Estimating the covariance of fragmented and other related types of functional data

Aurore Delaigle\(^1\), Peter Hall\(^{1,3}\), Wei Huang\(^1\) and Alois Kneip\(^2\)

\(^1\)ACEMS and School of Mathematics and Statistics, University of Melbourne, Australia
\(^2\)Department of Economics and Hausdorff Center for Mathematics, Universität Bonn, Germany
\(^3\)This work was initiated with Peter Hall during a visit of Kneip to the University Melbourne. Peter died on January 9, 2016.

Abstract: We consider the problem of estimating the covariance function of functional data which are only observed on a subset of their domain, such as fragments observed on small intervals or related types of functional data. We focus on situations where the data enable to compute the empirical covariance function or smooth versions of it only on a subset of its domain which contains a diagonal band. We show that estimating the covariance function consistently outside that subset is possible as long as the curves are sufficiently smooth. We establish conditions under which the covariance function is identifiable on its entire domain and propose a tensor product series approach for estimating it consistently. We derive asymptotic properties of our estimator and illustrate its finite sample properties on simulated and real data.

Keywords: identification, incomplete functional data, multiple fragments, orthogonal series, tensor product.

1 Introduction

We consider the problem of estimating the covariance function of partially observed independent and identically distributed functional data \(X_1, \ldots, X_n\) defined on an interval \(I\). We assume that the pairs of points \((s,t) \in S_0 \equiv I \times I\) at which we observe \(X_i(s)\) and \(X_i(t)\) (or a discrete and noisy version of them) for at least one \(i\), cover only a part of \(S_0\) which contains an area around the main diagonal of \(S_0\).

We are particularly interested in the case where only a fragment of each curve \(X_i\) is observed on a subinterval \(I_i\) of \(I\) \cite{delaigle2013, delaigle2016}. The left panel of Figure 1 illustrates an example. There, we are interested in
Figure 1: Left: subsample of 10 curve fragments (—) of the type considered in this paper, taken from a sample of size $n = 100$ of curve fragments $X_i(t)$, for $t \in \mathcal{I}_i \subset \mathcal{I} = [1, 100]$ and $i = 1, \ldots, n$; the long dotted lines show the 10 corresponding unobserved full curves. Right: scatterplot of points $(s, t) \in \mathcal{I} \times \mathcal{I}$ where at least one pair $(X_i(s), X_i(t))$ is observed, for $i = 1, \ldots, n$.

10 functional data $X_i(t)$ on a time interval $t \in \mathcal{I} = [1, 100]$, $i = 1, \ldots, 10$, but we only observe a fragment of each of them.

Estimating the covariance function on its entire domain $\mathcal{S}_0$ from such data is challenging; with fragmentary data, the pairs of points $(s, t) \in \mathcal{S}_0$ at which we observe $X_i(s)$ and $X_i(t)$ cover only a band around the main diagonal of $\mathcal{S}_0$, corresponding to the subset $\bigcup_{i=1}^{n} (\mathcal{I}_i \times \mathcal{I}_i)$. This is illustrated in the right panel of Figure 1. There we show those pairs of points for the full sample of $n = 100$ fragments generated as those in the left panel. When the functional data are fully observed, usual techniques for estimating a covariance function include the empirical covariance function and bivariate kernel or spline smoothers. See e.g. Rice and Silverman (1991) and Xiao et al. (2013); see also Ramsay and Silverman (2005), Ferraty and Vieu (2006) and Hsing and Eubank (2015) for general introductions to functional data. When only fragments are observed, those techniques can estimate the covariance function only on the subset where we have data, but not on the entire $\mathcal{S}_0$. Indeed, to estimate a covariance function at $(s, t) \in \mathcal{S}_0$, such methods require that at least some of the $(X_i(s), X_i(t))$’s be observed at or near $(s, t)$. However, in our setting, even when sample size increases, only a part of $\mathcal{S}_0$ is covered by observed data.
Our context is different from the one in Kraus (2015), Gromenko et al. (2017) and Liebl and Rameseder (2019). Those authors also consider fragmentary data, but in their case a sufficiently large proportion of the curves are either completely observed or have well scattered multiple observed fragments. In their setting, the points \((s, t)\) where at least one pair \((X_i(s), X_i(t))\) is observed entirely cover \(S_0\).

Our context is also markedly different from the so-called sparse functional data setting (see e.g. Yao et al., 2005 and Peng and Paul, 2009). There each curve \(X_i\) is observed at a sparse number of points which can roughly be located anywhere on the domain \(I\) of \(X_i\). As a result, the points \((s, t)\) where at least one \((X_i(s), X_i(t))\) is observed entirely cover \(S_0\). In Figure 2 we show sparse versions \(X_i(T_{ij})\) of the 10 curves \(X_1, \ldots, X_{10}\) in Figure 1 obtained by taking \(T_{ij} \sim U[1, 100]\), for \(j = 1, 2, 3\). The right panel of Figure 2 shows the \((s, t)\) pairs where at least one \((X_i(s), X_i(t))\) is observed, for a sample of 100 sparsely observed curves generated in the same way. As illustrated there, unlike in our context shown in Figure 1, here \(S_0\) is roughly covered by the pooled data even if each \(X_i\) is observed at only a few points.

In earlier work, James and Hastie (2001) developed a spline-based mixed model technique that they applied to a discretised version of fragments. However, their technique too is better suited for sparse functional data and generally does not provide consistent estimators for fragmentary data. See section 5.2 for an illustration.

To summarise, the problem of estimating a covariance function or matrix from incomplete data has received a lot of attention in the literature (see also Loh and Wainwright, 2012, Kolar and Xing, 2012, Lounici, 2014 and Cai and Zhang, 2016 for the related problem of estimating a covariance matrix from incomplete vectors). However, the existing techniques rely on the fact that, while individual data are incomplete, the pooled data from all individuals roughly cover the whole domain of the covariance function or matrix. This makes it possible to use interpolation techniques. Our context is markedly different: even when the observed \((X_i(s), X_i(t))\)'s are pooled, the corresponding \((s, t)\)'s only cover a band within \(S_0\). Thus, in order to recover the
While in the process of completing our project, we became aware of two approaches developed independently by Zhang and Chen (2017) and Descary and Panaretos (2019), based on matrix completion techniques applied to a discretised version of fragmentary data. Those approaches have similarities with ours, in the sense that the covariance function on $S_0$ we need to resort to extrapolation techniques.

Unsurprisingly, in our case the covariance function can be estimated on $S_0$ only if relatively strong structural assumptions are made. Fan et al. (2007) proposed a semiparametric estimator where the diagonal of the covariance function is estimated nonparametrically and the off-diagonal components are modeled parametrically. Laigle and Hall (2013) assumed that the curves located near each other behave similarly. Exploiting this fact they could extrapolate the fragments to the entire interval $I_i$ and obtain from there a covariance estimator. Delaigle and Hall (2016) suggested a related but more sophisticated method, assuming that a discretised version of the curves satisfies a low-order Markov assumption. In this work we propose more general conditions under which we can identify the covariance function on its entire domain from incompletely observed functional data such as fragmentary data, and suggest a way to estimate it consistently.

Figure 2: Left: subsample of 10 sparsely observed curves ($\circ$), markedly different from the fragment data considered in this paper, and taken from a sample of size $n = 100$ of sparsely observed curves $X_i(t)$, for $t \in I_i \subset I = [1, 100]$ and $i = 1, \ldots, n$; the long dotted lines show the 10 corresponding unobserved full curves. Right: scatterplot of points $(s, t) \in I \times I$ where at least one pair $(X_i(s), X_i(t))$ is observed, for $i = 1, \ldots, n$. 
authors also develop identifiability conditions under which an estimator of a discretised $K$, constructed on a band where we have data, can be extended to a consistent estimator of the discretised $K$ on the entire domain $S_0$. As we shall see in sections 3.1 and 5.2.1, compared to our results, these require stronger identifiability conditions and produce more noisy estimators when the data are of small size.

We define our setting of incomplete functional data and propose a nonparametric smooth covariance estimator based on a tensor product basis expansion in section 2. We define a subset $S$ of $S_0$ where we can compute an empirical covariance estimator from the incomplete data. Then we estimate the coefficients of the tensor product basis expansion in such a way that the tensor product estimator and the empirical estimator are as close as possible on $S$. For this technique to be valid, we assume that the coefficients on the subset $S$ can be extrapolated to the entire domain $S_0$. This can only be done if we can identify the covariance function on $S_0$ by knowing the covariance function on $S$. In section 3 we establish a set of more interpretable identifiability conditions and show that they hold if the unobserved complete curves are sufficiently smooth. We derive theoretical properties of our estimator in section 4. There we prove consistency and show that, under sufficient conditions, the convergence rate of our estimator on the entire set $S_0$, is identical to that of the empirical covariance estimator that can only be computed on the subset $S$. In section 5 we apply our estimator to simulated and real data. In the real data example, we illustrate the eigenfunctions obtained from our covariance estimator and show how the latter can be used to reconstruct the unobserved parts of the curves in our application. All proofs are derived in the supplementary file.
2 Model and methodology

2.1 Model and data

We are interested in estimating the covariance function \( K(s, t) = E\{X_i(s)X_i(t)\} - \mu(s)\mu(t) \) of i.i.d. curves \( X_1, \ldots, X_n \) defined on a compact interval \( \mathcal{I} = [a, b] \), where \( \mu(t) = E\{X_i(t)\} \) and \( s, t \in \mathcal{I} \). We assume that \( K \) and \( X \) are continuous. Instead of observing the entire curves, for \( i = 1, \ldots, n \), we observe \( X_i \) only on a subset of \( \mathcal{I} \). We are particularly interested in the case where only a fragment of each curve \( X_i \) is observed on an interval \( \mathcal{I}_i \subset \mathcal{I} \) and \( \mathcal{I}_i \) is independent of \( X_i \) (see e.g. Delaigle and Hall, 2013, Delaigle and Hall, 2016).

Our results also apply to more general data types. Indeed, all we require of the observed data is that there exists at least a subset \( S \subset S_0 \) such that

\[
\begin{align*}
\text{i) } & 0 < \sup \{\delta \mid ([t, t + \delta] \cap \mathcal{I}) \times ([t, t + \delta] \cap \mathcal{I}) \in S \text{ for all } t \in \mathcal{I}\}, \text{ i.e. } S \text{ is symmetric around the diagonal of } S_0; \\
\text{ii) } & \text{We have enough data on } S \text{ so that we can consistently estimate } K \text{ on } S \text{ by directly applying existing empirical methods (e.g. the empirical covariance estimator or smooth versions of this estimator).}
\end{align*}
\]

(2.1)

This is a weak condition satisfied by many types of data. For example, if each curve is observed only at two random points \( T_{i,1} \) and \( T_{i,2} \), then as long as the \( T_{i,j} \)'s roughly cover the interval \( \mathcal{I} \), Condition (2.1) is satisfied.

Our main goal is to deal with the case where we observe relatively short fragments of the curves (or discrete and/or noisy versions of them), as in Delaigle and Hall (2013) and Delaigle and Hall (2016). An example of short fragments is shown on the left panel of Figure 1; see also Figure A1 in Appendix A for an illustration of discretely observed fragments. Both examples satisfy Condition (2.1). For instance, for the data in Figure 1, we can take \( S \) to be the hull of the dark area on the right panel, since, from those data, we can compute the empirical covariance estimator of \( K(s, t) \) for \( (s, t) \) in that area.

With such data, it is not possible to estimate \( K \) outside a band \( S \) without making
structural assumptions that permit to identify $K$ on $S_0$; see section 3. The results and techniques we shall develop also apply when the data are richer, for example if we observe several fragments per curve or a combination of fragment(s) and sparse observation(s). However, the estimation procedure we develop in section 2.2 resorts to extrapolation techniques, which are the most beneficial in settings where $S$ is significantly smaller than $S_0$, such as the short fragments example. Therefore most of our discussion will be focused on the short fragment setting.

2.2 Method for estimating the covariance from incomplete data

Our idea is to construct a consistent estimator $\hat{K}$ of $K$ on $S$ and extrapolate $\hat{K}$ from $S$ to $S_0$ using a series approximation. Here $S$ is the set defined in (2.1); see section 5.1 for a construction of $S$ in practice for the fragmentary case. For example, in the case where we observe one or several fragments of each curve, at each $(s,t) \in S$ we can take $\hat{K}(s,t)$ to be the empirical estimator of the covariance between $X(s)$ and $X(t)$ computed from the curves observed at $s$ and $t$. This estimator is defined by

$$\hat{K}(s,t) = \frac{1}{m(s,t)} \sum_{i \in J(s,t)} \{X_i(s) - \bar{X}(s \mid s,t)\} \{X_i(t) - \bar{X}(t \mid s,t)\}, \quad (s,t) \in S, \tag{2.2}$$

where $\bar{X}(u \mid s,t) = \{m(s,t)\}^{-1} \sum_{i \in J(s,t)} X_i(u)$, $J(s,t)$ is the set of $i \in \{1,\ldots,n\}$ such that $\{s,t\} \subseteq I_i$, and $m(s,t) = \# J(s,t)$.

If instead we observe discrete and noisy versions $X_i(T_{ij}) + \epsilon_{ij}$ of fragments, where the $T_{ij}$’s are random variables taking values in $I_i \subset I$ and the $\epsilon_{ij}$’s represent independent noise of zero mean, then we can use a smooth estimator $\hat{K}$ of $K$ such as a spline (see section B.1 in the supplementary file) or a local linear estimator (see e.g. Kneip and Liebl, 2019).

Regardless of the type of data we have (see the discussion at the end of section 2.1) and the consistent estimator $\hat{K}$ we use to estimate $K$ on $S$, our approach is the same: after computing $\hat{K}$ on $S$, we extrapolate it from $S$ to $S_0$. For this we use a tensor
product basis decomposition of $K(s, t)$ for all $(s, t) \in S_0$, estimate the coefficients of that basis by minimising some distance between $K$ and $\hat{K}$ on $S$, and use those coefficients for all $(s, t) \in S_0$. Thus, we assume that the coefficients to be used for all $(s, t) \in S_0$ can be identified by considering only $(s, t) \in S$. We derive conditions under which this is satisfied in section 3.

For our extrapolation, we first write the tensor product decomposition of a general square-integrable function $M$ defined on $S_0$ and symmetric in the sense that $M(s, t) = M(t, s)$ for all $(s, t) \in S_0$. Let $\psi_1, \psi_2, \ldots$ be a sequence of linearly independent functions that define a complete basis of $L_2(I)$. For $M \in L_2(I \times I)$ we can write

$$M(s, t) = \sum_{j_1=1}^{\infty} \sum_{j_2=1}^{\infty} a_{j_1 j_2} \psi_{j_1}(s) \psi_{j_2}(t), \quad (2.3)$$

where the doubly infinite series on the right-hand side converges in $L_2$ to the left-hand side. In $(2.3)$ and below, the coefficients $a_{j_1 j_2}$ are assumed to be symmetric ($a_{j_1 j_2} = a_{j_2 j_1}$). This is necessary for $M$ to be symmetric.

Next, since in practice we cannot compute infinite series, we consider the $p$-dimensional approximation

$$M_p(s, t | A_p) = \sum_{j_1=1}^{p} \sum_{j_2=1}^{p} a_{j_1 j_2} \psi_{j_1}(s) \psi_{j_2}(t) \quad (2.4)$$

of $(2.3)$, where $A_p$ is the $p \times p$ matrix whose $(j_1, j_2)$th component, for $1 \leq j_1, j_2 \leq p$, is the value $a_{j_1 j_2}$. The $a_{j_1 j_2}$'s are constrained to satisfy $a_{j_1 j_2} = a_{j_2 j_1}$. Defined in this way, the function $M_p(s, t | A_p)$ is symmetric. It will also be semi-positive definite on $S_0$ if and only if $A_p$ can be expressed as $A_p = BB^T$, where $B$ is a $p \times p$ matrix.

Motivated by this, we let $\hat{A}_p$ be the $p \times p$ matrix that minimises the quantity

$$S_p(A_p) = \iint_S \left\{ \hat{K}(s, t) - M_p(s, t | A_p) \right\}^2 ds \, dt + \int_I \left\{ \hat{K}(t, t) - M_p(t, t | A_p) \right\}^2 dt \quad (2.5)$$

over the set $A = \{ A_p = BB^T, B \in \mathbb{R}^{p \times p} \}$. The first integral at $(2.5)$ calculates the squared difference between $\hat{K}$ and $M_p$ on the whole set $S$ and the second term
focuses on the diagonal of the whole domain (the diagonal plays an important role in covariance estimation; see for example Yao et al. [2005]).

The solution $\hat{A}_p$ is of the form $\hat{A}_p = \hat{B}\hat{B}^T$, where $\hat{B} \in \mathbb{R}^{p \times p}$. For a given matrix $\hat{A}_p$, the decomposition $\hat{A}_p = \hat{B}\hat{B}^T$ is not unique but this does not play any role in our procedure: any $\hat{B}$ leading to the same $\hat{A}_p$ leads to the same final estimator of $K$. Therefore, here and below $\hat{B}$ denotes any of the matrices leading to $\hat{A}_p$. For $1 \leq j_1, j_2 \leq p$, let $\hat{a}_{j_1,j_2}$ be the $(j_1,j_2)$th component of $\hat{A}_p$, let $\hat{b}_{j,k}$, $k = 1, \ldots, p$, denote the elements of the $j$th row of $\hat{B}$ and let $\tilde{v}_{k,p} = \sum_{j=1}^p \hat{b}_{j,k}\psi_j$. We define our estimator of $K(s,t)$ for all $(s,t) \in S_0$ by

$$\tilde{K}_p(s,t) = \sum_{j_1=1}^p \sum_{j_2=1}^p \hat{a}_{j_1,j_2} \psi_{j_1}(s) \psi_{j_2}(t) = \sum_{k=1}^p \tilde{v}_{k,p}(s)\tilde{v}_{k,p}(t). \quad (2.6)$$

By construction, $\tilde{K}_p$ is a covariance function. Since $\hat{B}$ is not unique, the $\tilde{v}_{k,p}$'s are not either but their sum is uniquely defined since the $\hat{a}_{j_1,j_2}$’s are unique.

Here, $p$ is a truncation parameter red that increases with $n$; its practical choice, as well as the details of implementation of our procedure, will be discussed in section 5.1. However we can already note here that $p$ is not a smoothing parameter. In particular, it is not used to construct an estimator of $K$ by smoothing, but rather to approximate as best as possible an estimator $\hat{K}$ of $K$, by a linear combination of the first $p$ elements of a tensor product basis. We shall see in section 4 that as long as $p \equiv p(n)$ is large enough compared to $n$, the truncation has no first order asymptotic effect on our estimator.

## 3 Identifiability

### 3.1 Introduction and basic setup

Since we only observe data on a set $S \subset S_0$, consistency of our estimator $\tilde{K}_p$ necessarily relies on the assumption that

the function $K$ on $S_0$ is uniquely determined by $K$ restricted to $S$. \quad (3.1)
If (3.1) fails then no method can be guaranteed to recover $K$ on all of $S_0$ from our data. In this sense, (3.1) is a weak condition; it is the least strong condition that is needed to ensure consistency. However, this condition is too high level to be useful, and in this section we derive sufficient lower level and more interpretable conditions for identifiability.

Our conditions will depend on the structure of the eigendecomposition of the covariance operator $K$. Let $\theta_1 \geq \theta_2 \geq \ldots$ denote the eigenvalues of $K$, and let $\phi_1, \phi_2, \ldots$ be the corresponding system of orthonormal eigenfunctions. By Mercer’s theorem we have

$$K(s, t) = \sum_{r=1}^{\infty} \theta_r \phi_r(s)\phi_r(t), \quad (3.2)$$

where the convergence is absolute and uniform in $s, t \in I$. Since we have assumed that $K$ is a continuous function of $s, t \in I$, the eigenfunctions are continuous too.

In section 3.2 we derive a theorem providing error bounds and sufficient conditions for identifiability of $K$. Then we show that these conditions can be expected to hold if the process $X$ is sufficiently smooth, where smoothness may be quantified directly through $X$ or through the eigenfunctions and the rate of decay of the eigenvalues. We will quantify this decay through the rate of convergence to zero of

$$\Theta_q := \left\{ \sup_{t \in I} \sum_{r=q+1}^{\infty} \theta_r \phi_r(t)^2 \right\}^{1/2}$$

as $q \to \infty$. Conditions of this type are quite natural for functional data, where theoretical studies are often based on assumptions on the sequence of eigenvalues. For example, it is common to assume that the sequence of eigenvalues decays sufficiently fast so that $\Theta_0 = \left\{ \sup_{t \in I} K(t, t) \right\}^{1/2} = \left[ \sup_{t \in I} \text{Var}\{X(t)\} \right]^{1/2}$ is finite. Moreover, as $q \to \infty$ we necessarily have $\Theta_q \to 0$ because any suitable series of continuous, orthonormal functions $\varphi_1, \varphi_2, \ldots$ in $L_2(I)$ and any series $\lambda_1 \geq \lambda_2 \ldots$ of positive numbers define a possible covariance function, only if, as $q \to \infty$ the values $\lambda_q$ converge to zero sufficiently fast such that $\sup_t |\sum_{r=1}^{\infty} \lambda_r \varphi_r(t)^2| < \infty$. Note too that
if \( D = \sup_{t \in I} \sup_{r=1,2,...} \phi_r(t)^2 < \infty \), then \( \Theta_q^2 \leq D \sum_{r=q+1}^{\infty} \theta_r \). This illustrates the fact that the behavior of \( \Theta_q \) as \( q \to \infty \) reflects the rate of decrease of the eigenvalues.

3.2 Main identifiability results

Our identifiability conditions are easiest understood by considering the case where the observations are in the form of fragments. There, since each curve \( X_i \) is observed only on an interval \( I_i \), the domain \( S \) defined at (2.1) essentially consists in a band around the main diagonal of \( S_0 \). Of course, the longer the \( I_i \)’s and the more they overlap, the larger \( S \) and the easier it is to estimate \( K \) from our data. More generally for settings that satisfy (2.1), consider a (non unique) partition of \( I \) into \( Q+1 \) of disjoint intervals, \( J_0, J_1, \ldots, J_Q \), constructed so that for all \( k = 1, \ldots, Q \), the cross-product \((J_{k-1} \cup J_k) \times (J_{k-1} \cup J_k) \) lies in \( S \). The difficulty of the problem can be measured by the size of \( Q \): the \( J_k \)’s are connected to the \( I_i \)’s in that, roughly speaking, \( Q \) can be taken to be smaller when the \( I_i \)’s are longer than when they are short.

We will see below that the smaller we can take \( Q \) in this construction, the easier it is to estimate \( K \) on \( S_0 \) from fragments. Figure 3 shows, for two simulated samples of fragments, a construction of such \( J_k \)’s and the corresponding squares \((J_{k-1} \cup J_k) \times (J_{k-1} \cup J_k) \) \( \subset S \). It also shows the domain \( S \) of values \((s,t)\) at which at least one fragment \((X_i(s), X_i(t))\) is observed. In the first case, the fragments are short, \( S \) is narrow around the diagonal, and we have no choice but taking \( Q \) large in order for the \( J_k \)’s to cover \( I = [0,1] \) and to have all the squares fit within \( S \). In the second case, the fragments are longer and we can take fewer and longer intervals \( J_k \).

Let \( X^J \) denote the fragment \( \{X(t), t \in J\} \), with \( J \) an interval where a curve \( X \) is observed. Reflecting the above illustration, below we show that the identification problem can be translated into one of prediction of \( X(t) \) for \( t \notin J_k \) from the fragment \( X^{J_k} \in L_2(J_k) \), for \( k = 1, \ldots, Q \), and that the difficulty of the problem increases with \( Q \). To understand why, note that identifiability of \( K \) requires that the correlation between \( X(t) \) and \( X(s) \) for two distant points \( t \) and \( s \) such that \((t,s) \notin S\) is uniquely
Figure 3: Illustration of the domain $S$ of points $(s,t)$ where at least one pair $(X_i(s), X_i(t))$ is observed, for a simulated sample of short fragments (first row) or of longer fragments (second row). The left panels show the observed fragments; the right panels show the corresponding $S$ (in red), and a construction of disjoint intervals $J_1, \ldots, J_Q$ such that $\bigcup_k J_k = \mathcal{I} = [0,1]$, and the corresponding squares $(J_{k-1} \cup J_k) \times (J_{k-1} \cup J_k) \subset S$.

determined by “local” correlations at the points $t^*, s^*$ such that $(t^*, s^*) \in S$. Now, correlations are linked to linear regressions, which is how we can relate the problem to one of linear prediction. We will need the following definition:

**Definition 3.1.** Let $J$ and $J^*$ be two subintervals of $\mathcal{I}$. For some $D > 0$ and $\epsilon \geq 0$ we say that the random variables $X(t)$, $t \in J^*$, are linearly $(D, \epsilon)$-predictable from $X^J$ if there exists an integrable function $\ell_t(u)$, $t \in J^*$, $u \in J$ which satisfies $\sup_{t \in J^*} \sup_{u \in J} |\ell_t(u)| < \infty$ and $\sup_{t \in J^*} \int_J |\ell_t(s)| \, ds \leq D$ and is such that

$$X(t) = \mu(t) + \int_J \ell_t(s) \{X(s) - \mu(s)\} \, ds + Z(t), \quad t \in J^*, \tag{3.3}$$

where for all $t \in J^*$, $Z(t)$ is a zero mean random variable such that $\text{Var}\{Z(t)\} \leq \epsilon^2$. 12
In other words, \((D, \epsilon)-\)predictability holds if, up to an error factor \(Z(t)\) whose variance is bounded by \(\epsilon^2\) for all \(t \in J^*\), we can predict \(X(t)\) linearly from \(X'\). As we shall see in the next theorem, the value of \(D\) plays an important role for bounding the accumulation of errors when predicting \(X(t)\) iteratively from consecutive small intervals where we have data. We do not require \(Z(t)\) to be uncorrelated with \(\int_{J} \ell_t(s) X(s) \, ds\).

Moreover, \((D, 0)-\)predictability means that \(X(t) = \mu(t) + \int_{J} \ell_t(s) \{X(s) - \mu(s)\} \, ds\).

The following theorem establishes a general link between \((D, \epsilon)-\)predictability and the error in reconstructing \(K\) on \(S_0\) from \(K\) on \(S\). Roughly speaking, it shows that the error in recovering \(K\) on \(S_0\) can be bounded by the accumulation of errors caused by predicting a fragment \(X^{J_0}\) using stepwise prediction at the intervals \(J_1, \ldots, J_Q\) described at the beginning of this section.

**Theorem 1.** Consider an integer \(Q\) and a sequence of disjoint subintervals \(J_0, J_1, \ldots, J_Q\) of \(I\) such that \(J_k \times J_k \subset S, k = 0, \ldots, Q\), and \(\bigcup_{k=0}^{Q} J_k = I\).

a) Assume that for some \(\epsilon_1 \geq 0, \ldots, \epsilon_Q \geq 0\) and for \(k = 1, \ldots, Q\), there exists \(D_k > 0\) such that the random variables \(X(t), t \in J_k\), are linearly \((D_k, \epsilon_k)-\)predictable from \(X^{J_{k^*}}\) for some interval \(J_{k^*}\), where \(0 \leq k^* \leq k - 1\) and \((J_k \cup J_{k^*}) \times (J_k \cup J_{k^*}) \subset S\). Then, for any continuous covariance function \(L\) defined on \(S_0\), such that \(L(t,s) = K(t,s)\) for all \((t,s) \in S\), we have

\[
\sup_{(t,s) \in S_0} |L(t,s) - K(t,s)| \leq \max_{j=1, \ldots, Q} C_j^\epsilon ,
\]

where \(C_0^\epsilon = 0\) and \(C_j^\epsilon = D_j C_{j-1}^\epsilon + 2\Theta_0 \epsilon j, j = 1, \ldots, Q\). \hspace{1cm} (3.4)

b) If the above condition is satisfied for any \(\epsilon_1 > 0, \ldots, \epsilon_Q > 0\), then for any continuous covariance function \(L\) defined on \(S_0\) such that \(L(t,s) = K(t,s)\) for all \((t,s) \in S\), we have \(\sup_{(t,s) \in S_0} |L(t,s) - K(t,s)| = 0\).

In the particular case where \(\epsilon := \epsilon_1 = \cdots = \epsilon_Q\) and \(D := D_2 = \cdots = D_Q\) (the value of \(D_1\) does not play any role), \((3.4)\) simplifies into \(\sup_{(t,s) \in S_0} |L(t,s) - K(t,s)| \leq 2\Theta_0 \epsilon \sum_{j=1}^{Q} D^{j-1}\). Theorem can be used to derive identifiability conditions that are
more concrete, easier to interpret and expressed in a form that is familiar in the functional data literature.

In what follows we consider particular types of processes for which we can show, using the theorem, that they lead to identifiability of $K$. As mentioned earlier, the key is for the process $X$ to be sufficiently smooth. This can be characterised by the fact that the eigenfunctions are sufficiently smooth and the eigenvalues decay sufficiently fast, or by directly imposing smoothness conditions on $X$. There, we will see that Theorem 1 can be used to derive conditions more general than those in Zhang and Chen (2017) and Descary and Panaretos (2019).

3.2.1 Finite number of nonzero eigenvalues.

Using the Karhunen-Loève decomposition, the process $X$ can be expressed as

$$X(t) - \mu(t) = \sum_{r=1}^{\infty} \xi_r \phi_r(t), \quad (3.5)$$

where, for $r = 1, 2, \ldots, \xi_1, \xi_2, \ldots$ are uncorrelated, zero mean random variables with $Var(\xi_r) = \theta_r$. In many important applications, the $X_i$’s are well represented by few functional principal components. Therefore it is common to assume that for some finite integer $q$, $\theta_{q+1} = \theta_{q+2} = \cdots = 0$ (and hence, $\Theta_{q+1} = 0$). This is one of the main assumptions imposed by Descary and Panaretos (2019) and we derive our first identifiability result from Theorem 1 under this assumption. In this case the process can be exactly described by $q$ principal components, and (3.5) simplifies to

$$X(t) - \mu(t) = \sum_{r=1}^{q} \xi_r \phi_r(t). \quad (3.6)$$

Another important condition assumed by Descary and Panaretos (2019) is that $\phi_1, \ldots, \phi_q$ are real analytic. In that setting, they showed that $K$ can be identified on $S_0$ from its values on a band around the diagonal. Our theorem can be used to establish the same result, but it can also be used to prove identifiability under more general conditions. For example, for an integer $Q$, let $J_0, J_1, \ldots, J_Q$ be a sequence of
disjoint intervals such that \( \bigcup_{k=0}^{Q} J_k = I \), and such that for any \( j = 1, \ldots, Q \) there exists \( j^* \in \{0, \ldots, j-1\} \) for which \((J_j \cup J_{j^*}) \times (J_{j^*} \cup J_j) \subset S\). Then for identifiability to hold, it suffices that

\[
\text{the } \phi_j \text{'s be linearly independent on the } J_k \text{'s; (3.7)}
\]

see section D.1 in the supplementary file for a proof.

Linear independence of the \( \phi_j \)'s on the \( J_k \)'s for \( k = 0, \ldots, Q - 1 \) can also be expressed in a matrix form. For example, it holds if, for \( k = 0, \ldots, Q - 1 \), the \( q \times q \) matrix \( \Phi^J_k \) of inner products

\[
\Phi^J_{q,rs} := \int_{J_k} \phi_r(t) \phi_s(t) \, dt, \quad r, s \in \{1, \ldots, q\},
\]

has full rank (and is thus invertible).

Condition (3.7) is satisfied if the \( \phi_j \)'s are real analytic on \( I \), but it includes much more general cases as it imposes a finite number of constraints on infinite dimensional functions. In particular, we do not need to assume that the \( \phi_j \)'s are linearly independent on all possible subintervals of \( I \) (which is satisfied if the \( \phi_j \)'s are real analytic); it suffices for them to be linearly independent on a finite number \( Q \) of subintervals.

Like us, Zhang and Chen (2017) did not require \( q \) to be finite and we will discuss their condition in the infinite case in the next section. Now, Theorem 1 can also be used to derive more general results than those in Zhang and Chen (2017) when \( q \) is finite. Indeed, Zhang and Chen's (2017) identifiability conditions rely on the fact that \( K(s, t), (s, t) \in \tilde{I} \times \tilde{I} \) for any \( \tilde{I} \subset I \) that is long than a small enough subinterval \( I_{ZC} \subset I \), has the same number of non-zero eigenvalues as \( K \) on \( I \times I \). This is equivalent to saying that the \( \phi_j \)'s are linearly independent on any subinterval of \( I \) that is longer than \( I_{ZC} \). This is similar to our condition at (3.7), except that, in our case, we do not need the \( \phi_j \)'s to be linearly independent on an infinite number of subintervals of \( I \). Being linearly independent on a finite number \( Q \) of subintervals is sufficient. Moreover, our identifiability condition holds even if some of the \( \phi_j \)'s vanish
in some parts of $\mathcal{I}$ which can be as long as half the length of $\mathcal{I}$; see for example model (1) in section 5.2. See also section D.1 in the supplementary file.

### 3.2.2 Infinite number of nonzero eigenvalues.

Theorem 1 can also be used to derive identifiability conditions for processes whose covariance operator has infinitely many nonzero eigenvalues. Proposition 1 below shows two results. First, in part $i)$ we show that, even if $K$ is not exactly identifiable on $\mathcal{S}$ from $K$ on $\mathcal{S}_0$, it is possible to approximate it with reasonably small error. Then, in part $ii)$ we give conditions under which we can identify $K$ on $\mathcal{S}$ from $K$ on $\mathcal{S}_0$.

**Proposition 1.** Consider an integer $Q$ and a sequence of disjoint subintervals $J_0, J_1, \ldots, J_Q$ of $\mathcal{I}$ with $\bigcup_{k=0}^{Q} J_k = \mathcal{I}$ such that for any $j = 1, \ldots, Q$ there exists a $j^* \in \{0, \ldots, j - 1\}$ for which $(J_j \cup J_{j^*}) \times (J_j \cup J_{j^*}) \subset \mathcal{S}$. Then

$i)$ For $k = 0, \ldots, Q - 1$, if there exists a positive integer $q$ such that the matrix $\Phi_{qJ_k}$ defined above (3.8) is invertible, then the random variables $X(t), t \in \mathcal{I}$, are linearly $(D_{q,J_k}, (D_{q,J_k} + 1)\Theta_q)$-predictable from $X^{J_k}$, where

$$D_{q,J_k} = \sup_{t \in \mathcal{I}} \int_{J_k} \left| \left( \phi_1(t), \ldots, \phi_q(t) \right) \left( \Phi_{qJ_k}^{-1} \right)^{-1} \left( \phi_1(s), \ldots, \phi_q(s) \right)^T \right| ds < \infty, \quad (3.9)$$

and with $\Theta_q$ and $\Theta_0$ as in section 3.7. Moreover, for any continuous covariance function $L$ defined on $\mathcal{S}_0$, such that $L(t,s) = K(t,s)$ for all $(t,s) \in \mathcal{S}$, we have $\sup_{(t,s) \in \mathcal{S}_0} |L(t,s) - K(t,s)| \leq \max_{j=1,\ldots,Q} C_j$, where $C_0 = 0$ and $C_j = D_{q,J_j} C_{j-1} + 2\Theta_0 \Theta_q (D_{q,J_j} + 1), \ j = 1, \ldots, Q$.

$ii)$ Additionally assume that the matrices $\Phi_{qJ_k}$ are invertible for all $k \in \{0, \ldots, Q - 1\}$ and all integers $q$, and suppose that

$$\lim_{q \to \infty} \max_{k=1,\ldots,Q-1} (D_{q,J_k} + 1)\Theta_q = 0. \quad (3.10)$$

Then for any continuous covariance function $L$ defined on $\mathcal{S}_0$ such that $L(t,s) = K(t,s)$ for all $(t,s) \in \mathcal{S}$, we have $\sup_{(t,s) \in \mathcal{S}_0} |L(t,s) - K(t,s)| = 0.$

16
A proof of Proposition 1. is given in the supplementary file. Proposition 1. ii is an immediate consequence of Proposition 1. i and Theorem b since (3.10) implies that for all $\epsilon > 0$ and all $k \in \{0, \ldots, Q - 1\}$ there exists a sufficiently large $q$ such that $(D^{q,J_k} + 1)\Theta_q \leq \epsilon$, $k = 0, \ldots, Q - 1$.

We deduce from Proposition 1. i that even if the covariance operator has an infinite number of nonzero eigenvalues, we can achieve a small approximation error. This holds if there exists an integer $q$ such that, for moderate values $D^{q,J_k}$, the first $q$ functional principal components approximate $X(t)$ with high accuracy in the sense that $\Theta_q$ is very small. In practice, it is frequently observed that the eigenvalues decrease rapidly and that a small number $q$ of functional principal components are able to explain a considerable part of the variability of $X$.

Proposition 1. ii gives a set of conditions under which we can identify $K$ on $S_0$ from $K$ on $S$ even if $K$ has an infinite number of nonzero eigenvalues. First, we need invertibility of all the $\Phi^{J_k}_q$’s. While it is difficult to assess how often this is satisfied, as discussed in section 3.2.1, this may not be a particularly restrictive condition (recall that it is automatically satisfied if the $\phi_k$’s are real analytic). The additional assumption (3.10), however, imposes a possibly strong condition on the rate of decay of the eigenvalues. If $q = 1$, then $\Phi^{J_k}_1 = \int_{J_k} \phi_1(t)^2 \, dt$, and $D^{1,J_k} = \sup_{t \in \Sigma} |\phi(t)| \int_{J_k} |\phi_1(s)| \, ds / \int_{J_k} \phi_1(s)^2 \, ds$. For general $q$ the constants $D^{q,J_k}$ are roughly proportional to the square root of the largest eigenvalue of $(\Phi^{J_k}_q)^{-1}$, which usually increases as $q$ increases and as the length of $J_k$ decreases. Generally, $D^{q,J_k} \to \infty$ as $q \to \infty$. Thus, (3.10) requires a sufficiently fast decrease of $\Theta_q$ as $q$ increases.

Recall that Zhang and Chen’s (2017) identifiability conditions also allow $q$ to increase to infinity as long as the eigenvalues decay fast enough. As for the case where $q$ is finite, they require the $\phi_j$’s to be linearly independent on any long enough subinterval. As discussed in section 3.2.1, this is similar to our condition that the matrices $\Phi^{J_k}_q$ are invertible for all $k \in \{0, \ldots, Q - 1\}$, except that in our case we only need the $\phi_j$’s to be linearly independent on a finite number $Q$ of subintervals. Moreover,
although they allow \( q \) to grow as \( n \) increases, they need the number \( q \) of nonzero eigenvalues to be bounded by a small constant times the size of their discretisation grid. This condition comes from the fact that they use matrix completion methods, and is not required in our case.

### 3.2.3 Very smooth sample functions.

In this section we provide a third type of characterisation of identifiability of \( K \) based on smoothness properties of the process \( X \). Specifically, we assume that \( X \) is almost surely real analytic and satisfies the following moment conditions:

\[
\begin{align*}
& \text{i) } X \text{ is infinitely differentiable and there exists } c > 0 \text{ such that, with probability 1, } X(t) = \sum_{r=0}^{\infty} \frac{1}{r!} X^{(r)}(s)(t - s)^r \text{ for all } s, t \in I \text{ with } |t - s| \leq c; \\
& \text{ii) } E \left\{ \sup_{t \in I, |t - s| \leq c} |X^{(r)}(t)|^2 \right\} < \infty \text{ for all } r = 0, 1, 2, \ldots \text{ and } s \in I; \\
& \text{iii) } \lim_{\nu \to \infty} \sup_{t \in I, |t - s| \leq c} E \left[ \left\{ \sum_{r=0}^{\nu} \frac{1}{r!} X^{(r)}(s)(t - s)^r \right\}^2 \right] = 0 \text{ for all } s \in I.
\end{align*}
\]

(3.11)

For simplicity of presentation, we also assume throughout this section that the process \( X \) is centred, i.e., \( E \{X(t)\} = \mu(t) = 0 \) for all \( t \in I \).

The following proposition shows that \( K \) is identifiable when (3.11) is satisfied. Its proof follows from Theorem 1b; see section D.3 in the supplementary file.

**Proposition 2.** If the centred process \( X \) satisfies (3.11), then, for any continuous covariance function \( L \) defined on \( S_0 \) with \( L(t, s) = K(t, s) \) for all \( (t, s) \in S \) we have

\[
\sup_{(t, s) \in S_0} |L(t, s) - K(t, s)| = 0.
\]

The assumption of the proposition may be satisfied for processes \( X \) generated by a nonlinear mechanism, for which it is difficult to characterise the rate of decay of eigenvalues. More precisely, suppose that for some (possibly infinite dimensional) random parameter \( \zeta_i \) taking values in some Banach space \( \mathcal{P} \), we have \( X_i(t) = G(\zeta_i, t) \) for all \( t \in I \), where \( G : \mathcal{P} \times I \to \mathbb{R} \) is such that \( G(\zeta_i, \cdot) \) is real analytic on \( I \) for any \( \zeta_i \in \mathcal{P} \). If expectations of derivatives satisfy (3.11), the corresponding covariance
function is identifiable. A very simple, special case consists in the model

\[ X_i(t) = \zeta_i \cdot g(\zeta_i t + \zeta_i^3), \quad t \in \mathcal{I}, \] (3.12)

where \( \zeta_i = (\zeta_i^1, \zeta_i^2, \zeta_i^3) \) is a random vector taking values in a compact subspace of \( \mathbb{R} \times \mathbb{R}_+ \times \mathbb{R} \), and where \( g : \mathbb{R} \to \mathbb{R} \) is a real analytic function on \( \mathbb{R} \) (e.g. a sine function, a polynomial, an exponential function, etc). Note that (3.12) is a finite dimensional, but nonlinear model, and the corresponding covariance operator will generally possess infinitely many nonzero eigenvalues. At the same time, for this model, condition (3.11) is fulfilled, and using Proposition 2 we conclude that the covariance function is identifiable on its domain \( \mathcal{S}_0 \).

4 Theoretical properties of estimator

In this section we derive asymptotic properties of our covariance estimator \( \hat{K}_p \) at (2.6), where the coefficient matrix \( A_p \) is estimated by minimising the criterion at (2.5). Recall that for some \( \delta > 0 \), the subset \( \mathcal{S} \) is defined at (2.1). The starting point of our analysis is the availability of a consistent estimator \( \hat{K}(t,s) \) of \( K(t,s) \) for \( (t,s) \in \mathcal{S} \).

Consistency of the empirical estimator \( \hat{K} \) at (2.2) is easily established if we assume that each curve \( X_i \) is observed on a random interval \( \mathcal{I}_i \) independent of \( X_i \), and that \( \inf_{t \in \mathcal{I}} \inf_{s \in \mathcal{I},|t-s| \leq \delta} P\{(t,s) \in \mathcal{I}_i \times \mathcal{I}_i \} > 0 \). Moreover, if there exists a constant \( C_0 < \infty \) such that \( E\{X_i(t)^4\} < C_0 \) for all \( t \in \mathcal{I} \), it is easy to prove that for all \( (t,s) \in \mathcal{S} \), we have \( E\{(\hat{K}(t,s) - K(t,s))^2\} \leq C_1 n^{-1} \) for some constant \( 0 < C_1 < \infty \). This implies that \( E\{\int_{\mathcal{S}} (\hat{K} - K)^2 \} = O(n^{-1}) \) and \( E\{\int_{\mathcal{I}} (\hat{K}(t,t) - K(t,t))^2 \} = O(n^{-1}) \).

However consistent estimators of \( K \) on \( \mathcal{S} \) can be obtained under more general settings, for example the one where each fragment is observed at random discrete points and with noise, as in Kneip and Liebl (2019). There, even as \( n \) increases, instead of the empirical covariance (2.2), some smoothing is required to obtain a consistent estimator of \( K \) on \( \mathcal{S} \). The resulting estimator \( \hat{K} \) of \( K \) converges at nonparametric rates which depend on several factors but which can be slower than \( n^{-1} \). Our es-
timator is applicable in these more general scenarios (see the discussion at the end of section 2.1). Therefore we do not need to restrict our consistency results to $\hat{K}$ at (2.2), nor to consider every possible data scheme and associated estimator $\hat{K}$. Instead we assume more generally that the incomplete curves are observed in whatever form (e.g. noisy or not, discrete or not, in the form of fragments or not, etc) from which we can compute a consistent estimator $\hat{K}$ on $\mathcal{S}$ that satisfies the following assumptions:

(i) $K$ is bounded and uniformly continuous on $\mathcal{S}_0$, (ii) the sequence $\psi_1, \psi_2, \ldots$ is a complete basis of $L_2(\mathcal{I})$, and (iii) for some $\kappa > 0$ the estimator $\hat{K}$ satisfies (4.1) $\int_{\mathcal{S}} (\hat{K} - K)^2 = O_p(n^{-\kappa})$ and $\int_{\mathcal{I}} \{\hat{K}(t,t) - K(t,t)\}^2 dt = O_p(n^{-\kappa})$.

The next theorem establishes conditions under which our estimator $\tilde{K}_p$ is consistent for $K$. It also establishes conditions under which it has the same convergence rate on $\mathcal{S}_0$ as $\hat{K}$ on $\mathcal{S}$. Since it is not possible to consistently estimate a non identifiable quantity, these conditions include identifiability conditions discussed earlier.

With respect to the dimension $p$ of our basis function approximation, asymptotics only rely on the condition that $p \equiv p(n)$ is large enough in order to guarantee a negligible approximation error (see Proposition 3 in the supplementary material).

The theorem also shows that if $K$ is not identifiable on $\mathcal{S}_0$ from $K$ on $\mathcal{S}$, but can be approximated with relatively small error, then asymptotically $\tilde{K}_p$ will also be a reasonable approximation to $K$ on $\mathcal{S}_0$.

**Theorem 2.** Assume that (4.1) holds, and for a fixed integer $Q$ consider a sequence of disjoint subintervals $J_0, J_1, \ldots, J_Q$ of $\mathcal{I}$ such that $\bigcup_{k=0}^{Q} J_k = \mathcal{I}$.

a) If, for $k = 0, \ldots, Q - 1$, the random variables $X(t)$, $t \in \mathcal{I}$, are linearly $(D_k, 0)$-predictable from $X_{J_k}$, where $D_k < \infty$, and there is a $k^* \in \{0, \ldots, k\}$ for which $(J_{k+1} \cup J_{k^*}) \times (J_{k+1} \cup J_{k^*}) \subset \mathcal{S}$, then as long as $p = p(n)$ is large enough,

$$\int_{\mathcal{S}_0} |\tilde{K}_p - K| = O_p(n^{-\kappa/2}).$$

(4.2)

b) If, for some $\epsilon_1 > 0, \ldots, \epsilon_Q > 0$, and for $k = 1, \ldots, Q$, there exists $D_k > 0$ such that the random variables $X(t)$, $t \in J_k$, are linearly $(D_k, \epsilon_k)$-predictable from
for some $0 \leq k^* \leq k - 1$ such that $(J_k \cup J_{k^*}) \times (J_k \cup J_{k^*}) \subset S$, then there exists a sequence of positive random variables $R_n = O_P(n^{-\kappa/2})$ such that for $p = p(n)$ large enough,

$$\int_{S_0} |\tilde{K}_p - K| \leq \max_{j=1,\ldots,Q} C_j^* \int_{S_0} ds \, dt + R_n,$$

(4.3)

where $C_0^*, \ldots, C_Q^*$ are defined as in Theorem 1.

c) If the condition in b) is satisfied for any $\epsilon_1 > 0, \ldots, \epsilon_Q > 0$, then as long as $p = p(n)$ is large enough,

$$\int_{S_0} |\tilde{K}_p - K| \to 0 \text{ as } n \to \infty.$$

(4.4)

Theorem 2.a follows the identifiability scenario described in section 3.2.1 where the eigenfunctions are all linearly independent on the $J_k$’s, for $k = 0, \ldots, Q - 1$. In that case the conditions are sufficiently strong to ensure that $\tilde{K}_p$ converges to $K$ on $S_0$ at the same rate of convergence, $n^{-\kappa/2}$, as $\hat{K}$ on $S$. Theorem 2.b shows that even if $K$ is not identifiable, the approximation error bounds derived in Theorem 1.a asymptotically also characterise the difference between $\tilde{K}_p$ and $K$.

Theorem 2.c establishes consistency (without convergence rates) of our estimator $\tilde{K}_p$ under milder conditions than Theorem 2.a. For example, it holds under the conditions of Proposition 1.ii or Proposition 2. Without stronger conditions, the convergence rates of $\tilde{K}_p$ may depend on the particular situation. Indeed, the error variable $R_n$ defined in Theorem 2.b may depend on $\epsilon_1, \ldots, \epsilon_Q$ and on the corresponding constants $D_1, \ldots, D_Q$. Although for fixed $\epsilon_1, \ldots, \epsilon_Q$ the rate of convergence of $R_n$ is always of order $n^{-\kappa/2}$, relevant constants may increase as $\epsilon_j, j = 1, \ldots, Q$, decrease, and $\int_{S_0} |\tilde{K}_p - K|$ may have a slower convergence rate than $n^{-\kappa/2}$.

**Remark 1.** Our theoretical results are presented in the case where the data are available in a continuous non discretised form. All our results hold also in the discretised case, where $S$ consists in a discrete set of $N$ points, as long as, asymptotically, the
resolution of the discretisation increases sufficiently fast, i.e. as long as $N$ grows sufficiently fast when $n \to \infty$.

5 Numerical illustrations

5.1 Computing the estimator in practice

In practice, for a given $p$ and a given set of basis functions $\psi_1, \ldots, \psi_p$, to compute the estimator $\hat{K}_p$, we need to minimise (2.5) with respect to $A_p = BB^T$. Since there is no analytic solution to this problem, we obtain the solution numerically using the quasi-Newton method in Matlab. Since (2.5) is convex in $A_p$ but not in $B$, minimising (2.5) numerically with respect to $B$ could lead to a local minimum so that we need an initial guess. We compute it analytically by minimising (2.5) with respect to $A_p$ without the semi-definite positive constraint. Then we turn the solution into a covariance matrix as described in section B.2 in the supplementary file.

In practice we found that our initial guess worked very well as long as we computed it with care. In particular, although theoretically we can choose $p$ as large as we want, in practice, the data are typically observed in a discretised form. This causes numerical issues with the initial guess if $p$ is too large. Specifically our initially guess requires to invert a matrix which is invertible only when $p \leq \sqrt{N}$, where $N$ is the number of points at which $S$ is discretised; see Remark 1 and section B.2 in the supplementary file. For each value of $p \leq \sqrt{N}$, we compute our initial estimator corresponding to that $p$ and find the corresponding $\hat{A}_p = \hat{B}_p \hat{B}_p^T$ that minimises (2.5) over the set $\mathcal{A} = \{A_p = BB^T, B \in \mathbb{R}^{p \times p}\}$. For each $p$, if the matrix we need to invert for computing the initial estimator is numerically ill conditioned, we add a ridge parameter to its diagonal (we take the smallest positive number that makes the matrix well conditioned as per Matlab’s default criterion). Finally, among the candidate values, $0, \ldots, \lfloor \sqrt{N} \rfloor$, of $p$, we choose the one which is even (see next two paragraphs) and minimises (2.5) with respect to $p$, where each $B$ there is replaced by...
$B_p$. In most cases, this gives $p = \lfloor \sqrt{N} \rfloor$ or $\lfloor \sqrt{N} \rfloor - 1$.

For the basis $\psi_1, \psi_2, \ldots$ used to construct $\tilde{K}_p$, one possibility would be to use the sine-cosine basis. On the interval $I = [a,b]$, and up to some normalising constants, this complete, orthonormal basis of $L_2(I)$ is defined by

$$
\psi_1(x) = 1, \\
\psi_{2j}(x) = \sin\{2\pi j(x-a)/(b-a)\} \text{ and } \psi_{2j+1}(x) = \cos\{2\pi j(x-a)/(b-a)\}, \text{ for } j = 1, \ldots
$$

A drawback of this basis is that it can produce very bad pointwise estimates close to the boundary. Indeed, since $\psi_i(a) = \psi_i(b)$ for $i = 1, \ldots$, it always produces functions $M_p(s,t)$ such that $M_p(a,a) = M_p(b,b) = M_p(a,b) = M_p(b,a)$. To avoid this boundary issue we rely on an overcomplete basis by adding a straight line to the basis. More precisely, up to some normalising constants, we use the basis functions

$$
\psi_1(x) = x, \\
\psi_2(x) = 1, \text{ and } \psi_{2j}(x) = \sin\{2\pi j(x-a)/(b-a)\}, \psi_{2j+2}(x) = \cos\{2\pi j(x-a)/(b-a)\}, \text{ for } j = 1, \ldots
$$

For any given (finite) $p$, the basis functions $\psi_1, \psi_2, \ldots, \psi_p$ are linearly independent functions on $I$. In our algorithmic implementation we orthonormalise them by the Gram-Schmidt procedure. The squared fitting error obtained when using $\psi_1, \psi_2, \ldots, \psi_p$ is necessarily smaller or equal to that obtained when using only the original Fourier basis $\psi_2, \ldots, \psi_p$. In particular we expect a much improved boundary behavior. We always take $p$ even so that sine and cosine functions at the same resolution are included in our expansion.

To compute $\tilde{K}_p$ we need to choose the region $S$ on which to compute $\tilde{K}$. This region $S$ depends on the estimator $\tilde{K}$ that we use, which itself depends on the type of incomplete data we observe (see the discussion at the end of section 2.1). For example, the observations could be continuous, discrete, noisy, etc. Depending on their nature, we would use the empirical estimator at (2.2), the kernel estimator of Yao et al. (2005), or other types of estimators that are consistent on a band $S$ around the diagonal of $S_0$. Here we describe how to do this in the fragment case which interests us the most, where each curve $X_i$ is observed on an interval $I_i$.

Moreover, we consider the scenario often encountered in practice where the curves
are only observed discretely, as in Descary and Panaretos (2019). The choice of $S$ is easiest described when $\hat{K}$ is the empirical estimator at (2.2). However, similar ideas can be used for other estimators and in other settings. For example, in sections 5.2 and 5.3 we use a spline estimator of $K$, which is obtained from the empirical estimator at (2.2) and uses the same $S$ as the one we describe below.

Following the discussion below (2.2), we take $S$ to be a subset of $D_0 \equiv \bigcup_{i=1}^{n} (I_i \times I_i)$ which consists of a band around the main diagonal plus areas off the diagonal which do not contain too few data points. Letting $\ell_i$ denote the length of $I_i$, we take the band around the diagonal to consist of all the points $(s,t)$ in $D_0$ such that $|s-t| \leq \bar{\ell}/2$, where $\bar{\ell} = n^{-1} \sum_{i=1}^{n} \ell_i$. We take the remaining points $(s,t)$ of $D_0$ only if there are enough fragments observed at both $s$ and $t$ from which to compute a reasonable estimator $\hat{K}(s,t)$, as follows.

Ideally, we would like to compute $\hat{K}(s,t)$ only when say, 10 fragments are observed at $s$ and $t$, that is when $m(s,t) \geq 10$, with $m$ defined below (2.2). However when $n$ is very small, most pairs $(s,t)$ do not satisfy this criterion so that $S$ is nearly empty. Taking this into account, outside the diagonal band constructed in the previous paragraph we keep all the points $(s,t)$ that are such that $m(s,t) \geq \min\{10, \max_{(s,t) \in D_0} [m(s,t)/20]\}$. In most practical situations, this rule amounts to removing the few points $(s,t)$ that are on the edge of the domain $D_0$ and for which very few fragments are observed, unless sample size is so small that very few fragments are observed at most points $(s,t)$.

5.2 Simulation results

5.2.1 Comparison with other nonparametric estimators

To examine the performance of our covariance function estimator in practice, for sample sizes $n = 50, 100$ and $500$ we generated data on a discrete grid $I_D = \{t_1 = 0 < t_2 < \ldots < t_G = 1\}$ of $G = 50$ equispaced points on the interval $I = [0, 1]$ for models (1) to (4) below. Specifically, for $i = 1, \ldots, n$, letting $I_{D,i} = I_i \cap I_D$, we generated
Figure 4: True covariance (first column), estimator $\tilde{K}_p$ (second column), $\tilde{K}_{DP}$ (third column) and $\tilde{K}_{ZC}$ (fourth column) corresponding to the sample that produced the median ISE out of 100 samples generated from model (2) with $\ell_i \sim U(0.1, 0.3)$ (row 1) and $\ell_i \sim U(0.4, 0.6)$ (row 2) when $n = 100$, and $\ell_i \sim U(0.4, 0.6)$ when $n = 500$ (row 3).

We took the $X_i(t_{ij})$'s to be normally distributed with zero mean and covariance $\text{cov}\{X_i(t_{ij}), X_i(t_{ik})\} = K(t_{ij}, t_{ik})$ and considered the following four models:

1. $K(s, t) = \sum_{i=1}^{3} \theta_i \phi_i(t)\phi_i(s)$ where $\theta_1 = 1$, $\theta_2 = 0.8$ and $\theta_3 = 0.3$, and $\phi_1(t) = \sqrt{5}(6t^2 - 6t + 1)$, $\phi_2(t) = \sqrt{2}\log(t + 0.5) \cdot 1\{t \leq 0.5\}$ and $\phi_3(t) = \sqrt{22}(252t^5 - 630t^4 + 560t^3 - 210t^2 + 30t - 1) \cdot 1\{t > 0.5\}$.

2. $K(s, t) = \sum_{i=1}^{4} 0.5^{i-1} \phi_i(t)\phi_i(s)$ where $\phi_1(t) = 1$, $\phi_2(t) = (2t - 1)\sqrt{3}$, $\phi_3(t) =$
Figure 5: True covariance (first column), estimator \( \tilde{K}_p \) (second column), \( \tilde{K}_{DP} \) (third column) and \( \tilde{K}_{ZC} \) (fourth column) corresponding to the sample that produced the median ISE out of 100 samples generated from model (1) with \( \ell_i \sim U(0.4, 0.6) \) when \( n = 50 \) (row 1), \( n = 100 \) (row 2) and \( n = 500 \) (row 3).

\[
(6t^2 - 6t + 1)\sqrt{5} \quad \text{and} \quad \phi_4(t) = (20t^3 - 30t^2 + 12t - 1)\sqrt{7};
\]

\[
(3) \quad K(s, t) = \sum_{i=1}^{2} 0.5^{i-1} \phi_i(t)\phi_i(s) \quad \text{where} \quad \phi_1(t) = \exp\{5(t - 0.5)\}/[1 + \exp\{5(t - 0.5)\}], \quad \phi_2(t) = (t - 0.5)^2/0.25;
\]

\[
(4) \quad K(s, t) = \sum_{i=1}^{50} i^{-2} \phi_i(t)\phi_i(s) \quad \text{where, for} \ k = 1, \ldots, 25, \ \phi_{2k-1}(t) = \sqrt{2} \sin(2k\pi t) \quad \text{and} \quad \phi_{2k}(t) = \sqrt{2} \cos(2k\pi t).
\]

For models (1) to (3), we took \((a_\ell, b_\ell) = (0.1, 0.3)\), resulting in short fragments, and \((a_\ell, b_\ell) = (0.4, 0.6)\), resulting in longer fragments. Model (4) has a relatively large number of non-zero eigenvalues and is harder to identify from very short fragments (see the discussion at the end of section 3.2.2). Therefore, there we took \((a_\ell, b_\ell) = (0.3, 0.5)\) and \((a_\ell, b_\ell) = (0.5, 0.7)\). Models (1) to (4) satisfy our identifia-
Table 1: \(10^3\times\text{Mean (standard deviation)}\) of 100 ISE values, computed on \(\mathcal{D} = \bigcup_{i=1}^{n}(\mathcal{I}_i \times \mathcal{I}_i) \equiv \mathcal{D}_0\), and on \(\mathcal{D} = \mathcal{S}_0 \setminus \mathcal{D}_0\), for three estimators of \(K\) from models (1) to (4): our estimator \(\tilde{K}_p\), Descary and Panaretos' (2019) estimator \(\tilde{K}_{DP}\) and Zhang and Chen's (2017) estimator \(\tilde{K}_{ZC}\).

<table>
<thead>
<tr>
<th>Model</th>
<th>(n)</th>
<th>(E(\ell))</th>
<th>(\tilde{K}_p)</th>
<th>(\tilde{K}_{DP})</th>
<th>(\tilde{K}_{ZC})</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>50</td>
<td>0.2</td>
<td>14.7(15.0)</td>
<td>32.6(12.5)</td>
<td>65.7(30.6)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.5</td>
<td>15.9(15.5)</td>
<td>7.01(4.22)</td>
<td>20.6(19.7)</td>
</tr>
<tr>
<td></td>
<td>100</td>
<td>0.2</td>
<td>9.53(3.05)</td>
<td>22.5(10.3)</td>
<td>10.0(5.31)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.5</td>
<td>7.98(3.86)</td>
<td>4.70(4.04)</td>
<td>10.8(5.83)</td>
</tr>
<tr>
<td></td>
<td>500</td>
<td>0.2</td>
<td>5.27(0.94)</td>
<td>19.3(6.39)</td>
<td>3.19(1.62)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.5</td>
<td>2.08(1.10)</td>
<td>1.42(1.19)</td>
<td>3.25(1.58)</td>
</tr>
</tbody>
</table>

For each combination of model, \(n\) and fragment length, we performed 100 Monte Carlo replications. That is, in each case we generated 100 samples and from each sample we computed our covariance estimator \(\tilde{K}_p\) as described in section 5.1. Since our data are discrete, we take \(\hat{K}\) in (2.5) to be a smooth weighted penalised cubic tensor product spline estimator described in section B.1 in the supplementary file.

In each case we also computed the estimators of Zhang and Chen (2017) and Descary and Panaretos (2019), which we refer to respectively as \(\tilde{K}_{ZC}\) and \(\tilde{K}_{DP}\). Both \(\tilde{K}_{ZC}\) and \(\tilde{K}_{DP}\) require to specify the number of non-zero eigenvalues used to construct
Figure 6: True covariance (first column), estimator $\tilde{K}_p$ (second column), $\tilde{K}_{DP}$ (third column) and $\tilde{K}_{ZC}$ (fourth column) corresponding to the sample that produced the median ISE out of 100 samples generated from model (4) when $n = 500$ and $\ell_i \sim U(0.5, 0.7)$.

the estimator. Descary and Panaretos (2019) provides a graphical way of choosing this number, but since extensive simulations require an automatic choice of all parameters, in their simulations they used the true number of nonzero eigenvalues instead of choosing it from the data. On the other hand, Zhang and Chen (2017) suggest to use a random subsampling cross-validation method. However, they do not provide details on how to obtain the subsamples and how many random splits to use. Therefore, in our simulation study, while our method was fully data-driven, for the other two estimators, in models (1) to (3), we used the true, non data-driven, number of nonzero eigenvalues. For model (4) however, the true number of nonzero eigenvalues exceeds the upper bounds those two methods have on that number. For $\tilde{K}_{ZC}$, we set the number as their upper bound. For $\tilde{K}_{DP}$, following Descary and Panaretos’ (2019) suggestions, we chose the number of nonzero eigenvalues by applying their visual method to one sample from each of the simulation scenarios, and used that number for all 100 simulated samples.

In Table [1] for each setting considered, we show $10^3 \times$ the mean and standard deviation of the integrated squared error, $\text{ISE} = \int_D \{\tilde{K}_p(s, t) - K(s, t)\}^2$, computed on the regions $D = D_0 \equiv \bigcup_{i=1}^n (I_i \times I_i)$ where the computation of standard covariance estimators from the raw data is possible, and $D = S_0 \setminus D_0$ where extrapolation to $\tilde{K}_p$ is needed. We also computed these integrated squared errors for $\tilde{K}_{DP}$ and $\tilde{K}_{ZC}$. The
results show that in most cases, our estimator outperformed $\tilde{K}_{DP}$ and $\tilde{K}_{ZC}$, especially when the sample size and the average fragment length $E(\ell)$ were small. As expected, all methods improved as sample size or the average fragment length increased.

To illustrate our results graphically, we show, for several settings and for all three methods ($\tilde{K}_p$, $\tilde{K}_{DP}$ and $\tilde{K}_{ZC}$), the estimated covariance corresponding to the sample that produced the median ISE value on the region $S_0 \setminus D_0$ where no data was observed. To highlight this region, we colour the covariance functions on that region only. In Figure 4, we show, for all the three methods, the covariance estimator corresponding to the median ISE value on the region $S_0 \setminus D_0$ obtained in model (2) when $n = 100$ and $\ell_i \sim U(0.1, 0.3)$ or $\ell_i \sim U(0.4, 0.6)$, or when $n = 500$ and $\ell_i \sim U(0.4, 0.6)$. This figure illustrates the fact that, as the average fragment length $E(\ell)$ or the sample size increases, all estimators improve. There, our estimator is visually more appealing in most of the cases, especially for smaller sample size and shorter fragments. The other two methods improve significantly as sample size increases. Descary and Panaretos’ (2019) estimator seems to capture the shape of the function best when $n = 500$, except near the boundary of the domain where it underestimates $K$.

In Figure 5, we show the estimated covariance functions corresponding to the median ISE value on the region $S_0 \setminus D_0$ for model (1) when $\ell_i \sim U(0.4, 0.6)$ and $n = 500, 100$ or 500. This figure illustrates the fact that the estimators improve as
the sample size increases. Although model (1) violates the identifiability conditions of Zhang and Chen (2017) and Descary and Panaretos (2019), we see that it can be estimated well. Zhang and Chen’s (2017) and Descary and Panaratos’ (2019) methods too perform well when the sample size and the fragment length are large enough. This is consistent with our theory which predicts that this model is identifiable.

In Figure 6 we present the estimators of the covariance function from model (4) which violates Zhang and Chen’s (2017) and Descary and Panaratos’ (2019) identifiability conditions but satisfies our weaker conditions. We show the estimators corresponding to the median ISE value on the region $S_0 \setminus D_0$, when $\ell_i \sim U(0.5, 0.7)$ and $n = 500$. Here too we can see that this model is identifiable and all three methods capture the true shape. Our estimator performs a bit better near the boundary of the domain.

Finally, in Figure 7 we show the estimators of the covariance function from model (3) corresponding to the median ISE value on the region $S_0 \setminus D_0$, when $\ell_i \sim U(0.1, 0.3)$ and $n = 500$. This model has two non-zero eigenvalues and real analytic eigenfunctions so that the identifiability conditions of all three methods are satisfied. In this example we can see that with a large sample size, even when the fragments are short, Zhang and Chen’s (2017) and Descary and Panaratos’ (2019) estimators seem to capture the true trend better than ours near the boundary of the domain. A problem with these estimators is that they are not smooth. As a result, they tend to be too noisy, which is why the numbers in Table 1 tend to be larger for those estimators than for our estimator. However, by inspecting Figure 7 we suspect that in this case, smoothing Zhang and Chen’s (2017) and Descary and Panaratos’ (2019) estimators would make them competitive even in terms of ISE.

Together with the results from Table 1, the graphs show that our method generally performs better than $\tilde{K}_{DP}$ and $\tilde{K}_{ZC}$, especially when the sample size and the fragment length are small. However, as the sample size and fragment length increase, $\tilde{K}_{DP}$ and $\tilde{K}_{ZC}$ become more competitive. When $n = 500$, they are usually comparable to our
method and can be even visually better if their identifiability conditions are satisfied, except that some smoothing maybe needed for denoising.

5.2.2 Comparison with parametric estimators

In problems where extrapolation is not needed, parametric estimators based on correct parametric assumptions usually perform significantly better than nonparametric ones, both in theory and in practice. However, when parametric assumptions are far from the truth, parametric estimators can perform poorly. To investigate these properties in our setting where extrapolation is needed, we considered three parametric estimators: one based on a correct parametric model, one based on a low dimensional spline approximation, and Delaigle and Hall’s (2016) estimator, $\tilde{K}_{DH}$, which makes a first order Markov assumption on the discretised functional data.

For the first, we assumed the right parametric form for the covariance. Specifically, we took $K(s,t) = \sum_{r=1}^{q} \theta_r \phi_r(s)\phi_r(t)$, where $\theta_1, \ldots, \theta_q$ are the unknown parameters to estimate and $q$ and the $\phi_r$’s are what was used to generate our data in models (1) to (4). Then, we estimated the vector of parameters $\Theta = (\theta_1, \ldots, \theta_q)$ by

$$\hat{\Theta} = \arg\min_{\Theta \in \mathbb{R}^q \text{ s.t. } \min_{1 \leq j \leq q} \theta_j \geq 0} \sum_{i=1}^{N} \left\{ \hat{K}(s_i, t_i) - \tilde{\phi}^\top(s_i, t_i)\Theta \right\}^2,$$

where $(s_1, t_1), \ldots, (s_N, t_N)$ are those used in section 5.2.1, $\tilde{\phi}^\top(s, t) = \{\phi_1(s)\phi_1(t), \ldots, \phi_q(s)\phi_q(t)\}$ and $\hat{K}$ is as at (2.2). We define the parametric estimator on the whole domain $S_0$ as $\tilde{K}_{True}(s, t) = \tilde{\phi}^\top(s, t)\hat{\Theta}$. Following the arguments from section 4, it can be proved that $\tilde{K}_{True}$ is a consistent estimator of $K$ on the whole $S_0$. In practice, we compute $\hat{\Theta}$ using the quasi-Newton algorithm. To generate an initial value for the algorithm, we first compute the solution $\tilde{\Theta}$ to (5.1) without the positivity constraint, i.e. $\tilde{\Theta} = (\tilde{\Phi}^\top\tilde{\Phi})^{-1}\tilde{\Phi}^\top\tilde{K}$, where $\tilde{\Phi} = \{\tilde{\phi}^\top(s_1, t_1), \ldots, \tilde{\phi}^\top(s_N, t_N)\}^\top \in \mathbb{R}^{N \times q}$ and $\tilde{K} = \{\tilde{K}(s_1, t_1), \ldots, \tilde{K}(s_N, t_N)\}^\top$. Then we set the initial values equal to $(\tilde{\Theta})_i = (\tilde{\Theta})_i1(\tilde{\Theta})_i \geq 0$, $i = 1, \ldots, q$.

For the model based on a low dimensional spline approximation, we followed the
approach of James and Hastie (2001). They used a mixed model for the full curves $X_i$ by taking, for $t \in I$, $X_i(t) = B^\top(t)(\lambda_0 + \gamma_i)$, where $B^\top(t) = (B_1(t), \ldots, B_{k+4}(t))^\top \in \mathbb{R}^{k+4}$ is a vector of cubic B-splines defined on $I$ with $k$ knots, for $k \geq 1$ a fixed integer. Here, $\lambda_0 \in \mathbb{R}$ is a constant and the $\gamma_i$'s are i.i.d. normal random vectors with mean zero and covariance $\Gamma$. Then $\gamma_0$ and $\Gamma$ are estimated by maximum likelihood (ML) under the constraint that $\Gamma$ is symmetric and positive definite, using an EM algorithm. This provides the estimators $\hat{\gamma}_0$ and $\hat{\Gamma}$. Finally, for $(s, t) \in S_0$ we estimate $K(s, t)$ by $\tilde{K}_{EM}(s, t) = B^\top(s) \hat{\Gamma} B(t)$.

Although this approach uses splines, it can be treated as a parametric model because their $k$ is usually small and fixed by the user. Unless the curves really come from this model, the approximation it provides is not guaranteed to be good. As highlighted by Delaigle and Hall (2013), although James and Hastie (2001) applied their technique to data that correspond to the fragment setting (see section 5.3), their method is better suited for sparse data. To improve its performance, instead choosing $k$ arbitrarily, in our simulation study, for each model we chose $k \in \{1, 3, 5, 8\}$ that minimised the mean integrated squared error of $\tilde{K}_{EM}$ on $S$ from one simulated sample of size $n = 100$ for which $E(\ell) = 0.5$. Then we used this $k$ for all simulated samples from that model (computing the best $k$ for each Monte Carlo replication would be too time consuming). This gave $k = 1, 1, 3, 8$ for models (1) to (4), respectively. We performed our simulations using the MATLAB code provided by James and Sugar (2003).

To compare our nonparametric estimator $\tilde{K}_p$ with those three parametric estimators, for each Monte Carlo replication from each combination of model, $n$ and fragment length described in section 5.2.1, we calculated $\tilde{K}_{True}$, $\tilde{K}_{EM}$ and $\tilde{K}_{DH}$. Table 2 summarises the means and standard deviations of the 100 ISE values of these four estimators for each configuration. Figures 8 to 11 show, for the same settings as Figures 5 to 7, the estimators $\tilde{K}_{True}$, $\tilde{K}_{EM}$ and $\tilde{K}_{DH}$ corresponding to their median ISE value.
As expected, the parametric method $\tilde{K}_{\text{True}}$ using the true model generally performed better than our nonparametric $\tilde{K}_p$, except in some cases where the parameters were difficult to estimate in small sample sizes; compare for example $\tilde{K}_{\text{True}}$ and $\tilde{K}_p$ for model (2) in the table. Compare also the second column of Figure 8 which shows $\tilde{K}_{\text{True}}$ for model (2) when $n = 100$, and the second column of Figure 4 which shows $\tilde{K}_p$. As expected too, the estimator $\tilde{K}_{\text{EM}}$, which is best suited for sparse functional data than for data in the form of fragments (see Delaigle and Hall, 2013 and Delaigle and Hall, 2016), generally did not perform well and had significantly poorer performance than the other estimators we considered, for example estimating highly pronounced upward trends by downward ones.

The performance of the estimator $\tilde{K}_{\text{DH}}$ depended on the complexity of the covariance function. In cases where the covariance function was decaying smoothly near the boundary of its domain, or was relatively flat in that area, this estimator performed quite well, sometimes even better than $\tilde{K}_{\text{True}}$; see for example the results for model (2) in Table 2 and Figure 8. This is because, as can also be seen from Figures 9 to 11 regardless of the shape of the true covariance function near the boundary of its domain, this estimator has a pronounced tendency to produce estimators that are quite flat in that extrapolation area. Therefore, it is particularly suitable for smoothly decaying or relatively flat cases like models (2) or (3), but not so much so for those with pronounced upward trends near the boundary of the domain like models (1) or (4).

Comparing these results to those in section 5.2.1, we see that when we expect the covariance function to smoothly decay near the boundary of its domain, Delaigle and Hall’s (2016) method could be recommended. Otherwise, when sample size and fragment length are small our method seems to outperform the other methods, at least in most cases we considered, both visually and in terms of the ISE values. When the sample size reaches 500, Zhang and Chen’s (2017) and Descary and Panaretos’ (2019) methods seem to become comparable to ours visually, except that some smoothing maybe needed for denoising. An advantage of those two approaches is their com-
Computational speed because they make use of simple and low-rank matrix operations. Therefore, when sample size is as large as 500, if having a smooth estimator is not crucial, Zhang and Chen’s (2017) and Descary and Panaretos’ (2019) methods could be the methods of choice.

We assessed the computation speed of our nonparametric method using MATLAB on a 3.4GHz computer with 16GB memory running Windows. For various combinations of \( n \) and \( \mathbb{E}(\ell) \) considered in our simulations, we recorded the computational time over 10 data sets from each of models (1) to (4) (i.e. 60 data sets in total). Then for each model, we took the average over the 10 computational times for each method and each combination \((n, \mathbb{E}(\ell))\), and this is what we show in Table 3. In particular, we measured the average computational times for calculating \( \tilde{K}_p \) with \( k \) and \( \lambda \) for \( \tilde{K}_s \) chosen by leave-one-curve-out cross-validation from 25 potential combinations.

<table>
<thead>
<tr>
<th>Model</th>
<th>( n )</th>
<th>( \mathbb{E}(\ell) )</th>
<th>( \mathbb{E}(\ell) )</th>
<th>( \mathbb{E}(\ell) )</th>
<th>( \mathbb{E}(\ell) )</th>
<th>( \mathbb{E}(\ell) )</th>
<th>( \mathbb{E}(\ell) )</th>
<th>( \mathbb{E}(\ell) )</th>
<th>( \mathbb{E}(\ell) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>50</td>
<td>0.2</td>
<td>14.7(15.0)</td>
<td>32.6(12.5)</td>
<td>7.83(9.52)</td>
<td>82.7(10.04)</td>
<td>446(1889)</td>
<td>259(590)</td>
<td>15.8(7.43)</td>
</tr>
<tr>
<td>2</td>
<td>50</td>
<td>0.2</td>
<td>206(144)</td>
<td>95.3(62.4)</td>
<td>180(107)</td>
<td>537(513)</td>
<td>4153(37959)</td>
<td>1179(108093)</td>
<td>70.4(37.2)</td>
</tr>
<tr>
<td>3</td>
<td>50</td>
<td>0.2</td>
<td>20.5(10.9)</td>
<td>13.1(8.3)</td>
<td>16.0(21.2)</td>
<td>12.3(15.5)</td>
<td>1003(340665)</td>
<td>2946(532083)</td>
<td>12.9(6.60)</td>
</tr>
<tr>
<td>4</td>
<td>50</td>
<td>0.2</td>
<td>17.7(20.4)</td>
<td>13.8(9.02)</td>
<td>15.1(22.0)</td>
<td>4.05(5.14)</td>
<td>397(1555)</td>
<td>27.5(20.6)</td>
<td>7.1(5.71)</td>
</tr>
</tbody>
</table>
Figure 8: True covariance (first column), $\tilde{K}_{\text{True}}$ (second column), $\tilde{K}_{\text{EM}}$ (third column) and $\tilde{K}_{\text{DH}}$ (fourth column) corresponding to the sample that produced the median ISE out of 100 samples generated from model (2) with $\ell_i \sim U(0.1, 0.3)$ (row 1) and $\ell_i \sim U(0.4, 0.6)$ (row 2) when $n = 100$, and $\ell_i \sim U(0.4, 0.6)$ (row 2) when $n = 500$ (row 3).

$(k, \lambda) \in \{1, 3, 5, 7, 9\} \times \{\exp(-10), \exp(-5), 1, \exp(1), \exp(2)\}$.

Table 3: Average computational times (in minutes) over models (1) to (4) computed in each case from 10 data sets.

<table>
<thead>
<tr>
<th>$n$</th>
<th>$E(\ell)$</th>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
<th>$E(\ell)$</th>
<th>(4)</th>
</tr>
</thead>
<tbody>
<tr>
<td>50</td>
<td>0.2</td>
<td>0.38</td>
<td>0.35</td>
<td>0.35</td>
<td>0.4</td>
<td>1.48</td>
</tr>
<tr>
<td></td>
<td>0.5</td>
<td>1.70</td>
<td>1.53</td>
<td>1.72</td>
<td>0.6</td>
<td>1.77</td>
</tr>
<tr>
<td></td>
<td>0.5</td>
<td>0.63</td>
<td>0.64</td>
<td>0.63</td>
<td>0.4</td>
<td>2.15</td>
</tr>
<tr>
<td>100</td>
<td>0.5</td>
<td>2.68</td>
<td>2.25</td>
<td>2.22</td>
<td>0.6</td>
<td>2.75</td>
</tr>
<tr>
<td></td>
<td>0.2</td>
<td>11.4</td>
<td>11.1</td>
<td>11.1</td>
<td>0.4</td>
<td>23.0</td>
</tr>
<tr>
<td>500</td>
<td>0.2</td>
<td>28.7</td>
<td>26.9</td>
<td>26.6</td>
<td>0.4</td>
<td>27.0</td>
</tr>
</tbody>
</table>

35
Figure 9: True covariance (first column), $\tilde{K}_{\text{True}}$ (second column), $\tilde{K}_{\text{EM}}$ (third column) and $\tilde{K}_{\text{DH}}$ (fourth column) corresponding to the sample that produced the median ISE out of 100 samples generated from model (1) with $\ell_i \sim U(0.4, 0.6)$ when $n = 50$ (row 1), $n = 100$ (row 2) and $n = 500$ (row 3).

Figure 10: True covariance (first column), $\tilde{K}_{\text{True}}$ (second column), $\tilde{K}_{\text{EM}}$ (third column) and $\tilde{K}_{\text{DH}}$ (fourth column) corresponding to the sample that produced the median ISE out of 100 samples generated from model (4) when $n = 500$ and $\ell_i \sim U(0.5, 0.7)$.
Figure 11: True covariance (first column), $\tilde{K}_{\text{True}}$ (second column), $\tilde{K}_{\text{EM}}$ (third column) and $\tilde{K}_{\text{DH}}$ (fourth column) corresponding to the sample that produced the median ISE out of 100 samples generated from model (3) with $\ell_i \sim U(0.1, 0.3)$ when $n = 500$.

Figure 12: Partially observed spine BMD curves for 130 females. Each fragment corresponds to one female.

5.3 Real data analysis

We applied our method as well as Descary and Panaratos’ (2019), Delaigle and Hall’s (2016) and James and Hastie’s (2001) methods to estimate the covariance function of the spine bone mineral density (BMD) data described in Bachrach et al. (1999). We considered the $n = 130$ females in the study whose BMD measurement (playing the role of $X$ in this case) was taken at various ages $t$ in the interval $I = [9.5, 21]$. In this study, for each female, the BMD was measured for only a short period of time, so that for $i = 1, \ldots, n$, $X_i(t)$ was observed only on an interval $I_i \subset I$ whose average length was about 15.64% of the length of $I$. As in our simulations, the data in this example were observed at a sequence of discrete nearby points, that is we observed discretised fragments; see Figure 12 where, for each female, we connect
the observed points.

In order to be able to apply directly Descary and Panaratos’ (2019) estimator, which is designed for data observed on equispaced grids, we took a discrete grid $I_D = \{t_1 = 9.5 < t_2 < \ldots < t_G = 21\}$ of $G = 116$ equispaced points on $I$ (corresponding essentially to the pooled observations). For each $i = 1, \ldots, n$, for $t \in I_i \cap I_D$, we obtained $X_i(t)$ by linear interpolation of the observations for that individual.

Figure 13 shows estimators of the covariance function computed from the equispaced data using four methods: our estimator $\tilde{K}_p$ with $\tilde{K}$ computed by spline as in section B.1 in the supplementary file, Descary and Panaratos’ (2019) estimator $\tilde{K}_{DP}$ with rank $q = 4$ chosen by the graphical way they proposed, Delaigle and Hall’s (2016) estimator $\tilde{K}_{DH}$ and James and Hastie’s (2001) estimator $\tilde{K}_{EM}$. We chose the number of knots $k$ for James and Hastie’s (2001) method as described on page 32 from the
set \( \{0, 1, 2, 3, 4, 5, 8\} \). We use the same colour code as in the simulation section, that is we gray out the region \( \bigcup_{i=1}^{n} (I_i \times I_i) \) where a standard covariance estimator can be computed, and colour the region where an extrapolation technique is required.

We can see from Figure 13 that James and Hastie’s (2001) estimator is very different from the other estimators in both the shape of the surface and the range of the values. \( \tilde{K}_{DP} \) is somewhat similar to \( \tilde{K}_p \) but is more noisy. Finally, Delaigle and Hall’s (2016) estimator \( \tilde{K}_{DH} \) is also somewhat similar to \( \tilde{K}_p \), but the assumed Markov property makes that estimator rather flat in the extrapolation area, a feature already encountered in the simulations section.

These findings are reflected in Figure 14, where we show, for the four covariance estimators, the two eigenfunctions corresponding to the two largest eigenvalues. They explain 88.97% and 10.87% (\( \tilde{K}_p \)), 62.46% and 30.01% (\( \tilde{K}_{DP} \)), 67.52% and 32.48% (\( \tilde{K}_{EM} \)), and 93.36% and 4% (\( \tilde{K}_{DH} \)) of the variability of the data. The first eigenfunction of \( \tilde{K}_p \) indicates that females who grow fast (resp., slowly) before age 14 and achieve a higher (resp., lower) BMD than the mean until that age, then tend to become more average by the end of their puberty. The second eigenfunction mainly seems to capture additional effects occurring in late puberty/early adulthood (after age 16). As for the covariance estimators, the eigenfunctions of \( \tilde{K}_{EM} \) are very different from those of \( \tilde{K}_p \), \( \tilde{K}_{DP} \) and \( \tilde{K}_{DH} \), which behave somewhat similarly to each other except that those of \( \tilde{K}_{DP} \) are much more noisy and the first one of \( \tilde{K}_{DH} \) is very flat.

These features are also reflected in Figure 15, where we use the estimated eigenfunctions to reconstruct the complete trajectories of the individual functions \( X_i(t) \) observed only subintervals \( I_i \) of \( I \). If we can reasonably assume that the data are Gaussian, then a simple way to approximate the values of \( X_i(t) \) for \( t \in I, t \notin I_i \), is to rely on the PACE methodology developed by Yao et al. (2005). A somewhat more sophisticated approach, which does not rely on the Gaussian assumption, is described by Kneip and Liebl (2017). They present an algorithm to approximate the best possible linear operator leading to the smallest possible mean square reconstruction error.
Figure 14: Eigenfunctions of female growth data, corresponding to the largest eigenvalue (left) and the second largest eigenvalue (right), computed from \( \tilde{K}_p \) (---), \( \tilde{K}_{DP} \) (-- -), \( \tilde{K}_{DH} \) (· · ·) and \( \tilde{K}_{EM} \) (- - -).

Figure 15: Reconstructed curves for the spine BMD data based on our estimator \( \tilde{K}_p \) (top left), \( \tilde{K}_{DP} \) (top right), \( \tilde{K}_{DH} \) (bottom left) and \( \tilde{K}_{EM} \) (bottom right).

For reconstructing \( X_i(t) \) for \( t \in \mathcal{I} \setminus \mathcal{I}_i \), the procedure requires an estimator of \( K(t, s) \) for all \( t, s \in \mathcal{I} \), and we applied it with each of the four covariance estimators. The resulting recovered curves from \( \tilde{K}_p \) in Figure 15 reasonably show an increasing pattern with peaks around 16 years old. The reconstructed curves obtained when using
instead $\tilde{K}_\text{DP}$ and $\tilde{K}_\text{EM}$ look somewhat similar but less realistic. The flat aspect of the first eigenfunction of $\tilde{K}_\text{DH}$ for $t > 14$ is reflected in the reconstructed curves, which tend to be very flat in that area too. The fact that its second eigenfunction explains only 4% of the variability of the data makes all reconstructed curves based on $\tilde{K}_\text{DH}$ look very similar to each other.

**Acknowledgement**

Delaigle’s research was supported by a Future Fellowship (FT130100098) and a Discovery Project (DP170102434) of the Australian Research Council. Huang’s research was supported by the Australian Research Council Centre of Excellence for Mathematical and Statistical Frontiers (ACEMS).

**References**


42