{m'=1}^{M_i} K\left(\frac{T_{m'}^{(i)} - t}{h}\right)\right]^{-1}, \quad 1 \leq m \leq M_i.$$

Herein, K is a nonnegative, bounded kernel with the support in $[-1, 1]$.

4.1. Adaptive optimal mean estimation

With finite samples it may happen that $\widehat{X}_t^{(i)}$ is degenerate, that means $W_m^{(i)}(t) = 0$ for all $1 \leq m \leq M_i$. In such a case the i-th curve will be dropped for the mean and covariance estimations. With kernel smoothing, in the case of common design, the number of degenerate estimates $\widehat{X}_t^{(i)}$ is either equal to N or to zero. In the independent design case, this number could be any integer between 0 and N. A suitable bandwidth rule should be penalizing for the number of curves which are not considered for the estimation. In the following, we propose a natural way to penalize which adapts to the sparse and dense regimes. Moreover, the two types of designs are automatically handled. For this purpose, let $\mathbf{1}\{\cdot\}$ denote the indicator function and define

$$w_i(t; h) = 1 \quad \text{if} \quad \sum_{m=1}^{M_i} \mathbf{1}\{|T_m^{(i)} - t| \leq h\} \geq 1, \quad \text{and} \quad w_i(t; h) = 0 \text{ otherwise,} \tag{13}$$

and let

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

Our adaptive mean function estimator is

$$\widehat{\mu}_N^*(t) = \widehat{\mu}_N(t; h_\mu^*) \quad \text{with} \quad \widehat{\mu}_N(t; h) = \frac{1}{\mathcal{W}_N(t; h)} \sum_{i=1}^{N} w_i(t; h)\widehat{X}_t^{(i)}, \tag{14}$$

where h_μ^* is a suitable bandwidth defined below. The mean estimator $\widehat{\mu}_N(t; h)$ is a version of that defined in (2) which takes into account that some trajectories have no observation times between $t - h_\mu^*$ and $t + h_\mu^*$. The normalization of the mean estimator by $\mathcal{W}_N(t; h)$ is also implicitly used in the definition of the estimators proposed by Li and Hsing (2010) and Zhang and Wang (2016).

To introduce our bandwidth rule, for any $h > 0$, $\alpha > 0$, let

$$c_i(t; h) = \sum_{m=1}^{M_i} \left|W_m^{(i)}(t; h)\right|, \qquad c_i(t; h, \alpha) = \sum_{m=1}^{M_i} \left|(T_m^{(i)} - t)/h\right|^\alpha \left|W_m^{(i)}(t; h)\right|, \tag{15}$$

and

$$\overline{C}(t;h,\alpha)=\frac{1}{\mathcal{W}_N(t;h)}\sum_{i=1}^{N}w_i(t;h)c_i(t;h)c_i(t;h,\alpha).$$

In (15), $W_m^{(i)}(t;h)$ can be the weights corresponding to local polynomial smoothing. With the NW estimator, all the $W_m^{(i)}(t;h)$ are nonnegative, and the $c_i(t;h)$ are equal to 1. Moreover,

$$\overline{C}(t;h,\alpha)\approx\int|u|^{\alpha}K(u)du. \tag{16}$$

The details for (16) are provided in the Supplement. Using the equivalent kernels idea, see Section 3.2.2 in Fan and Gijbels (1996), the same approximation could be used in the case of local linear estimators. The accuracy of the approximation (16) could be high since it involves the $T_m^{(i)}$ to be close to t for all the curves with $w_i(t;h)=1$. Next, using the rule $0/0=0$, let

$$\mathcal{N}_i(t;h)=\frac{w_i(t;h)}{\max_{1\le m\le M_i}|W_m^{(i)}(t;h)|}\quad\text{and}\quad\mathcal{N}_\mu(t;h)=\left[\frac{1}{\mathcal{W}_N^2(t;h)}\sum_{i=1}^{N}w_i(t;h)\frac{c_i(t;h)}{\mathcal{N}_i(t;h)}\right]^{-1}. \tag{17}$$

With the NW estimator, $\mathcal{N}_\mu(t;h)$ is equal to $\mathcal{W}_N(t;h)$ times the harmonic mean of $\mathcal{N}_i(t;h)$, over the curves with $w_i(t;h)=1$. Moreover, under the mild condition (23) below, $\mathcal{N}_i(t;h)=O_{\mathbb{P}}(\mathfrak{m}h)$.

Let $\mathcal{H}_N$ be a bandwidth range. We define the bandwidth for computing $\widehat{\mu}_N^*(t)$ such that it minimizes the mean squared difference between $\widehat{\mu}_N(t;h)$ and $\widetilde{\mu}_N(t)$. This leads us to define the optimal bandwidth

$$h_\mu^*=h_\mu^*(t)=\arg\min_{h\in\mathcal{H}_N}\mathcal{R}_\mu(t;h), \tag{18}$$

with,

$$\mathcal{R}_\mu(t;h)=q_1^2h^{2\widehat{H}_t}+\frac{q_2^2}{\mathcal{N}_\mu(t;h)}+q_3^2\left[\frac{1}{\mathcal{W}_N(t;h)}-\frac{1}{N}\right], \tag{19}$$

and

$$q_1^2=\overline{C}(t;h,2\widehat{H}_t)\widehat{L}_t^2\,,\qquad q_2^2=\sigma_{\max}^2\,,\qquad q_3^2=\mathrm{Var}(X_t),$$

where $\sigma_{\max}$ is a bound for the function $\sigma(t,x)$ in (4) and $\widehat{L}_t$ is an estimate of the Hölder constant L_t from (H2). In Section 5, we propose a simple procedure to build $\widehat{L}_t$ based on $\widehat{H}_t$ and the preliminary nonparametric estimates $\widetilde{X}$ of the sample paths used for $\widehat{H}_t$. We show in the Appendix that $\mathcal{R}_\mu(t;h)/2$ is a sharp bound for $\mathbb{E}_{M,T}\left[\{\widehat{\mu}_N(t;h)-\widetilde{\mu}_N(t)\}^2\right]$. The minimization of $\mathcal{R}_\mu(t;h)$ can be easily performed on a grid of h values in the range $\mathcal{H}_N$.

The bandwidth rule (18) could be used with both independent and common design. With common design, the $T_m^{(i)}\equiv T_m$ and $W_m^{(i)}(t;h)\equiv W_m(t;h)$ no longer depend on i and the solution h_μ^* will always be a value in the set of h such that $\mathcal{W}_N(t;h)=N$. Moreover, for the NW estimator, whenever $\mathcal{W}_N(t;h)=N$,

$$\overline{C}(t;h,2\widehat{H}_t)=\sum_{m=1}^{\mathfrak{m}}|(T_m-t)/h|^{2\widehat{H}_t}W_m(t;h)\quad\text{and}\quad\mathcal{N}_\mu^{-1}(t;h)=N^{-1}\max_{1\le m\le\mathfrak{m}}W_m(t;h). \tag{20}$$

In a data-driven way, h_μ^* automatically chooses between interpolation and smoothing.

The following result states that our estimator $\widehat{\mu}_N^*(t)$ achieves the best rates one can expect. We assume

$$N\mathfrak{m}\min\mathcal{H}_N/\log(N\mathfrak{m})\to\infty\quad\text{and}\quad\max\mathcal{H}_N\to0, \tag{21}$$

a minimal condition for the bandwidth range. For simplicity, we also assume that

$$\limsup_{N,\mathfrak{m}\to\infty}\{\log(N)/\log(\mathfrak{m})\} < \infty, \tag{22}$$

a technical condition which is realistic in applications. Moreover, we impose the following mild technical condition in the independent design case:

$$\exists c_L, C_U > 0 \text{ such that } c_L \leq M_i \mathfrak{m}^{-1} \leq C_U, \quad \text{for all } N \text{ and } 1 \leq i \leq N. \tag{23}$$

With a common design where $M_i \equiv \mathfrak{m}$ and the $T_1^{(i)}, \ldots, T_{\mathfrak{m}}^{(i)}$ are not changing with i, we suppose that:

$$\exists C_U \geq 1 \text{ such that } \max_{1\leq m\leq \mathfrak{m}-1}\{T_{m+1}^{(i)} - T_m^{(i)}\} \leq C_U \min_{1\leq m\leq \mathfrak{m}-1}\{T_{m+1}^{(i)} - T_m^{(i)}\}. \tag{24}$$

THEOREM 2. *Assume the conditions of Theorem 1, and assume (21), (22) hold true. Assume that $T_m^{(i)}$ are either independently drawn, with a Hölder continuous density which is bounded away from zero and (23) holds true, or $T_m^{(i)}$ are the points of a common design satisfying (24). Then,*

$$h_\mu^* \sim (N\mathfrak{m})^{-\frac{1}{1+2H_t}},$$

and the estimator $\widehat{\mu}_N^*(t) = \widehat{\mu}_N(t; h_\mu^*)$ defined by (14) and (18) satisfies

$$\widehat{\mu}_N^*(t) - \widetilde{\mu}_N(t) = O_{\mathbb{P}}\left((N\mathfrak{m})^{-\frac{H_t}{1+2H_t}}\right) \quad \text{and} \quad \widehat{\mu}_N^*(t) - \mu(t) = O_{\mathbb{P}}\left((N\mathfrak{m})^{-\frac{H_t}{1+2H_t}} + N^{-1/2}\right),$$

in the independent design case. Meanwhile, with the common design,

$$\widehat{\mu}_N^*(t) - \mu(t) = O_{\mathbb{P}}\left(\max\left\{(N\mathfrak{m})^{-\frac{H_t}{1+2H_t}}, \mathfrak{m}^{-H_t}\right\} + N^{-1/2}\right) = O_{\mathbb{P}}\left(\mathfrak{m}^{-H_t} + N^{-1/2}\right).$$

The rates of $\widehat{\mu}_N^*(t)$ are the best one could expect in view of the results of Cai and Yuan (2011). The difference between the common and independent designs comes from the fact that, in order to avoid degenerate mean estimator, the bandwidth cannot decrease faster than $\mathfrak{m}^{-1}$.

4.2. Adaptive covariance function estimates

For any $s, t \in \mathcal{T}$, $s \neq t$, define

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

with $w_i(s;h)$ and $w_i(t;h)$ as in (13). Our adaptive covariance function estimator is

$$\widehat{\Gamma}_N^*(s,t) = \widehat{\Gamma}_N(s,t;h_\Gamma^*) \quad \text{with} \quad \widehat{\Gamma}_N(s,t;h) = \widehat{\gamma}_N(s,t;h) - \widehat{\mu}_N(s;h)\widehat{\mu}_N(t;h), \tag{25}$$

where $\widehat{\mu}_N(s;h)$, $\widehat{\mu}_N(t;h)$ are defined according to (14), and

$$\widehat{\gamma}_N(s,t;h) = \frac{1}{\mathcal{W}_N(s,t;h)} \sum_{i=1}^N w_i(s,t;h)\widehat{X}_s^{(i)}\widehat{X}_t^{(i)}. \tag{26}$$

Here, $\widehat{X}_s^{(i)}$ and $\widehat{X}_t^{(i)}$ are the NW estimators built with some suitable bandwidth h_Γ^* which is defined below. This covariance function estimator is a practical version of that defined in (3). The

normalization of the covariance estimator by $\mathcal{W}_N(s,t;h)$ is also implicitly used in the definition of the estimators proposed by Li and Hsing (2010) and Zhang and Wang (2016).

We define the bandwidth for computing $\widehat{\gamma}_N(s,t;h)$, and eventually $\widehat{\Gamma}^*_N(s,t)$, such that it minimizes the mean squared difference between $\widehat{\gamma}_N(s,t;h)$ and the unfeasible estimator

$$\widetilde{\gamma}_N(s,t) = N^{-1}\sum_{i=1}^{N} X_s^{(i)} X_t^{(i)},$$

of $\mathbb{E}(X_sX_t)$. To this aim, we define modified versions of $\mathcal{N}_i(t;h)$ and $\mathcal{N}_\mu(t;h)$, see (17), taking into account only the curves with $w_i(s,t;h)=1$:

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

This idea leads us to define the optimal bandwidth, in some range $\mathcal{H}_N$, as

$$h^*_\Gamma = h^*_\Gamma(s,t) = h^*_\Gamma(t,s) = \arg\min_{h\in\mathcal{H}_N}\{\mathcal{R}_\Gamma(s|t;h)+\mathcal{R}_\Gamma(t|s;h)\}, \tag{27}$$

with

$$\mathcal{R}_\Gamma(t|s;h) = \mathfrak{q}_1^2(t|s)h^{2\widehat{H}_t} + \frac{\mathfrak{q}_2^2(t|s)}{\mathcal{N}_\Gamma(t|s;h)} + \mathfrak{q}_3^2\left[\frac{1}{\mathcal{W}_N(s,t;h)}-\frac{1}{N}\right]. \tag{28}$$

The $\mathfrak{q}_\ell$, $1\le l\le 3$, are defined by:

$$\mathfrak{q}_1^2(t|s) = 2\mathbb{E}(X_s^2)\overline{\mathfrak{C}}(t|s;h,2\widehat{H}_t)\widehat{L}_t^2, \qquad \mathfrak{q}_2^2(t|s) = \sigma^2_{\max}\,\mathbb{E}(X_s^2), \qquad \mathfrak{q}_3^2 = \frac{\mathrm{Var}(X_sX_t)}{2},$$

where

$$\overline{\mathfrak{C}}(t|s;h,\alpha) = \frac{\sum_{i=1}^{N} w_i(s,t;h)c_i(t;h,\alpha)}{\mathcal{W}_N(s,t;h)} \approx \int |u|^\alpha K(u)du, \tag{29}$$

and the approximation (29) is valid with NW or local linear estimators. The details for (28) are provided in the Supplement.

We show in the Supplement that the function of h minimized in (27) is a sharp bound for $\mathbb{E}_{M,T}[\{\widehat{\gamma}_N(s,t;h)-\widetilde{\gamma}_N(s,t)\}^2]/2$. The sum of the first two terms in the expressions of $\mathcal{R}_\Gamma(s|t;h)$ and $\mathcal{R}_\Gamma(t|s;h)$ represents the quadratic risk of our estimator of $\mathbb{E}(X_sX_t)$ compared to the unfeasible one based on the true values $X_s^{(i)}X_t^{(i)}$ from the curves yielding non-degenerate estimates $\widehat{X}_s^{(i)}\widehat{X}_t^{(i)}$. The third term in (28) penalizes for the number of curves which are dropped when calculating our estimator. The minimization in (27) can be performed on a grid of values h.

Like for the mean function, the definition (27) can be used with both independent and common design. Indeed, with common design, h^*_Γ will always be a value in $\mathcal{H}_N$ such that $\mathcal{W}_N(s,t;h)=N$. In a completely data-driven way, h^*_Γ will choose between interpolation and smoothing.

THEOREM 3. *Let $s\neq t$. Assume $N\{\mathfrak{m}\min\mathcal{H}_N\}^2/\log^2(N\mathfrak{m})\to\infty$, $\sup_{t\in\mathcal{T}}\mathbb{E}(X_t^4)<\infty$, and the conditions of Theorem 2 hold true. Let $H(s,t)=\min\{H_s,H_t\}$. Then*

$$h^*_\Gamma \sim \max\left\{(N\mathfrak{m}^2)^{-\frac{1}{2\{H(s,t)+1\}}},\ (N\mathfrak{m})^{-\frac{1}{2H(s,t)+1}}\right\},$$

and the estimator $\widehat{\Gamma}^*_N(s,t)=\widehat{\Gamma}^*_N(s,t;h^*_\Gamma)$ defined by (25) and (27) satisfies

$$\widehat{\Gamma}^*_N(s,t)-\Gamma(s,t) = O_{\mathbb{P}}\left((N\mathfrak{m}^2)^{-\frac{H(s,t)}{2\{H(s,t)+1\}}}+(N\mathfrak{m})^{-\frac{H(s,t)}{2H(s,t)+1}}+N^{-1/2}\right),$$

in the independent design case. Meanwhile with the common design,

$$\widehat{\Gamma}_N^*(s,t) - \Gamma(s,t) = O_{\mathbb{P}}\left(\mathfrak{m}^{-H(s,t)} + N^{-1/2}\right).$$

In view of the results of Cai and Yuan (2010), the rate achieved by $\widehat{\Gamma}_N^*(s,t)$ is the best one could expect in the case of common design. However, with independent design,

$$\mathfrak{m}^{2H(s,t)} \ll N \quad \text{if and only if} \quad N^{-1/2} \ll (N\mathfrak{m})^{-\frac{H(s,t)}{2H(s,t)+1}} \ll (N\mathfrak{m}^2)^{-\frac{H(s,t)}{2\{H(s,t)+1\}}},$$

and thus the rate of $\widehat{\Gamma}_N^*(s,t)$ is slower than one may expect. Even if this rate seems sub-optimal in the sparse case, with respect to the minimax rate for the $\mathbb{L}^2$-risk obtained by Cai and Yuan (2010), we conjecture that $\widehat{\Gamma}_N^*(s,t)$ achieves the optimal pointwise rate. We let for future work the clarification of this subtle aspect.

4.3. The estimator on the diagonal of the covariance function

As mentioned in (3), we propose to use the estimator of $\Gamma(s,t)$ defined in (26) only outside a diagonal set $\mathcal{D}$. It remains to give a data-driven rule for choosing $\mathcal{D}$ shrinking to the diagonal segment $\{(s,s) : s \in \mathcal{T}\}$, and to propose an estimator for the covariance function on the diagonal set. Let us fix $t \in \mathcal{T}$, and consider $\mathfrak{d}_t \leq \Delta_*/2$, with Δ_* like in Theorem 1. Under mild moment assumptions, we show in the Appendix that, a constant C exists such that

$$\mathbb{E}\left[\left(\widetilde{\Gamma}_N(t-\mathfrak{u}_1, t+\mathfrak{u}_2) - \widetilde{\Gamma}_N(t,t)\right)^2\right] \leq C\mathfrak{d}_t^{2H_t}, \qquad \forall\, 0 \leq \mathfrak{u}_1, \mathfrak{u}_2 \leq \mathfrak{d}_t. \tag{30}$$

On the other hand, for proving Theorem 3, we need a bandwidth smaller than $|s-t|/2$ for a kernel with support in $[-1,1]$. Taking into account these aspects, our estimator of $\Gamma_N(t-\mathfrak{u}_1, t+\mathfrak{u}_2)$ and $\Gamma_N(t+\mathfrak{u}_2, t-\mathfrak{u}_1)$, for $0 \leq \mathfrak{u}_1, \mathfrak{u}_2 \leq \mathfrak{d}_t$, is defined as

$$\widehat{\Gamma}_N(t-\mathfrak{u}_1, t+\mathfrak{u}_2) = \widehat{\Gamma}_N(t+\mathfrak{u}_2, t-\mathfrak{u}_1) = \widehat{\Gamma}_N(t-\mathfrak{d}_t, t+\mathfrak{d}_t).$$

The quantity $\mathfrak{d}_t$ can be the smallest value d which is larger than the bandwidth $h_\Gamma^*(t-d, t+d)$ defined in (27). In practice, one can simply consider the points $(t-d, t+d)$ on a grid, for decreasing values of d. Then $\mathfrak{d}_t$ is the smallest value of d for which $d \geq h_\Gamma^*(t-d, t+d)$.

4.4. Implementation aspects

The risks $\mathcal{R}_\mu$ and $\mathcal{R}_\Gamma$ defined in (19) and (28), respectively, depend on the second order moments of X_t and X_sX_t, on L_t^2, and the conditional variance bound $\sigma_{\max}^2$. For the second order moments, we simply use empirical moments with X replaced by $\widetilde{X}$. To obtain the presmoothed curves $\widetilde{X}$ introduced in Section 3.2, we use the NW estimator using the bandwidth defined in Bertin (2004) and the triangular kernel $K(t) = (1-|t|)\,\mathbf{1}_{[-1,1]}(t)$.

In view of (H2), with $[t_1, t_3] \subset \mathcal{O}_*(t)$ and $t_3 - t_1 = \Delta_*/2$, if t_2 is the midpoint of $[t_1, t_3]$,

$$L_t^2 \approx \frac{\theta(t_2,t_3)}{|t_3-t_2|^{2H_t}} \approx \frac{\theta(t_1,t_2)}{|t_2-t_1|^{2H_t}}.$$

Given the estimate $\widehat{H}_t$ and estimates $\widehat{\theta}(t_2,t_3)$ and $\widehat{\theta}(t_1,t_2)$ as in (12), we then define the estimate

$$\widehat{L}_t^2 \approx \frac{1}{2}\left(\frac{\widehat{\theta}(t_2,t_3)}{|t_3-t_2|^{2\widehat{H}_t}} + \frac{\widehat{\theta}(t_1,t_2)}{|t_2-t_1|^{2\widehat{H}_t}}\right). \tag{31}$$

To estimate the conditional variance bound, let us first consider the case where $\sigma^2(t,x)$ does not depend on x. In this case, one can compute

$$\widehat{\sigma}^2(t) = \frac{1}{N}\sum_{i=1}^{N} \frac{1}{2|\mathcal{S}_i|} \sum_{m\in\mathcal{S}_i} \left[Y_m^{(i)} - Y_{m-1}^{(i)}\right]^2,$$

where $\mathcal{S}_i$ is a subset of indices m for the $i-$th trajectory, and $|\mathcal{S}_i|$ denotes its cardinal. When the variance of the errors is considered constant, $\mathcal{S}_i$ can be the set $\{2,3,\dots,M_i\}$. When the variance depends on t, one could define $\mathcal{S}_i$ as the set of indices corresponding to the K_0 values $T_m^{(i)}$ closest to t. The theory allows a choice such as $K_0 = \lfloor \widehat{\mathfrak{m}} \exp\left(-\{\log\log\widehat{\mathfrak{m}}\}^2\right)\rfloor$. Then $\sigma^2_{\max}$ could be $\max_{t\in\mathcal{T}} \widehat{\sigma}^2(t)$, and this choice was used in our empirical investigation.

5. Empirical study

To investigate the finite sample properties of our adaptive nonparametric estimators of the mean and covariance function, we proceed to an extensive simulation study. We first introduce a general class of zero mean processes satisfying (H2). Next, we use the functions $u \mapsto H_u$ and $u \mapsto L_u$ estimated from a real dataset to choose a process in this class. Finally, we add the estimated mean function from the real data, and thus define the simulated data generating process.

5.1. A general class of Gaussian processes with predefined local regularity

We first consider the class of multifractional Brownian motion (MfBm) processes. See, *e.g.*, Balança (2015) and the references therein for the formal definitions and the properties of this large class of Gaussian processes. A MfBm, say $(W(t))_{t\geq 0}$, 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

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

To make the MfBm class even more general, we consider a deterministic time deformation. Here, the time deformation is defined by $t \mapsto A(t) \geq 0$, a strictly increasing, continuously differentiable function defined on $[0,\infty)$. Moreover, the derivative $A'(t)$ is strictly positive on any compact interval. Let $A^{-1}(\cdot)$ denote the inverse of $A(\cdot)$, and let

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

We consider the MfBm $(W_{A,t})_{t\geq 0}$ with Hurst index function $H_{A,t}$. Given the Hurst index function H and time deformation function A, the process we consider is

$$X(t) = W_A(A(t)), \qquad t\geq 0, \tag{32}$$

with covariance function

$$C_A(s,t) = \mathbb{E}\left[X(s)X(t)\right] = D(H_s,H_t)\left[A(s)^{H_s+H_t} + A(t)^{H_s+H_t} - |A(t)-A(s)|^{H_s+H_t}\right]. \tag{33}$$

LEMMA 1. *Assume $t \mapsto H_t \in (0,1)$ is twice continuously differentiable, and $t \mapsto L_t > 0$ is continuous, $t\geq 0$. Then X defined in (32) satisfies condition (H2) with local regularity H_t and Hölder constant L_t, provided that, for some $A(0)\geq 0$, the time deformation is*

$$A(t) = A(0) + \int_0^t L_s^{1/H_s} ds, \qquad t\geq 0.$$

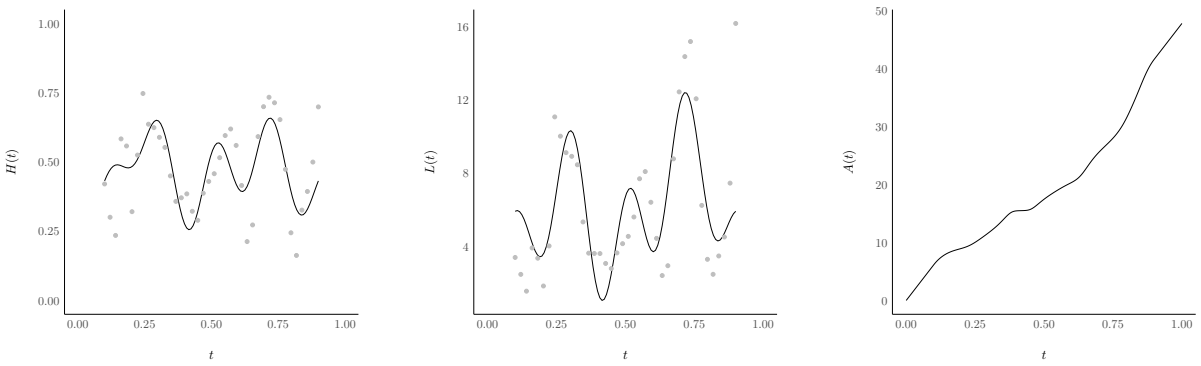
5.2. Simulation design

Our simulation study is based on the Household Active Power Consumption dataset which was sourced from the UC Irvine Machine Learning Repository (https://archive.ics.uci.edu/ml/datasets/Individual+household+electric+power+consumption). This dataset contains diverse energy related features gathered in a house located near Paris, every minute between December 2006 and November 2010. In total, its represents around 2 million data points. Here, we focus on the daily voltage and we only consider the days without missing values in the measurements. The extracted dataset contains 708 voltage curves with an uniform common design with 1440 points, normalized such that $\mathcal{T} = [0, 1]$. We aim to simulate datasets using the data generating process defined in Section 5.1, with an Hurst index function H_t and a time deformation function A_t estimated on the Power Consumption dataset, to which we add a mean curve also fitted to the real dataset. For the fitted mean curve, we consider the model

$$\mu(t) = \beta_0 t + \sqrt{2} \sum_{1 \leq k \leq 50} \left\{ \beta_{1,k} \cos(2k\pi t) + \beta_{2,k} \sin(2k\pi t) \right\}, \quad t \in [0, 1].$$

The coefficients β obtained by LASSO regression with the `R` package `glmnet`. The outcomes are given by the 1440 values of the empirical mean of the 708 curves, and t on the regular grid of 1440 points. The regularity of the mean function is controlled using the penalty parameter s.

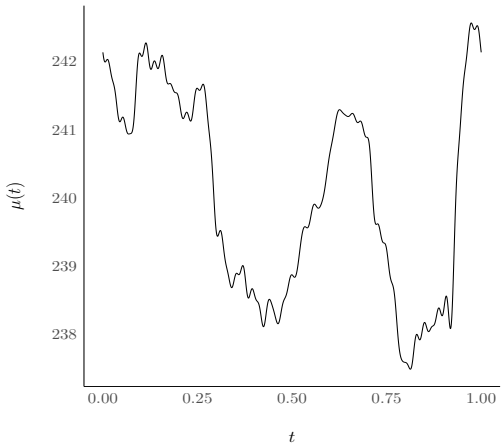
For the estimation of the Hurst index function H_t and Hölder constant function L_t using the Power Consumption dataset, we apply (12) and (31), respectively. The estimated values of H_t and L_t are smoothed using few functions from the Fourier basis. The resulting smoothed functions H and L are plotted in Figures 1a and 1b. The time deformation function $A(t)$ is then estimated using Lemma 1, the result is in Figure 1c. Using these quantities, we estimate the covariance $C_A(\cdot, \cdot)$ of a MfBm process, as defined in (33). Finally, to prevent each curve to start from a same point, we add a random shift $X(0) \sim \mathcal{N}(0, \varpi^2)$. The final covariance of the process is thus given by $\Gamma(s, t) = \varpi^2 + C_A(s, t)$ for all $s, t \in \mathcal{T}$. Next, we generate samples of independent paths $X^{(i)}$ from the Gaussian process characterized by μ and Γ. Finally, to obtain simulated functional data, we add a Gaussian noise of variance σ^2 at any observation time $T_m^{(i)}$.



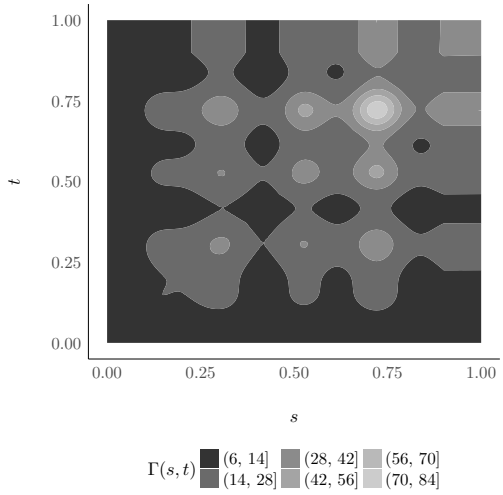
(a) Estimation of $H(t)$ (b) Estimation of $L(t)$ (c) Estimation of $A(t)$

Fig. 1: Estimation of the different quantities for the data generating process.

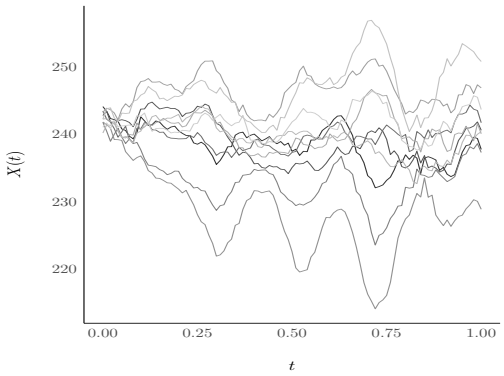
We consider eight experiments, each of them replicated 500 times. For each experiment, except specifically specified, we consider $N \in \{50, 100, 200\}$, $\mathfrak{m} \in \{20, 30, 40, 50\}$ and that the number of points per curve M_i has a Poisson distribution with mean $\mathfrak{m}$. In *Experiment 1*, we assume that the distribution of the sampling points is random uniform in $\mathcal{T}$, the standard deviation of the noise is $\sigma = 0.5$, the regularity of the mean function is $s = \exp(-6)$, the number of Fourier



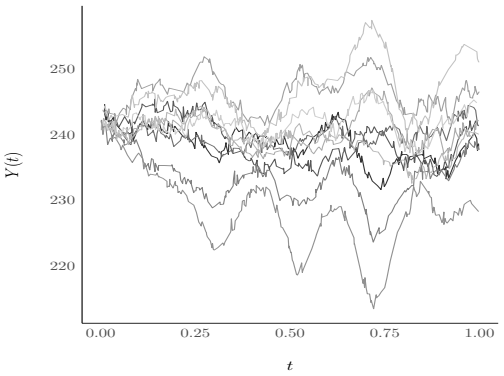
(a) Mean curve $\mu(\cdot)$



(b) Covariance surface $\Gamma(\cdot,\cdot)$



(c) Curves $X^{(i)}$



(d) Noisy curves $Y^{(i)}$

Fig. 2: Description of the simulated dataset

basis functions for the estimation of H_t and L_t is 9, and $\varpi = 2.5$. All the other experiments are designed starting from *Experiment 1* and modifying one parameter at a time. The mean and covariance functions corresponding to *Experiment 1* are plotted in Figures 2a and Figure 2b, respectively. A random sample of ten curves $X^{(i)}$ generated according to *Experiment 1* are plotted on Figure 2c, while their noisy versions $Y^{(i)}$ are in Figure 2d.

In *Experiment 2* and *Experiment 3*, we consider $\sigma = 0.25$ and $\sigma = 1$, respectively. We set $s = \exp(-3)$ for *Experiment 4* resulting in a smoother mean function μ. We used only 7 functions in the Fourier basis in *Experiment 5*, that is a smoother estimation of H_t and L_t. For *Experiment 6*, the distribution of the sampling points is a mixture of beta distributions $0.5\mathcal{B}(1,2) + 0.5\mathcal{B}(2,1)$. For *Experiment 7*, we set $\varpi = 1$. Finally, in *Experiment 8*, we apply our approach to the case of differentiable trajectories that we obtain by integrating the sample paths generated as in *Experiment 1*. The results from *Experiment 1* are presented below, those of the other seven experiments, as well as some additional implementation details, are provided in the Supplementary Material. An implementation of the method used in all experiments is available as a `R` package on Github at the URL adress: `https://github.com/StevenGolovkine/funestim`.

5.3. Mean estimation

For the adaptive estimation of the mean curve, we first compute $\widehat{H}_t$, according to (12), on a uniform grid of 20 points t_2 between 0.2 and 0.8, with $t_3 - t_1 = \Delta_* = \min(\log(\widehat{\mathfrak{m}})^{-1.1}, 0.2)$. The local regularity being a local property, we constrain Δ_* to sufficiently small values. For each value of the 20 estimates $\widehat{H}_t$, we compute the optimal bandwidths h^*_μ by minimization with respect to h over a geometric grid $\mathcal{H}_N$ of 151 points. We then estimate the mean function on 101 regularly spaced points in $[0, 1]$. The 101 bandwidth values used for our estimator are then obtained from the 20 optimal bandwidths h^*_μ by linear interpolation.

Our mean estimator, denoted $\widehat{\mu}_{GKP}$, is compared to that of Cai and Yuan (2011), denoted $\widehat{\mu}_{CY}$, and Zhang and Wang (2016), denoted $\widehat{\mu}_{ZW}$. To compute $\widehat{\mu}_{CY}$, we use the `smooth.splines` function in the `R` package `stats`, with the $M_1 + \ldots + M_N$ data points $(Y_m^{(i)}, T_m^{(i)})$. To obtain $\widehat{\mu}_{ZW}$, we use the `R` package `fdapace`, see Carroll et al. (2021). To compare the estimators, we use the integrated squared error (ISE) risk. For any $\varepsilon \in [0, 1)$, if f and g are real-valued functions defined on $[0, 1]$, let

$$\mathrm{ISE}_\varepsilon(f, g) = \int_{[\varepsilon, 1-\varepsilon]} \{f(t) - g(t)\}^2 dt.$$

The integral is approximated by the trapezoidal rule with an equidistant grid. For each configuration $(N, \mathfrak{m})$, and each of the 500 samples, we compute compute the ratios

$$\frac{\mathrm{ISE}_\varepsilon(\widehat{\mu}_{GKP}, \mu)}{\mathrm{ISE}_\varepsilon(\widehat{\mu}_{CY}, \mu)} \quad \text{and} \quad \frac{\mathrm{ISE}_\varepsilon(\widehat{\mu}_{GKP}, \mu)}{\mathrm{ISE}_\varepsilon(\widehat{\mu}_{ZW}, \mu)},$$

and compare them to 1.

The results for the ISE_0 ratios obtained in *Experiment 1* are plotted in the Figure 3, on a logarithmic scale. To account for a possible boundary effect, we also computed the $\mathrm{ISE}_{0.05}$ ratios, for which the results are similar, and reported in the Supplement. Our mean function estimator reveals good performance. Except some cases where $N\mathfrak{m}$ is large, our estimator outperforms the competitors. In that cases, the three estimators have similar performance. The fact that the advantage of our estimator wanes when $N\mathfrak{m}$ is large could be explained by the fact that the approaches of Cai and Yuan (2011) and Zhang and Wang (2016) smooth over the pooled observations $(Y_m^{(i)}, T_m^{(i)})$. Similar conclusions are drawn from the *Experiments 2* to *7*. In the setup with a more regular mean function (*Experiment 4*), the advantage of our estimator diminishes.

5.4. Covariance estimation

For the adaptive estimation of the covariance function, we use the estimates $\widehat{H}_t$ computed for the mean function on the grid of 20 points between 0.2 and 0.8. For each of the 190 pairs (s, t), $s < t$, on the grid, we compute the optimal bandwidths $h^*_\Gamma(s, t)$ by minimization over a logarithmic grid of 41 points. We then estimate the covariance on a 101×101 regular grid. The 101×101 bandwidth values used for our estimator are obtained from the 190 optimal bandwidths $h^*_\Gamma(s, t)$ by symmetry and linear interpolation.

Our covariance estimator, denoted $\widehat{\Gamma}_{GKP}$, is compared to that of Cai and Yuan (2010), denoted $\widehat{\Gamma}_{CY}$, and from Zhang and Wang (2016), denoted $\widehat{\Gamma}_{ZW}$. We compute $\widehat{\Gamma}_{CY}$ using the `R` package `ssfcov`, see Cai and Yuan (2010). For $\widehat{\Gamma}_{ZW}$, we use the `R` package `fdapace`, see Carroll et al. (2021). To compare the accuracy of the estimators, we use the 2-dimensional ISE risk. For any $\varepsilon \in [0, 1)$, if f and g are real-valued functions defined on $[0, 1] \times [0, 1]$, let

$$\text{2-ISE}_\varepsilon(f, g) = \int_{[\varepsilon, 1-\varepsilon]} \int_{[\varepsilon, 1-\varepsilon]} \{f(s, t) - g(s, t)\}^2 ds dt,$$

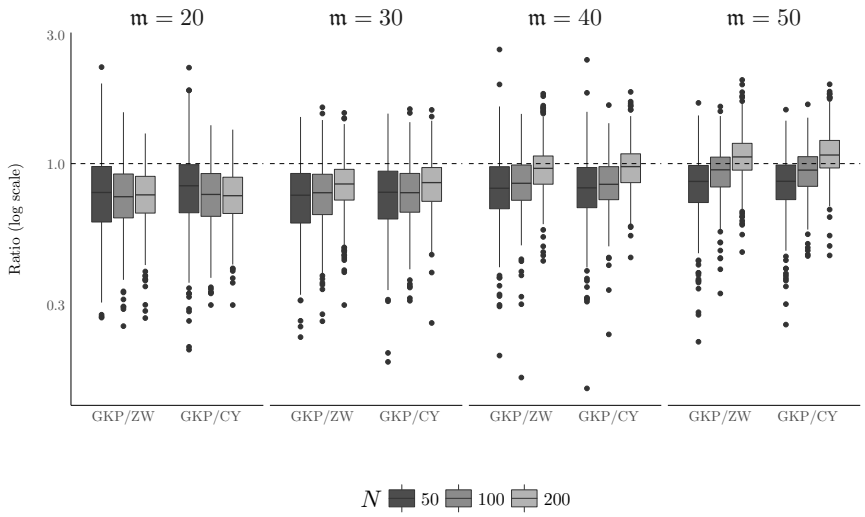


Fig. 3: Results for the estimation of μ in *Experiment 1*. The ratios are computed using ISE_0.

and the integral is approximated by the trapezoidal rule. For each configuration $(N, \mathfrak{m})$, and each replication, we compute the 2-ISE_ε's with respect to the true covariance function Γ. We then compute the ratios

$$\frac{\text{2-ISE}_\varepsilon(\widehat{\Gamma}_{GKP}, \Gamma)}{\text{2-ISE}_\varepsilon(\widehat{\Gamma}_{CY}, \Gamma)} \quad \text{and} \quad \frac{\text{2-ISE}_\varepsilon(\widehat{\Gamma}_{GKP}, \Gamma)}{\text{2-ISE}_\varepsilon(\widehat{\Gamma}_{ZW}, \Gamma)}.$$

The results on the ratios obtained with 2-ISE_0 in *Experiment 1* are plotted in the Figure 4, on a logarithmic scale. Those obtained with 2-$\text{ISE}_{0.05}$, presented in the Supplement, are similar. Our estimator shows better accuracy for estimating Γ than $\widehat{\Gamma}_{ZW}$ and $\widehat{\Gamma}_{CY}$ in all cases considered. The advantage of our approach increases with N.

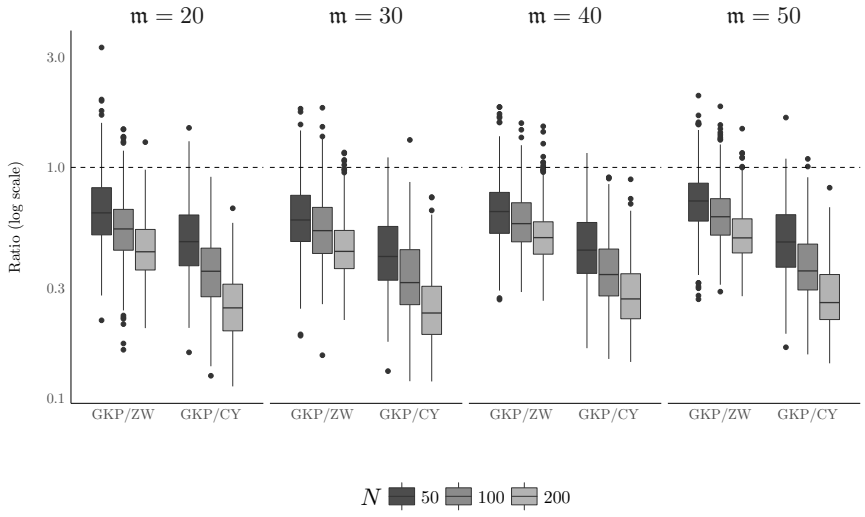


Fig. 4: Estimation of Γ in *Experiment 1*. The ratios are computed using 2-ISE_0.

5.5. Eigenvalues estimation

We also consider the estimation of the eigenvalues of Γ based on *Experiment 1*. We only compare our results with the ones obtained from Zhang and Wang (2016)'s covariance estimator, because it was more accurate, and it can be computed much faster than the one of Cai and Yuan (2010). The true eigenvalues $\lambda_i, i = 1, 2, \ldots$ are estimated using the eigendecomposition of the true covariance function Γ computed on the 101×101 grid. The eigenvalues $\widehat{\lambda}_{i,GKP}$ and $\widehat{\lambda}_{i,ZW}$, $i = 1, 2, \ldots$ are estimated using the eigendecomposition of the covariance matrices $\widehat{\Gamma}_{GKP}$ and $\widehat{\Gamma}_{ZW}$, respectively. The logarithm of the eigenvalues computed using the true covariance Γ are plotted in the Figure 5. We compare the estimations using the log-ratios

$$\log(|\lambda_i - \widehat{\lambda}_{i,GKP}|) - \log(|\lambda_i - \widehat{\lambda}_{i,ZW}|), \quad i \in \{1, \ldots, 5\}. \tag{34}$$

The results obtained for the estimation of the eigenvalues are plotted in the Figure 6, on the logarithmic scale. Our estimators reveals better performance compared in most of the cases.

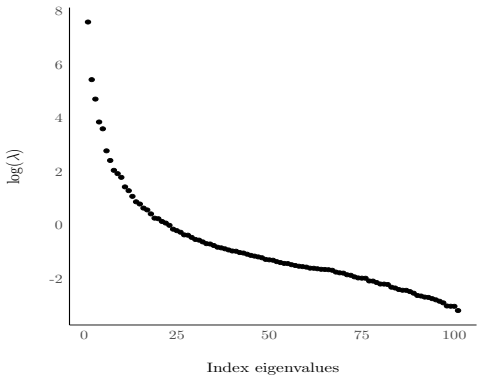


Fig. 5: Eigenvalues derived from the eigendecompostion of $\Gamma(s, t)$.

6. Discussion and conclusions

We propose new nonparametric estimators for the mean and covariance functions. They are built using a novel "smoothing first, then estimate" strategy based on univariate kernel smoothing. The main novelty comes from the fact that the optimal bandwidths are selected by minimization of suitable penalized quadratic risks. The penalized risks for the mean and the covariance functions are quite similar, could be easily built from data, and be optimized on a grid of bandwidths. What distinguishes them from the usual sum between the squared bias and the variance, is a penalty for the fact that not all the curves have enough observation points to be included in the final estimator. Removing curves from the nonparametric estimators of the mean and covariance functions is an aspect which characterizes practically all smoothing-based approaches. Indeed, to entirely benefit from the replication feature of functional data, one has to determine the amount of smoothing for the mean and covariance estimation using all the curves. In this case, some curves could present too few observation points and thus will be dropped. This is more likely to happen in the so-called sparse regime. To our best knowledge, our bandwidth choice is the first attempt to explicitly account for this aspect. We thus build estimators which achieve optimal rates of convergence in a completely adaptive, data-driven way. The theoretical results are derived under very mild conditions. In particular, the curves could be observed with heteroscedastic errors at discrete observations points. These points could be common to all curves or they could

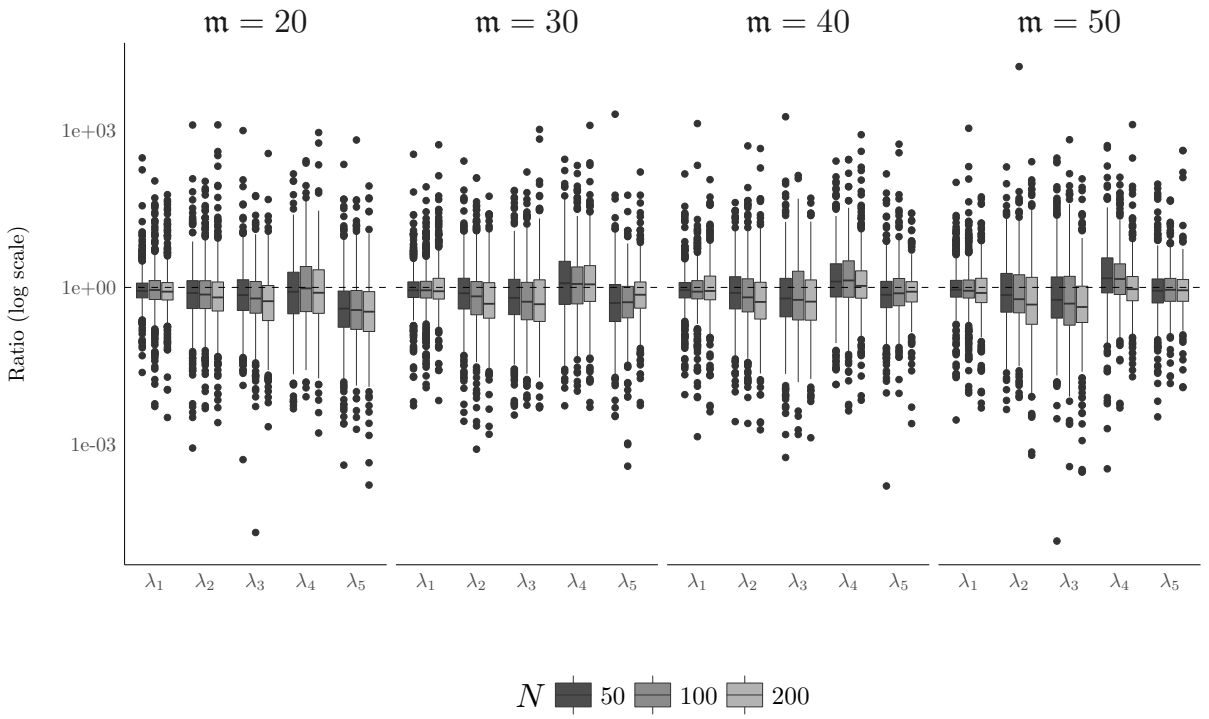


Fig. 6: Estimation of the eigenvalues λ for *Experiment 9*: log-ratios defined as in (34)

change randomly from one curve to another. In the case of the common observation points, our procedure automatically chooses between smoothing and interpolation, the latter being known to be rate optimal, but is not necessarily the best solution with finite samples.

Our nonparametric estimation approach relies on a probabilistic concept of local regularity for the sample paths of the process generating the curves. In some common examples, this local regularity is related to the polynomial decrease rate of the eigenvalues of the covariance operator, a characteristic of the data generating process widely used in the literature and usually supposed to be known. The local regularity also determines the regularity of the trajectories, the usual concept used in nonparametric regression. It is well-known that the optimal rates, in the minimax sense, for estimating the mean and covariance functions depend on the regularity of the paths. Moreover, the so-called sparse and dense regimes in functional data analysis, are defined using the regularity of the trajectories, which usually is supposed to be known. We therefore consider a simple estimator of the local regularity of the process and we use it to build our penalized quadratic risk. Applied to real data, the local regularity estimator reveals that the regularity of the trajectories could be quite far from what is usually assumed in the literature. However, in some applications assuming smooth trajectories seems reasonable. The mean and covariance functions estimation approach based on local regularity extends to such situations. In the Supplement, we provide empirical evidence of good performance of our mean function estimator with differentiable trajectories.

Our method performs well in simulations and outperforms the main competitors when the mean and covariance functions have a regularity close to that of the trajectories. The approach is still satisfactory when the mean or the covariance functions are more regular than the trajectories. The reason is that, in some sense, our nonparametric estimators are as close as possible to the

empirical mean and covariance, respectively, which are the ideal estimators if the trajectories were observed at any point without error. In the case where the mean and covariance function are smoother than the trajectories, our penalized quadratic risk should be built using the mean or covariance functions' regularity instead of the trajectories' regularity. However, for now, the estimation of the regularity of the mean or covariance function remains an open problem.

A. Technical details and proofs

PROOF (DETAILS ON (19)). To explain $\mathcal{R}_\mu(t;h)$ in the case of non differentiable paths, let

$$\widetilde{\mu}_W(t;h) = \frac{1}{\mathcal{W}_N(t;h)} \sum_{i=1}^{N} w_i(t;h) X_t^{(i)},$$

be the unfeasible estimator of $\mu(\cdot)$ using only the curves for which $\widehat{X}_t^{(i)}$ is well-defined. In the following, we write w_i and $\mathcal{W}_N$ instead of $w_i(t;h)$ and $\mathcal{W}_N(t;h)$, respectively. By (6),

$$\begin{aligned}
\mathbb{E}_{M,T}\left[\{\widetilde{\mu}_N(t) - \widehat{\mu}_N(t;h)\}^2\right] &= \mathbb{E}_{M,T}\left[\left\{\widetilde{\mu}_N(t) - \widetilde{\mu}_W - \frac{1}{\mathcal{W}_N}\sum_{i=1}^{N} w_i\left(B_t^{(i)} + V_t^{(i)}\right)\right\}^2\right] \\
&\leq 2\mathbb{E}_{M,T}\left[\{\widetilde{\mu}_N(t) - \widetilde{\mu}_W(t;h)\}^2\right] + 2\mathbb{E}_{M,T}\left[\left\{\frac{1}{\mathcal{W}_N}\sum_{i=1}^{N} w_i\left(B_t^{(i)} + V_t^{(i)}\right)\right\}^2\right] =: 2E_1 + 2E_2.
\end{aligned}$$

Since

$$\widetilde{\mu}_W(t;h) - \widetilde{\mu}_N(t) = \frac{1}{\mathcal{W}_N}\sum_{i=1}^{N}\left\{X_t^{(i)} - \mu(t)\right\}\left\{w_i - \frac{\mathcal{W}_N}{N}\right\},$$

the trajectories of X are drawn independently, and independently of the M_i and the $T_m^{(i)}$, we have

$$E_1 = \frac{\mathrm{Var}(X_t)}{\mathcal{W}_N^2}\sum_{i=1}^{N}\left\{w_i - \frac{\mathcal{W}_N}{N}\right\}^2 = q_3^2\left\{\frac{1}{\mathcal{W}_N} - \frac{1}{N}\right\}.$$

For E_2, let us first look at the bias part. By Theorem 1, there exists $\varrho > 1$ such that the probability of the event $\{|\widehat{H}_t - H_t| > \log^{-\varrho}(\mathfrak{m})\}$ is exponentially small. Hence, by (22) and (23), we have $h^{2\widehat{H}_t} = h^{2H_t}\{1 + o_{\mathbb{P}}(1)\}$, uniformly over the range $\mathcal{H}_N$. Next, by (H2),

$$\mathbb{E}_{M,T}\left(\left\{X^{(i)}(T_m^{(i)}) - X_t^{(i)}\right\}^2\right) = \mathbb{E}\left(\left\{X^{(i)}(T_m^{(i)}) - X_t^{(i)}\right\}^2 \mid \mathcal{T}_{obs}^{(i)}\right) = \{1 + o_{\mathbb{P}}(1)\}L_t^2\left|(T_m^{(i)} - t)/h\right|^{2H_t}.$$

Similarly to (8) and (10), for $\widehat{X}_t^{(i)}$ the NW estimator and $\overline{C}$ defined in (20), we then have

$$\begin{aligned}
&\mathbb{E}_{M,T}\left[\left\{\frac{1}{\mathcal{W}_N}\sum_{i=1}^{N} w_i B_t^{(i)}\right\}^2\right] \\
&\quad\leq L_t^2 h^{2H_t}\{1 + o_{\mathbb{P}}(1)\} \times \frac{1}{\mathcal{W}_N}\sum_{i=1}^{N} w_i\left\{\sum_{m=1}^{M_i} W_m^{(i)}(t) \times \sum_{m=1}^{M_i}\left|\frac{T_m^{(i)} - t}{h}\right|^{2H_t} W_m^{(i)}(t)\right\} \\
&\quad= L_t^2 h^{2\widehat{H}_t} \times \overline{C}(t;h,2\widehat{H}_t) \times \{1 + o_{\mathbb{P}}(1)\} = L_t^2 h^{2\widehat{H}_t} \times \int |u|^{2\widehat{H}_t} K(u)du \times \{1 + o_{\mathbb{P}}(1)\}.
\end{aligned}$$

Using the equivalent kernels, see Section 3.2.2 in Fan and Gijbels (1996), the bound on the last line of the last display could be extended to the case of local linear estimators.

To complete the bound for E_2, note that by construction, $\mathbb{E}_{M,T}\left\{V_t^{(i)}B_t^{(i)}\right\}=0$,

$$\mathbb{E}_{M,T}\left\{V_t^{(i)}B_t^{(j)}\right\}=\mathbb{E}_{M,T}\left\{V_t^{(i)}V_t^{(j)}\right\}=0,\quad \forall 1\leq i\neq j\leq N.$$

The variance part in E_2 can be bounded as in (7). Up to negligible terms, for the NW estimator,

$$E_2\leq h^{2\widehat{H}_t}L_t^2\overline{C}(t;h,2\widehat{H}_t)+\frac{\sigma^2_{\max}}{\mathcal{W}_N^2}\sum_{i=1}^N w_i\mathcal{N}_i^{-1}(t;h)=q_1^2h^{2\widehat{H}_t}+q_2^2\mathcal{N}_\mu^{-1}(t;h).$$

Proof (Theorem 2). First, note that if $\mathcal{W}_N(t;h)=0$, then necessarily $\mathcal{N}_\mu(t;h)=0$. Moreover, it will be shown below that $\inf_{h\in\mathcal{H}_N}\mathbb{E}[\mathcal{W}_N(t;h)]$ stays away from zero, and $\mathcal{W}_N(t;h)$ uniformly concentrates to $\mathbb{E}[\mathcal{W}_N(t;h)$ with high probability. We therefore, in the following, work on the event $\{\inf_{h\in\mathcal{H}_N}\mathcal{W}_N(t;h)\geq 1\}$. First, let us prove that a constant $C>0$ exists such that

$$0\leq \mathcal{W}_N(t;h)^{-1}-N^{-1}\leq C\max\left\{h^{2H_t},\ \mathcal{N}_\mu^{-1}(t;h)\right\}\{1+r_N(h)\},\tag{A.1}$$

with $\sup_{h\in\mathcal{H}_N}|r_N(h)|=o_{\mathbb{P}}(1)$. Property (A.1) is implied by the following ones: two constants $\mathfrak{c}_1,\mathfrak{c}_2>0$ exist such that

$$\inf_{h\in\mathcal{H}_N}\frac{\mathcal{W}_N(t;h)}{\min\{1,\mathfrak{m}h\}}\geq \mathfrak{c}_1N\{1+o_{\mathbb{P}}(1)\}.\tag{A.2}$$

and

$$\inf_{h\in\mathcal{H}_N}\frac{N\mathfrak{m}h}{\mathcal{N}_\mu(t;h)}\geq \mathfrak{c}_2\{1+o_{\mathbb{P}}(1)\},\tag{A.3}$$

Indeed, (A.2) and (A.3) imply

$$\begin{aligned}\mathcal{W}_N(t;h)^{-1}-N^{-1}&\leq \max\left\{0,\ \mathfrak{c}_1^{-1}(N\min\{1,\mathfrak{m}h\})^{-1}-N^{-1}\right\}\{1+o_{\mathbb{P}}(1)\}\\&\leq \mathfrak{c}_1^{-1}\mathfrak{c}_2\max\left\{h^{2H_t},\ \mathcal{N}_\mu(t;h)^{-1}\right\}\{1+o_{\mathbb{P}}(1)\},\end{aligned}$$

with the $o_{\mathbb{P}}(1)$ terms uniform with respect to $h\in\mathcal{H}_N$. The detailed justification of (A.2) and (A.3) is provided in the Supplement. Let us provide a brief insight on how these properties are obtained. Here, $\mathcal{W}_N(t;h)$ is a Binomial variable with N trials and the success probability a non decreasing function of h. The property (A.2) follows by suitably bounding $\mathbb{E}[\mathcal{W}_N(t;h)]$ and using Chernoff's inequality on a grid of points in $\mathcal{H}_N$. The uniformity with respect to all $h\in\mathcal{H}_N$ is obtained using the monotonicity of $\mathcal{W}_N(t;h)$ with respect to h. For (A.3), by definition, $\mathcal{W}_N(t;h)\mathcal{N}_\mu(t;h)^{-1}$ is the mean over the curves with $w_i=1$ of the $\max_{1\leq m\leq M_i}|W_m^{(i)}(t;h)|$. Then (A.3) will be obtained using a positive lower bound for the kernel K on a sub-interval of the support, Cauchy-Schwarz inequality and Chernoff's inequality. Finally, to complete the proof in the independent design case, it suffices first to notice that from above, we have

$$\min\left\{h^{2H_t}+\mathcal{N}_\mu^{-1}(t;h)\right\}=O_{\mathbb{P}}\left(h^{2H_t}+(N\mathfrak{m}h)^{-1}\right),\tag{A.4}$$

which is minimized by h with the rate $(N\mathfrak{m})^{-1/\{2H_t+1\}}$. The details on (A.4) are also provided in the Supplement. Next, by (22) and (23), uniformly over $h\in\mathcal{H}_N$, we have $h^{2\widehat{H}_t}=h^{2H_t}\{1+o_{\mathbb{P}}(1)\}$. The rate of $\widehat{\mu}_N^*(t)-\widetilde{\mu}(t)$ follows. For the rate of $\widehat{\mu}_N^*(t)-\mu(t)$, we simply add the rate of $\widetilde{\mu}_N(t)-\mu(t)$.

With a common design, $\mathcal{W}_N(t;h)$ can only take the values 0 or N. Thus the penalty introduced by $\mathcal{W}_N(t;h)^{-1} - N^{-1}$ constrains the bandwidth to be greater than or equal to the lengths of the intervals $[T_m^{(i)}, T_{m+1}^{(i)}]$ including t. By condition (24), this means that the rate of convergence of $\widehat{\mu}_N^*(t) - \widetilde{\mu}_N(t)$ could not be faster than $O_{\mathbb{P}}\left(\mathfrak{m}^{-2H_t}\right)$. This aspect is automatically included in the definition of $\mathcal{R}_\mu(t;h)$ because, under the constraint $\mathfrak{m}h \geq c_L/C_U$,

$$\min\left\{h^{2H_t} + \mathcal{N}_\mu^{-1}(t;h)\right\} = O_{\mathbb{P}}\left(\left\{h^{2H_t} + (N\mathfrak{m}h)^{-1} + N^{-1}\right\}\right) = O_{\mathbb{P}}\left(\mathfrak{m}^{-2H_t}\right).$$

Finally, H_t can be replaced by $\widehat{H}_t$ using again $h^{2\widehat{H}_t} = h^{2H_t}\{1 + o_{\mathbb{P}}(1)\}$.

The proof of Theorem 3 follows the lines of that of the proof of Theorem 2 and is left to the Supplementary Material.

Proof (Equation (30)). Let

$$\widetilde{D}_t(\mathfrak{u}_1, \mathfrak{u}_2) = \widetilde{\Gamma}_N(t - \mathfrak{u}_1, t + \mathfrak{u}_2) - \widetilde{\Gamma}_N(t, t).$$

To derive the bound in (30), we use the assumptions: $\sup_{t\in\mathcal{T}} \mathbb{E}(X_t^4) < \infty$ and a constant c exists such that

$$\mathbb{E}(\{X_s - X_t\}^4) \leq c\mathbb{E}^2(\{X_s - X_t\}^2), \quad \forall s, t \in \mathcal{T}. \tag{A.5}$$

We then have

$$\mathbb{E}[\widetilde{D}_t(\mathfrak{u}_1, \mathfrak{u}_2)^2] \leq 2\mathbb{E}\left[\left(\frac{1}{N}\sum_{i=1}^N \left(\{X_t^{(i)}\}^2 - X_{t-\mathfrak{u}_1}^{(i)} X_{t+\mathfrak{u}_2}^{(i)}\right)\right)^2\right]$$
$$+ 2\mathbb{E}\left[\left(\left\{\frac{1}{N}\sum_{i=1}^N X_t^{(i)}\right\}^2 - \left\{\frac{1}{N}\sum_{i=1}^N X_{t-\mathfrak{u}_1}^{(i)}\right\}\left\{\frac{1}{N}\sum_{i=1}^N X_{t+\mathfrak{u}_2}^{(i)}\right\}\right)^2\right] =: 2D_1 + 2D_2.$$

By (H2), (A.5), and Jensen and Cauchy-Schwarz inequalities, a constant C_1 exists, depending on L_t, S, and c appearing in (A.5), such that

$$D_1 \leq \mathbb{E}\left[\left\{X_t^2 - X_{t-\mathfrak{u}_1} X_{t+\mathfrak{u}_2}\right\}^2\right] \leq C_1 \mathfrak{d}_t^{2H_t},$$

provided $0 \leq \mathfrak{u}_1, \mathfrak{u}_1 \leq \mathfrak{d}_t \leq \Delta_*/2$. On the other hand, by similar arguments,

$$D_2 \leq 2\mathbb{E}^{1/2}\left[\left(\frac{1}{N}\sum_{i=1}^N \left\{X_t^{(i)} - X_{t-\mathfrak{u}_1}^{(i)}\right\}\right)^4\right] \mathbb{E}^{1/2}\left[\left(\frac{1}{N}\sum_{i=1}^N X_t^{(i)}\right)^4\right]$$
$$+ 2\mathbb{E}^{1/2}\left[\left(\frac{1}{N}\sum_{i=1}^N X_{t-\mathfrak{u}_1}^{(i)}\right)^4\right] \mathbb{E}^{1/2}\left[\left(\frac{1}{N}\sum_{i=1}^N \left\{X_t^{(i)} - X_{t+\mathfrak{u}_2}^{(i)}\right\}\right)^4\right] \leq C_2 \mathfrak{d}_t^{2H_t},$$

for some constant C_2. Gathering facts, we deduce that (30).

Proof (Lemma 1). By construction, $\mathbb{E}\left[W_A(A(t))\right] = 0$, and the covariance function of X is

$$\mathrm{Cov}_A(s,t) = D(H_s, H_t)\left[A(s)^{H_s+H_t} + A(t)^{H_s+H_t} - |A(t) - A(s)|^{H_s+H_t}\right], \quad s, t \geq 0.$$

Moreover, for any t and $u, v \in \mathcal{O}_*(t)$, we have

$$\mathbb{E}\left[(X_u - X_v)^2\right] \approx \{A'(t)\}^{2H_t}|u-v|^{2H_t}. \tag{A.6}$$

The formal proof of (A.6) is provided in the Supplement. To match condition (H2), it suffices to define $A(\cdot)$ such that $\{A'(t)\}^{H_t} = L_t$, and thus $A(t) = \int_0^t L_s^{1/H_s} ds$.

Acknowledgements

The authors thank Groupe Renault and the ANRT (French National Association for Research and Technology) for their financial support via the CIFRE convention No. 2017/1116. S. Golovkine was partially supported by Science Foundation Ireland under Grant No. 19/FFP/7002 and co-funded under the European Regional Development Fund. V. Patilea gratefully acknowledges support from the Joint Research Initiative "Models and mathematical processing of very large data" under the aegis of Risk Foundation, in partnership with MEDIAMETRIE and GENES, France, and from the grant of the Romanian Ministry of Research, Innovation and Digitization, CNCS/CCCDI – UEFISCDI, project number PN-III-P4-ID-PCE-2020-1112, within PNCDI III.

Supplementary Material

In the Supplementary Material, we provide additional technical arguments and simulation results. In Section A and B, we provide technical details on some proofs and equations presented above. In Section C, we prove Theorem 3. Additional simulation results are gathered in Section D.

References

Balança, P. (2015) Some sample path properties of multifractional Brownian motion. *Stochastic Processes Appl.*, **125**, 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**, 345 – 366.

Bertin, K. (2004) Minimax exact constant in sup-norm for nonparametric regression with random design. *Journal of Statistical Planning and Inference*, **123**, 225–242.

Blanke, D. and Vial, C. (2014) Global smoothness estimation of a Gaussian process from general sequence designs. *Electron. J. Stat.*, **8**, 1152–1187.

Cai, T. and Yuan, M. (2010) Nonparametric covariance function estimation for functional and longitudinal data. *University of Pennsylvania and Georgia Institute of Technology*.

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

— (2016) Minimax and adaptive estimation of covariance operator for random variables observed on a lattice graph. *J. Amer. Statist. Assoc.*, **111**, 253–265.

Carroll, C., Gajardo, A., Chen, Y., Dai, X., Fan, J., Hadjipantelis, P. Z., Han, K., Ji, H., Mueller, H.-G. and Wang, J.-L. (2021) *fdapace: Functional Data Analysis and Empirical Dynamics*. R package version 0.5.6.

Fan, J. and Gijbels, I. (1996) *Local polynomial modelling and its applications*. Monographs on statistics and applied probability. London: Chapman & Hall.

Gaïffas, S. (2007) Sharp estimation in sup norm with random design. *Statist. Probab. Lett.*, **77**, 782–794.

Goldenshluger, A. and Lepski, O. (2011) Bandwidth selection in kernel density estimation: Oracle inequalities and adaptive minimax optimality. *Ann. Statist.*, **39**, 1608–1632.

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**.

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

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

Ramsay, J. and Silverman, B. W. (2005) *Functional Data Analysis*. Springer Series in Statistics. New York: Springer-Verlag, 2 edn.

Revuz, D. and Yor, M. (2013) *Continuous Martingales and Brownian Motion*. Springer Science & Business Media.

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

Wong, R. K. and Zhang, X. (2019) Nonparametric operator-regularized covariance function estimation for functional data. *Comput. Statist. Data Anal.*, **131**, 131 – 144.

Zhang, J.-T. and Chen, J. (2007) Statistical inferences for functional data. *Ann. Statist.*, **35**, 1052–1079.

Zhang, X. and Wang, J.-L. (2016) From sparse to dense functional data and beyond. *Ann. Statist.*, **44**, 2281–2321.

— (2018) Optimal weighting schemes for longitudinal and functional data. *Statist. Probab. Lett.*, **138**, 165 – 170.