Mean and Covariance Estimation for Functional Snippets

Zhenhua Lin*
Department of Statistics and Applied Probability
National University of Singapore
and
Jane-Ling Wang†
Department of Statistics
University of California at Davis

Abstract

We consider estimation of mean and covariance functions of functional snippets, which are short segments of functions possibly observed irregularly on an individual specific subinterval that is much shorter than the entire study interval. Estimation of the covariance function for functional snippets is challenging since information for the far off-diagonal regions of the covariance structure is completely missing. We address this difficulty by decomposing the covariance function into a variance function component and a correlation function component. The variance function can be effectively estimated nonparametrically, while the correlation part is modeled parametrically, possibly with an increasing number of parameters, to handle the missing information in the far off-diagonal regions. Both theoretical analysis and numerical simulations suggest that this hybrid strategy is effective. In addition, we propose a new estimator for the variance of measurement errors and analyze its asymptotic properties. This estimator is required for the estimation of the variance function from noisy measurements.

*stalz@nus.edu.sg. Research supported by NIH ECHO grant (5UG3OD023313-03).

†janelwang@ucdavis.edu. Research supported by NIH ECHO grant (5UG3OD023313-03) and NSF (15-12975 and 19-14917).

Keywords: Functional data analysis, functional principal component analysis, sparse functional data, variance function, correlation function.

1 Introduction

Functional data are random functions on a common domain, e.g., an interval $\mathcal{T} \subset \mathbb{R}$. In reality they can only be observed on a discrete schedule, possibly intermittently, which leads to an incomplete data problem. Luckily, by now this problem has largely been resolved (Rice and Wu, 2001; Yao et al., 2005; Li and Hsing, 2010; Zhang and Wang, 2016) and there is a large literature on the analysis of functional data. For a more comprehensive treatment readers are referred to the monographs by Ramsay and Silverman (2005), Ferraty and Vieu (2006), Hsing and Eubank (2015) and Kokoszka and Reimherr (2017), and a review paper by Wang et al. (2016).

In this paper, we address a different type of incomplete data, which occurs frequently in longitudinal studies when subjects enter the study at random time and are followed for a short period within the domain $\mathcal{T} = [a, b] \subset \mathbb{R}$. Specifically, we focus on functional data with the following property: each function X_i is only observed on a subject-specific interval $O_i = [A_i, B_i] \subset [a, b]$, and

(S) there exists an absolute constant δ such that $0 < \delta < 1$ and $B_i - A_i \leq \delta(b - a)$ for all $i = 1, 2, \ldots$.

As a result, the design of support points (Yao et al., 2005) where one has information about the covariance function $\mathcal{C}(s, t)$ is incomplete in the sense that there are no design points in the off-diagonal region, $\mathcal{T}_\delta^c = \{(s, t) : | s-t | > \delta(b-a), s, t \in [a, b]\}$. This is mathematically characterized by

$$\left(\bigcup_i [A_i, B_i]^2 \right) \cap \mathcal{T}_\delta^c = \emptyset. \tag{1}$$

Consequently, local smoothing methods, such as PACE (Yao et al., 2005), that are interpolation methods fail to produce a consistent estimate of the covariance function in the off-diagonal region as the problem requires data extrapolation.

An example is the spinal bone mineral density data collected from 423 subjects ranging in age from 8.8 to 26.2 years (Bachrach et al., 1999). The design plot for the covariance function, as shown in Figure 1, indicates that all of the design points fall within a narrow band around the diagonal area but the domain of interest $[8.8, 26.2]$ is much larger than this band. The cause of this phenomenon is that each individual trajectory is only recorded in an individual specific subinterval that is much shorter than the span of the study. For the spinal bone mineral density data, the span (length of interval between the first measurement and the last one) for each individual is no larger than 4.3 years, while the span for the study is about 17 years. Data with this characteristic, mathematically described by (S) or (1), are called *functional snippets* in this paper, analogous to the longitudinal snippets studied in Dawson and Müller (2018). As it turns out, functional snippets are quite common in longitudinal studies (Raudenbush and Chan, 1992; Galbraith et al., 2017) and require extrapolation methods to handle. Usually, this is not an issue for parametric approaches, such as linear mixed-effects models, but requires a thoughtful plan for non- and semi-parametric approaches.

Functional fragments (Liebl, 2013; Kraus, 2015; Kraus and Stefanucci, 2019; Kneip and Liebl, 2019+; Liebl and Rameseder, 2019), like functional snippets, are also partially observed functional data and have been studied broadly in the literature. However, for data investigated in these works as functional fragments, the span of a single individual domain $[A_i, B_i]$ can be nearly as large as the span $[a, b]$ of the study, making them distinctively different from functional snippets. Such data, collectively referred to as "nonsnippet functional data" in this paper, often satisfy the following condition:

(F) for any $\epsilon \in (0, 1)$, $\lim_n \Pr\{B_{i_n} - A_{i_n} > (1 - \epsilon)(b - a)\} > 0$ for a strictly increasing sequence $\{i_n\}_{n=1}^{\infty}$.

For instance, Kneip and Liebl (2019+) assumed that $\Pr([A_i, B_i]^2 = [a, b]^2) > 0$, which

implies that design points and local information are still available in the off-diagonal region $\mathcal{T}_\delta^c$. In other words, for non-snippet functional data and for each $(s,t) \in [a,b]^2$, one has $\Pr\{(s,t) \in \bigcup_{i=1}^n [A_i, B_i]^2\} > 0$ for sufficiently large n, contrasting with (1) for functional snippets. Other related works by Gellar et al. (2014); Goldberg et al. (2014); Gromenko et al. (2017); Stefanucci et al. (2018) on partially observed functional data, although do not explicitly discuss the design, require condition (F) for their proposed methodologies and theory. All of them can be handled with a proper interpolation method, which is fundamentally different from the extrapolation methods needed for functional snippets.

The analysis of functional snippets is more challenging than non-snippet functional data, since information in the far off-diagonal regions of the covariance structure is completely missing for functional snippets according to (1). Delaigle and Hall (2016) addressed this challenge by assuming that the underlying functional data are Markov processes, which is only valid at the discrete level, as pointed out by Descary and Panaretos (2019). Zhang and Chen (2018) and Descary and Panaretos (2019) used matrix completion methods to handle functional snippets, but their approaches require modifications to handle longitudinally recorded snippets that are sampled at random design points, and their theory does not cover random designs. Delaigle et al. (2019) proposed to numerically extrapolate an estimate, such as PACE (Yao et al., 2005), from the diagonal region to the entire domain via basis expansion. In this paper, we propose a divide-and-conquer strategy to analyze (longitudinal) functional snippets with a focus on the mean and covariance estimation. Once the covariance function has been estimated, functional principal component analysis can be performed through the spectral decomposition of the covariance operator.

Specifically, we divide the covariance function into two components, the variance function and the correlation function. The former can be estimated via classic kernel smoothing, while the latter is modeled parametrically with a potentially diverging number of parame-

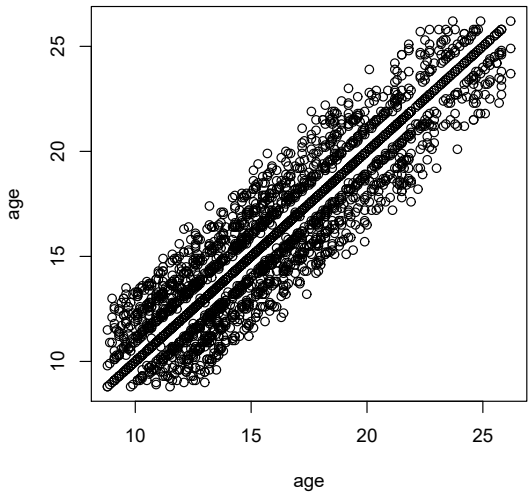

Figure 1: The design of covariance function from spinal bone mineral density data.

ters. The principle behind this idea is to nonparametrically estimate the unknown components for which sufficient information is available while parameterizing the component with missing pieces. Since the correlation structure is usually much more structured than the covariance surface and it is possible to estimate the correlation structure nonparametrically within the diagonal band, a parametric correlation model can be selected from candidate models in existing literature and this usually works quite well to fit the unknown correlation structure.

Compared to the aforementioned works, our proposal enjoys at least two advantages. First, it can be applied to all types of designs, either sparsely/densely or regularly/irregularly observed snippets. Second, our approach is simple thanks to the parametric structure of the correlation structure, and yet powerful due to the potential to accommodate growing dimension of parameters and nonparametric variance component. We stress that, our semi-

parametric and divide-and-conquer strategy is fundamentally different from the penalized basis expansion approach that is adopted in the recent paper by Lin et al. (2019) where the covariance function is represented by an analytic basis and the basis coefficients are estimated via penalized least squares. Numerical comparison of these two methods is provided in Section 5.

This divide-and-conquer approach has been explored in Fan et al. (2007) and Fan and Wu (2008) to model the covariance structure of time-varying random noise in a varying-coefficient partially linear model. We demonstrate here that a similar strategy can overcome the challenge of the missing data issue in functional snippets and further allow the dimension of the correlation function to grow to infinity. In addition, we take into account the measurement error in the observed data, which is an important component in functional data analysis but is of less interest in a partially linear model and thus not considered in Fan et al. (2007) and Fan and Wu (2008). The presence of measurement errors complicates the estimation of the variance function, as they are entangled together along the diagonal direction of the covariance surface. Consequently, the estimation procedure for the variance function in Fan et al. (2007) and Fan and Wu (2008) does not apply. While it is possible to estimate the error variance using the approach in Yao et al. (2005) and Liu and Müller (2009), these methods require a pilot estimate of the covariance function in the diagonal area, which involves two-dimensional smoothing, and thus are not efficient. A key contribution of this paper is a new estimator for the error variance in Section 3 that is simple and easy to compute. It improves upon the estimators in Yao et al. (2005) and Liu and Müller (2009), as demonstrated through theoretical analysis and numerical studies; see Section 4 and 5 for details.

2 Mean and Covariance Function Estimation

Let X be a second-order random process defined on an interval $\mathcal{T} \subset \mathbb{R}$ with mean function $\mu(t) = \mathbb{E}X(t)$, and covariance function $\mathcal{C}(s,t) = \text{cov}(X(s), X(t))$. Without loss of generality, we assume $\mathcal{T} = [0,1]$ in the sequel.

Suppose $\{X_1, \ldots, X_n\}$ is an independent random sample of X, where n is the sample size. In practice, functional data are rarely fully observed. Instead, they are often noisily recorded at some discrete points on $\mathcal{T}$. To accommodate this practice, we assume that each X_i is only measured at m_i points $T_{i1}, \ldots, T_{im_i}$, and the observed data are $Y_{ij} = X_i(T_{ij}) + \varepsilon_{ij}$ for $j = 1, \ldots, m_i$, where ε_{ij} represents the homoscedastic random noise such that $\mathbb{E}\varepsilon_{ij} = 0$ and $\mathbb{E}\varepsilon_{ij}^2 = \sigma_0^2$. This homoscedasticity assumption can be relaxed to accommodate heteroscedastic noise; see Section 3 for details. To further elaborate the functional snippets characterized by (S), we assume that the ith subject is only available to be studied between time $O_i - \delta/2$ and $O_i + \delta/2$, where the variable $O_i \in [\delta/2, 1-\delta/2]$, called reference time in this paper, is specific to each subject and is modeled as identically and independently distributed (i.i.d.) random variables. We then assume that, $T_{i1}, \ldots, T_{im_i}$ are i.i.d., conditional on O_i. These assumptions reflect the reality of many data collection processes when subjects enter a study at random time $O_i - \delta/2$ and are followed for a fixed period of time. Such a sampling plan, termed accelerated longitudinal design, has the advantage to expand the time range of interest in a short period of time as compared to a single cohort longitudinal design study.

2.1 Mean Function

Even though only functional snippets are observed rather than a full curve, smoothing approaches such as Yao et al. (2005) can be applied to estimate the mean function μ,

since for each t, there is positive probability that some design points fall into a small neighborhood of t. Here, we adopt a ridged version of the local linear smoothing method in Zhang and Wang (2016), as follows.

Let K be a kernel function and h_μ a bandwidth, and define $K_{h_\mu}(u) = h_\mu^{-1}K(u/h_\mu)$. The non-ridged local linear estimate of μ is given by $\tilde{\mu}(t) = \hat{b}_0$ with

$$(\hat{b}_0, \hat{b}_1) = \underset{(b_0,b_1)\in\mathbb{R}^2}{\arg\min} \sum_{i=1}^{n} w_i \sum_{j=1}^{m_i} K_{h_\mu}(T_{ij} - t)\{Y_{ij} - b_0 - b_1(T_{ij} - t)\}^2,$$

where $w_i \geq 0$ are weight such that $\sum_{i=1}^n m_i w_i = 1$. For the optimal choice of weight, readers are referred to Zhang and Wang (2018). It can be shown that $\tilde{\mu}(t) = (R_0 S_2 - R_1 S_1)/(S_0 S_2 - S_1^2)$, where

$$S_r = \sum_{i=1}^{n} w_i \sum_{j=1}^{m_i} K_{h_\mu}(T_{ij} - t)\{(T_{ij} - t)/h_\mu\}^r,$$
$$R_r = \sum_{i=1}^{n} w_i \sum_{j=1}^{m_i} K_{h_\mu}(T_{ij} - t)\{(T_{ij} - t)/h_\mu\}^r Y_{ij}.$$

Although $\tilde{\mu}$ behaves well most of the time, for a finite sample, there is positive probability that $S_0 S_2 - S_1^2 = 0$, hence $\tilde{\mu}$ may become undefined. This minor issue can be addressed by ridging, a regularization technique used by Fan (1993) with details in Seifert and Gasser (1996) and Hall and Marron (1997). The basic idea is to add a small positive constant to the denominator of $\tilde{\mu}$ when $S_0 S_2 - S_1^2$ falls below a threshold. More specifically, the ridged version of $\tilde{\mu}(t)$ is given by

$$\hat{\mu}(t) = \frac{R_0 S_2 - R_1 S_1}{S_0 S_2 - S_1^2 + \Delta 1_{\{|S_0 S_2 - S_1^2| < \Delta\}}}, \tag{2}$$

where Δ is a sufficiently small constant depending on n and $m_1, \ldots, m_n$. A convenient choice here is $\Delta = (nm)^{-2}$, where $m = n^{-1}\sum_{i=1}^n m_i$.

The tuning parameter h_μ could be selected via the following κ-fold cross-validation procedure. Let κ be a positive integer, e.g., $\kappa = 5$, and $\{\mathcal{P}_1, \ldots, \mathcal{P}_\kappa\}$ be a roughly even random partition of the set $\{1, \ldots, n\}$. For a set $\mathcal{H}$ of candidate values for h_μ, we choose one from it such that the following cross-validation error

$$\mathrm{CV}(h) = \sum_{k=1}^{\kappa} \sum_{i \in \mathcal{P}_k} \sum_{j=1}^{m_i} \{Y_{ij} - \hat{\mu}_{h,-k}(T_{ij})\}^2 \tag{3}$$

is minimized, where $\hat{\mu}_{h,-k}$ is the estimator in (2) with $h_\mu = h$ and subjects in $\mathcal{P}_k$ excluded.

2.2 Covariance Function

Estimation of the covariance function $\mathcal{C}(s,t)$ for functional snippets is considerably more challenging. As we have pointed out in Section 1, local information in the far off-diagonal region, $|s-t| > \delta$, is completely missing. To tackle this challenge, we first observe that the covariance function can be decomposed into two parts, a variance function and a correlation structure, i.e., $\mathcal{C}(s,t) = \sigma_X(s)\sigma_X(t)\rho(s,t)$, where $\sigma_X^2(\cdot)$ is the variance function of X, or more precisely, $\sigma_X^2(t) = \mathbb{E}\{X(t) - \mu(t)\}^2$, and $\rho(\cdot,\cdot)$ is the correlation function. Like the mean function μ, the variance function can be well estimated via local linear smoothing even in the case of functional snippets. The real difficulty stems from the estimation of the correlation structure, which we propose to model parametrically. At first glance, a parametric model might be restrictive. However, with a nonparametric variance component and a large number of parameters, the model will often still be sufficiently flexible to capture the covariance structure of the data. Indeed, in our simulation studies that are presented in Section 5, we demonstrate that even with a single parameter, the proposed model often yields good performance when sample size is limited. As an additional flexibility, our parametric model does not require the low-rank assumption and hence is able to model truly infinitely-dimensional functional data. This trade of the low-rank assumption with

the proposed parametric assumption seems worthwhile, especially because we allow the dimension of the parameters to increase with the sample size. The increasing dimension of the parameter essentially puts the methodology in the nonparametric paradigm.

To estimate $\sigma_X^2(\cdot)$, we first note that the PACE method in Yao et al. (2005) can still be used to estimate $\mathcal{C}(s,t)$ on the band $\mathcal{T}_\delta^2 = \{(s,t) \in \mathcal{T} \times \mathcal{T} : |s-t| < \delta\}$ that includes the diagonal, although not on the full domain $\mathcal{T} \times \mathcal{T}$. Since $\sigma_X^2(t) = \mathcal{C}(t,t)$, the PACE estimate $\tilde{\mathcal{C}}$ for $\mathcal{C}$ on the diagonal gives rise to an estimate of $\sigma_X^2(t)$. However, this method requires two-dimensional smoothing, which is cumbersome and computationally less efficient. In addition, it has the convergence rate of a two-dimensional smoother, which is suboptimal for a target $\sigma_X^2(t)$ that is a one-dimensional function. Here we propose a simpler approach that only requires one-dimensional smoothing, based on the observation that the quantity $\varsigma^2(t) \equiv \mathbb{E}\{Y(t) - \mu(t)\}^2 = \sigma_X^2(t) + \sigma_0^2$ can be estimated by local linear smoothing on the observations $\{Y_{ij} - \hat{\mu}(T_{ij})\}^2$. More specifically, the non-ridged local linear estimate of $\varsigma^2(t)$, denoted by $\tilde{\varsigma}^2(t)$, is $\hat{b}_0$ with

$$(\hat{b}_0, \hat{b}_1) = \underset{(b_0,b_1)\in\mathbb{R}^2}{\arg\min} \sum_{i=1}^{n} w_i \sum_{j=1}^{m_i} K_{h_\sigma}(T_{ij} - t)[\{Y_{ij} - \hat{\mu}(T_{ij})\}^2 - b_0 - b_1(T_{ij} - t)]^2,$$

where h_σ is the bandwidth to be selected by a cross-validation procedure similar to (3). As with the ridged estimate of the mean function in (2), to circumvent the positive probability of being undefined for $\tilde{\varsigma}^2$, we adopt the ridged version of $\tilde{\varsigma}^2$ as the estimate for ς^2, denoted by $\hat{\varsigma}^2$. Then our estimate of $\sigma_X^2(t)$ is $\hat{\sigma}_X^2(t) = \hat{\varsigma}^2(t) - \hat{\sigma}_0^2$, where $\hat{\sigma}_0^2$ is a new estimate of σ_0^2, to be defined in the next section, that has a convergence rate of a one-dimensional smoother. Because $\hat{\varsigma}^2(t)$ also has a one-dimensional convergence, the resulting estimate of $\hat{\sigma}_X^2(t)$ has a one-dimensional convergence rate.

For the correlation function ρ, we assume that ρ is indexed by a d_n-dimensional vector of parameters, denoted by $\theta \in \mathbb{R}^{d_n}$. Here, the dimension of parameters is allowed to grow

with the sample size at a certain rate; see Section 4 for details. Some popular parametric families for correlation function are listed below.

1. Power exponential:

$$\rho_\theta(s,t) = \exp\left\{-\frac{|s-t|^{\theta_1}}{\theta_2^{\theta_1}}\right\}, \quad 0 < \theta_1 \leq 2,\ \theta_2 > 0.$$

2. Rational quadratic (Cauchy):

$$\rho_\theta(s,t) = \left\{1 + \frac{|s-t|^2}{\theta_2^2}\right\}^{-\theta_1}, \quad \theta_1, \theta_2 > 0.$$

3. Matérn:

$$\rho_\theta(s,t) = \frac{1}{\Gamma(\theta_1)2^{\theta_1-1}}\left(\sqrt{2\theta_1}\frac{|s-t|}{\theta_2}\right)^{\theta_1} B_{\theta_1}\left(\sqrt{2\theta_1}\frac{|s-t|}{\theta_2}\right), \quad \theta_1, \theta_2 > 0, \tag{4}$$

with $B_\theta(\cdot)$ being the modified Bessel function of the second kind of order θ.

Note that if $\rho_1, \ldots, \rho_p$ are correlation functions, then $\sum_{k=1}^p v_k \rho_k$ is also a correlation function if $\sum_{k=1}^p v_k = 1$ and $v_k \geq 0$ for all k. Therefore, a fairly flexible class of correlation functions can be constructed from several relatively simple classes by this convex combination. We point out here that, even when one adopts a stationary correlation function, the resulting covariance can be non-stationary due to a nonparametric and hence often non-stationary variance component.

Given the estimate $\hat{\sigma}_X^2(t)$, the parameter θ can be effectively estimated using the following least squares criterion, i.e., $\hat{\theta} = \arg\min_\theta \hat{Q}_n(\theta)$ with

$$\hat{Q}_n(\theta) = \sum_{i=1}^n \frac{1}{m_i(m_i-1)} \sum_{1 \leq j \neq l \leq m_i} \{\hat{\sigma}_X(T_{ij})\hat{\sigma}_X(T_{il})\rho_\theta(T_{ij}, T_{il}) - C_{ijl}\}^2,$$

where $C_{ijl} = \{Y_{ij} - \hat{\mu}(T_{ij})\}\{Y_{il} - \hat{\mu}(T_{il})\}$ is the raw covariance of subject i at two different measurement times, T_{ij} and T_{il}.

3 Estimation of Noise Variance

The estimation of σ_0^2 received relatively little attention in the literature. For sparse functional data, the PACE estimator $2|\mathcal{T}|^{-1}\int_{\mathcal{T}}\{\hat{\varsigma}(t)-\hat{\mathcal{C}}(t,t)\mathrm{d}t\}$ proposed in Yao et al. (2005) is a popular option. However, the PACE estimator can be negative in some cases. Liu and Müller (2009) refined this PACE estimator by first fitting the observed data using the PACE estimator and then estimating σ_0^2 by cross-validated residual sum of squares; see appendix A.1 of Liu and Müller (2009) for details. These methods require an estimate of the covariance function, which we do not have here before we obtain an estimate of σ_0^2. Moreover, the estimate $\hat{\mathcal{C}}(t,t)$ in both methods is obtained by two-dimensional local linear smoothing as detailed in Yao et al. (2005), which is computationally costly and leads to a slower (two-dimensional) convergence rate of these estimators. To resolve this conundrum, we propose the following new estimator that does not require estimation of the covariance function or any other parameters such as the mean function.

For a bandwidth $h_0 > 0$, define the quantities

$$A_0 = \mathbb{E}[\{\mathcal{C}(T_1,T_1)+\mu(T_1)\mu(T_1)+\sigma_0^2\}1_{|T_1-T_2|<h_0}],$$
$$A_1 = \mathbb{E}[\{\mathcal{C}(T_1,T_1)+\mu(T_1)\mu(T_1)\}1_{|T_1-T_2|<h_0}],$$

and

$$B = \mathbb{E}1_{|T_1-T_2|<h_0},$$

where T_1 and T_2 denote two design points from the same generic subject. From the above definition, we immediately see that $A_0 = A_1 + B\sigma_0^2$. Also, these quantities seem easy to estimate. For example, A_0 and B can be straightforwardly estimated respectively by

$$\hat{A}_0 = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{m_i(m_i-1)}\sum_{j\neq l}Y_{ij}^2 1_{|T_{ij}-T_{il}|<h_0} \tag{5}$$

and

$$\hat{B} = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{m_i(m_i-1)} \sum_{j\neq l} 1_{|T_{ij}-T_{il}|<h_0}. \tag{6}$$

This motivates us to estimate σ_0^2 via estimation of A_0, A_1 and B.

It remains to estimate A_1, which cannot be estimated using information along the diagonal only, due to the presence of random noise. Instead, we shall explore the smoothness of the covariance function and observe that if T_1 is close to T_2, say $|T_1 - T_2| < h_0$, then $\mathcal{C}(T_1, T_1) \approx \mathcal{C}(T_1, T_2)$ and

$$A_1 \approx A_2 = \mathbb{E}[\{\mathcal{C}(T_1, T_2) + \mu(T_1)\mu(T_2)\}1_{|T_1-T_2|<h_0}].$$

Indeed, we show in Lemma 5 that $A_1 = A_2 + O(h_0^3)$. Therefore, it is sensible to use A_2 as a surrogate of A_1. The former can be effectively estimated by

$$\hat{A}_2 = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{m_i(m_i-1)} \sum_{j\neq l} Y_{ij}Y_{il}1_{|T_{ij}-T_{il}|<h_0}, \tag{7}$$

and we set $\hat{A}_1 = \hat{A}_2$. Finally, the estimate of σ_0^2 is given by

$$\hat{\sigma}_0^2 = (\hat{A}_0 - \hat{A}_1)/\hat{B}. \tag{8}$$

To choose h_0, motivated by the convergence rate stated in Theorem 1 of the next section, we suggest the following empirical rule, $h_0 = 0.29\hat{\delta}\|\hat{\varsigma}\|_2(nm^2)^{-1/5}$, for sparse functional snippets, where $\hat{\delta} = \max_{1\leq i\leq n}\max_{1\leq j,l\leq m_i}|T_{ij} - T_{il}|$ acts as an estimate for δ, $m = n^{-1}\sum_{i=1}^{n} m_i$ represents the average number of measurements per curve, $\hat{\varsigma}^2(t)$ is the estimate of $\varsigma^2(t) = \sigma_X^2(t) + \sigma_0^2$ defined in Section 2, and $\|\hat{\varsigma}\|_2^2 = \int \hat{\varsigma}^2(t)\mathrm{d}t$ represents the overall variability of the data. The coefficient 0.29 is determined by a method described in the appendix. If this rule yields a value of h_0 that makes the neighborhood $\mathcal{N}(h_0) = \{(T_{ij}, T_{il}) : |T_{ij} - T_{il}| < h_0, i = 1, \ldots, n, 1 \leq j \neq l \leq m_i\}$ empty or contain too few

points, then we recommend to choose the minimal value of h_0 such that $\mathcal{N}(h_0)$ contains at least $10^{-1}\sum_{i=1}^{n} m_i(m_i-1)$ points. In this way, we ensure that a substantial fraction of the observed data are used for estimation of the variance σ_0^2. This rule is found to be very effective in practice; see Section 5 for its numerical performance.

Compared to Yao et al. (2005) and Liu and Müller (2009), the proposed estimate (8) is simple and easy to compute. Indeed, it can be computed much faster since it does not require the costly computation of $\hat{\mathcal{C}}$. More importantly, the ingredients $\hat{A}_0$, $\hat{A}_1 = \hat{A}_2$ and $\hat{B}$ for our estimator are obtained by one-dimensional smoothing, with the term $1_{|T_{ij}-T_{il}|<h_0}$ in (5)–(7) acting as a local constant smoother. Consequently, as we show in Section 4, our estimator enjoys an asymptotic convergence rate that is faster than the one from a two-dimensional local linear smoother. In addition, the proposed estimate is always nonnegative, in contrast to the one in Yao et al. (2005). This is seen by the following derivation:

$$\begin{aligned}\hat{A}_1 &= \frac{1}{n}\sum_{i=1}^{n}\frac{1}{m_i(m_i-1)}\sum_{j\neq l} Y_{ij}Y_{il}1_{|T_{ij}-T_{il}|<h_0} \leq \frac{1}{n}\sum_{i=1}^{n}\frac{1}{m_i(m_i-1)}\sum_{j\neq l}\frac{Y_{ij}^2+Y_{il}^2}{2}1_{|T_{ij}-T_{il}|<h_0} \\ &= \frac{1}{n}\sum_{i=1}^{n}\frac{1}{m_i(m_i-1)}\sum_{j\neq l} Y_{ij}^2 1_{|T_{ij}-T_{il}|<h_0} = \hat{A}_0. \end{aligned} \tag{9}$$

Remark: The above discussion assumes that the noise is homoscedastic, i.e., its variance is identical for all t. As an extension, it is possible to modify the above procedure to account for heteroscedastic noise, as follows. With intuition and rationale similar to the homoscedastic case, we define

$$\begin{aligned}\hat{A}_0(t) &= \frac{1}{n}\sum_{i=1}^{n}\frac{1}{m_i(m_i-1)}\sum_{j\neq l} Y_{ij}^2 1_{|T_{ij}-t|<h_0}1_{|T_{il}-t|<h_0}, \\ \hat{A}_1(t) &= \frac{1}{n}\sum_{i=1}^{n}\frac{1}{m_i(m_i-1)}\sum_{j\neq l} Y_{ij}Y_{il} 1_{|T_{ij}-t|<h_0}1_{|T_{il}-t|<h_0},\end{aligned}$$

$$\hat{B}(t) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{m_i(m_i-1)} \sum_{j\neq l} 1_{|T_{ij}-t|<h_0} 1_{|T_{il}-t|<h_0},$$

and let

$$\hat{\sigma}_0^2(t) = \{\hat{A}_0(t) - \hat{A}_1(t)\}/\hat{B}(t)$$

be the estimate of $\sigma_0^2(t)$ which is the variance of the noise at $t \in \mathcal{T}$. Like the derivation in (9), one can also show that this estimator is nonnegative.

4 Theoretical Properties

For clarity of exposition, we assume throughout this section that all the m_i have the same rate m, i.e., $m_i = m$, where the sampling rate m may tend to infinity. We emphasize that parallel asymptotic results can be derived without this assumption by replacing m with $\frac{1}{n}\sum_{i=1}^{n} m_i$. Note that the theory to be presented below applies to both the case that m is bounded by a constant, i.e., $m \leq m_0$ for some $m_0 < \infty$, and the case that m diverges to ∞ as $n \to \infty$.

We assume that the reference time O_i is identically and independently distributed (i.i.d.) sampled from a density f_O, and $T_{i1}, \ldots, T_{im_i}$ are i.i.d., conditional on O_i. The i.i.d. assumptions can be relaxed to accommodate heterogeneous distributions and weak dependence, at the cost of much more complicated analysis and heavy technicalities. As such relaxation does not provide further insight into our problem, we decide not to pursue it in the paper. The following conditions about O_i and other quantities are needed for our theoretical development.

(A1) The density f_O of each O_i satisfies $f_O(u) > 0$ for all $u \in [\delta/2, 1-\delta/2]$, and the conditional density $f_{T|O}$ of T_{ij} given O_i satisfies $f_{T|O}(t|u) = f_0(t-u+\delta/2) > 0$ for

a fixed function f_0 and for all $u \in [\delta/2, 1-\delta/2]$ and $t \in [u-\delta/2, u+\delta/2]$. Also, the derivative $\frac{\mathrm{d}}{\mathrm{d}t} f_0(t)$ is Lipschitz continuous on $[0, \delta]$.

(A2) The second derivatives of μ and $\mathcal{C}$ are continuous and hence bounded on $\mathcal{T}$ and $\mathcal{T} \times \mathcal{T}$, respectively.

(A3) $\mathbb{E}\|X\|^4 < \infty$ and $\mathbb{E}\varepsilon^4 < \infty$.

In the above, the condition (A1) characterizes the design points for functional snippets and can be relaxed, while the regularity conditions (A2) and (A3) are common in the literature, e.g., in Zhang and Wang (2016). According to Scheuerer (2010), (A2) also implies that the sample paths of X are continuously differentiable and hence Lipschitz continuous almost surely. Let L_X be the best Lipschitz constant of X, i.e., $L_X = \inf\{C \in \mathbb{R} : |X(s) - X(t)| \leq C|s-t| \text{ for all } s, t \in \mathcal{T}\}$. We will see shortly that a moment condition on L_X allows us to derive a rather sharp bound for the convergence rate of $\hat{\sigma}_0^2$. For the bandwidth h_0, we require the following condition:

(H1) $h_0 \to 0$ and $nm^2 h_0 \to \infty$.

The following result gives the asymptotic rate of the estimator $\hat{\sigma}_0^2$. The proof is straightforward once we have Lemma 5, which is given in the appendix.

Theorem 1. *Assume the conditions (A1)–(A3) and (H1) hold.*

(a) $(\hat{\sigma}_0^2 - \sigma_0^2)^2 = O_P(h_0^4 + n^{-1} + n^{-1}m^{-2}h_0^{-1})$. *With the optimal choice* $h_0 \asymp (nm^2)^{-1/5}$, $(\hat{\sigma}_0^2 - \sigma_0^2)^2 = O_P(n^{-4/5}m^{-8/5} + n^{-1})$.

(b) *If in addition* $\mathbb{E}L_X^4 < \infty$, *then* $(\hat{\sigma}_0^2 - \sigma_0^2)^2 = O_P(h_0^4 + n^{-1}m^{-1} + n^{-1}m^{-2}h_0^{-1})$. *With the optimal choice* $h_0 \asymp (nm^2)^{-1/5}$, $(\hat{\sigma}_0^2 - \sigma_0^2)^2 = O_P(n^{-4/5}m^{-8/5} + n^{-1}m^{-1})$.

If we define $\hat{\sigma}_0^2 = (\hat{A}_0 - \hat{A}_1)/(\hat{B} + \Delta)$ with $\Delta = (nm)^{-2}h_0$, the ridged version of (8), then in the above theorem, $(\hat{\sigma}_0^2 - \sigma_0^2)^2$ can be replaced with $\mathbb{E}(\hat{\sigma}_0^2 - \sigma_0^2)^2$ and $O_P(\cdot)$ can be replaced with $O(\cdot)$, respectively. For comparison, under conditions stronger than (A1)–(A3), the rate derived in Yao et al. (2005) for the PACE estimator is at best $(\hat{\sigma}_0^2 - \sigma_0^2)^2 = O_P(n^{-1/2})$. This rate was improved by Paul and Peng (2011) to $\mathbb{E}(\hat{\sigma}_0^2 - \sigma_0^2)^2 = O(n^{-1} + n^{-4/5}m^{-4/5} + n^{-2/3}m^{-4/3})$. Our estimator clearly enjoys a faster convergence rate, in addition to its computational efficiency. The rate in part (b) of Theorem 1 has little room for improvement, since when n is finite but $m \to \infty$, the rate is optimal, i.e., $\mathbb{E}(\hat{\sigma}_0^2 - \sigma_0^2)^2 = O(m^{-1})$. When m is finite but $n \to \infty$ in the sparse design, we obtain $\mathbb{E}(\hat{\sigma}_0^2 - \sigma_0^2)^2 = O(n^{-4/5})$, in contrast to the rate $O_P(n^{-2/3})$ for the PACE estimator according to Paul and Peng (2011).

To study the properties of $\hat{\mu}(t)$ and $\hat{\sigma}^2(t)$, we shall assume

(B1) the kernel K is a symmetric and Lipschitz continuous density function supported on $[-1, 1]$.

Also, the bandwidth h_μ and h_σ are assumed to meet the following conditions.

(H2) $h_\mu \to 0$ and $nmh_\mu \to \infty$.

(H3) $h_\sigma \to 0$ and $nmh_\sigma \to \infty$.

The choice of these bandwidths depends on the interplay of the sampling rate m and sample size n. The optimal choice is given in the following condition.

(H4) If $m \lesssim n^{1/4}$, then $h_\mu \asymp h_\sigma \asymp (nm)^{-1/5}$, where the notation $a_n \lesssim b_n$ means $\lim_{n\to\infty} a_n/b_n < \infty$. Otherwise, $\max\{h_\mu, h_\sigma\} \asymp n^{-1/4}$. Also, $h_0 \asymp (nm^2)^{-1/5}$.

The asymptotic convergence rates for $\hat{\mu}$ and $\hat{\sigma}_X^2$ are given in the following theorem, whose proof can be obtained by adapting the proof of Proposition 1 in Lin and Yao (2020+)

and hence is omitted. It shows that both $\hat{\mu}$ and $\hat{\sigma}_X^2$ have the same rate, which is hardly surprising since they are both obtained by a one-dimensional local linear smoothing technique. Note that our results generalize those in Fan et al. (2007) and Fan and Wu (2008) by taking the measurement errors and the order of the sampling rate m into account in the theoretical analysis. In addition, our $\mathcal{L}^2$ convergence rates of these estimators complement the asymptotic normality results in Fan et al. (2007) and Fan and Wu (2008).

Theorem 2. *Suppose the conditions (A1)–(A3) hold.*

(a) *With additional conditions (B1) and (H2),* $\mathbb{E}\|\hat{\mu} - \mu\|^2 = O(h_\mu^4 + n^{-1} + n^{-1}m^{-1}h_\mu^{-1})$. *With the choice of bandwidth* h_μ *in (H4),* $\mathbb{E}\|\hat{\mu} - \mu\|^2 = O\left((nm)^{-4/5} + n^{-1}\right)$.

(b) *With additional conditions (B1) and (H1)–(H3),* $\mathbb{E}\|\hat{\sigma}_X^2 - \sigma_X^2\|^2 = O(h_\sigma^4 + h_\mu^4 + h_0^4 + n^{-1} + n^{-1}m^{-1}h_\sigma^{-1} + n^{-1}m^{-1}h_\mu^{-1} + n^{-1}m^{-2}h_0^{-1})$. *With the choice of bandwidth in (H4),* $\mathbb{E}\|\hat{\sigma}_X^2 - \sigma_X^2\|^2 = O\left((nm)^{-4/5} + n^{-1}\right)$.

To derive the asymptotic properties of $\hat{\mathcal{C}}(s,t) = \hat{\sigma}_X(s)\rho_{\hat{\theta}}(s,t)\hat{\sigma}_X(t)$, we need the convergence rate of $\hat{\theta}$. Define

$$Q(\theta) = \mathbb{E}\{\sigma_X(T_{11})\sigma_X(T_{12})\rho_\theta(T_{11},T_{12}) - [Y_{11} - \mu(T_{11})][Y_{12} - \mu(T_{12})]\}^2,$$

and assume the following conditions.

(B2) $\rho_\theta(s,t)$ is twice continuously differentiable with respect to s and t. Furthermore, the first three derivatives of $\rho_\theta(s,t)$ with respect to θ are uniformly bounded for all θ, s, t, d_n.

(B3) $\lambda_{\min}\left(\frac{\partial^2 Q}{\partial\theta^2}\,|_{\theta=\theta_0}\right) > c_0 d_n^{-\tau}$ for some $c_0 > 0$ and $\tau \geq 0$, where θ_0 denotes the true value of θ, and $\lambda_{\min}(\cdot)$ denotes the smallest eigenvalue of a matrix.

(B4) $\mathbb{E}\sup_t \|X(t)\|^{4+\epsilon_0} < \infty$ for some $\epsilon_0 > 0$ and $\mathbb{E}\varepsilon^4 < \infty$.

Note that in the condition (B3), we allow the smallest eigenvalue of the Hessian $\frac{\partial^2 Q}{\partial \theta^2}$ to decay with d_n. This, departing from the assumption in Fan and Wu (2008) of fixed dimension on the parameter θ, enables us to incorporate the case that ρ_θ is constructed from the aforementioned convex combination of a *diverging* number of correlation functions, e.g., $\rho_\theta(s,t) = d_n^{-1}\sum_{j=1}^{d_n}\rho_{\theta_j}(s,t)$, where $\tau = 1$ if all components ρ_{θ_j} satisfy (B2) uniformly. The condition (B4), although it is slightly stronger than (A3), is often required in functional data analysis, e.g., in Li and Hsing (2010) and Zhang and Wang (2016) for the derivation of uniform convergence rates for $\hat{\mu}$. Such uniform rates are required to bound $\partial\hat{Q}_n/\partial\theta$ sharply in our development, which is critical to establish the following rate for $\hat{\theta}$.

Proposition 3. *Suppose the conditions (A1)–(A2) and (B1)–(B4) hold. If* $d_n = o(n^{1/(4+4\tau)})$, *then with the choice of bandwidth in (H4),* $\|\hat{\theta} - \theta_0\|^2 = O_P(d_n^{2\tau+1}/n)$.

The above result suggests that the estimation quality of $\hat{\theta}$ depends on the dimension of parameters, sample size and singularity of the Hessian matrix at $\theta = \theta_0$, measured by the constant τ in condition (B3). In practice, a few parameters are often sufficient for an adequate fit. In such cases, the dimension d_n might not grow with sample size, i.e., $d_n = O(1)$, and we obtain a parametric rate for $\hat{\theta}$. Now we are ready to state our main theorem that establishes the convergence rate for $\hat{\mathcal{C}}$ in the Hilbert-Schmidt norm $\|\cdot\|_{HS}$, which follows immediately from the above results.

Theorem 4. *Under the same conditions of Proposition 3, we have* $\|\hat{\mathcal{C}} - \mathcal{C}\|_{HS}^2 = O_P(h_\sigma^4 + h_\mu^4 + h_0^4 + n^{-1} + n^{-1}m^{-1}h_\sigma^{-1} + n^{-1}m^{-1}h_\mu^{-1} + n^{-1}m^{-2}h_0^{-1} + d_n^{2\tau+1}n^{-1})$. *With the choice of bandwidth in (H4),* $\|\hat{\mathcal{C}} - \mathcal{C}\|_{HS}^2 = O_P\left((nm)^{-4/5} + d_n^{2\tau+1}n^{-1}\right)$.

In practice, a fully nonparametric approach like local regression to estimating the correlation structure is inefficient, in particular when data are snippets. On the other hand, a

parametric method with a fixed number of parameters might be restrictive when the sample size is large. One way to overcome such a dilemma is to allow the family of parametric models to grow with the sample size. As a working assumption, one might consider that the correlation function ρ falls into $\mathcal{F}_n$, a d_n-dimensional family of models for correlation functions, when the sample size is n. Here, the dimension typically grows with the sample size. For example, one might consider a d_n-Fourier basis family:

$$\kappa_\theta(s,t) = \frac{1}{\psi(s)\psi(t)} \sum_{j=1}^{d_n} \theta_j \phi_j(s)\phi_j(t), \quad \theta_1, \ldots, \theta_{d_n} \geq 0 \text{ and } \sum_{j=1}^{d_n} \theta_j = 1, \tag{10}$$

where $\psi(t) = \left(\sum_{j=1}^{d_n} \theta_j \phi_j^2(t)\right)^{1/2}$ and $\phi_1, \ldots$ are fixed orthonormal Fourier basis functions defined on $\mathcal{T}$. The theoretical result in Theorem 4 applies to this setting by explicitly accounting for the impact of the dimension d_n on the convergence rate.

5 Simulation Studies

To evaluate the numerical performance of the proposed estimators, we generated $X(\cdot)$ from a Gaussian process. Three different covariance functions were considered, namely,

I. $\mathcal{C}(s,t) = \sigma_X(s)\rho_\theta(s,t)\sigma_X(t)$ with the variance function $\sigma_X^2(t) = \sqrt{t}e^{-(t-0.1)^2/10} + 1$ and the Matérn correlation function $\rho_{\theta=(0.5,1)}$,

II. $\mathcal{C}(s,t) = \sum_{k=1}^{50} 2k^{-\lambda}\phi_k(s)\phi_k(t)$ with $\lambda = 2$ and Fourier basis functions $\phi_k(t) = \sqrt{2}\sin(2k\pi t)$, and

III. $\mathcal{C}(s,t) = \sum_{1\leq j,k\leq 5} c_{jk}\phi_j(s)\phi_k(t)$ with $c_{jk} = e^{-|j-k|}/5$.

Two different sample sizes $n = 50$ and $n = 200$ were considered to illustrate the behavior of the estimators under a small sample size and a relatively large sample size. We set the domain $\mathcal{T} = [0,1]$ and $\delta = 0.25$.

To evaluate the impact of the mean function, we also considered two different mean functions, $\mu_1(t) = 2t^2\cos(2\pi t)$ and $\mu_2(t) = e^t/2$. We found that the results are not sensitive to the mean function, and thus focus only on the case $\mu_1(t)$ in this section; the results for the case $\mu(t) = e^t/2$ are provided in Supplementary Material. In addition, to evaluate the impact of the design, we considered two design schemes. In the first scheme, that is referred to as the sparse design, each curve was sparsely sampled at 4 points on average to mimic the scenario of the data application in Section 6. In the second scheme, that is referred to as the dense design, each snippet was recorded in a dense ($m_1 = \cdots = m_n = 26$) and regular grid of an individual specific subinterval of length δ. As the focus of the paper is on sparse snippets, we report the results for the sparse design below. The results for dense snippets are reported in Supplementary Material.

To assess the performance of the estimators for the noise variance σ_0^2, we considered different noise levels $\sigma_0^2 = 0, 0.1, 0.25, 0.5$, varying from no noise to large noise. For example, when $\sigma_0^2 = 0.5$, the signal-to-noise ratio $\mathbb{E}\|X - \mu\|^2/\mathrm{Var}(\varepsilon)$ is about 2. The performance of $\hat{\sigma}_0^2$ is assessed by the root mean squared error (RMSE), defined by

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}|\hat{\sigma}_0^2 - \sigma_0^2|^2},$$

where N is the number of independent simulation replicates, which we set to 100. For the purpose of comparison, we also computed the PACE estimate of Yao et al. (2005) and the estimate proposed by Liu and Müller (2009), denoted by LM, using the fdapace R package (Chen et al., 2020) that is available in the comprehensive R archive network (CRAN). The bandwidth h_μ and h_σ, as well as those in Yao et al. (2005) and Liu and Müller (2009), were selected by five-fold cross-validation. The tuning parameter h_0 was selected by the empirical rule $h_0 = 0.29\hat{\delta}\|\hat{\varsigma}\|_2(nm^2)^{-1/5}$ that is described in Section 3. The simulation results are summarized in Table 1 for the sparse design with mean function

μ_1, as well as Tables S.1–S.3 for the dense design and mean function μ_2 in Supplementary Material, where SNPT denotes our method proposed in Section 3. We observe that in almost all cases, SNPT performs significantly better than the other two methods. The results also demonstrate the effectiveness of the proposed empirical selection rule for the tuning parameter h_0.

To evaluate the performance of the estimators for the covariance structure, we considered two levels of signal-to-noise ratio (SNR), namely, SNR = 2 and SNR = 4. The performance of estimators for the variance function and the covariance function is evaluated by the root mean integrated squared error (RMISE), defined by

$$\text{RMISE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\int_{\mathcal{T}} |\hat{\sigma}_X^2(t) - \sigma_X^2(t)|^2 \mathrm{d}t}$$

for the variance function and

$$\text{RMISE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\int_{\mathcal{T}}\int_{\mathcal{T}} |\hat{\mathcal{C}}(s,t) - \mathcal{C}(s,t)|^2 \mathrm{d}s\mathrm{d}t}$$

for the covariance function. We compared four methods. The first two, denoted by SNPTM and SNPTF, are our semi-parametric approach with the correlation given in (4) and (10), respectively. For the SNPTF method, the dimension d_n of (10) is selected via five-fold cross-validation. It is noted that SNPTM and SNPTF yield the same estimates of the variance function but different estimates of the correlation structure. The third one, denoted by PFBE (penalized Fourier basis expansion), is the method proposed by Lin et al. (2019), and the last one, denoted by PACE, is the approach invented by Yao et al. (2005).

For the estimation of the variance function $\sigma_X^2(t)$, the results are summarized in Table 2 for the sparse design and mean function μ_1, and also in Tables S.4–S.6 in Supplementary Material for the dense design and mean function μ_2. In these tables, the results of SNPTF

are not reported since they are the same as the results of SNPTM. We observe that, in all cases, SNPTM and PFBE substantially outperform PACE. For the dense design, the methods SNPTM and PFBE yield comparable results. The SNPTM method performs better than PFBE when $n = 200$ in most cases, except in the setting III which favors the PFBE method. This suggests that the SNPTM method, which adopts the local linear smoothing strategy combined with our estimator for the variance of the noise, generally converges faster as the sample size grows.

For the estimation of the covariance function $\mathcal{C}$, we summarize the results in Table 3 for the sparse design and mean function μ_1, and in Tables S.7–S.9 in Supplementary Material for the dense design and mean function μ_2. As expected, in all cases, SNPTM, SNPTF and PFBE substantially outperform PACE, since PACE is not designed to process functional snippets. Among the estimators SNPTM, SNPTF and PFBE, in the setting I, SNPTM outperforms the others since in this case the model is correctly specified for SNPTM, in the setting II, SNPTF is the best since the model is correctly specified for SNPTF, and in the setting III, PFBE has a favorable performance. Although there is no universally best estimator, overall these three estimators have comparable performance. To select a method in practice, one can first produce a scatter plot of the raw covariance function. If the function appears to decay monotonically as the point (s, t) moves away from the diagonal, then SNPT with a monotonic decaying correlation such as SNPTM is recommended. Otherwise, SNPT with a general correlation structure such as SNPTF or the PFBE approach might be adopted.

Table 1: RMSE and their standard errors for $\hat{\sigma}_0^2$ under the sparse design and μ_1

			method		
Cov	n	σ_0^2	SNPT	PACE	LM
I	50	0	0.012 (0.009)	0.144 (0.166)	0.129 (0.203)
		0.1	0.029 (0.038)	0.129 (0.146)	0.186 (0.197)
		0.25	0.050 (0.056)	0.147 (0.185)	0.117 (0.125)
		0.5	0.100 (0.135)	0.181 (0.195)	0.157 (0.131)
	200	0	0.009 (0.005)	0.080 (0.103)	0.073 (0.077)
		0.1	0.017 (0.019)	0.091 (0.098)	0.144 (0.150)
		0.25	0.032 (0.038)	0.086 (0.097)	0.093 (0.127)
		0.5	0.049 (0.064)	0.098 (0.118)	0.165 (0.106)
II	50	0	0.036 (0.030)	0.252 (0.245)	0.219 (0.255)
		0.1	0.047 (0.052)	0.254 (0.285)	0.237 (0.255)
		0.25	0.087 (0.133)	0.241 (0.244)	0.159 (0.151)
		0.5	0.128 (0.202)	0.238 (0.260)	0.126 (0.134)
	200	0	0.024 (0.015)	0.177 (0.172)	0.192 (0.200)
		0.1	0.027 (0.027)	0.185 (0.179)	0.176 (0.174)
		0.25	0.042 (0.050)	0.177 (0.177)	0.097 (0.097)
		0.5	0.071 (0.084)	0.174 (0.182)	0.124 (0.089)
III	50	0	0.004 (0.004)	0.099 (0.103)	0.028 (0.064)
		0.1	0.024 (0.029)	0.102 (0.106)	0.099 (0.127)
		0.25	0.049 (0.063)	0.093 (0.109)	0.077 (0.080)
		0.5	0.094 (0.130)	0.113 (0.146)	0.172 (0.128)
	200	0	0.002 (0.002)	0.065 (0.077)	0.009 (0.023)
		0.1	0.010 (0.012)	0.066 (0.067)	0.049 (0.075)
		0.25	0.027 (0.033)	0.068 (0.071)	0.069 (0.067)
		0.5	0.059 (0.071)	0.067 (0.073)	0.163 (0.091)

Table 2: RMISE and their standard errors for $\hat{\sigma}_X^2(t)$ under the sparse design and μ_1

			method		
Cov	SNR	n	SNPTM	PFBE	PACE
I	2	50	0.535 (0.218)	0.518 (0.211)	2.133 (1.536)
		200	0.339 (0.130)	0.330 (0.118)	1.344 (1.126)
	4	50	0.531 (0.199)	0.517 (0.229)	1.845 (1.461)
		200	0.313 (0.136)	0.334 (0.127)	1.151 (0.952)
II	2	50	0.775 (0.396)	0.743 (0.214)	2.602 (1.747)
		200	0.509 (0.163)	0.530 (0.141)	1.699 (1.045)
	4	50	0.768 (0.303)	0.734 (0.351)	2.510 (1.578)
		200	0.471 (0.162)	0.507 (0.149)	1.515 (1.056)
III	2	50	0.633 (0.201)	0.592 (0.136)	1.478 (1.052)
		200	0.376 (0.133)	0.392 (0.107)	1.178 (0.700)
	4	50	0.592 (0.208)	0.586 (0.158)	1.428 (1.166)
		200	0.350 (0.139)	0.385 (0.114)	0.923 (0.451)

Table 3: RMISE and their standard errors for $\hat{\mathcal{C}}$ under the sparse design and μ_1

			method			
Cov	SNR	n	SNPTM	SNPTF	PFBE	PACE
I	2	50	0.339 (0.101)	0.441 (0.158)	0.399 (0.156)	1.470 (0.808)
		200	0.235 (0.092)	0.359 (0.089)	0.295 (0.101)	1.044 (0.625)
	4	50	0.315 (0.093)	0.424 (0.135)	0.371 (0.143)	1.348 (0.809)
		200	0.225 (0.084)	0.341 (0.090)	0.254 (0.097)	0.902 (0.513)
II	2	50	0.556 (0.119)	0.521 (0.183)	0.541 (0.160)	2.061 (1.061)
		200	0.474 (0.068)	0.436 (0.132)	0.465 (0.101)	1.625 (0.632)
	4	50	0.536 (0.126)	0.472 (0.148)	0.517 (0.139)	2.014 (0.868)
		200	0.457 (0.063)	0.419 (0.133)	0.431 (0.112)	1.543 (0.604)
III	2	50	0.503 (0.090)	0.511 (0.154)	0.491 (0.130)	1.248 (0.650)
		200	0.473 (0.041)	0.439 (0.092)	0.366 (0.052)	1.136 (0.439)
	4	50	0.493 (0.075)	0.499 (0.120)	0.487 (0.122)	1.217 (0.727)
		200	0.469 (0.055)	0.423 (0.087)	0.358 (0.063)	0.997 (0.316)

6 Application

We applied the proposed method to analyze the longitudinal data that was collected and detailed in Bachrach et al. (1999). It consists of longitudinal measurements of spinal bone mineral density for 423 healthy subjects. The measurement for each individual was observed annually for up to 4 years. Among 423 subjects, we focused on $n = 280$ subjects ranging in age from 8.8 to 26.2 years who completed at least 2 measurements. A plot for the design of the covariance function is given in Figure 1, while a scatter plot for the raw covariance surface is given in Figure 2. The raw covariance surface seems to decay rapidly to zero as design points move away from the diagonal. This motivated us to estimate the covariance structure with a Matérn correlation function. This method is referred to as SNPTM. In addition, we also used the more flexible d_n-Fourier basis family to see whether a better fit can be achieved, where $d_n = 2$ was selected by Akaike information criterion (AIC). Such approach is denoted by SNPTF.

The estimated variance of the measurement error is 1.5×10^{-3} by the method proposed in Section 3, 10^{-6} by PACE and 7.8×10^{-7} by LM, respectively. The estimates of the covariance surface are depicted in Figure 3. We observe that, the estimates produced by SNPTM and SNPTF are similar in the diagonal region, while visibly differ in the off-diagonal region. For this dataset, the upward off-diagonal parts of the estimated covariance surface by SNPTF seem artificial, so we recommend the SNPTM estimate for this data. For the PACE estimate, due to the missing data in the off-diagonal region and insufficient observations at two ends of the diagonal region, it suffers from significant boundary effect.

The mean function estimated by SNPTM[1] shown in the left panel of Figure 4 and found similar to its counterpart in Lin et al. (2019), suggests that the spinal bone mineral density

[1]SNPTM, SNPTF and PACE use the same method to estimate the mean function.

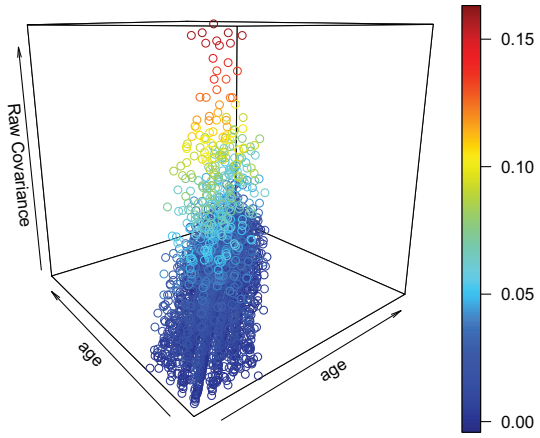

Figure 2: Scatter plot of the raw covariance function of the spinal bone mineral density data.

increases rapidly from age 9 to age 16, and then slows down afterward. The mineral density has the largest variation around age 14, indicated by the variance function estimated by SNPTM[2] and shown in the middle panel of Figure 4. As a comparison, the PACE estimate, shown in the right panel of Figure 4, suffers from the boundary effect that is passed from the PACE estimate of the covariance function, because the PACE method estimates the variance function by the diagonal of the estimated covariance function.

[2]SNPTM and SNPTF use the same method to estimate the variance function.

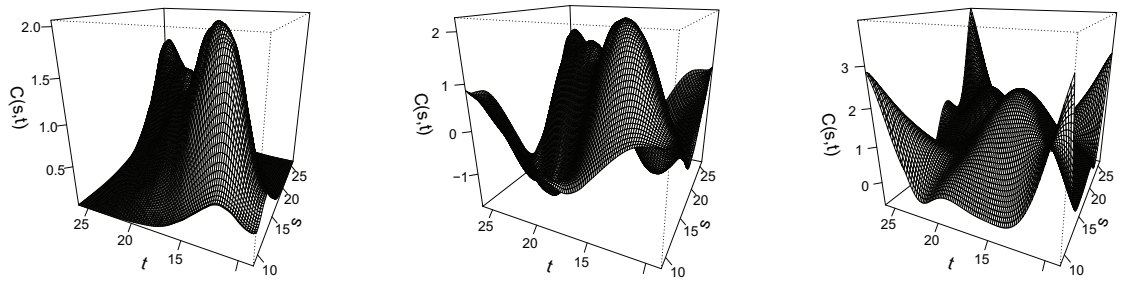

Figure 3: The estimated covariance functions by SNPTM (left), SNPTF (middle) and PACE (right). The z-axis is scaled by 10^{-2} for visualization.

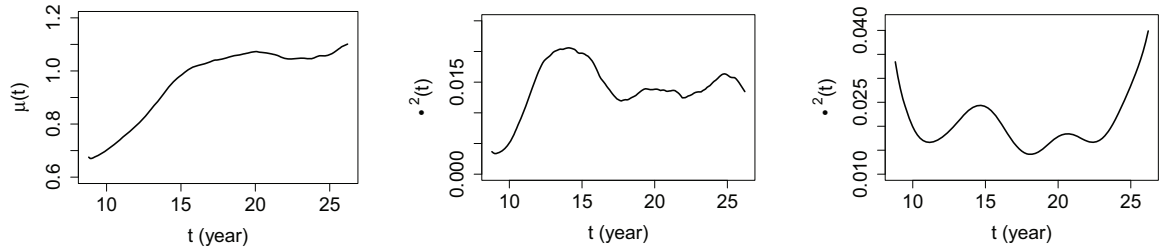

Figure 4: The estimated mean function (left), the estimated variance function by SNPTM and SNPTF (middle), and the estimated variance function by PACE (right).

7 Concluding Remarks

In this paper, we consider the mean and covariance estimation for functional snippets. The estimation of the mean function is still an interpolation problem so previous approaches based on local smoothing methods still work, except that the theory needs a little adjustment to reflect the new design of functional snippets. However, the estimation of the covariance function is quite different because it is now an extrapolation problem rather an interpolation problem, so previous approaches based on local smoothing do not work anymore. We propose a hybrid approach that leverages the available information and structure of the correlation in the diagonal band to estimate the correlation function parametrically but the variance function nonparametrically. Because the dimension of the parameters can grow with the sample size, the approach is very flexible and can be made nearly nonparametric for the final covariance estimate.

An interesting feature of the algorithm is that it reverses the order of estimation for the variance components, compared to existing approaches for non-snippets functional data, by first estimating the noise variance σ_0, then estimating the variance function $\sigma_X^2(t)$, followed by the fitting of the correlation function. The estimation of the covariance function is performed at the very end when all other components have been estimated. The proposed approach differs substantially from traditional approaches, such as PACE (Yao et al., 2005), which estimate the covariance function first, from there the variance function is obtained as a byproduct through the diagonal elements of the covariance estimate, while the noise variance is estimated at the very end. The new procedure to estimate the noise variance is both simpler and better than the PACE estimates. Thus, even if the data are non-snippet types, one can use the new method proposed in Section 3 to estimate the noise variance.

We emphasize that, although the proposed method targets functional snippets, it is also

applicable to functional fragments or functional data in which each curve consists of multiple disjoint snippets. In addition, the theory presented in Section 4 can be slightly modified to accommodate such data. In contrast, methods designed for nonsnippet functional data are generally not applicable to functional snippets, due to the reasons discussed in Section 1. In practice, one might distinguish between functional snippets and nonsnippets by the design plot like Figure 1. If the support points cover the entire region, then the data are of the nonsnippet type. Otherwise they are functional snippets. However, there might be some case that it is unclear whether the entire region is fully covered by support points, especially when data are sparsely observed. In such situation, snippet-based methods, such as the proposed one, is a safer option.

Reliable estimates of the mean and covariance functions are fundamental to the analysis of functional data. They are also the building blocks of functional regression methods and functional hypothesis test procedures. The proposed estimators for the mean and covariance of functional snippets together provide a stepping stone to future study on regression and inference that are specific to functional snippets.

Supplementary Material

The online supplementary material contains additional simulation results, as well as information for implementation of the proposed method in the `R` package `mcfda`[3].

[3]https://github.com/linulysses/mcfda.

Appendix

Selection of h_0

The constant 0.29 in the empirical rule $h_0 = 0.29\hat{\delta}\|\hat{\varsigma}\|_2(nm^2)^{-1/5}$ presented in Section 3 was determined by optimizing $\sum\{\hat{h} - c\hat{\delta}\|\hat{\varsigma}\|_2(nm^2)^{-1/5}\}^2$ over $c \in \mathbb{R}$, where the summation is taken over the combinations of various parameters. Specifically, for each tuple $(n, m, \delta, \sigma_0^2, \mathcal{C})$, we generated a batch of $G = 100$ independent datasets of n centered Gaussian snippets with the covariance function $\mathcal{C}$. Each snippet was recorded at m random points from a random subinterval of length δ in $[0, 1]$. For each batch of datasets, we found $\hat{h}$ to minimize $\sum_{r=1}^{G}\{\hat{\sigma}_{0,r}^2(\hat{h}) - \sigma_0^2\}^2$, where $\hat{\sigma}_{0,r}^2(\hat{h})$ is the estimate of σ_0^2 based the rth dataset in the batch and by using the proposed method with the bandwidth $\hat{h}$. We also obtained the quantities $\hat{\delta} = G^{-1}\sum_{r=1}^{G}\hat{\delta}_r$ and $\|\hat{\varsigma}\|_2 = G^{-1}\sum_{r=1}^{G}\|\hat{\varsigma}_r\|_2$, where $\hat{\delta}_r$ and $\hat{\varsigma}_r$ are the estimate of δ and ς based on the rth dataset in the batch, respectively. In this way, we obtain a collection $\mathscr{H}$ of vectors $(\hat{h}, n, m, \hat{\delta}, \|\hat{\varsigma}\|_2)$. Finally, we found $c = 0.29$ to minimize $\sum\{\hat{h} - c\hat{\delta}\|\hat{\varsigma}\|_2(nm^2)^{-1/5}\}^2$, where the summation is taken over the collection $\mathscr{H}$.

In the above process, the covariance function $\mathcal{C}$ was taken from a collection composed by 1) covariance functions whose correlation part is the correlation function listed in Section 2 with various values of the parameters and whose variance functions are exponential functions, squared sin/cos functions and positive polynomials, 2) covariance functions $\mathcal{C}(s, t) = a\min\{s, t\}$ with various values of $a > 0$, 3) covariance functions $\mathcal{C}(s, t) = \sum_{k=1}^{K} ak^{-\lambda}\phi_k(s)\phi_k(t)$ with various values of $a > 0$, $\lambda > 0$ and $K \geq 1$, where the functions ϕ_k are the Fourier basis functions described in Section 5, and 4) covariance functions $\mathcal{C}(s, t) = \sum_{1\leq j,k\leq K} ae^{-b|j-k|}$ with various choices of $a > 0$, $b > 0$ and $K \geq 1$.

Technical Lemmas

Lemma 5.

(a) *Under conditions (A1)–(A2), one has* $A_2 = A_1 + O(h_0^3)$.

(b) *With condition (A1),* $\mathbb{E}(\hat{B} - B)^2 = O(n^{-1}m^{-2}h_0 + n^{-1}m^{-1}h_0^2)$.

(c) *Under conditions (A1)–(A3),* $\mathbb{E}\{(\hat{A}_0 - \hat{A}_1) - (A_0 - A_1)\}^2 = O(h_0^6 + n^{-1}m^{-2}h_0 + n^{-1}h_0^2)$. *If* $\mathbb{E}L_X^4 < \infty$ *is also assumed, then* $\mathbb{E}\{(\hat{A}_0 - \hat{A}_1) - (A_0 - A_1)\}^2 = O(h_0^6 + n^{-1}m^{-2}h_0 + n^{-1}m^{-1}h_0^2)$.

Proof. To show $A_2 = A_1 + O(h_0^3)$ in part (a), we define $\mathcal{T}_{h_0,\delta} = \{(s,t,u) : u \in [\delta/2, 1-\delta/2], u-\delta/2 \le s,t \le u+\delta/2, |s-t| < h_0\}$ and $g(s,t,u) = \{\mathcal{C}(s,t)+\mu(s)\mu(t)\}f_{T|O}(s|u)f_{T|O}(t|u)f_O(u)$. Let g_s be the partial derivative of g with respect to s. Then, g_s is Lipschitz continuous given condition (A1) and (A2). With t^* denoting a real number satisfying $\min(s,t) \le t^* \le \max(s,t)$, one has

$$\begin{aligned} A_2 &= \iiint_{\mathcal{T}_{h_0,\delta}} [g(t,t,u) + g_s(t,t,u)(s-t) + \{g_s(t^*,t,u) - g_s(t,t,u)\}(s-t)^2\}]\mathrm{d}s\mathrm{d}t\mathrm{d}u \\ &= A_1 + \iiint_{\mathcal{T}_{h_0,\delta}} g_s(t,t,u)(s-t)\mathrm{d}s\mathrm{d}t\mathrm{d}u + O(h_0^3) = A_1 + O(h_0^3), \end{aligned}$$

where the last equality is obtained by observing that

$$\begin{aligned} \iiint_{\mathcal{T}_{h_0,\delta}} g_s(t,t,u)(s-t)\mathrm{d}s\mathrm{d}t\mathrm{d}u =& \int_{\delta/2}^{1-\delta/2}\int_{u-\delta/2+h_0}^{u+\delta/2-h_0}\int_{t-h_0}^{t+h_0} g_s(t,t,u)(s-t)\mathrm{d}s\mathrm{d}t\mathrm{d}u \\ &+ \int_{\delta/2}^{1-\delta/2}\int_{u-\delta/2}^{u-\delta/2+h_0}\int_{\max(u-\delta/2,t-h_0)}^{\min(u+\delta/2,t+h_0)} g_s(t,t,u)(s-t)\mathrm{d}s\mathrm{d}t\mathrm{d}u \\ &+ \int_{\delta/2}^{1-\delta/2}\int_{u+\delta/2-h_0}^{u+\delta/2}\int_{\max(u-\delta/2,t-h_0)}^{\min(u+\delta/2,t+h_0)} g_s(t,t,u)(s-t)\mathrm{d}s\mathrm{d}t\mathrm{d}u \\ =& 0 + O(h_0^3) + O(h_0^3) = O(h_0^3). \end{aligned}$$

For part (b), it is seen that $\mathbb{E}\hat{B} = B$ and

$$\mathbb{E}(\hat{B} - B)^2 = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\frac{1}{m(m-1)}\sum_{j\neq l}1_{|T_{ij}-T_{il}|<h_0} - B\right]^2$$
$$= \frac{1}{n}\mathbb{E}\left[\frac{1}{m(m-1)}\sum_{j\neq l}1_{|T_{ij}-T_{il}|<h_0} - B\right]^2. \tag{11}$$

Now we first observe that $\mathbb{E}(1_{|T_{ij}-T_{il}|<h_0} \mid O_i) = B$, since

$$\begin{aligned}\mathbb{E}(1_{|T_{ij}-T_{il}|<h_0} \mid O_i) &= \iint_{\substack{|s-t|<h_0\\ O_i-\delta/2\le s,t\le O_i+\delta/2}} f_{T|O}(s|O_i)f_{T|O}(t|O_i)\mathrm{d}s\mathrm{d}t\\ &= \iint_{\substack{|s-t|<h_0\\ O_i-\delta/2\le s,t\le O_i+\delta/2}} f_0(s-O_i+\delta/2)f_0(t-O_i+\delta/2)\mathrm{d}s\mathrm{d}t\\ &= \iint_{\substack{|s-t|<h_0\\ 0\le s,t\le\delta}} f_0(s)f_0(t)\mathrm{d}s\mathrm{d}t\end{aligned}$$

and

$$B = \mathbb{E}1_{|T_{ij}-T_{il}|<h_0} = \mathbb{E}\mathbb{E}(1_{|T_{ij}-T_{il}|<h_0} \mid O_i) = \iint_{\substack{|s-t|<h_0\\ 0\le s,t\le\delta}} f_0(s)f_0(t)\mathrm{d}s\mathrm{d}t.$$

Therefore, if j, l, p, q are all distinct, then

$$\begin{aligned}&\mathbb{E}\{(1_{|T_{ij}-T_{il}|<h_0} - B)(1_{|T_{ip}-T_{iq}|<h_0} - B)\}\\ &= \mathbb{E}\mathbb{E}\{(1_{|T_{ij}-T_{il}|<h_0} - B)(1_{|T_{ip}-T_{iq}|<h_0} - B) \mid O_i\}\\ &= \mathbb{E}\{\mathbb{E}(1_{|T_{ij}-T_{il}|<h_0} - B \mid O_i)\mathbb{E}(1_{|T_{ip}-T_{iq}|<h_0} - B \mid O_i)\} = 0.\end{aligned}$$

It is relatively straightforward to show that if $j = p$ but $l \neq q$ or $j = q$ but $l \neq p$, then $\mathbb{E}\{(1_{|T_{ij}-T_{il}|<h_0} - B)(1_{|T_{ip}-T_{iq}|<h_0} - B)\} = O(h_0^2)$, and if $j = p$ and $l = q$ or $j = q$ and $l = p$, then $\mathbb{E}\{(1_{|T_{ij}-T_{il}|<h_0} - B)(1_{|T_{ip}-T_{iq}|<h_0} - B)\} = O(h_0)$. Assembling the above results, one has

$$\mathbb{E}\left[\frac{1}{m(m-1)}\sum_{j\neq l}1_{|T_{ij}-T_{il}|<h_0} - B\right]^2 = O(m^{-2}h_0 + m^{-1}h_0^2),$$

which together with (11) implies the conclusion of part (b).

For part (c), with the aid of part (a), it is straightforward to see that

$$\mathbb{E}\{(\hat{A}_0 - \hat{A}_1) - (A_0 - A_1)\} = O(h_0^3). \tag{12}$$

Now we shall calculate the variance of $\hat{A}_0 - \hat{A}_1$. With definition $E_0 = \mathbb{E}(Y_{ij} - Y_{il})^2 1_{|T_{ij}-T_{il}|<h_0}$, one derives

$$\begin{aligned}
&\mathrm{Var}(\hat{A}_0 - \hat{A}_1) \\
&= \mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^{n} \frac{1}{m(m-1)} \sum_{j\neq l} \frac{(Y_{ij} - Y_{il})^2}{2} 1_{|T_{ij}-T_{il}|<h_0}\right) \\
&= \frac{1}{4n}\mathrm{Var}\left(\frac{1}{m(m-1)} \sum_{j\neq l} (Y_{ij} - Y_{il})^2 1_{|T_{ij}-T_{il}|<h_0}\right) \\
&= \frac{1}{4n}\left(\frac{1}{m^2(m-1)^2} \sum_{j\neq l}\sum_{p\neq q} \mathbb{E}\{(Y_{ij} - Y_{il})^2 1_{|T_{ij}-T_{il}|<h_0} - E_0\}\{(Y_{ip} - Y_{iq})^2 1_{|T_{ip}-T_{iq}|<h_0} - E_0\}\right) \\
&\equiv \frac{1}{4n}\left(\frac{1}{m^2(m-1)^2} \sum_{j\neq l}\sum_{p\neq q} V(j,l,p,q)\right).
\end{aligned} \tag{13}$$

Below we derive bounds for the term $V(j,l,p,q)$.

- Case 1: j, l, p and q are all distinct. In this case, via straightforward computation, one can show that $V(j,l,p,q) = \mathbb{E}\{(Y_{ij} - Y_{il})^2 1_{|T_{ij}-T_{il}|<h_0}\}\{(Y_{ip} - Y_{iq})^2 1_{|T_{ip}-T_{iq}|<h_0}\} - E_0^2 = O(h_0^2)$.

- Case 2: $j = p$ but $l \neq q$ or $j = q$ but $l \neq p$. Similar to Case 1, one has $V(j,l,p,q) = O(h_0^2)$.

- Case 3: $j = p$ and $l = q$ or $j = q$ and $l = p$. In this case,

$$V(j,l,p,q) = \mathbb{E}\{(Y_{ij} - Y_{il})^4 1_{|T_{ij}-T_{il}|<h_0}\} - E_0^2 = O(h_0).$$

Based on the above bounds, we have $\mathrm{Var}(\hat{A}_0 - \hat{A}_1) = O(n^{-1}h_0^2 + n^{-1}m^{-1}h_0^2 + n^{-1}m^{-2}h_0) = O(n^{-1}h_0^2 + n^{-1}m^{-2}h_0)$. Together with the bias given in (12), this implies the first statement of part (c).

For the second statement of part (c), we observe that with condition $\mathbb{E}L_X^4 < \infty$, the bound in Case 1 can be sharpened in the following way. First, we see that

$$E_0 = \mathbb{E}\{X_i(T_{ij}) - X_i(T_{il})\}^2 1_{|T_{ij}-T_{il}|<h_0} + \mathbb{E}(\varepsilon_{ij} - \varepsilon_{il})^2 1_{|T_{ij}-T_{il}|<h_0} = E_1 + 2\sigma_0^2 B,$$

where $E_1 = \mathbb{E}\{X_i(T_{ij}) - X_i(T_{il})\}^2 1_{|T_{ij}-T_{il}|<h_0}$. Then, we decompose $V(j,l,p,q)$ into $I_1 + I_2 + I_3 + I_4$, where

$$\begin{aligned}
I_1 &= \mathbb{E}[\{X(T_{ij}) - X(T_{il})\}^2 1_{|T_{ij}-T_{il}|<h_0} - E_1][\{X(T_{ip}) - X(T_{iq})\}^2 1_{|T_{ip}-T_{iq}|<h_0} - E_1],\\
I_2 &= \mathbb{E}[\{X(T_{ij}) - X(T_{il})\}^2 1_{|T_{ij}-T_{il}|<h_0} - E_1][(\varepsilon_{ip} - \varepsilon_{iq})^2 1_{|T_{ip}-T_{iq}|<h_0} - 2\sigma_0^2 B],\\
I_3 &= \mathbb{E}[(\varepsilon_{ij} - \varepsilon_{il})^2 1_{|T_{ij}-T_{il}|<h_0} - 2\sigma_0^2 B][\{X(T_{ip}) - X(T_{iq})\}^2 1_{|T_{ip}-T_{iq}|<h_0} - E_1],\\
I_4 &= \mathbb{E}[(\varepsilon_{ij} - \varepsilon_{il})^2 1_{|T_{ij}-T_{il}|<h_0} - 2\sigma_0^2 B][(\varepsilon_{ip} - \varepsilon_{iq})^2 1_{|T_{ip}-T_{iq}|<h_0} - 2\sigma_0^2 B].
\end{aligned}$$

For I_2, one can show that

$$\begin{aligned}
I_2 &= \mathbb{E}\mathbb{E}\left([\{X(T_{ij}) - X(T_{il})\}^2 1_{|T_{ij}-T_{il}|<h_0} - E_1][(\varepsilon_{ip} - \varepsilon_{iq})^2 1_{|T_{ip}-T_{iq}|<h_0} - 2\sigma_0^2 B] \mid O_i\right)\\
&= \mathbb{E}\left(\mathbb{E}[\{X(T_{ij}) - X(T_{il})\}^2 1_{|T_{ij}-T_{il}|<h_0} - E_1 \mid O_i]\mathbb{E}[(\varepsilon_{ip} - \varepsilon_{iq})^2 1_{|T_{ip}-T_{iq}|<h_0} - 2\sigma_0^2 B \mid O_i]\right)\\
&= 0,
\end{aligned}$$

where the first equality is due to the assumption that $T_{i1}, \ldots, T_{im}$ are i.i.d. conditional on O_i, and the second one is based on the following observation

$$\mathbb{E}[(\varepsilon_{ip} - \varepsilon_{iq})^2 1_{|T_{ip}-T_{iq}|<h_0} - 2\sigma_0^2 B \mid O_i] = 2\sigma_0^2 \mathbb{E}(1_{|T_{ip}-T_{iq}|<h_0} \mid O_i) - 2\sigma_0^2 B = 2\sigma_0^2 B - 2\sigma_0^2 B = 0,$$

where we recall that $\mathbb{E}(1_{|T_{ij}-T_{il}|<h_0} \mid O_i) = B$. Similarly, $I_3 = 0$ and $I_4 = 0$. For I_1, one can show that

$$|I_1| = |\mathbb{E}[\{X(T_{ij}) - X(T_{il})\}^2 1_{|T_{ij}-T_{il}|<h_0} - E_1][\{X(T_{ip}) - X(T_{iq})\}^2 1_{|T_{ip}-T_{iq}|<h_0} - E_1]|$$

$$
\begin{aligned}
&= |\mathbb{E}[\{X(T_{ij}) - X(T_{il})\}^2 1_{|T_{ij}-T_{il}|<h_0} \{X(T_{ip}) - X(T_{iq})\}^2 1_{|T_{ip}-T_{iq}|<h_0}] - E_1^2| \\
&\leq \mathbb{E}(L_X^4 |T_{ij} - T_{il}|^2 |T_{ip} - T_{iq}|^2 1_{|T_{ij}-T_{il}|<h_0} 1_{|T_{ip}-T_{iq}|<h_0}) + E_1^2 \\
&\leq h_0^4 \mathbb{E} L_X^4 \mathbb{E} 1_{|T_{ij}-T_{il}|<h_0} 1_{|T_{ip}-T_{iq}|<h_0} + E_1^2 \\
&= O(h_0^6) + E_1^2,
\end{aligned}
$$

where the first inequality is due to the Lipschitz continuity property of sample paths. Again, based on such continuity property, one has $E_1 = \mathbb{E}\{X_i(T_{ij}) - X_i(T_{il})\}^2 1_{|T_{ij}-T_{il}|<h_0} \leq \mathbb{E} L_X^2 |T_{ij} - T_{il}|^2 1_{|T_{ij}-T_{il}|<h_0} \leq h_0^2 \mathbb{E} L_X^2 \mathbb{E} 1_{|T_{ij}-T_{il}|<h_0} = O(h_0^3)$. Therefore, we conclude that $I_1 = O(h_0^6)$. Together with $I_2 = I_3 = I_4 = 0$, this implies that $V(j,l,p,q) = O(h_0^6)$. It further indicates that $\mathrm{Var}(\hat{A}_0 - \hat{A}_1) = O(n^{-1}h_0^6 + n^{-1}m^{-1}h_0^2 + n^{-1}m^{-2}h_0)$. Combined with the bias term in (12), this implies the second statement of part (c). □

Proofs of Main Results

Proof of Proposition 3. For the moment, we assume $\mu \equiv 0$. Denote

$$
Q_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{m(m-1)} \sum_{1\leq j\neq l\leq m} \{\sigma_X(T_{ij})\sigma_X(T_{il})\rho_\theta(T_{ij}, T_{il}) - C_{ijl}\}^2.
$$

Now we show that

$$
\left\| \frac{\partial \hat{Q}_n}{\partial \theta} - \frac{\partial Q_n}{\partial \theta} \right\| = O_P\left(\sqrt{\frac{d_n a_n \log n}{n}} \right), \tag{14}
$$

where $a_n = (\log n)\{(nm)^{-4/5} + n^{-1}\}$. First, we observe that

$$
\frac{\partial \hat{Q}_n}{\partial \theta} - \frac{\partial Q_n}{\partial \theta} = I_1 + I_2 + I_3
$$

with

$$
I_1 = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{m(m-1)} \sum_{1\leq j\neq l\leq m} 2\{\sigma_X(T_{ij})\sigma_X(T_{il})\rho_\theta(T_{ij}, T_{il}) - C_{ijl}\} \times
$$

$$\{\hat{\sigma}_X(T_{ij})\hat{\sigma}_X(T_{il}) - \sigma_X(T_{ij})\sigma_X(T_{il})\}\frac{\partial \rho_\theta(T_{ij}, T_{il})}{\partial \theta},$$

$$I_2 = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{m(m-1)} \sum_{1\le j\neq l\le m} 2\{\hat{\sigma}_X(T_{ij})\hat{\sigma}_X(T_{il}) - \sigma_X(T_{ij})\sigma_X(T_{il})\}\rho_\theta(T_{ij}, T_{il})\times$$
$$\sigma_X(T_{ij})\sigma_X(T_{il})\frac{\partial \rho_\theta(T_{ij}, T_{il})}{\partial \theta},$$

$$I_3 = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{m(m-1)} \sum_{1\le j\neq l\le m} 2\{\hat{\sigma}_X(T_{ij})\hat{\sigma}_X(T_{il}) - \sigma_X(T_{ij})\sigma_X(T_{il})\}\rho_\theta(T_{ij}, T_{il})\times$$
$$\{\hat{\sigma}_X(T_{ij})\hat{\sigma}_X(T_{il}) - \sigma_X(T_{ij})\sigma_X(T_{il})\}\frac{\partial \rho_\theta(T_{ij}, T_{il})}{\partial \theta}.$$

To derive the rate for I_1, we define

$$G = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{m(m-1)} \sum_{1\le j\neq l\le m} 2\{\sigma_X(T_{ij})\sigma_X(T_{il})\rho_\theta(T_{ij}, T_{il}) - C_{ijl}\} \equiv \frac{1}{n}\sum_{i=1}^{n} G_i.$$

It can be verified that $\mathbb{E}G_i = 0$, and also $\mathbb{E}G_i^2 < \infty$ given condition (A3) and (B2). We view each G_i as a random linear functional from the space $\Lambda_0 = \{f \in C^2(\mathcal{T}) : \|f\|_\infty \le 1\}$, i.e.,

$$G_i(f) \mapsto \frac{1}{m(m-1)} \sum_{1\le j\neq l\le m} 2\{\sigma_X(T_{ij})\sigma_X(T_{il})\rho_\theta(T_{ij}, T_{il}) - C_{ijl}\}f(T_{ij}, T_{il}),$$

where $f \in \Lambda_0$. Then we follow the same lines of the argument for Lemma 2 of Severini and Wong (1992) to establish that $\sqrt{n}G$ converges to a Gaussian element on the Banach space $C(\Lambda_0)$ of continuous functions on Λ_0 with the sup norm. On the other hand, using the same technique of Zhang and Wang (2016) for the uniform convergence of the local linear estimator for the mean function, we can show that $\sup_t |\hat{\sigma}_X(t) - \sigma_X(t)| = O_P(\sqrt{a_n})$, and hence $\sup_{s,t} |\hat{\sigma}_X(s)\hat{\sigma}_X(t) - \sigma_X(s)\sigma_X(t)| = O_P(\sqrt{a_n})$. By condition (B2) that $\partial\rho_\theta(s,t)/\partial\theta_j$ is uniformly bounded for all j, we can deduce that, for sufficiently large n, with probability tending to one, the function $(a_n \log n)^{-1/2} f_j$ with $f_j : (s,t) \mapsto \{\hat{\sigma}_X(s)\hat{\sigma}_X(t) -$

$\sigma_X(s)\sigma_X(t)\}\partial\rho_\theta(s,t)/\partial\theta_j$ falls into Λ_0 for all j. Therefore,

$$\left\|\sqrt{n}G\left(\frac{f_j}{\sqrt{a_n\log n}}\right)\right\| \le \|\sqrt{n}G\| \left\|\frac{f_j}{\sqrt{a_n\log n}}\right\| = O_P(1),$$

where O_P is uniform for all j. Noting that $I_1 = (Gf_1,\ldots,Gf_{d_n})^{\mathrm{T}}$, one can deduce from the above that

$$\|I_1\| \le \sqrt{\sum_{j=1}^{d_n}\|Gf_j\|^2} \le \sqrt{d_n}\max_{1\le j\le d_n}\|Gf_j\| = O_P\left(\sqrt{\frac{d_n a_n\log n}{n}}\right).$$

When $\mu \neq 0$, an argument similar to the above can also be applied to handle extra terms induced by the discrepancy between $\hat{\mu}$ and μ, so that we still obtain the same rate as the above. Similar argument applies to I_2, and we have $I_2 = O_P(\sqrt{d_n a_n\log n}/\sqrt{n})$. It is easy to see that I_3 is dominated by the other terms. Together, we establish (14). It is seen that $\|\partial Q_n/\partial\theta\mid_{\theta=\theta_0}\| = O_P(\sqrt{d_n/n})$. Thus, we have

$$\begin{aligned}\left\|\frac{\partial\hat{Q}_n}{\partial\theta}\mid_{\theta=\theta_0}\right\| &\le \left\|\frac{\partial Q_n}{\partial\theta}\mid_{\theta=\theta_0}\right\| + \left\|\left(\frac{\partial\hat{Q}_n}{\partial\theta}-\frac{\partial Q_n}{\partial\theta}\right)\mid_{\theta=\theta_0}\right\| \\ &= O_P\left(\sqrt{\frac{d_n}{n}}+\sqrt{\frac{d_n a_n\log n}{n}}\right) = O_P\left(\sqrt{\frac{d_n}{n}}\right),\end{aligned}$$

Straightforward but somewhat tedious calculation can show that

$$\left\|\frac{\partial^2\hat{Q}_n}{\partial\theta^2}\mid_{\theta=\theta_0} - \frac{\partial^2 Q}{\partial\theta^2}\mid_{\theta=\theta_0}\right\| = O_P\left(\frac{d_n}{\sqrt{n}}+d_n\sqrt{a_n}\right) = O_P\left(d_n\sqrt{a_n}\right)$$

and

$$\sup_\theta\left|\sum_{|\alpha|=3} v^\alpha\frac{\partial^\alpha\hat{Q}_n(\theta)}{\alpha!}\right| = O_P\left(d_n^{3/2}\|v\|^3\right).$$

Now let $\eta_n = \sqrt{d_n^{1+2\tau}/n}$. By Taylor expansion,

$$D(u) \equiv \hat{Q}_n(\theta_0+\eta_n u) - \hat{Q}_n(\theta_0)$$

$$
\begin{aligned}
&= \eta_n \left(\frac{\partial \hat{Q}_n}{\partial \theta} \mid_{\theta=\theta_0} \right)^{\mathrm{T}} u + \eta_n^2 u^{\mathrm{T}} \left(\frac{\partial^2 \hat{Q}_n}{\partial \theta^2} \mid_{\theta=\theta_0} \right) u + \eta_n^3 \sum_{|\alpha|=3} u^\alpha \frac{\partial^\alpha \hat{Q}_n}{\alpha!} \mid_{\theta=\theta^*} \\
&= O_P \left(\eta_n \sqrt{\frac{d_n}{n}} \right) \|u\| + \eta_n^2 \lambda_{\min} \left(\frac{\partial^2 Q}{\partial \theta^2} \mid_{\theta=\theta_0} \right) \|u\|^2 + O_P \left(\eta_n^3 d_n^{3/2} \right) \|u\|^3 \\
&\geq O_P \left(d_n^{1+\tau} n^{-1} \right) \|u\| + c_0 d^{1+\tau} n^{-1} \|u\|^2 + o_P(d^{1+\tau} n^{-1}) \|u\|^3 > 0
\end{aligned}
$$

for some constant $c_0 > 0$ and if $\|u\| = c$ for a sufficiently large absolute constant $c > 0$. Thus, $\|\hat{\theta} - \theta_0\| = O_P(\eta_n) = O_P(n^{-1/2} d_n^{\tau+1/2})$. □

References

Bachrach, L. K., Hastie, T., Wang, M.-C., Narasimhan, B., and Marcus, R. (1999), "Bone mineral acquisition in healthy Asian, hispanic, black, and Caucasian youth: A longitudinal study," *The Journal of Clinical Endocrinology & Metabolism*, 84, 4702–4712.

Chen, Y., Carroll, C., Dai, X., Fan, J., Hadjipantelis, P. Z., Han, K., Ji, H., Müller, H.-G., and 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.

Dawson, M. and Müller, H.-G. (2018), "Dynamic modeling of conditional quantile trajectories, with application to longitudinal snippet data," *Journal of the American Statistical Association*, 113, 1612–1624.

Delaigle, A. and Hall, P. (2016), "Approximating fragmented functional data by segments of Markov chains," *Biometrika*, 103, 779–799.

Delaigle, A., Hall, P., Huang, W., and Kneip, A. (2019), "Estimating the covariance of frag-

mented and other related types of functional data," *Journal of the American Statistical Association*, to appear.

Descary, M.-H. and Panaretos, V. M. (2019), "Functional data analysis by matrix completion," *The Annals of Statistics*, 47, 1–38.

Fan, J. (1993), "Local linear regression smoothers and their minimax efficiencies," *The Annals of Statistics*, 21, 196–216.

Fan, J., Huang, T., and Li, R. (2007), "Analysis of longitudinal data with semiparametric estimation of covariance function," *Journal of the American Statistical Association*, 102, 632–641.

Fan, J. and Wu, Y. (2008), "Semiparametric estimation of covariance matrices for longitudinal data," *Journal of American Statistical Association*, 103, 1520–1533.

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

Galbraith, S., Bowden, J., and Mander, A. (2017), "Accelerated longitudinal designs: an overview of modelling, power, costs and handling missing data," *Statistical Methods in Medical Research*, 26, 374–398.

Gellar, J. E., Colantuoni, E., Needham, D. M., and Crainiceanu, C. M. (2014), "Variable-domain functional regression for modeling ICU data," *Journal of the American Statistical Association*, 109, 1425–1439.

Goldberg, Y., Ritov, Y., and Mandelbaum, A. (2014), "Predicting the continuation of a function with applications to call center data," *Journal of Statistical Planning and Inference*, 147, 53–65.

Gromenko, O., Kokoszka, P., and Sojka, J. (2017), "Evaluation of the cooling trend in the ionosphere using functional regression with incomplete curves," *The Annals of Applied Statistics*, 11, 898–918.

Hall, P. and Marron, J. S. (1997), "On the shrinkage of local linear curve estimators," *Statistics and Computing*, 516, 11–17.

Hsing, T. and Eubank, R. (2015), *Theoretical Foundations of Functional Data Analysis, with an Introduction to Linear Operators*, Wiley.

Kneip, A. and Liebl, D. (2019+), "On the optimal reconstruction of partially observed functional data," *The Annals of Statistics*, to appear.

Kokoszka, P. and Reimherr, M. (2017), *Introduction to Functional Data Analysis*, Chapman and Hall/CRC.

Kraus, D. (2015), "Components and completion of partially observed functional data," *Journal of Royal Statistical Society: Series B (Statistical Methodology)*, 77, 777–801.

Kraus, D. and Stefanucci, M. (2019), "Classification of functional fragments by regularized linear classifiers with domain selection," *Biometrika*, 106, 161–180.

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

Liebl, D. (2013), "Modeling and forecasting electricity spot prices: A functional data perspective," *The Annals of Applied Statistics*, 7, 1562–1592.

Liebl, D. and Rameseder, S. (2019), "Partially observed functional data: The case of systematically missing parts," *Computational Statistics & Data Analysis*, 131, 104–115.

Lin, Z., Wang, J.-L., and Zhong, Q. (2019), "Basis expansions for functional snippets," *arxiv*.

Lin, Z. and Yao, F. (2020+), "Functional regression on manifold with contamination," *Biometrika*, to appear.

Liu, B. and Müller, H.-G. (2009), "Estimating derivatives for samples of sparsely observed functions, with application to online auction dynamics," *Journal of the American Statistical Association*, 104, 704–717.

Paul, D. and Peng, J. (2011), "Principal components analysis for sparsely observed correlated functional data using a kernel smoothing approach," *Electronic Journal of Statistics*, 5, 1960–2003.

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

Raudenbush, S. W. and Chan, W.-S. (1992), "Growth Curve Analysis in Accelerated Longitudinal Designs," *Journal of Research in Crime and Delinquency*, 29, 387–411.

Rice, J. A. and Wu, C. O. (2001), "Nonparametric Mixed Effects Models for Unequally Sampled Noisy Curves," *Biometrics*, 57, 253–259.

Scheuerer, M. (2010), "Regularity of the sample paths of a general second order random field," *Stochastic Processes and their Applications*, 120, 1879–1897.

Seifert, B. and Gasser, T. (1996), "Finite-Sample Variance of Local Polynomials: Analysis and Solutions," *Journal of the American Statistical Association*, 91, 267–275.

Severini, T. A. and Wong, W. H. (1992), "Profile Likelihood and Conditionally Parametric Models," *The Annals of Statistics*, 20, 1768–1802.

Stefanucci, M., Sangalli, L. M., and Brutti, P. (2018), "PCA-based discrimination of partially observed functional data, with an application to AneuRisk65 data set," *Statistica Neerlandica*, 72, 246–264.

Wang, J.-L., Chiou, J.-M., and Müller, H.-G. (2016), "Review of functional data analysis," *Annual Review of Statistics and Its Application*, 3, 257–295.

Yao, F., Müller, H.-G., and Wang, J.-L. (2005), "Functional Data Analysis for Sparse Longitudinal Data," *Journal of the American Statistical Association*, 100, 577–590.

Zhang, A. and Chen, K. (2018), "Nonparametric covariance estimation for mixed longitudinal studies, with applications in midlife women's health," *arXiv*.

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

— (2018), "Optimal weighting schemes for longitudinal and functional data," *Statistics & Probability Letters*, 138, 165–170.