Biometrika (2021), **108**, 3, *pp.* 709–726
Printed in Great Britain
doi: 10.1093/biomet/asaa088
Advance Access publication 21 May 2021

Basis expansions for functional snippets

BY ZHENHUA LIN

Department of Statistics and Applied Probability, National University of Singapore, 6 Science Drive, 117546, Singapore
linz@nus.edu.sg

JANE-LING WANG

Department of Statistics, University of California, One Shields Avenue, Davis, California 95616, U.S.A.
janelwang@ucdavis.edu

AND QIXIAN ZHONG

Department of Mathematical Sciences, Tsinghua University, Beijing 100084, China
zqx15@mails.tsinghua.edu.cn

SUMMARY

Estimation of mean and covariance functions is fundamental for functional data analysis. While this topic has been studied extensively in the literature, a key assumption is that there are enough data in the domain of interest to estimate both the mean and covariance functions. We investigate mean and covariance estimation for functional snippets in which observations from a subject are available only in an interval of length strictly, and often much, shorter than the length of the whole interval of interest. For such a sampling plan, no data is available for direct estimation of the off-diagonal region of the covariance function. We tackle this challenge via a basis representation of the covariance function. The proposed estimator enjoys a convergence rate that is adaptive to the smoothness of the underlying covariance function, and has superior finite-sample performance in simulation studies.

Some key words: Covariance estimation; Fourier series; Functional fragment; Legendre polynomial; Longitudinal data; Penalized estimation; Sequential compactness.

1. INTRODUCTION

Nowadays functional data are commonly encountered in practice, due to the advances in modern science and technology that enhance data collection and processing capabilities. Both unsupervised learning, such as dimension reduction via functional principal component analysis (Rao, 1958; Hall & Hosseini-Nasab, 2009; Mas & Ruymgaart, 2015), and supervised learning, such as functional regression (Müller & Stadtmüller, 2005; Ferraty & Vieu, 2006; Hall & Horowitz, 2007; Müller & Yao, 2008; Kong et al., 2016), are well studied in the literature. For a comprehensive treatment of these subjects, we recommend the monographs by Ramsay & Silverman (2005), Ferraty & Vieu (2006), Horváth & Kokoszka (2012), Hsing & Eubank (2015)

and Kokoszka & Reimherr (2017), and the review papers Wang et al. (2016) and Aneiros et al. (2019).

Critical to the statistical analysis of such data is the estimation of the mean and covariance functions, since they are the foundation of the aforementioned unsupervised and supervised learning tasks. For example, covariance estimation is a critical step to functional principal component analysis, as illustrated in § 5. In reality, functions can only be recorded at a set of discrete points on the domain of the functions, where this set may vary among subjects and the measurements may contain noise. Estimation of mean and covariance functions in this context has been extensively studied by Rice & Silverman (1991), Cardot (2000), James et al. (2000), Yao et al. (2005b), Cai & Yuan (2010, 2011), Li & Hsing (2010) and Zhang & Wang (2016), among many others. In addition to the discrete nature of observed functional data, subjects often stay in the study only for a subject-specific period that is much shorter than the span of the whole study. This brings challenges to covariance estimation.

For illustration, and without loss of generality, we assume that the domain of the functional data $X(t)$ is the unit interval $\mathcal{T} = [0, 1]$ and each subject only stays in the study for a period of length $\delta < 1$. Data with these characteristics are termed functional snippets in this article, in analogy to the longitudinal snippets analysed in Dawson & Müller (2018). For such data, there is no information in the off-diagonal region $\mathcal{T}_\delta^c := \{(s, t) \in [0, 1]^2 : |s - t| > \delta\}$ of the covariance function $\mathrm{cov}\{X(s), X(t)\}$, and therefore there is no local information available for estimating the covariance function in this region. Mathematically, this amounts to

$$\mathrm{pr}\{(\cup_{i=1}^n [A_i, B_i]^2) \cap \mathcal{T}_\delta^c = \emptyset\} = 1, \tag{1}$$

for some $\delta > 0$ and for all n, where $[A_i, B_i]$ denotes the subinterval on which X_i is observed. In § 5 we illustrate such a situation for the bone mineral density data, where the band with available data is prominent and narrow. Estimating the covariance function for this type of data is therefore an extrapolation problem. Methods based on interpolation, such as local smoothing methods (Yao et al., 2005a; Li & Hsing, 2010), fail to yield a consistent estimate of the covariance function in the off-diagonal region.

Functional snippets were previously studied by Delaigle & Hall (2013, 2016), Descary & Panaretos (2018, 2019) and Zhang & Chen (2020a,b) under the term fragments or fragmentary functional data. For instance, Delaigle & Hall (2016) proposed approximating snippets by segments of Markov chains. This method is only valid at the discrete level, as explained in Descary & Panaretos (2019). To analyse functional snippets, Descary & Panaretos (2019) and Zhang & Chen (2020b) used matrix completion techniques that innately work on a common grid. These approaches require modification when snippets are recorded on random and irregular points, which are often encountered in applications. Yet, published theoretical analyses focus on the regular and dense design.

Fragments and other terms, such as censored functional data, incomplete functional data and partially observed functional data, have also been used to refer to fragments that are not functional snippets (Liebl, 2013; Gellar et al., 2014; Goldberg et al., 2014; Kraus, 2015; Gromenko et al., 2017; Mojirsheibani & Shaw, 2018; Stefanucci et al., 2018; Kraus & Stefanucci, 2019; Liebl & Rameseder, 2019; Kneip & Liebl, 2020). For example, Kneip & Liebl (2020) assumed that $\mathrm{pr}([A_i, B_i]^2 = [0, 1]^2) > 0$, and consequently information and design points for the off-diagonal region $\mathcal{T}_\delta^c$ are still available. In these works, the problem of recovering the covariance function is often formulated as an interpolation problem. In contrast, information for that region is completely missing for functional snippets characterized by (1), which significantly elevates the difficulty of statistical analysis. Because of this fundamental difference between these two types of data, we

adopt the term functional snippets to distinguish them from the fragments that are not snippets and other partially observed functional data.

As an extrapolation problem, estimating the covariance function for functional snippets requires additional identifiability assumptions, for which the minimal one is the trivial condition that the covariance function in the observable band $\mathcal{T}_\delta$ determines the covariance function on the entire domain. This minimal identifiability is a high-level concept (Delaigle et al., 2020), and consequently existing works attempt to find some specialized conditions that imply the above minimal identifiability assumption. For instance, Descary & Panaretos (2019) assumed that the covariance function is an analytic function, and Delaigle et al. (2020) proposed the linear predictability assumption, which assumes that the values of the process $X(t)$ on a subinterval can be linearly predicted by the values of the same process on another subinterval.

In contrast to the aforementioned approaches, which impose a particular assumption on the process $X(t)$ itself or on its covariance function, we define identifiability through a family $\mathcal{C}$ in which the covariance function resides, and term such a family $\mathcal{T}_\delta$-identifiable in the case that, if any two members from the family $\mathcal{C}$ are identical on the diagonal region $\mathcal{T}_\delta$, then they are equal everywhere. This $\mathcal{T}_\delta$-identifiability is the same as the above minimal identifiability except that we make the reference to a family explicit. The family $\mathcal{C}$ is comparable to the traditional parameter space or model, so our definition of identifiability is in line with the conventional statistical concept of identifiability that is imposed on the model. This concept of identifiability is rather general and encompasses the aforementioned identifiability assumptions as special cases. For example, the class of analytic functions considered in Descary & Panaretos (2019) and the class of covariance functions associated with linearly predictable random processes are $\mathcal{T}_\delta$-identifiable families; see Examples 1 and 4 for details. The primary reason that we adopt this minimal identifiability is that our method and theory to be developed in § 2 and § 3 apply to all $\mathcal{T}_\delta$-identifiable families under some regularity conditions.

Under the umbrella of $\mathcal{T}_\delta$-identifiability, we propose to approach functional snippets from the perspective of basis expansion. The main idea is to represent the covariance function by basis functions composed from tensor products of analytic orthonormal functions defined on $\mathcal{T}$. Basis functions, in particular spline basis functions, have been extensively explored in both nonparametric smoothing and functional data analysis by Wahba (1990), Rice & Wu (2001), Wood (2003), Ramsay & Silverman (2005) and Crambes et al. (2009), among many others. However, they are not suited for the extrapolation problem of functional snippets, as these bases are local. Unlike spline bases that are controlled by knots, analytic bases are global, in the sense that they are independent of local information such as knots or design points and are completely determined by their values on a countably infinite subset of the interval $\mathcal{T}$. This feature of analytic bases allows information to pass from the diagonal region to the off-diagonal region along the basis functions. Consequently, the missing pieces of the covariance function can then be inferred from the data available in the diagonal region when the covariance function is from a $\mathcal{T}_\delta$-identifiable class. In contrast, this is generally impossible for B-spline or other local bases.

In addition to the minimal identifiability assumption, the consistency of the proposed estimator requires extra regularity conditions to overcome the challenges of extrapolation. One regularity condition that we identified is the bounded sequential compactness of the family $\mathcal{C}$ of covariance functions under consideration. This new concept, developed in § 3.3, essentially controls the complexity of the family $\mathcal{C}$ and enables us to establish the consistency and convergence rate of the proposed estimator in a nonparametric extrapolation setting. This condition is mild; for example,

all families of functions that are uniformly bounded and Lipschitz continuous with a common Lipschitz constant are boundedly sequentially compact, as shown in § 3.3. Such a regularity condition, not seen in the literature, is not required for interpolation and thus intrinsically separates nonparametric extrapolation from interpolation.

When our work was completed, we became aware of a related piece of work that was independently developed by Delaigle et al. (2020). Although that work also uses a basis expansion approach, it is substantially different from ours. First, it focuses more on development of identifiability conditions while ours is on methodological development and theoretical analysis of the proposed estimators. Second, the method of Delaigle et al. (2020) extrapolates a pilot estimate from the diagonal region to the entire region by basis expansion without regularization. This may lead to excessive variability of the estimator in the off-diagonal region. In contrast, our method estimates the basis coefficients directly from data with penalized least squares and does not require a pilot estimate. Third, the convergence rate established in that work hinges on, and thus is limited by, the convergence rate of the pilot estimate, while our analysis gives an explicit rate that is adaptive to the smoothness of the underlying covariance function. Finally, the work of Delaigle et al. (2020) heuristically includes the identity function into the Fourier basis to handle nonperiodicity, while ours adopts the Fourier extension technique that is well established in numerical analysis.

2. METHODOLOGY

2.1. *Mean function*

Let $\{X(t) : t \in \mathcal{T}\}$ be a second-order stochastic process on a compact interval $\mathcal{T} \subset \mathbb{R}$, which without loss of generality is taken to be $[0, 1]$. The mean and covariance functions of X are defined as $\mu_0(t) = E\{X(t)\}$ and $\gamma_0(s, t) = \mathrm{cov}\{X(s), X(t)\}$, respectively. The observed functions $X_1, \ldots, X_n$ are statistically modelled as independent and identically distributed realizations of X. In practice, each realization X_i is only recorded at m_i subject-specific time-points $T_{i1}, \ldots, T_{im_i}$ with measurement errors. More precisely, for functional snippets the observed data are pairs (T_{ij}, Y_{ij}), where

$$Y_{ij} = X_i(T_{ij}) + \varepsilon_{ij} \qquad (i = 1, \ldots, n,\ j = 1, \ldots, m_i).$$

Here, ε_{ij} is the random noise with mean zero and unknown variance σ^2, and there is a constant $\delta \in (0, 1)$ for which $|T_{ij} - T_{ik}| \leqslant \delta$ for all i, j and k. The focus of this paper is to estimate the mean and covariance functions of X using these data pairs (T_{ij}, Y_{ij}).

Although functional snippets pose a challenge for covariance estimation, they usually do not obstruct mean estimation, since data for the estimation are likely available across the whole domain of interest. In light of this observation, traditional methods such as local linear smoothing (Yao et al., 2005b; Li & Hsing, 2010) can be employed. Below we adopt a different approach based on analytic basis expansions. The advantage of this approach is its computational efficiency and adaptivity to the regularity of the underlying mean function μ_0; see also § 3.2.

Let $\Phi = \{\phi_1, \ldots\}$ be a complete orthonormal basis of $L^2(0, 1)$ that consists of squared integrable functions defined on the interval $[0, 1]$. When $\mu_0 \in L^2(0, 1)$, it can be represented by the series $\mu_0(t) = \sum_{k=1}^{\infty} a_k \phi_k(t)$ in terms of the basis Φ, where $a_k = \int_0^1 \mu_0(t)\phi_k(t)\,\mathrm{d}t$. In practice, one often approximates such a series by its first $q > 0$ leading terms, where q is a tuning parameter controlling the approximation quality. The coefficients $a_1, \ldots, a_q$ are then estimated from data by penalized least squares. Specifically, with the notation $\Phi_q(t) = \{\phi_1(t), \ldots, \phi_q(t)\}^{\mathrm{T}} \in \mathbb{R}^q$ and

$A_0 = (a_1, \ldots, a_q)^{\mathrm{T}}$, the estimator of A_0 is given by

$$\hat{A} = \underset{A \in \mathbb{R}^q}{\arg\min} \Big[\sum_{i=1}^{n} v_i \sum_{j=1}^{m_i} \{Y_{ij} - A^{\mathrm{T}} \Phi_q(T_{ij})\}^2 + \rho H(A^{\mathrm{T}} \Phi_q) \Big], \tag{2}$$

and μ_0 is estimated by $\hat{\mu}(t) = \hat{A}^{\mathrm{T}} \Phi_q(t)$, where the weights $v_i > 0$ satisfy $\sum_{i=1}^{n} v_i m_i = 1$, $H(\cdot)$ represents the roughness penalty, and ρ is a tuning parameter that provides a trade-off between the fidelity to the data and the smoothness of the estimate. There are two commonly used schemes for the weights, equal weight per observation and equal weight per subject, for which the weights v_i are $1/(\sum_{i=1}^{n} m_i)$ and $1/(nm_i)$, respectively. These, and also alternative weight schemes, are discussed in Zhang & Wang (2016, 2018).

The penalty term in (2) is introduced to prevent excessive variability of the estimator when a large number of basis functions are required to adequately approximate μ_0, and when the sample size is not sufficiently large. In the asymptotic analysis of $\hat{\mu}$ in § 3, we will see that this penalty term does not affect the convergence rate of $\hat{\mu}$ when the tuning parameter ρ is not too large. In our study, the roughness penalty is $H(g) = \int_0^1 \{g^{(2)}(t)\}^2 \, \mathrm{d}t$, where $g^{(2)}$ is the second derivative of g. The choices of q and ρ are discussed in § 4.1.

2.2. *Covariance function*

Since functional snippets do not provide any direct information for the off-diagonal region, the only way to recover the covariance in the off-diagonal region is to infer it from the diagonal region. The following definition formulates this basic requirement for identifiability.

DEFINITION 1. *A family $\mathcal{C}$ of covariance functions is called a $\mathcal{T}_\delta$-identifiable family if $\gamma_1, \gamma_2 \in \mathcal{C}$ and $\gamma_1(s,t) = \gamma_2(s,t)$ for all $(s,t) \in \mathcal{T}_\delta$ imply that $\gamma_1(s,t) = \gamma_2(s,t)$ for all $(s,t) \in \mathcal{T}^2$.*

Intuitively, we consider a family $\mathcal{C}$ of covariance functions and require the covariance functions to be uniquely identified within the family $\mathcal{C}$ by their values on the diagonal region. Below we provide four examples to illustrate the ubiquitousness of $\mathcal{T}_\delta$-identifiable families.

Example 1 (Analytic functions). A function is analytic if it can be locally represented by a convergent power series. By Corollary 1.2.7 of Krantz & Parks (2002), if two analytic functions agree on $\mathcal{T}_\delta$ then they are identical on $[0,1]^2$. Thus, the family of analytic functions is a $\mathcal{T}_\delta$-identifiable family, as observed by Descary & Panaretos (2019), who also provided an elegant example to demonstrate that the space of infinitely differentiable functions is not $\mathcal{T}_\delta$-identifiable.

Example 2 (Sobolev sandwich families). For any $0 < \epsilon < \delta$, consider the family of continuous functions that belong to a two-dimensional Sobolev space on $\mathcal{T}_\epsilon$ and are analytic elsewhere. Such functions have an r-times differentiable diagonal component sandwiched between two analytic off-diagonal pieces. The family is $\mathcal{T}_\delta$-identifiable, because the values of such functions on the off-diagonal region are fully determined by the values on the uncountable set $\mathcal{T}_\epsilon^{\mathrm{c}} \cap \mathcal{T}_\delta \subset \mathcal{T}_\delta$ according to Corollary 1.2.7 of Krantz & Parks (2002). This family contains functions with derivatives only up to a finite order.

Example 3 (Semiparametric families). Consider the family of functions of the form $g(s)h(s,t)g(t)$, where g is a function from a nonparametric class $\mathcal{G}$ and h is from a parametric class $\mathcal{H}$ of correlation functions. This family, considered in Lin & Wang (2020), is generally

$\mathcal{T}_\delta$-identifiable, as long as both $\mathcal{G}$ and $\mathcal{H}$ are identifiable, for instance when $\mathcal{G}$ is a Sobolev space and $\mathcal{H}$ is the class of Matérn correlation functions. No analyticity is assumed for this family.

Example 4 (Linearly predictable families). For a random process X defined on $\mathcal{T}$, we say that the snippet $\{X(t)\}_{t\in I^*}$ on the subinterval $I^* \subset \mathcal{T}$ is linearly (B,ϵ)-predictable (Delaigle et al., 2020) from another subinterval $I \subset \mathcal{T}$ if, for all $t \in I^*$, there exists an integrable function $L_t(s)$ defined on I such that $\sup_{t\in I^*}\sup_{s\in I}|L_t(s)| < \infty$, $\sup_{t\in I^*}\int_I |L_t(s)|\,\mathrm{d}s < B$ and $X(t) = \mu(t) + \int_I L_t(s)\{X(s)-\mu(s)\}\,\mathrm{d}s + Z(t)$, where, for all $t \in I^*$, $Z(t)$ is a zero-mean random variable such that $E\{Z^2(t)\} \leqslant \epsilon^2$. Fix an integer $h > 0$ and a partition $I_0,\ldots,I_h$ of $\mathcal{T}$ such that $I_j \times I_j \subset \mathcal{T}_\delta$. Consider only the class $\mathscr{X}$ of random processes X whose snippet $\{X(t)\}_{t\in I_j}$ is linearly (B_j,ϵ_j)-predictable from I_{j^*} for some $0 \leqslant j^* \leqslant j-1$ and all $\epsilon_j > 0$, and for all $j = 1,\ldots,h$. Let $\mathcal{L}$ be the collection of covariance functions of random processes in $\mathscr{X}$. By Delaigle et al. (2020), each member in $\mathcal{L}$ is identifiable within $\mathcal{L}$ from its values on the diagonal region $\mathcal{T}_\delta$, and thus $\mathcal{L}$ is $\mathcal{T}_\delta$-identifiable.

With the $\mathcal{T}_\delta$-identifiability of the family $\mathcal{C}$, it is now possible to infer the off-diagonal region by the information contained in the raw covariance $\Gamma_{ijk} = \{Y_{ij}-\hat{\mu}(T_{ij})\}\{Y_{ik}-\hat{\mu}(T_{ik})\}$ available only in the diagonal region. To this end, we propose to transport information from the diagonal region to the off-diagonal region through the basis functions $\phi_k \otimes \phi_l$ with $(\phi_k \otimes \phi_l)(s,t) = \phi_k(s)\phi_l(t)$ for $s,t \in \mathcal{T}$, by approximating γ_0 with

$$\gamma_{C_0}(s,t) = \sum_{1\leqslant k,l\leqslant p} c_{kl}\phi_k(s)\varphi_l(t), \qquad (s,t) \in [0,1]^2, \tag{3}$$

where $c_{kl} = \iint \gamma_0(s,t)\phi_k(s)\phi_l(t)\,\mathrm{d}s\,\mathrm{d}t$, C_0 is the matrix of coefficients c_{kl} and $p \geqslant 1$ is an integer. There are countless bases that can serve in (3); however, if we choose an analytic basis Φ, then their values in the diagonal region completely determine their values in the off-diagonal region. When such a representation of the covariance function γ_0 is adopted and the unknown coefficients c_{kl} are estimated from data, the information contained in the estimated coefficients extends from the diagonal region to the off-diagonal region through the analyticity of the basis.

To estimate the coefficients c_{kl} from data, we adopt the idea of penalized least squares, where the squared loss of a given function γ is measured by the sum of weighted squared errors $\sum_{i=1}^n w_i \sum_{1\leqslant j\neq k\leqslant m_i}\{\Gamma_{ijl} - \gamma(T_{ij},T_{il})\}^2$, where $w_i > 0$ are weights satisfying $\sum_{i=1}^n m_i(m_i-1)w_i = 1$, while the roughness penalty is given by $J(\gamma) = \iint \left\{\left(\partial^2\gamma/\partial s^2\right)^2/2 + \left(\partial^2\gamma/\partial s\partial t\right)^2 + \left(\partial^2\gamma/\partial t^2\right)^2/2\right\}\mathrm{d}s\,\mathrm{d}t$. The estimator $\hat{\gamma}(s,t)$ of $\gamma_0(s,t)$ is then taken as $\hat{\gamma}(s,t) = \Phi_p^{\mathrm{T}}(s)\hat{C}\Phi_p(t)$, with

$$\hat{C} = \underset{C:\,\gamma_C\in\mathcal{C}}{\arg\min} \sum_{i=1}^n w_i \sum_{1\leqslant j\neq l\leqslant m_i}\{\Gamma_{ijl} - \gamma_C(T_{ij},T_{il})\}^2 + \lambda J(\gamma_C), \tag{4}$$

where γ_C is defined in (3) with C_0 replaced by C, and λ is a tuning parameter that provides a trade-off between the fidelity to the data and the smoothness of the estimate. A numerical method to solve the constraint optimization (4) is detailed in the Supplementary Material.

Similar to (2), the penalty term in (4) is introduced to overcome excessive variability of an estimator when a large number of basis functions are required while the sample size is relatively small. It does not affect the convergence rate of $\hat{\gamma}$ when the tuning parameter λ is not too large. The choices of p and λ are discussed in § 4.1. For the weights w_i, Zhang & Wang (2016) discussed

several weighting schemes, including $w_i = 1/\{\sum_{i=1}^n m_i(m_i - 1)\}$ and $w_i = 1/\{nm_i(m_i - 1)\}$. An optimal weighting scheme was proposed in Zhang & Wang (2018); we refer to this paper for further details.

3. Theory

3.1. *Analytic basis*

While all complete orthonormal bases can be used for the proposed estimator in (2), an analytic basis is preferred for the estimator in (4). For a clean presentation, we exclusively consider analytic bases $\Phi = \{\phi_1, \ldots\}$ that work for both (2) and (4). In this paper, a basis is called an analytic (α, β)-basis if its basis functions are all analytic and satisfy the following property: for some constants $\alpha, \beta \geqslant 0$, there exists a constant c such that $\|\phi_k\|_\infty \leqslant ck^\alpha$ and $\|\phi_k^{(r)}\|_{L^2} \leqslant ck^{\beta r}$ for $r = 1, 2$ and all $k = 1, \ldots$. Here, $\|\phi_k\|_\infty$ denotes the supremum norm of ϕ_k, defined as $\sup_{t\in[0,1]} |\phi_k(t)|$, and $\phi_k^{(r)}$ represents the rth derivative of ϕ_k.

Different bases lead to different convergence rates of the approximation to μ_0 and γ_0. For the mean function μ_0, when using the first q basis functions $\phi_1, \ldots, \phi_q$, the approximation error is quantified by $\mathcal{E}(\mu_0, \Phi, q) = \big\|\mu_0 - \sum_{k=1}^q a_k\phi_k\big\|_{L^2}$, where we recall that $a_k = \int_0^1 \mu_0(t)\phi_k(t)\,\mathrm{d}t$. The convergence rate of the error $\mathcal{E}(\mu_0, \Phi, q)$, denoted by $\tau_q = \tau_q(\mu_0, \Phi)$, signifies the approximation power of the basis Φ for μ_0. Similarly, the approximation error for γ_0 is measured by $\mathcal{E}(\gamma_0, \Phi, p) = \big\|\gamma_0 - \sum_{k=1}^p \sum_{l=1}^p c_{kl}\phi_k \otimes \phi_l\big\|_{L^2}$, where the L^2 norm of a function $\gamma(s,t)$ is defined by $\|\gamma\|_{L^2} = \{\int_0^1 \int_0^1 \gamma^2(s,t)\,\mathrm{d}s\,\mathrm{d}t\}^{1/2}$. The convergence rate of $\mathcal{E}(\gamma_0, \Phi, p)$ is denoted by $\kappa_p = \kappa_p(\gamma_0, \Phi)$. Below we discuss two examples of analytic (α, β)-bases.

Example 5 (Fourier basis). Fourier basis functions, defined by $\phi_1(t) = 1$, $\phi_{2k}(t) = \cos(2k\pi t)$ and $\phi_{2k+1}(t) = \sin(2k\pi t)$ for $k \geqslant 1$, constitute a complete orthonormal basis of $L^2(\mathcal{T})$ for $\mathcal{T} = [0, 1]$. It is also an analytic $(0, 1)$-basis. When μ_0 is periodic on $\mathcal{T}$ and belongs to the Sobolev space $\mathscr{H}^r(\mathcal{T})$, see Appendix A.11.a and A.11.d of Canuto et al. (2006) for the definition, then, according to equation (5.8.4) of Canuto et al. (2006) one has $\tau_q = O(q^{-r})$. Similarly, if γ_0 is periodic and belongs to $\mathscr{H}^r(\mathcal{T}^2)$, then $\kappa_p = O(p^{-r})$.

Example 6 (Legendre polynomials). The canonical Legendre polynomial $P_k(t)$ of degree k is defined on $[-1, 1]$ by $P_k(t) = 2^{-k}g_k^{(k)}(t)/k!$, with $g_k(t) = (t^2 - 1)^k$ and $g_k^{(k)}$ denoting the kth derivative of the function g_k. These polynomials are orthogonal in $L^2(-1, 1)$ and can be turned into an orthonormal basis of $L^2(\mathcal{T})$. One can show that the Legendre basis is an analytic $(1/2, 1)$-basis. According to equation (5.8.11) of Canuto et al. (2006), one has $\tau_q = O(q^{-r})$ and $\kappa_p = O(p^{-r})$ when μ_0 belongs to $\mathscr{H}^r(\mathcal{T})$ and γ_0 belongs to $\mathscr{H}^r(\mathcal{T}^2)$, respectively.

3.2. *Mean function*

As functional snippets are often sparsely recorded, in the sense that $m_i \leqslant m_0 < \infty$ for all $i = 1, \ldots, n$ and some $m_0 > 0$, in this article we focus on theoretical analysis tailored to this scenario. For simplicity, we assume an identical number of observations and identical weight for each trajectory, i.e., $m_1 = \cdots = m_n = m$, $v_1 = \cdots = v_n$ and $w_1 = \cdots = w_n$. The results for a general number m_i of observations and general weight schemes can be derived in a similar fashion.

For functional snippets we shall assume that the observations from a subject scatter randomly in a subject-specific time interval whose length is δ and whose middle point is called the reference

time in this paper. We further assume that the reference time R_i of the ith subject is independently and identically distributed in the interval $[\delta/2, 1-\delta/2]$, and the observed time-points $T_{i1}, \ldots, T_{im_i}$, conditional on R_i, are independently and identically distributed in the interval $[R_i-\delta/2, R_i+\delta/2]$.

To study the estimator $\hat{\mu}$, we make the following assumptions.

Assumption 1. There exist $0 < \mathfrak{c}_1 \leqslant \mathfrak{c}_2 < \infty$ such that the density $f_R(s)$ of the reference time R satisfies $\mathfrak{c}_1 \leqslant f_R(s) \leqslant \mathfrak{c}_2$ for any $s \in [\delta/2, 1-\delta/2]$. There exist $0 < \mathfrak{c}_3 \leqslant \mathfrak{c}_4 < \infty$ such that the conditional density $f_{T|R}(t \mid s)$ of the observed time T satisfies $\mathfrak{c}_3 \leqslant f_{T|R}(t \mid s) \leqslant \mathfrak{c}_4$ for any given reference time $s \in [\delta/2, 1-\delta/2]$ and $t \in [s-\delta/2, s+\delta/2]$.

Assumption 2. We have that $E\{\|X\|_{L^2}^2\} \leqslant \mathfrak{c}_5 < \infty$ for some constant $\mathfrak{c}_5 > 0$.

Assumption 3. We have that $q^{2\alpha+2}/n \to 0$ and $\rho/(n^{-1/2}q^{\alpha-4\beta-1/2}) \to 0$.

Assumption 1 requires the density of the reference time and conditional densities of the time-points to be bounded away from zero and infinity. This also guarantees that the marginal probability density of the time-points T_{ij} is bounded away from zero and infinity. Assumption 2 is mild and Assumption 3 facilitates the convergence rate, where the dimension q can grow with n. In the following, we use $a_n \asymp b_n$ to denote $0 < \lim_{n\to\infty} a_n/b_n < \infty$.

THEOREM 1. *If Φ is a (α, β)-basis, Assumptions 1–3 imply that*

$$\|\hat{\mu} - \mu_0\|_{L^2}^2 = O_P\left(\frac{q^{2\alpha+1}}{n} + \tau_q^2\right), \tag{5}$$

where τ_q is the convergence rate of $\mathcal{E}(\mu_0, \Phi, q)$ defined in § 3.1.

Under Assumption 3, the tuning parameter ρ does not have direct impact on the asymptotic rate of $\hat{\mu}$. We also observe that in (5), the term $q^{2\alpha+1}n^{-1}$ specifies the estimation error using a finite sample, while τ_q is the deterministic approximation error for using only the first $q < \infty$ basis functions. The latter term depends on the smoothness of μ_0. Intuitively, it is easier to approximate smooth functions with basis functions. For a given number of basis functions, smoother functions generally yield smaller approximation errors. As discussed in Examples 5 and 6, when μ_0 belongs to the Sobolev space $\mathscr{H}^r(0,1)$, i.e., μ_0 is r times differentiable, we have $\tau_q = O(q^{-r})$. This leads to the following convergence rate.

COROLLARY 1. *Suppose that $\mu_0^{(r)}$ exists and satisfies $\|\mu_0^{(r)}\|_{L^2} < \infty$ for some $r \geqslant 1$. Assume that Assumptions 1–3 hold.*

(i) *If Φ is the Fourier basis and μ_0 is periodic, then $\|\hat{\mu} - \mu_0\|_{L^2}^2 = O_P(n^{-2r/(2r+1)})$ with the choice $q \asymp n^{1/(2r+1)}$.*
(ii) *If Φ is the Legendre basis, then $\|\hat{\mu} - \mu_0\|_{L^2}^2 = O_P(n^{-r/(r+1)})$ with the choice $q \asymp n^{1/(2r+2)}$.*

For $r = 2$ and a periodic function μ_0, the convergence rate for $\hat{\mu}$ is $n^{-2/5}$ and $n^{-1/3}$, respectively, for an estimator based on Fourier and Legendre bases. We can see that the convergence rate is faster for a Fourier basis. This is because, although they are both (α, β)-bases, $\alpha = 1/2$ for the Legendre basis is larger than $\alpha = 0$ for the Fourier basis. According to (5), a larger value of α leads to a slower rate. Indeed, α controls the growth rate of the extrema of basis functions. Fourier basis functions are uniformly bounded between -1 and 1. In contrast, high-order Legendre basis

functions tend to have large extrema that amplify variability. This limits the number of basis functions for estimation and thus causes a slower convergence rate for the Legendre basis. When μ_0 is nonperiodic, the classic Fourier basis suffers from the so-called Gibbs phenomenon which, however, can be substantially alleviated by the Fourier extension technique; see the Supplementary Material for more details.

3.3. *Covariance function*

In § 2 we assumed γ_0 to reside in a $\mathcal{T}_\delta$-identifiable family $\mathcal{C}$ in order to meet a basic criterion of identifiability. To study the asymptotic properties of the covariance estimator, we require the family $\mathcal{C}$ to satisfy an additional regularity condition as described below. Let $\mathcal{F}$ be the space of real-valued functions defined on $\mathcal{T}^2$ endowed with the product topology. In this topology, a sequence of functions $\{f_k\}$ converges to a limit f if and only if $\lim_{k\to\infty} f_k(s,t) = f(s,t)$ for all $(s,t) \in \mathcal{T}^2$.

DEFINITION 2. *A subset $\mathcal{S}$ of $\mathcal{F}$ is called a boundedly sequentially compact family if every sequence $\{f_k\} \subset \mathcal{S}$ that is bounded in the L^2 norm, i.e., $\sup_k \|f_k\|_{L^2} < \infty$, has a subsequence converging to a limit in $\mathcal{S}$ in the product topology.*

The concept of bounded sequential compactness is closely related to the topological concept of sequential compactness. Specifically, a subset $\mathcal{S} \subset \mathcal{F}$ is sequentially compact if every sequence in $\mathcal{S}$ has a subsequence that converges to a limit in $\mathcal{S}$ in the topology of $\mathcal{F}$. Sequential compactness is stronger than bounded sequential compactness and thus implies the latter. However, when the subset $\mathcal{S}$ is uniformly bounded in the L^2 norm, i.e., $\sup_{f\in\mathcal{S}} \|f\|_{L^2} < \infty$, then the two concepts coincide. If all functions in $\mathcal{S}$ are bounded by a common constant and are Lipschitz continuous with a common Lipschitz constant, then $\mathcal{S}$ is a boundedly sequentially compact family. Such a family is locally equicontinuous, and also the set $\{f(s,t) : f \in \mathcal{S}\}$ is bounded for all $s,t \in \mathcal{T}$. Then the claim follows from the Arzelà–Ascoli theorem (Chapter 7, Remmert, 1997). Also, the product of two boundedly sequentially compact families of which the functions are uniformly bounded in the L^2 norm is also a boundedly sequentially compact family. This property is useful for constructing new boundedly sequentially compact families from existing ones; see Example 8 for an illustration. The following proposition, of which the proof is trivial or already discussed in the above and thus is omitted, summarizes the aforementioned properties of bounded sequential compactness and its connections to sequential compactness.

PROPOSITION 1. *Let $\mathcal{F}$ be the collection of real-valued functions defined on $\mathcal{T}^2$ and endowed with the product topology. Let $\mathcal{S}$ be a subset of $\mathcal{F}$.*

(i) *If $\mathcal{S}$ is sequentially compact, then it is boundedly sequentially compact.*
(ii) *If $\mathcal{S}$ is a boundedly sequentially compact family and $\sup_{f\in\mathcal{S}} \|f\|_{L^2} < \infty$, then $\mathcal{S}$ is sequentially compact.*
(iii) *If $\sup_{f\in\mathcal{S}} \|f\|_\infty < \infty$ and $\sup_{f\in\mathcal{S}} \sup_{x\neq y} |f(x)-f(y)|/\|x-y\|_2 < \infty$, then $\mathcal{S}$ is a boundedly sequentially compact family.*
(iv) *If both $\mathcal{S}$ and $\mathcal{Q} \subset \mathcal{F}$ are sequentially compact, then the family $\mathcal{SQ} = \{fg : f \in \mathcal{S}, g \in \mathcal{Q}\}$ is sequentially compact. Consequently, if $\mathcal{S}$ and $\mathcal{Q}$ are boundedly sequentially compact and satisfy $\sup_{f\in\mathcal{S}\cup\mathcal{Q}} \|f\|_{L^2} < \infty$, then $\mathcal{SQ}$ is also boundedly sequentially compact.*

The following examples utilize Proposition 1 to exhibit boundedly sequentially compact families and illustrate their abundance.

Example 7 (Bounded Sobolev sandwich families). Let $M_1, M_2 > 0$ be fixed, but potentially arbitrarily large constants. Let $\mathcal{S}(M_1, M_2)$ be the subfamily of the $\mathcal{T}_\delta$-identifiable family introduced in Example 2 such that, if $f \in \mathcal{S}(M_1, M_2)$, then $\|f\|_\infty \leqslant M_1$ and the Lipschitz constant of f is bounded by M_2, where the Lipschitz constant of f is defined as $\sup_{x \neq y} |f(x) - f(y)|/\|x - y\|_2$. By Proposition 1(iii), $\mathcal{S}(M_1, M_2)$ is a boundedly sequentially compact family.

Example 8 (Boundedly sequentially compact semiparametric families). Let $\mathcal{H}$ be a family of covariance functions indexed by a parameter θ in a compact space $\Theta \subset \mathbb{R}^d$ for some $d > 0$. If each $f_\theta \in \mathcal{H}$ is Lipschitz continuous and has a Lipschitz constant continuous in θ, i.e., $|f_\theta(s) - f_\theta(y)| \leqslant \ell_\theta \|x - y\|_2$ for $x, y \in \mathbb{R}^2$ and ℓ_θ is continuous in θ on Θ, then $\mathcal{H}$ is a boundedly sequentially compact family. The continuity of ℓ_θ and compactness of $\mathcal{T}$ and Θ imply that $\|f\|_\infty \leqslant M_1$ and $\sup_{x \neq y} |f(x) - f(y)|/\|x - y\|_2 \leqslant M_2$ for some constants $M_1, M_2 > 0$. Then the claim follows from Proposition 1(iii). Similar reasoning shows that the family $\mathcal{G} = \{g \otimes g : \|g\|_\infty < M_3$ and $\|g'\|_\infty < M_4\}$ is a boundedly sequentially compact family of which functions are uniformly bounded in the L^2 norm, where $(g \otimes g)(s, t) = g(s)g(t)$ and $M_3, M_4 > 0$ are constants. According to Proposition 1(iv), the semiparametric family $\{gh : g \in \mathcal{G}, h \in \mathcal{H}\}$ is boundedly sequentially compact.

The construction in the above example can be used to derive boundedly sequentially compact subfamilies of the $\mathcal{T}_\delta$-identifiable families introduced in Examples 1 and 4. In the following we shall assume that the family $\mathcal{C}$ under consideration is boundedly sequentially compact. The above examples suggest that this regularity condition, essentially controlling the complexity of the family, holds for any family of functions that are collectively bounded and Lipschitz continuous, and thus is mild. Formally, we shall assume the following conditions.

Assumption 4. The covariance function γ_0 belongs to a $\mathcal{T}_\delta$-identifiable boundedly sequentially compact family $\mathcal{C}$.

Assumption 5. The random function X satisfies $E\{\|X\|_{L^2}^4\} < \infty$.

Assumption 6. We have that $p^{8\alpha+4}/n \to 0$ and $\lambda/(n^{-1/2}p^{2\alpha-4\beta-3/2}) \to 0$ as $n \to \infty$.

Since a $\mathcal{T}_\delta$-identifiable boundedly sequentially compact family such as $\mathcal{S}(M_1, M_2)$ in Example 7 may contain covariance functions of infinite rank, the theory developed below applies to functional snippets of infinite dimension. To avoid entanglement with the error from mean function estimation, we shall assume that μ_0 is known in the following discussion, noting that the case that μ_0 is unknown can also be covered, but requires a much more involved presentation and tedious technical details, and thus is not pursued here. The following result establishes the convergence rate of the proposed estimator for any class of analytic (α, β)-bases.

THEOREM 2. *If Φ is an analytic (α, β)-basis, under Assumptions 1, 4–6 and $m \geqslant 2$, we have*

$$\|\hat{\gamma} - \gamma_0\|_{L^2}^2 = O_P\left(\frac{p^{4\alpha+2}}{n} + \kappa_p^2\right), \tag{6}$$

where κ_p is the convergence rate of $\mathcal{E}(\gamma_0, \Phi, p)$ defined in § 3.1.

With the condition of Assumption 6 on λ, the tuning parameter λ does not affect the asymptotic rate of $\hat{\gamma}$. As in the case of the mean function, the rate in (6) contains two components, the estimation error $p^{4\alpha+2}n^{-1}$ stemming from the finiteness of the sample, and the approximation

bias κ_p attributed to the finiteness of the number of basis functions being used in the estimation. When the Fourier basis or Legendre basis is used, we have the following convergence rate for r times differentiable covariance functions.

COROLLARY 2. *Suppose that Assumptions 1 and 4–6 hold, $m \geqslant 2$, and γ_0 belongs to the Sobolev space $\mathscr{H}^r(\mathcal{T}^2)$ for some $r \geqslant 1$.*

(i) *If Φ is the Fourier basis and γ_0 is periodic, then with $p \asymp n^{1/(2r+2)}$, one has $\|\hat{\gamma} - \gamma_0\|_{L^2}^2 = O_P\big(n^{-r/(r+1)}\big)$.*

(ii) *If Φ is the Legendre basis, then with $p \asymp n^{\min\{1/(2r+4),1/8\}}$, one has $\|\hat{\gamma} - \gamma_0\|_{L^2}^2 = O_P\big(n^{-\min\{r/(r+2),r/4\}}\big)$.*

When $r = 2$ and γ_0 is periodic, the convergence rate for $\hat{\gamma}$ is $n^{-1/3}$ and $n^{-1/4}$, respectively, for the estimator based on the Fourier and Legendre bases. The reason behind this observation is similar to the case of the mean function: high-order Legendre basis functions tend to have large extrema that amplify variability, which limits the number of basis functions for estimation and leads to a relatively slower rate. For a nonperiodic γ_0, the Fourier extension technique can be used to alleviate the Gibbs phenomenon suffered by the Fourier basis; see the Supplementary Material for more details.

Corollaries 1 and 2 show that the proposed analytic basis expansion approach automatically adapts to the smoothness of μ_0 and γ_0. In particular, when μ_0 or γ_0 is smooth, i.e., has infinite order of differentiability, our estimators enjoy a near-parametric rate. This contrasts with the local polynomial smoothing method and the B-spline basis approach, for which the convergence rate is limited by the order of the polynomials or B-spline basis functions used in the estimation, even when μ_0 or γ_0 might have a higher order of smoothness. In practice, it is not easy to determine the right order for these methods, since the mean and covariance functions and their smoothness are unknown.

Remark 1. Delaigle et al. (2020) adopted a two-stage procedure to estimate the covariance function, where in the first stage a pilot estimate is constructed only in the diagonal region $\mathcal{T}_\delta$. This pilot estimate can be obtained by a smoother, for example, Yao et al. (2005a). At the second stage, this pilot estimate is numerically extrapolated by basis expansion without penalization to arrive at an estimate of the entire covariance function. The advantage of this two-stage approach is that the convergence rate of the final estimator is immediately available and inherited from the convergence rate of the pilot estimator, since the basis approximation error is negligible when a sufficiently large number of basis functions are used. The drawback is that the convergence rate (Theorem 2, Delaigle et al., 2020) is then limited by the pilot estimate. For instance, if a local linear smoother is adopted to produce the pilot estimate, then the convergence rate of the final estimator is the same as the local linear smoother, even though the true covariance function might have a higher-order degree of smoothness. In contrast, our approach estimates the basis coefficients directly from the data, and thus is able to automatically exploit the high-order smoothness of the true covariance. This also allows us to establish an explicit convergence rate without reference to a pilot estimate. The convergence theories in Delaigle et al. (2020) and our paper are based on incomparable sets of conditions, and thus do not imply each other.

One of the theoretical novelties in Delaigle et al. (2020) is the approximation error bound in their Theorem 2 to quantify the approximation errors when the covariance function is not identifiable. This is a very appealing feature and upon the suggestion of a referee we address the case that γ_0 is not in the $\mathcal{T}_\delta$-identifiable boundedly sequentially compact family $\mathcal{C}$, but can be well

approximated by a member of the family. Specifically, let $\tilde{\gamma} \in \mathcal{C}$ and assume $\|\gamma_0 - \tilde{\gamma}\|_{L^2} \leqslant \eta$ for some constant $\eta \geqslant 0$. Below we show that, when the model of γ_0 is misspecified, the estimation quality also depends on the degree of misspecification that is quantified by η.

THEOREM 3. *Suppose that Φ is an analytic (α, β)-basis, $m \geqslant 2$, and Assumptions 1 and 4–6 hold. If there exists $\tilde{\gamma} \in \mathcal{C}$ such that $\|\tilde{\gamma} - \gamma_0\|_{L^2} \leqslant \eta$, then we have*

$$\|\hat{\gamma} - \gamma_0\|_{L^2}^2 = O_P\left(\frac{p^{4\alpha+2}}{n} + \kappa_p^2 + \eta^2\right),$$

where κ_p is the convergence rate of $\mathcal{E}(\gamma_0, \Phi, p)$ defined in § 3.1.

COROLLARY 3. *Suppose that the conditions of Theorem 3 hold, and γ_0 belongs to the Sobolev space $\mathscr{H}^r(\mathcal{T}^2)$ for some $r \geqslant 1$.*

(i) *If Φ is the Fourier basis and γ_0 is periodic, then with $p \asymp n^{1/(2r+2)}$, one has $\|\hat{\gamma} - \gamma_0\|_{L^2}^2 = O_P\big(n^{-r/(r+1)} + \eta^2\big)$.*

(ii) *If Φ is the Legendre basis, then with $p \asymp n^{\min\{1/(2r+4), 1/8\}}$, one has $\|\hat{\gamma} - \gamma_0\|_{L^2}^2 = O_P\big(n^{-\min\{r/(r+2), r/4\}} + \eta^2\big)$.*

Remark 2. Although for simplicity we focus on functional snippets where each observed trajectory consists of only a single snippet of equal width, without any modification, the proposed estimation procedure in § 2 is applicable to more general cases, including the case that snippets have random and different widths of span and/or the case that each trajectory is composed of multiple pieces of snippets. The theory developed in this section can also accommodate such cases with a slight modification of Assumption 1, as follows. For random width δ, we require $\mathrm{pr}(\delta > \delta_0) > 0$ for some constant $\delta_0 \in (0, 1)$. To model multiple pieces of snippets per trajectory, one can introduce multiple reference time-points, one for each piece, i.e., for the jth piece within a single trajectory there is a reference time $R^{(j)}$. Without loss of generality, we assume $R^{(1)} < R^{(2)} < \cdots < R^{(S)}$, where the potentially random quantity $S \geqslant 1$ denotes the number of snippets per trajectory. Then, our theories are still valid if Assumption 1 holds for the first and last pieces, i.e., for the pieces indexed by the reference time-points $R^{(1)}$ and $R^{(S)}$.

4. NUMERICAL STUDIES

4.1. *Computational details*

To compute the estimator in (2), one needs to determine a set of basis functions. We recommend the Fourier basis, for it is often computationally stabler than Legendre polynomials and other polynomial bases. To handle nonperiodic data, we incorporate the technique of Fourier extension, which seems less explored in statistics, to overcome the Gibbs phenomenon (Zygmund, 2003); see the Supplementary Material for a brief introduction. It requires selection of an additional tuning parameter, the extension margin ζ. Through extensive numerical experiments, we found that the results are often not sensitive to ζ when it is not too large and too small. As a rule of thumb, we recommend ζ to be one tenth of the span of the study. If computational capacity allows, a data-driven value for ζ can also be selected via cross-validation.

To select the other two tuning parameters q and ρ, we adopt a K-fold cross-validation procedure with $K = 5$, as follows. Let Ξ and Θ be sets of candidate values for q and ρ,

respectively. We choose a pair $(q, \rho) \in \Xi \times \Theta$ that minimizes the validation error $\text{CV}(q, \rho) = \sum_{k=1}^{K} \sum_{i \in \mathcal{P}_k} \sum_{j=1}^{m_i} \{Y_{ij} - \hat{\mu}_{-k}^{q,\rho}(T_{ij})\}^2$, where $\mathcal{P}_1, \ldots, \mathcal{P}_K$ form a roughly even random partition of $\{1, \ldots, n\}$, and $\hat{\mu}_{-k}^{q,\rho}$ is the estimator for μ_0 when there are q basis functions, the penalization parameter is ρ, and only subjects with indices in $\{1, \ldots, n\} \backslash \mathcal{P}_k$ are used in estimation.

To compute the covariance estimator in (4), we choose a boundedly sequentially compact family $\mathcal{C}$, and for our simulation studies we adopt the family exhibited in Example 7 with large values of $M_1 = 100$ and $M_2 = 100$, since it is a large family that allows us to reduce model bias while identifying the covariance function. The matrix C in (4) should be positive definite, which makes the optimization challenging. We tackle the issue of positive definiteness via the geometric Newton method by realizing that the space of symmetric positive definite matrices is a Riemannian manifold when it is endowed with the easy-to-compute log-Cholesky metric (Lin, 2019); see the Supplementary Material for details. In contrast, Delaigle et al. (2020) reparameterized C by its Cholesky factor B, which is a lower triangular matrix satisfying $C = BB^{\mathrm{T}}$, and then turned it into a nonconstrained optimization problem. This approach suffers from numerical instability, as the Cholesky decomposition $C = BB^{\mathrm{T}}$ is not unique.

For the other two parameters p and λ, we adopt the following selection procedure. Let Ξ and Θ be sets of candidate values for p and λ, respectively. We choose a pair $(p, \lambda) \in \Xi \times \Theta$ that minimizes the error

$$\text{ERR}(p, \lambda) = \sum_{1 \leqslant j, l \leqslant G, |t_j - t_l| \leqslant \hat{\delta}} \{\check{\gamma}(t_j, t_l) - \hat{\gamma}^{p,\lambda}(t_j, t_l)\}^2, \tag{7}$$

where $t_1, \ldots, t_G$ are G equally spaced points on the domain $\mathcal{T}$ for some $G > 0$, $\hat{\delta} = \max\{|T_{ij} - T_{il}| : 1 \leqslant i \leqslant n, 1 \leqslant j, l \leqslant m_i\}$ is an estimate for δ, $\check{\gamma}$ is a pilot estimate for γ_0 on the diagonal region $\mathcal{T}_{\hat{\delta}}$ and can be computed by a smoother (Yao et al., 2005a; Chen et al., 2020), and $\hat{\gamma}^{p,\lambda}$ is the estimator with p basis functions and the penalization parameter λ. Unlike the selection of q and ρ for estimating the mean function, we use (7) instead of the cross-validation error that would be computed from the raw observations $\Gamma_{ijl} = \{Y_{ij} - \hat{\mu}(T_{ij})\}\{Y_{il} - \hat{\mu}(T_{il})\}$. This is because the raw observations are too noisy and often result in substantial variability in the cross-validation procedure. In contrast, by utilizing a smoothed pilot estimate $\check{\gamma}$, we not only denoise the raw observations, but also better leverage the information available in the diagonal region through the equally spaced grid $(t_1, \ldots, t_G)$. In Delaigle et al. (2020), the pilot estimator $\check{\gamma}$ is used to directly estimate the basis coefficients, instead of selecting tuning parameters. The simulation studies in the next subsection demonstrate that our strategy is preferable in most cases.

4.2. *Monte Carlo simulations*

We now illustrate the numerical performance of the proposed approach using the Fourier basis. For the mean function we consider two scenarios, $\mu_1(t) = \sum_{k=1}^{9} (-1)^k 1.2^{-k} \phi_k(t)$ and $\mu_2(t) = 2t$, where ϕ_k is the Fourier basis function defined in Example 5. The former is a periodic function while the latter is nonperiodic. For the covariance function, we consider the following cases:

I. the periodic covariance function $\gamma_1(s, t) = \sum_{1 \leqslant j, k \leqslant l} c_{jk} \phi_j(s) \phi_k(t)$ for $l = 5$, where $c_{jk} = 2^{-|j-k|-5/2}$ if $j \neq k$ and 1.5^{1-j} if $j = k$;

II. the nonperiodic and nonsmooth covariance function $\gamma_2(s, t) = \sqrt{v(s)v(t)}/2\mathrm{e}^{-|s-t|^2}$ with the correlation function $\mathrm{e}^{-|s-t|^2}$ and the variance function $v(t) = \{1 + \int_0^t (1 + \lfloor 4.5x \rfloor)\, \mathrm{d}x\}/\sqrt{2}$, where $\lfloor 4.5x \rfloor$ denotes the integer part of $4.5x$;

Table 1. MISE *of the proposed estimator for the mean function*

		$n = 50$		$n = 150$		$n = 450$	
Setting	Method	$\delta = 0.25$	$\delta = 0.75$	$\delta = 0.25$	$\delta = 0.75$	$\delta = 0.25$	$\delta = 0.75$
μ_1	FE	3.27(3.19)	1.94(1.27)	1.10(0.59)	0.68(0.44)	0.35(0.20)	0.25(0.16)
	NFE	3.18(3.09)	1.89(1.26)	1.05(0.57)	0.65(0.41)	0.34(0.19)	0.23(0.16)
μ_2	FE	2.74(3.20)	1.53(0.95)	0.99(0.73)	0.66(0.50)	0.42(0.28)	0.27(0.20)
	NFE	3.20(3.03)	2.00(0.75)	1.61(0.60)	1.20(0.37)	0.89(0.31)	0.77(0.21)

The MISE and their Monte Carlo standard errors in this table are scaled by 10 for a clean presentation. FE, the proposed estimator with Fourier extension; NFE, the method without the Fourier extension.

III. the periodic covariance function γ_3 that is the same as γ_1 except that $l = 30$;
IV. the one-rank covariance function $\gamma_4(s,t) = 0.4\varphi(s)\varphi(t)$ with $\varphi(t) = 0.3f_{0.3,0.05}(t) + 0.7f_{0.7,0.05}(t)$, where $f_{a,b}$ is the probability density of the normal distribution with mean a and standard deviation b.

The covariance functions in Cases I, III and IV fall into the family $\mathcal{C}$ we chose in § 4.1, while the one in Case II does not, since the function γ_2 is nonsmooth and falls outside the chosen family $\mathcal{C}$. The covariance functions in the last two cases, γ_3 and γ_4, although having different ranks, require a large number of Fourier basis functions for a good approximation. All of the last three covariance functions represent challenging cases for our approach.

In evaluating the performance of the mean estimate, the covariance function is fixed to be γ_1, while the mean function is fixed to be μ_1 when evaluating the covariance function. This strategy avoids the bias from covariance influencing the estimation of the mean function, and vice versa.

The estimation quality is measured by the empirical mean integrated squared error based on $N = 100$ independent simulation replicates. For the mean estimator $\hat{\mu}$, the mean integrated squared error is defined by $\text{MISE} = \frac{1}{N}\sum_{k=1}^{N}\int\{\hat{\mu}_k(t)-\mu_0(t)\}^2\,\mathrm{d}t$, and for the covariance estimator $\hat{\gamma}$ it is defined by $\text{MISE} = \frac{1}{N}\sum_{k=1}^{N}\iint\{\hat{\gamma}_k(s,t)-\gamma_0(s,t)\}^2\,\mathrm{d}s\,\mathrm{d}t$, where $\hat{\mu}_k$ and $\hat{\gamma}_k$ are estimators in the kth simulation replicate. The tuning parameters q, p, ρ, λ and the extension margin are selected by the procedures described in § 4.1.

In all replicates, the reference times R_i are sampled from a uniform distribution on $[\delta/2, 1-\delta/2]$. The numbers of observations m_i are independently sampled from the distribution $2 + \text{Po}(3)$. The measurement noise variables ε_{ij} are independently sampled from a centred Gaussian distribution with variance σ^2, where the noise level σ^2 is set to make the signal-to-noise ratio $E\{\|X - \mu_0\|_{L^2}^2\}/\sigma^2 = 4$. We consider three sample sizes, $n = 50, 150, 500$, and two different values of δ, $\delta = 0.25, 0.75$, representing short snippets and long snippets, respectively.

The results are summarized in Tables 1 and 2 for the mean and covariance functions, respectively. As expected, in all settings, the performance of the estimators improves as n or δ increases. Also, if the function to be estimated is periodic, like μ_1 and γ_1, Fourier extension leads to slightly reduced estimation quality. However, if the function is nonperiodic, like μ_2 and γ_2, then the estimators with Fourier extension considerably outperform those without the extension, especially for the mean function or when the sample size is large. This demonstrates that Fourier extension is a rather effective technique that complements the Fourier basis for nonparametric smoothing, and might deserve further investigation in the framework of statistical methodology.

As a comparison, we implement the estimators $\hat{\gamma}_{\text{DP}}$, $\hat{\gamma}_{\text{ZC}}$ and $\hat{\gamma}_{\text{DHHK}}$ of Descary & Panaretos (2019), Zhang & Chen (2020b) and Delaigle et al. (2020), respectively, where the first two are obtained by extrapolating the raw covariance function in the region $\mathcal{T}_\delta$ using matrix completion techniques; see the Supplementary Material for implementation details.

Table 2 summarizes the numerical results, while figures in the Supplementary Material provide a visual comparison of the four methods. Based on these results, we summarize the findings

Table 2. MISE *of the estimators for the covariance function*

Setting	n	δ	FE	NFE	$\hat{\gamma}_{\text{DHHK}}$	$\hat{\gamma}_{\text{DP}}$	$\hat{\gamma}_{\text{ZC}}$
γ_1	50	0.25	6.40(3.90)	6.32(3.85)	10.65(4.10)	9.64(5.12)	8.22(4.66)
		0.75	6.04(3.17)	5.81(3.91)	9.46(3.59)	7.58(4.89)	8.11(3.93)
	150	0.25	3.60(2.20)	3.44(2.05)	8.94(3.13)	8.53(5.17)	7.71(4.21)
		0.75	3.38(2.01)	3.11(2.25)	5.91(2.99)	4.42(3.36)	5.64(3.18)
	450	0.25	2.22(1.21)	2.09(1.12)	7.61(3.06)	7.67(4.25)	6.68(3.01)
		0.75	1.52(1.39)	1.42(1.38)	3.53(2.65)	2.31(1.79)	3.11(2.18)
γ_2	50	0.25	5.40(3.50)	5.51(3.73)	7.88(4.89)	6.35(3.49)	5.83(2.82)
		0.75	4.09(3.03)	4.16(3.13)	4.90(2.93)	4.41(3.10)	4.13(2.75)
	150	0.25	2.65(1.47)	2.74(1.43)	4.32(4.15)	3.73(4.40)	3.07(3.33)
		0.75	2.35(2.02)	2.41(2.59)	2.48(2.18)	2.33(2.78)	2.35(2.70)
	450	0.25	1.56(1.02)	1.68(1.07)	2.47(2.16)	2.00(1.64)	1.58(1.03)
		0.75	1.16(0.93)	1.29(0.94)	1.12(1.17)	0.98(0.91)	1.08(1.00)
γ_3	50	0.25	7.91(5.62)	7.85(5.78)	12.81(7.72)	11.55(6.88)	10.14(4.98)
		0.75	6.85(3.74)	6.53(3.45)	11.51(8.13)	8.61(4.61)	7.94(3.66)
	150	0.25	5.25(2.88)	4.93(2.82)	9.92(5.85)	9.26(5.56)	8.37(3.83)
		0.75	4.81(3.45)	4.42(3.61)	7.61(5.31)	6.14(3.82)	6.09(3.81)
	450	0.25	3.50(2.13)	3.22(1.88)	8.46(4.52)	8.14(4.22)	8.06(4.63)
		0.75	2.18(0.98)	1.79(0.73)	4.52(3.02)	2.88(1.23)	3.35(1.50)
γ_4	50	0.25	10.64(5.27)	10.54(5.12)	13.28(7.81)	10.13(5.77)	12.47(6.85)
		0.75	7.50(2.04)	7.45(2.07)	8.14(2.24)	7.84(2.04)	8.08(2.24)
	150	0.25	8.57(2.15)	8.51(2.19)	8.97(3.72)	8.40(4.88)	9.19(5.88)
		0.75	6.93(1.36)	6.89(1.41)	7.17(1.33)	7.23(1.23)	6.67(1.30)
	450	0.25	7.05(1.20)	6.97(1.16)	8.19(2.92)	7.25(4.53)	7.90(3.82)
		0.75	6.47(0.67)	6.41(0.68)	6.84(0.94)	6.24(0.93)	6.26(0.97)

The MISE and their Monte Carlo standard errors in this table are scaled by 10 for a clean presentation. FE, the proposed estimator with Fourier extension; NFE, the method without the Fourier extension.

below for each case. For Case I the proposed method substantially outperforms the competing procedures. This is not surprising, since the true covariance function γ_1 favours our method. In Case II all methods have similar performance. Our method is slightly better when the sample size is small, while the matrix completion methods have slightly better performance when the sample size is large. In Case III our method has superior performance in all settings. This could be attributed to the fact that the true covariance function is still representable by a finite number of Fourier basis functions, although the number of Fourier basis functions in such a representation is large. Since Case IV is a rather challenging case, all methods have similar, but deteriorating performance, and there is no clear winner. In summary, the performance of our method dominates the estimator $\hat{\gamma}_{\text{DHHK}}$ which also depends on Fourier basis expansion. In addition, we find that the approach of Delaigle et al. (2020) tends to produce excessive variability, especially in the off-diagonal region. The detailed discussion in Remark 1 provides an explanation for these observations. The proposed method outperforms the matrix completion methods when the sample size is small or a small number of basis functions are sufficient to approximate the true covariance function. In other scenarios, like Cases II and IV with a large sample size, the performance of our estimator is nearly as good as the matrix completion methods.

5. Applications

We present an application to spinal bone mineral density in this section. A second application to systolic blood pressure is presented in the Supplementary Material.

In the study of Bachrach et al. (1999), 423 individuals with ages ranging from 8 to 27 were examined for their longitudinal spinal bone mineral density. The bone density of each individual

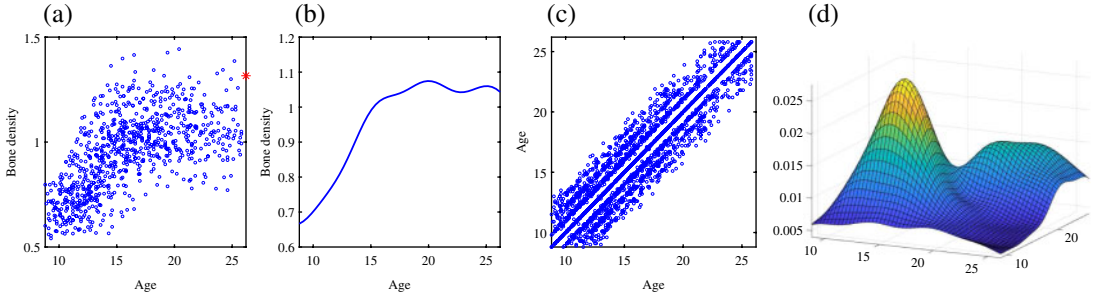

Fig. 1. (a) Spinal bone mineral density data. (b) Estimated mean function. (c) Empirical design plot of the covariance structure of the spinal bone mineral density. (d) Estimated covariance function, using the proposed method with Fourier basis and nonperiodic extension.

was irregularly recorded in four consecutive years, at most once for each year. This resulted in functional snippets with measurements spanning at most four years for all subjects. In our study, individuals who have only one measurement are excluded, since they do not carry information to the covariance structure. This results in a total of 280 individuals who have at least two measurements and whose ages range from 8.8 to 26.2.

We are interested in the mean and covariance structure of the mineral density across the age spectrum. The covariance structure enables us to derive the principal components. Figure 1(c) depicts the empirical design of the covariance function, underscoring the nature of these data as a collection of snippets: there are no data available to directly infer the off-diagonal region of the covariance structure, and the design time-points are irregular. This feature renders techniques based on matrix completion less appropriate since they require a regular design for the measurement time-points. In contrast, our method is able to accommodate this irregularity.

The mineral density data and the estimated mean function are displayed in Figs. 1(a) and 1(b), respectively; the marked rightmost point in Fig. 1(a) is removed to avoid the boundary effect. We observe that the mean density starts with a low level, rises rapidly before age 16, then increases relatively slowly to a peak at age 20, and finally levels off after age 20. This indicates that the spinal bone mineral accumulates fast during adolescence, during which rapid physical growth and psychological changes occur, and then remains at a stable high level in the early 20s.

The estimated covariance surface is shown in Fig. 1(d), which suggests larger variability of the data around the age of 17. It also indicates that the covariance of the longitudinal mineral density at different ages decays drastically as ages become more distant. The first three principal components based on the estimated covariance function in the entire domain are shown in Fig. 2. These principal components account for 69.6%, 22.1% and 3.6% of the variance of the data, respectively, and altogether they account for over 95% of the total variation of the data. The first principal component, which explains nearly 70% of the variation, shows that the highest variation in the bone density trajectories corresponds to overall growth that is consistently either above or below the mean curve, with the difference from the mean most prominent around age 15. Those with positive first principal component scores have bone densities consistently above the mean function, with a surge at age 15, and vice versa for those with negative scores. The second principal component reflects the contrast of growth before and after age 17.5. Those with positive second scores have above average bone densities up to age 17.5, but then drop below the average afterwards. The third principal component reflects the random fluctuation of the bone growth.

ACKNOWLEDGEMENT

We thank the editor, associate editor and three reviewers for their detailed and constructive comments that helped us to substantially improve the paper. Lin's research was supported by the

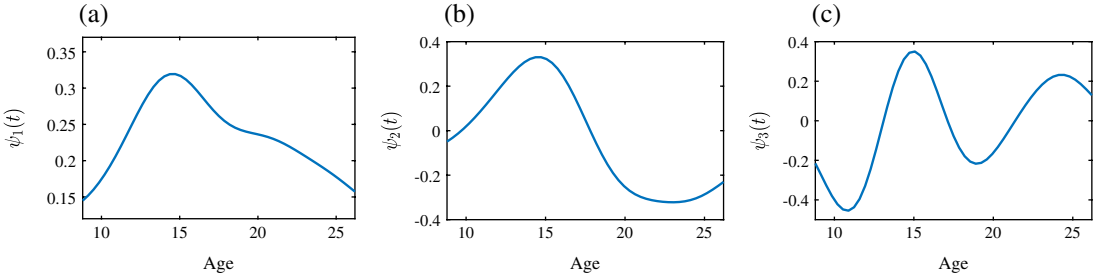

Fig. 2. The first three principal components of the spinal bone density data.

National Institutes of Health (5UG3OD023313-03 and 4UH3-OD023313-04), and the National University of Singapore (R-155-001-217-133). Wang was supported by the National Science Foundation (15-12975 and 19-14917) and National Institutes of Health (4UH3-OD023313-04). Zhong was supported by the National Science Foundation of China (NSFC11771241 and NSFC11931001) and Chinese Government Scholarship (CSC201806210163). Zhong also thanks Professor Ying Yang at Tsinghua University for valuable advice and kind support.

Supplementary material

Supplementary Material available at *Biometrika* online contains additional implementation details, the second data application and technical proofs of the theorems in the article.

References

Aneiros, G., Cao, R., Fraiman, R., Genest, C. & Vieu, P. (2019). Recent advances in functional data analysis and high-dimensional statistics. *J. Multivar. Anal.* **170**, 3–9.

Bachrach, L. K., Hastie, T., Wang, M.-C., Narasimhan, B. & Marcus, R. (1999). Bone mineral acquisition in healthy asian, hispanic, black, and caucasian youth: A longitudinal study. *J. Clin. Endocrin. Metab.* **84**, 4702–12.

Cai, T. & Yuan, M. (2010). Nonparametric covariance function estimation for functional and longitudinal data. Tech. rep., University of Pennsylvania.

Cai, T. & Yuan, M. (2011). Optimal estimation of the mean function based on discretely sampled functional data: Phase transition. *Ann. Statist.* **39**, 2330–55.

Canuto, C., Hussaini, M. Y., Quarteroni, A. & Zang, T. A. (2006). *Spectral Methods: Fundamentals in Single Domains*. Berlin: Springer.

Cardot, H. (2000). Nonparametric estimation of smoothed principal components analysis of sampled noisy functions. *J. Nonparam. Statist.* **12**, 503–38.

Chen, Y., Carroll, C., Dai, X., Fan, J., Hadjipantelis, P. Z., Han, K., Ji, H., Müller, H.-G. & Wang, J.-L. (2020). *fdapace: Functional Data Analysis and Empirical Dynamics*. R package version 0.5.2, available at https://CRAN.R-project.org/package=fdapace.

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

Dawson, M. & Müller, H.-G. (2018). Dynamic modeling of conditional quantile trajectories, with application to longitudinal snippet data. *J. Am. Statist. Assoc.* **113**, 1612–24.

Delaigle, A. & Hall, P. (2013). Classification using censored functional data. *J. Am. Statist. Assoc.* **108**, 1269–83.

Delaigle, A. & Hall, P. (2016). Approximating fragmented functional data by segments of Markov chains. *Biometrika* **103**, 779–99.

Delaigle, A., Hall, P., Huang, W. & Kneip, A. (2020). Estimating the covariance of fragmented and other related types of functional data. *J. Am. Statist. Assoc.* to appear, DOI: 10.1080/01621459.2020.1723597.

Descary, M.-H. & Panaretos, V. M. (2018). Recovering covariance from functional fragments. *arxiv:*1708.02491v3.

Descary, M.-H. & Panaretos, V. M. (2019). Recovering covariance from functional fragments. *Biometrika* **106**, 145–60.

Ferraty, F. & Vieu, P. (2006). *Nonparametric Functional Data Analysis: Theory and Practice*. New York: Springer.

Gellar, J. E., Colantuoni, E., Needham, D. M. & Crainiceanu, C. M. (2014). Variable-domain functional regression for modeling ICU data. *J. Am. Statist. Assoc.* **109**, 1425–39.

Goldberg, Y., Ritov, Y. & Mandelbaum, A. (2014). Predicting the continuation of a function with applications to call center data. *J. Statist. Plan. Infer.* **147**, 53–65.

GROMENKO, O., KOKOSZKA, P. & SOJKA, J. (2017). Evaluation of the cooling trend in the ionosphere using functional regression with incomplete curves. *Ann. Appl. Statist.* **11**, 898–918.
HALL, P. & HOROWITZ, J. L. (2007). Methodology and convergence rates for functional linear regression. *Ann. Statist.* **35**, 70–91.
HALL, P. & HOSSEINI-NASAB, M. (2009). Theory for high-order bounds in functional principal components analysis. *Math. Proc. Camb. Phil. Soc.* **146**, 225–56.
HORVÁTH, L. & KOKOSZKA, P. (2012). *Inference for Functional Data with Applications*. New York: Springer.
HSING, T. & EUBANK, R. (2015). *Theoretical Foundations of Functional Data Analysis, with an Introduction to Linear Operators*. Chichester: Wiley.
JAMES, G. M., HASTIE, T. J. & SUGAR, C. A. (2000). Principal component models for sparse functional data. *Biometrika* **87**, 587–602.
KNEIP, A. & LIEBL, D. (2020). On the optimal reconstruction of partially observed functional data. *Ann. Statist.* **48**, 1692–717.
KOKOSZKA, P. & REIMHERR, M. (2017). *Introduction to Functional Data Analysis*. London: Chapman and Hall/CRC.
KONG, D., XUE, K., YAO, F. & ZHANG, H. H. (2016). Partially functional linear regression in high dimensions. *Biometrika* **103**, 147–59.
KRANTZ, S. G. & PARKS, H. R. (2002). *A Primer of Real Analytic Functions*. New York: Springer.
KRAUS, D. (2015). Components and completion of partially observed functional data. *J. R. Statist. Soc.* B **77**, 777–801.
KRAUS, D. & STEFANUCCI, M. (2019). Classification of functional fragments by regularized linear classifiers with domain selection. *Biometrika* **106**, 161–80.
LI, Y. & HSING, T. (2010). Uniform convergence rates for nonparametric regression and principal component analysis in functional/longitudinal data. *Ann. Statist.* **38**, 3321–51.
LIEBL, D. (2013). Modeling and forecasting electricity spot prices: A functional data perspective. *Ann. Appl. Statist.* **7**, 1562–92.
LIEBL, D. & RAMESEDER, S. (2019). Partially observed functional data: The case of systematically missing parts. *Comp. Statist. Data Anal.* **131**, 104–15.
LIN, Z. (2019). Riemannian geometry of symmetric positive definite matrices via Cholesky decomposition. *SIAM J. Matrix Anal. Appl.* **40**, 1353–70.
LIN, Z. & WANG, J.-L. (2020). Mean and covariance estimation for functional snippets. *J. Am. Statist. Assoc.* to appear, DOI: 10.1080/01621459.2020.1777138.
MAS, A. & RUYMGAART, F. (2015). High-dimensional principal projections. *Complex Anal. Oper. Theory* **9**, 35–63.
MOJIRSHEIBANI, M. & SHAW, C. (2018). Classification with incomplete functional covariates. *Statist. Prob. Lett.* **139**, 40–6.
MÜLLER, H.-G. & STADTMÜLLER, U. (2005). Generalized functional linear models. *Ann. Statist.* **33**, 774–805.
MÜLLER, H.-G. & YAO, F. (2008). Functional additive models. *J. Am. Statist. Assoc.* **103**, 1534–44.
RAMSAY, J. O. & SILVERMAN, B. W. (2005). *Functional Data Analysis*, 2nd ed. New York: Springer.
RAO, C. R. (1958). Some statistical methods for comparison of growth curves. *Biometrics* **14**, 1–17.
REMMERT, R. (1997). *Classical Topics in Complex Function Theory*. New York: Springer.
RICE, J. A. & SILVERMAN, B. W. (1991). Estimating the mean and covariance structure nonparametrically when the data are curves. *J. R. Statist. Soc.* B **53**, 233–43.
RICE, J. A. & WU, C. O. (2001). Nonparametric mixed effects models for unequally sampled noisy curves. *Biometrics* **57**, 253–9.
STEFANUCCI, M., SANGALLI, L. M. & BRUTTI, P. (2018). PCA-based discrimination of partially observed functional data, with an application to aneurisk65 data set. *Statist. Neerlandica* **72**, 246–64.
WAHBA, G. (1990). *Spline Models for Observational Data*. Philadelphia: Society for Industrial and Applied Mathematics.
WANG, J.-L., CHIOU, J.-M. & MÜLLER, H.-G. (2016). Review of functional data analysis. *Annu. Rev. Statist. Appl.* **3**, 257–95.
WOOD, S. N. (2003). Thin plate regression splines. *J. R. Statist. Soc.* B **65**, 95–114.
YAO, F., MÜLLER, H.-G. & WANG, J.-L. (2005a). Functional data analysis for sparse longitudinal data. *J. Am. Statist. Assoc.* **100**, 577–90.
YAO, F., MÜLLER, H.-G. & WANG, J.-L. (2005b). Functional linear regression analysis for longitudinal data. *Ann. Statist.* **33**, 2873–903.
ZHANG, A. & CHEN, K. (2020a). Nonparametric covariance estimation for mixed longitudinal studies, with applications in midlife women's health. *arXiv:*1711.00101v3.
ZHANG, A. & CHEN, K. (2020b). Nonparametric covariance estimation for mixed longitudinal studies, with applications in midlife women's health. *Statistica Sinica* to appear, DOI:10.5705/ss.202019.0219.
ZHANG, X. & WANG, J.-L. (2016). From sparse to dense functional data and beyond. *Ann. Statist.* **44**, 2281–321.
ZHANG, X. & WANG, J.-L. (2018). Optimal weighting schemes for longitudinal and functional data. *Statist. Prob. Lett.* **138**, 165–70.
ZYGMUND, A. (2003). *Trigonometric Series*. Cambridge: Cambridge University Press.

[*Received on* 14 *September* 2019. *Editorial decision on* 10 *September* 2020]