Nonparametric Operator-Regularized Covariance Function Estimation for Functional Data

Raymond K. W. Wong[a,*], Xiaoke Zhang[b]

[a] *Department of Statistics, Texas A&M University*
[b] *Department of Statistics, George Washington University*

Abstract

In functional data analysis (FDA), the covariance function is fundamental not only as a critical quantity for understanding elementary aspects of functional data but also as an indispensable ingredient for many advanced FDA methods. A new class of nonparametric covariance function estimators in terms of various spectral regularizations of an operator associated with a reproducing kernel Hilbert space is developed. Despite their nonparametric nature, the covariance estimators are automatically positive semi-definite, which is an essential property of covariance functions, via a one-step procedure. An unconventional representer theorem is established to provide a finite dimensional representation for this class of covariance estimators based on data, although the solutions are searched over infinite dimensional functional spaces. To further achieve a low-rank representation, another desirable property, e.g., for dimension reduction and easy interpretation, the trace-norm regularization is particularly studied, under which an efficient algorithm is developed based on the accelerated proximal gradient method. The outstanding practical performance of the trace-norm-regularized covariance estimator is demonstrated by a simulation study and the analysis of a traffic dataset. Under both fixed and random designs, an excellent rate of convergence is established for a broad class of operator-regularized covariance function estimators, which generalizes both the trace-norm-regularized covariance estimator and other popular alternatives.

Keywords:

Functional data analysis, low-rank estimation, reproducing kernel Hilbert space, spectral

*Corresponding author. Department of Statistics, Texas A&M University, College Station, Texas 77843, U.S.A.
Email address: raywong@stat.tamu.edu (Raymond K. W. Wong)

regularization.

1. Introduction

In recent decades, functional data analysis (FDA) has received substantial attention and become increasingly important especially with the advent of the "Big Data" era. Representative monographs on FDA include Ramsay and Silverman (2005), Ferraty and Vieu (2006), Horváth and Kokoszka (2012) and Hsing and Eubank (2015). Typically functional data are collected from n curves $\{X_i : i = 1, \ldots, n\}$ that are regarded as independent copies of a real-valued L^2 stochastic process X defined on a compact domain $\mathcal{T}$ with mean function $\mu_0(t) = \mathbb{E}\{X(t)\}, t \in \mathcal{T}$, and covariance function $C_0(s,t) = \mathrm{cov}\{X(s), X(t)\}, s, t \in \mathcal{T}$. In reality, due to discrete recording and the presence of noise, the data are often represented by $\{(T_{ij}, Y_{ij}) : i = 1, \ldots, n; j = 1, \ldots m_i\}$, where m_i is the number of observations from the i-th curve X_i, and Y_{ij} is the noisy observation from X_i measured at the discrete time point T_{ij}, i.e., $Y_{ij} = X_i(T_{ij}) + \varepsilon_{ij}$. Here $\{\varepsilon_{ij} : i = 1, \ldots, n; j = 1, \ldots m_i\}$ are independent errors with zero mean and finite variance σ^2. For simplicity and without loss of generality we assume $m_i = m \geq 2$ for all i, and m is nonrandom but may vary over n.

Among various population quantities, the covariance function C_0 is fundamental in FDA. Generally C_0 has two major roles. It is not only an important quantity that characterizes the temporal dependency (Yao et al., 2005a; Li and Hsing, 2010; Zhang and Wang, 2016), but also a building block for many advanced approaches in FDA, e.g., functional principal component analysis (FPCA). Hence the FDA literature that involves covariance function estimation may accordingly be categorized into two types depending on the role of C_0. As for the estimation of C_0, a variety of nonparametric methods have been proposed, such as local polynomial smoothing (Li and Hsing, 2010; Zhang and Wang, 2016), B-splines (James et al., 2000; Rice and Wu, 2001), penalized splines (Goldsmith et al., 2011; Xiao et al., 2013), and smoothing splines (Rice and Silverman, 1991). In this article we adopt the framework of a reproducing kernel Hilbert space (RKHS), which has recently gained increasing popularity in FDA (e.g., Cai and Yuan, 2010; Avery et al., 2014; Zhu et al., 2014; Wang and Ruppert, 2015; Wong et al., 2017).

Positive semi-definiteness is an essential characteristic of covariance functions. Therefore, a valid

covariance estimator is usually desired to be positive semi-definite, especially when this feature is indispensable in subsequent analyses, such as correlation estimation (see Section 5) and functional linear regression (Yao et al., 2005b). Meanwhile, having a low rank is another appealing feature of a covariance estimator. First, a low rank can encourage dimension reduction, facilitate simple interpretations (e.g., of FPCA), and alleviate computational and storage burdens. Moreover, a low rank is often technically needed in trajectory prediction, functional linear regression and some other FDA methods (e.g., Yao et al., 2005a; Chiou, 2012; Yao et al., 2005b; Li et al., 2013; Jiang et al., 2016).

A majority of existing FDA methods cannot directly produce a covariance estimator that is positive semi-definite or of low rank. In the literature there are roughly two types of indirect approaches that always involve multi-step procedures, apart from tuning parameter selection. The first type begins with a constraint-free covariance function estimator, then followed by a reconstruction step, e.g., via FPCA and truncation. See Hall and Vial (2006) and Poskitt and Sengarapillai (2013) for instances. For the other type, a small number of eigenvalues and eigenfunctions are first estimated, e.g., by fitting a mixed-effects model (James et al., 2000; Paul and Peng, 2009), and then a positive semi-definite covariance function estimator of low rank can be reconstructed in terms of eigen-decomposition. These methods, however, are unfavorable since they may not only complicate the theoretical analysis of the final estimator, but also make computation unstable due to the non-smooth truncation. Therefore, in this article, we aim to develop a coherent "one-step" procedure that can automatically produce both a positive semi-definite and a low-rank covariance function estimator.

To achieve this goal, we propose a novel class of tensor product RKHS covariance estimators via a variety of spectral regularizations of an operator. The estimation framework respects the semi-positive structure of covariance functions by imposing a constraint, so the resulting estimator automatically inherits this characteristic. The spectral regularizations generalize the popular Hilbert-Schmidt penalty (e.g., Cai and Yuan, 2010), and can easily enable low-rank representations, e.g., when the trace-norm penalty is used. Given any penalty, the corresponding covariance estimator is obtained by a single step, which can reduce the computational and theoretical complexities

of the aforementioned indirect approaches.

We establish a crucial representer theorem that provides a finite dimensional representation for this class of covariance estimators based on data, although the solutions of the corresponding optimizations are searched over infinite dimensional functional spaces. Compared with its classical counterparts (e.g., Wahba, 1990; Cai and Yuan, 2010), this representer theorem is unconventional and technically innovative due to the semi-positivity constraint and a wide range of regularizations. A byproduct of the theorem is an explicit expression of the L^2 eigen-decomposition admitted by each covariance estimator, which avoids numerical approximations commonly needed in FPCA due to discretization.

To encourage low rank, we particularly focus on the trace-norm regularization in our algorithmic development although the underlying strategy could be applied to other convex regularizations. The corresponding objective function is convex but non-differentiable. We develop an efficient algorithm based on the representer theorem and accelerated proximal gradient method (Beck and Teboulle, 2009). The numerical performance of the resulting estimator is shown in a simulation study to be the best among popular alternatives with respect to rank reduction, estimation accuracy, and computational stability. Its applicability is convincingly illustrated in the analysis of a traffic dataset. Note that, asymptotically, the use of trace-norm regularization does not rule out the case when C_0 is of high or infinite rank. Irrespective of the true rank, our estimator is consistent with the optimal convergence rate, up to some order of $\log n$, as implied by the theoretical results below.

Despite the lack of a closed form due to the semi-positivity constraint and possibly non-differentiable penalties, we develop the empirical L^2 rate of convergence for a class of operator-regularized covariance function estimators in the tensor product Sobolev-Hilbert spaces. The asymptotic results are broad since they hold for both fixed and random designs, and incorporate a variety of spectral regularizations, where the trace-norm and Hilbert-Schmidt regularizations are both special cases. Generally, the rate is comparable to the optimal rate of standard two-dimensional nonparametric smoothers. If X is additionally periodic, we can improve the result significantly and achieve the optimal one-dimensional nonparametric rate, up to some order of $\log n$, which is comparable to the minimax rate obtained by Cai and Yuan (2010) and the L^2

rate achieved by Paul and Peng (2009) for sparse functional data. In contrast to these two pioneer works, our objective function is not necessarily differentiable, which thus requires different technical treatments. Our theoretical results are established in terms of empirical processes techniques. The success of the proofs depends on the upper bound of the entropy for tensor product Sobolev-Hilbert spaces, which is the first appearance in the FDA literature to our best knowledge.

To summarize, the main contribution of this article is three-fold. First, we propose a new and broad class of RKHS covariance estimators via a variety of spectral regularizations of an operator. The resulting estimator is automatically positive semi-definite through a one-step procedure. Additionally, low-rank estimation is encouraged when a proper penalty is chosen, e.g., the trace-norm penalty. Second, we establish an unconventional representer theorem that provides a finite dimensional representation for the covariance estimator. This theorem makes the estimation procedure practically computable and facilitates our algorithmic development. Lastly, we develop the asymptotic results for a broad class of covariance function estimators in tensor product Sobolev-Hilbert spaces, which hold for both fixed and random designs, and incorporate a variety of spectral regularizations. For periodic functional spaces, in particular, the estimators can be shown to achieve a one-dimensional rate for a two-dimensional target, based on a new entropy upper bound.

The rest of the article is organized as follows. The proposed methodology is presented in Section 2 and computational issues are discussed in Section 3. The empirical performance of the trace-norm-regularized estimator is evaluated by a simulation study in Section 4 and a real data application is given in Section 5. Section 6 provides theoretical results and the article is concluded with Section 7. Additional materials, including technical details, more simulation results and further algorithmic descriptions, are provided in separate supplementary material. An R package "rkhscovfun" based on this article is available at `https://github.com/raymondkww/rkhscovfun`.

2. Methodology

In the same vein as penalized splines (e.g., Pearce and Wand, 2006) and smoothing splines (e.g., Wahba, 1990; Eggermont and LaRiccia, 2009; Gu, 2013), we suppose that the sample path of X belongs to a RKHS $\mathcal{H}(K)$ equipped with an inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}(K)}$ and corresponding norm

$\|\cdot\|_{\mathcal{H}(K)}$, which is associated with a continuous and square-integrable reproducing kernel $K(\cdot,\cdot)$ defined on $\mathcal{T}\times\mathcal{T}$. A key property of K is the so-called reproducing property:

$$\langle K(t,\cdot), f(\cdot)\rangle_{\mathcal{H}(K)} = f(t), \quad \text{for any } t\in\mathcal{T} \text{ and } f\in\mathcal{H}(K).$$

Moreover, K also uniquely determines both $\langle\cdot,\cdot\rangle_{\mathcal{H}(K)}$ and $\|\cdot\|_{\mathcal{H}(K)}$. A canonical example of RKHS is the r-th order Sobolev-Hilbert space on $\mathcal{T}=[0,1]$:

$$\mathcal{W}^r = \{g : g^{(v)}, v=0,\ldots,r-1, \text{ are absolutely continuous; } g^{(r)}\in L^2([0,1])\},$$

equipped with the squared norm

$$\|g\|^2 = \sum_{v=0}^{r-1}\left\{\int_0^1 g^{(v)}(t)dt\right\}^2 + \int_0^1\left\{g^{(r)}(t)\right\}^2 dt.$$

Under the assumption $\mathbb{E}\|X\|^2_{\mathcal{H}(K)}<\infty$, Cai and Yuan (2010) showed that $C_0\in\mathcal{H}(K\otimes K)$, where $\mathcal{H}(K\otimes K)$ is the tensor product RKHS equipped with the norm $\|\cdot\|_{\mathcal{H}(K\otimes K)}$ and the reproducing kernel

$$K\otimes K((s_1,t_1),(s_2,t_2)) = K(s_1,s_2)K(t_1,t_2), \quad s_1,s_2,t_1,t_2\in\mathcal{T}.$$

This motivates us to adopt a tensor product RKHS modeling of C_0. With slight abuse of notation, we hereafter also use the notation $\otimes$ to denote the tensor product of functions, i.e., $f\otimes g(s,t) = f(s)g(t)$.

2.1. Spectral decomposition on RKHS

We first introduce spectral decomposition in RKHS and then define a variety of spectral regularizations which we will use to obtain a class of covariance function estimators.

For a bivariate function $C(\cdot,\cdot)$ on $\mathcal{T}\times\mathcal{T}$, define its transpose, denoted by $C^\top$, as $C^\top(s,t) = C(t,s)$ for any $s,t\in\mathcal{T}$. Due to the symmetry of covariance functions, we focus on the space $\mathcal{S}(K) = \{C\in\mathcal{H}(K\otimes K) : C = C^\top\}$. For any $C\in\mathcal{S}(K)$, define its self-adjoint operator $\mathcal{C}_C : \mathcal{H}(K)\to\mathcal{H}(K)$ by

$$(\mathcal{C}_C f)(s) = \langle C(s,\cdot), f(\cdot)\rangle_{\mathcal{H}(K)}, \qquad \text{for any } f\in\mathcal{H}(K) \text{ and } s\in\mathcal{T}. \tag{1}$$

Note that $\|C\|_{\mathcal{H}(K\otimes K)} < \infty$ since $C \in \mathcal{S}(K)$ and that the Hilbert-Schmidt norm of $\mathcal{C}_C$ coincides with $\|C\|_{\mathcal{H}(K\otimes K)}$. Therefore, $\mathcal{C}_C$ is a Hilbert-Schmidt operator and hence admits a spectral decomposition. In Section 2.2, we will define a penalty function based on this spectral decomposition.

Note that the spectral decomposition of $\mathcal{C}_C$ is different from that of the Hilbert-Schmidt integral operator $\mathcal{L}_C : L^2(\mathcal{T}) \to L^2(\mathcal{T})$ as is often used in the FDA literature:

$$(\mathcal{L}_C f)(s) = \langle C(s,\cdot), f(\cdot)\rangle_{L^2(\mathcal{T})} = \int_{\mathcal{T}} C(s,t)f(t)\,dt, \quad \text{for any } f \in L^2(\mathcal{T}) \text{ and } s \in \mathcal{T}. \tag{2}$$

There are two reasons why we adopt $\mathcal{C}_C$ instead of $\mathcal{L}_C$. First, $\mathcal{C}_C$ is more aligned with the RKHS modeling of X, especially when the inner product of $\mathcal{H}(K)$ is chosen to mimic the physical reality. Many examples can be found in the work on L-splines, e.g., Chapter 4.5 in Gu (2013) and Chapter 21 of Ramsay and Silverman (2005). Second, using $\mathcal{C}_C$ enables a finite dimensional representation of our proposed covariance estimators as in Theorem 1 below, and thus simplifies practical computation.

2.2. Spectrally regularized covariance estimator

For any $C \in \mathcal{S}(K)$, let $\tau_1(C), \tau_2(C), \ldots$ be the eigenvalues corresponding to the spectral decomposition of $\mathcal{C}_C$ such that $|\tau_1(C)| \geq |\tau_2(C)| \geq \cdots$. We propose the following covariance estimator:

$$\hat{C} = \underset{C\in\mathcal{S}^+(K)}{\arg\min} \left\{\ell(C) + \lambda\Psi(C)\right\}, \tag{3}$$

where $\mathcal{S}^+(K) = \{C \in \mathcal{S}(K) : \langle \mathcal{C}_C f, f\rangle_{\mathcal{H}(K)} \geq 0, \text{ for all } f \in \mathcal{H}(K)\}$ contains all positive semi-definite functions in $\mathcal{H}(K \otimes K)$, ℓ is a convex and smooth loss function that depends on C through $\{C(T_{ij}, T_{ik}) : i = 1,\ldots,n; j,k = 1,\ldots,m\}$, $\lambda > 0$ is a tuning parameter, and $\Psi(C) = \sum_{k\geq 1}\psi(|\tau_k(C)|)$ with ψ being a non-decreasing penalty function satisfying $\psi(0) = 0$ (Abernethy et al., 2009).

Obviously, the covariance estimator $\hat{C}$ is obtained by one step for a given λ. Moreover, regardless of the form of Ψ, the estimator is always positive semi-definite since the solution to the minimization (3) is searched only within $\mathcal{S}^+(K)$.

The choice of ψ, and thus Ψ, is broad. Below we list a few interesting forms and briefly discuss their effects on the estimator.

Example 1 (Rank regularization)**.** If $\psi(\tau) = I(\tau \neq 0)$ where $I(\cdot)$ is the indicator function, $\Psi(C)$ is the rank of the operator $\mathcal{C}_C$. This penalty obviously encourages a low-rank solution. However, the minimization (3) is now difficult owing to its non-convexity, and over-fitting may occur since no regularizations are imposed on non-zero eigenvalues.

Example 2 (Hilbert-Schmidt-norm regularization)**.** If $\psi(\tau) = \tau^2$, $\Psi(C)$ becomes the squared Hilbert-Schmidt norm of the operator $\mathcal{C}_C$, which equals $\|C\|^2_{\mathcal{H}(K\otimes K)}$. Similar to the ℓ_2-norm regularization for vectors, the Hilbert-Schmidt-norm regularization ensures the convexity of the objective function in (3), but does not encourage sparsity in eigenvalues, so the resulting covariance estimator is usually of high rank. Cai and Yuan (2010) used this regularization to estimate C_0 under the RKHS framework, but did not impose the constraint $C \in \mathcal{S}^+(K)$, so neither positive semi-definiteness nor low rank can be guaranteed for their estimator.

Example 3 (Trace-norm regularization)**.** If $\psi(\tau) = \tau$, $\Psi(C)$ is the trace norm of $\mathcal{C}_C$. Similar to the celebrated ℓ_1-regularization for vectors and the trace-norm regularization for matrices, the trace-norm penalty Ψ for operators not only promotes the sparsity of eigenvalues and hence low-rank solutions, but also regularizes non-zero eigenvalues. The minimization (3) now becomes convex but nondifferentiable, which allows leveraging recent developments in non-smooth convex optimizations (e.g., Beck and Teboulle, 2009) to achieve feasible computations. See Section 3 for more details.

2.3. Representer theorem

Since commonly used $\mathcal{H}(K)$, e.g., $\mathcal{W}^r$, are infinite dimensional, solving (3) is typically an infinite dimensional optimization problem. Therefore, it is not obvious whether $\hat{C}$ can be computed in practice. To answer this question, we establish a representer theorem that provides a finite dimensional representation of $\hat{C}$. This theorem holds for the entire class of estimators in (3).

Write $N = nm$ and $(\tilde{T}_1, \ldots, \tilde{T}_N) = (T_{11}, \ldots, T_{1m}, T_{21}, \ldots, T_{2m}, \ldots, T_{n1}, \ldots, T_{nm})$.

Theorem 1 (Representer theorem)**.** *If the solution set of (3) is not empty, then there always exists a solution lying in the space* $\mathcal{K} \otimes \mathcal{K} = \mathrm{span}\{K(\cdot, \tilde{T}_i) \otimes K(\cdot, \tilde{T}_j) : i, j = 1, \ldots, N\}$*, where* $\mathcal{K} = \mathrm{span}\{K(\cdot, \tilde{T}_i) : i = 1, \ldots, N\}$*. Moreover, the solution takes the form:*

$$C(s,t) = z(s)^\top A\, z(t), \tag{4}$$

where A *is an* $N \times N$ *symmetric matrix and* $z(\cdot) = (K(\cdot, \tilde{T}_1), \ldots, K(\cdot, \tilde{T}_N))^\top$*.*

Classical representer theorems (e.g., Wahba, 1990), as adopted in Cai and Yuan (2010), do not cover the scenario addressed by Theorem 1 due to the semi-positivity constraint and a wide choice of regularizations, e.g., the trace-norm regularization. To show this theorem, we significantly utilize the fact that the spectral analysis is based on the RKHS geometry. This is the main reason for using the operator $\mathcal{C}_C$ in (1) instead of $\mathcal{L}_C$ in (2). We remark that the conclusion of Theorem 1 also holds when the semi-positivity is not imposed, i.e., $\mathcal{S}^+(K)$ is replaced by $\mathcal{S}(K)$ in (3). In Section 4, this fact will be used to compute unconstrained estimators for comparison.

At a first glance, a significant number of scalar parameters, i.e., $(N+1)N/2$, is involved in the solution (4). However, if a low-rank inducing penalty, such as the trace-norm regularization, is used, the resulting estimator is often of low rank, which will benefit computation and storage of its estimation, and subsequent uses.

2.4. Parametrization

By Theorem 1, we are able to parametrize the solution to (3) in terms of a finite dimensional representation since it suffices to merely focus on covariance functions of the form $C(\cdot,\cdot) = \sum_{i=1}^N \sum_{j=1}^N A_{ij} K(\cdot, \tilde{T}_i) \otimes K(\cdot, \tilde{T}_j)$. The eigenvalues of the operator $\mathcal{C}_C$, $\{\tau_j(C) : j \geq 1\}$, are the eigenvalues of the matrix $B = M^\top A M$, where M is any $N \times q$ matrix such that $MM^\top = \tilde{K} = [K(\tilde{T}_i, \tilde{T}_j)]_{1 \leq i,j \leq N}$ with $q = \mathrm{rank}(\tilde{K})$. The matrix M provides a representation based on an orthonormal basis (of $\mathcal{K}$) $\{v_1, \ldots, v_q\}$:

$$K(\cdot, \tilde{T}_i) \otimes K(\cdot, \tilde{T}_j) = \sum_{k=1}^q \sum_{l=1}^q M_{ik} M_{jl} v_k \otimes v_l, \qquad 1 \leq i, j \leq N.$$

Therefore $C = \sum_{k=1}^{q}\sum_{l=1}^{q} B_{kl} v_k \otimes v_l$. As $C(s,t) = \langle K(\cdot,s), \mathcal{C}_C K(\cdot,t)\rangle_{\mathcal{H}(K)}$, we have $[C(\tilde{T}_i, \tilde{T}_j)]_{1\le i,j\le N} = MBM^\top$. Moreover, write $M = (M_1^\top, \ldots, M_n^\top)^\top$, where $\{M_i : i = 1, \ldots, n\}$ are $m \times q$ matrices. Then the loss function depends on C through $[C(T_{ij}, T_{ik})]_{1\le j,k\le m} = M_i B M_i^\top$ for $i = 1, \ldots, n$. Compared with A, the new parametrization B is unique even when $\{T_{ij} : i = 1, \ldots, n; j = 1, \ldots, m\}$ are not all unique. Now (3) can be rewritten as

$$\arg\min_{B \in \mathcal{S}_q^+} \left\{ \tilde{\ell}(B) + \lambda \tilde{\Psi}(B) \right\}, \tag{5}$$

where $\mathcal{S}_q^+$ is the set of all $q \times q$ positive semi-definite matrices, $\tilde{\ell}(B) = \ell(\sum_{k=1}^{q}\sum_{l=1}^{q} B_{kl} v_k \otimes v_l)$, and $\tilde{\Psi}(B) = \sum_{k=1}^{q} \psi(|\xi_k(B)|)$ with $\xi_1(B), \ldots, \xi_q(B)$ being the eigenvalues of the matrix B such that $|\xi_1(B)| \ge \cdots \ge |\xi_q(B)|$. Conversely, with the new parametrization B, we can represent $C(s,t) = z(s)^\top (M^+)^\top B M^+ z(t)$ where M^+ is the Moore-Penrose pseudoinverse of M. Consequently, solving (3) is equivalent to solving (5), a finite-dimensional optimization.

2.5. Explicit expression of L^2 eigen-decomposition

By Mercer's theorem, we can represent an arbitrary covariance function $C \in \mathcal{S}^+(K)$ in terms of the typical spectral decomposition via the L^2 inner product, i.e., $C(s,t) = \sum_{k\ge 1} \zeta_k \phi_k(s)\phi_k(t)$, where $\{\phi_k : k \ge 1\}$ are the L^2 eigenfunctions and $\{\zeta_k : k \ge 1\}$ are the corresponding L^2 eigenvalues. This eigen-decomposition is a key component of FPCA among other FDA methods. In the literature, approximate computations are commonly involved where the eigen-decomposition is based on the discretized covariance function estimator (e.g., Rice and Silverman, 1991). In contrast, due to Theorem 1, our covariance estimator leads to an explicit expression of this eigen-decomposition so that such computational complication can be avoided.

Following the notation in Sections 2.3 and 2.4, let $Q = [\int_{\mathcal{T}} K(s,\tilde{T}_i)K(s,\tilde{T}_j)\,ds]_{1\le i,j\le N} = MRM^\top$ where $R = [\int_{\mathcal{T}} v_k(s)v_l(s)\,ds]_{1\le k,l\le q} = M^+Q(M^+)^\top$. Similar to Lemma 3 of Cai and Yuan (2010), once $\hat{B}$ is obtained from (5), the L^2 eigenfunctions of $\hat{C} = \sum_{k=1}^{q}\sum_{l=1}^{q} \hat{B}_{kl} v_k \otimes v_l$ can be expressed as $\hat{\phi}_k(\cdot) = U_k^\top z(\cdot), k = 1, \ldots, n$, where U_k is the k-th column of $U = (M^+)^\top R^{-1/2} V$ and V is the eigenvectors of $R^{1/2}\hat{B}R^{1/2}$. The L^2 eigenvalues of $\hat{C}$ coincide with those of $R^{1/2}\hat{B}R^{1/2}$, and the number of nonzero eigenvalues is exactly the rank of $\hat{B}$.

3. Computational issues

To achieve a desirable low-rank covariance estimator, in this section we only focus on the trace-norm regularization and develop Algorithm 1 below. A similar algorithm can be obtained for the Hilbert-Schmidt-norm regularization, which will be implemented in Section 4, and its details are given in Section S1 of the supplementary material.

3.1. Algorithm

By (5), with the trace-norm regularization in (3), it is equivalent to solving

$$\arg\min_{B\in\mathcal{S}_q^+}\left\{\tilde{\ell}(B)+\lambda\|B\|_*\right\}, \tag{6}$$

where $\|\cdot\|_*$ represents the typical trace norm for matrices. We can also rewrite (6) as

$$\arg\min_{B\in\mathcal{S}_q}\left\{\tilde{\ell}(B)+\lambda h(B)\right\}, \quad \text{where} \quad h(B)=\begin{cases}\|B\|_*, & B\in\mathcal{S}_q^+\\ \infty, & B\notin\mathcal{S}_q^+\end{cases}. \tag{7}$$

Here $\mathcal{S}_q$ represents the set of all $q\times q$ matrices.

The objective function in (7) is the sum of a smooth function $\tilde{\ell}(\cdot)$ and a non-smooth function $\lambda h(\cdot)$. A popular approach to such optimizations is the accelerated proximal gradient (APG) method (Beck and Teboulle, 2009). To apply this method, define an operator $\mathrm{svec}:\mathcal{S}_q\to\mathbb{R}^{q(q+1)/2}$ by

$$\mathrm{svec}(B)=[B_{11},\sqrt{2}B_{21},\ldots,\sqrt{2}B_{q1},B_{22},\sqrt{2}B_{32},\ldots,\sqrt{2}B_{q2},\ldots,B_{qq}]^\top,$$

for any $B=[B_{ij}]_{1\le i,j\le q}\in\mathcal{S}_q$. This operator provides an isometry between $\mathcal{S}_q$ and $\mathbb{R}^{q(q+1)/2}$. Denote its inverse by svec^{-1} and write $\check{\ell}(b)=\tilde{\ell}(\mathrm{svec}^{-1}(b))$ for any $b\in\mathbb{R}^{q(q+1)/2}$. The APG algorithm of our case involves the proximal operator $\mathrm{prox}_\nu:\mathbb{R}^{q(q+1)/2}\to\mathbb{R}^{q(q+1)/2}$ defined by

$$\begin{aligned}\mathrm{prox}_\nu(b)&=\arg\min_{d\in\mathbb{R}^{q(q+1)/2}}\left\{\frac{1}{2}\|d-b\|_E^2+\nu h(\mathrm{svec}^{-1}(d))\right\}\\&=\mathrm{svec}\left[\arg\min_{D\in\mathcal{S}_q^+}\left\{\frac{1}{2}\|D-B\|_F^2+\nu\|D\|_*\right\}\right],\end{aligned}$$

for any $b \in \mathbb{R}^{q(q+1)/2}$ and $\nu > 0$. Here $\|\cdot\|_E$ and $\|\cdot\|_F$ represent the Euclidean norm and Frobenius norm respectively. The following proposition states the closed-form solution of this proximal operator.

Proposition 1. *For any $\nu > 0$ and $b \in \mathbb{R}^{q(q+1)/2}$ with eigen-decomposition $\mathrm{svec}^{-1}(b) = P\mathrm{diag}(\tilde{b})P^\top$,*

$$\mathrm{prox}_\nu(b) = \mathrm{svec}(P\mathrm{diag}(\tilde{c})P^\top),$$

where $\tilde{c} = (g_\nu(\tilde{b}_1), \ldots, g_\nu(\tilde{b}_q))^\top$ and $\tilde{b} = (\tilde{b}_1, \ldots, \tilde{b}_q)^\top$. Here $g_\nu(x) = (x - \nu)_+$ for any $x \in \mathbb{R}$.

Due to this closed-form solution, we can avoid an inner numerical optimization within every iteration of the APG algorithm. The proof uses the same technique as in the proof of Lemma 1 in Mazumder et al. (2010), and is thus omitted.

The standard APG method requires the knowledge of the Lipschitz constant of $\nabla\check{\ell}$, which directly relates to the step size in each iteration of the algorithm. For many choices of the loss function ℓ, the corresponding Lipschitz constant of $\nabla\check{\ell}$ is difficult to obtain. Moreover, even when the Lipschitz constant is known (e.g., the choice of ℓ described in Section 3.2), the algorithm usually suffers from conservative step sizes (Becker et al., 2011). Hence the APG with backtracking steps is usually preferred. Following the suggestions of Becker et al. (2011), we adopt a modified version of the APG method and develop Algorithm 1.

Modifying lines 7–8 in Algorithm 1 will result in other variants of proximal gradient methods. More discussions are in Section 5.2 of Becker et al. (2011). For convergence properties of the APG method, we refer interested readers to Beck and Teboulle (2009).

3.2. A choice of ℓ

The choice of ℓ is broad, but hereafter we only focus on the following quadratic loss:

$$\ell(C) = \frac{1}{nm(m-1)} \sum_{i=1}^{n} \sum_{1 \le j \ne k \le m} \{Z_{ijk} - C(T_{ij}, T_{ik})\}^2, \tag{8}$$

Algorithm 1: The APG algorithm with backtracking for trace-norm-regularized covariance estimation

Input: $B_0 \in \mathcal{S}_q^+$, $\hat{L} > 0$, $\eta > 1$, $\alpha < 1$

$b_0 \leftarrow \mathrm{svec}(B_0)$, $\bar{b}_0 \leftarrow b_0$, $\theta_{-1} \leftarrow +\infty$, $L_{-1} \leftarrow \hat{L}$
for $k = 0, 1, 2, \ldots$ **do**
  $L_k \leftarrow \alpha L_{k-1}$
  repeat
    $\theta_k \leftarrow 2/[1 + \{1 + 4L_k/(L_{k-1}\theta_{k-1}^2)\}^{1/2}]$
    $e_k \leftarrow (1 - \theta_k)b_k + \theta_k \bar{b}_k$
    $b_{k+1} \leftarrow \mathrm{prox}_{\lambda/L_k}(e_k - \nabla\check{\ell}(e_k)/L_k)$
    $\bar{b}_{k+1} \leftarrow \{b_{k+1} - (1 - \theta_k)b_k\}/\theta_k$
    $\hat{L} \leftarrow 2|(e_k - b_{k+1})^\top\{\nabla\check{\ell}(b_{k+1}) - \nabla\check{\ell}(e_k)\}|/\|b_{k+1} - e_k\|_E^2$
    if $L_k \geq \hat{L}$ **then**
      break
    $L_k \leftarrow \max\{\eta L_k, \hat{L}\}$
  until *convergence*;

where $Z_{ijk} = \{Y_{ij} - \hat{\mu}(T_{ij})\}\{Y_{ik} - \hat{\mu}(T_{ik})\}$ and $\hat{\mu}$ is an estimator of the mean function μ_0. Since $[C(T_{ij}, T_{ik})]_{1\leq j,k\leq m} = M_i B M_i^\top$ as shown in Section 2.4, $\ell(C)$ becomes

$$\begin{aligned}\tilde{\ell}(B) &= \frac{1}{nm(m-1)}\sum_{i=1}^{n}\|\rho(Z_i - M_i B M_i^\top)\|_F^2 \\ &= \frac{1}{2}\mathrm{vec}(B)^\top\nabla^2\tilde{\ell}(B)\mathrm{vec}(B) - \left(\sum_{i=1}^{n} M_i^\top Z_i M_i\right)^\top \mathrm{vec}(B),\end{aligned}$$

up to an additive constant independent of B. Here $Z_i = [Z_{ijk}]_{1\leq j,k\leq m}$, ρ is an operator setting the diagonal entries of its input to zero, and

$$\nabla^2\tilde{\ell}(B) = \frac{2}{nm(m-1)}\sum_{i=1}^{n}(M_i^\top \otimes M_i^\top)\mathrm{diag}\{\mathrm{vec}(\tilde{I})\}(M_i \otimes M_i),$$

with $\tilde{I} \in \mathbb{R}^{q\times q}$ consisting of elements $\tilde{I}_{ij} = I(i \neq j)$. By simple derivations, one can obtain the closed-form expressions of $\check{\ell}$ and $\nabla\check{\ell}$ as required in Algorithm 1, which we omit here.

4. Numerical experiments

Numerical experiments were conducted to illustrate the practical performance of the proposed methodology. We generated $\{X_i : i = 1, \ldots, n\}$ from a Gaussian process with $\mu_0(t) = 3\sin\{3\pi(t + 0.5)\} + 2t^3$ and $C_0(s,t) = \sum_{k=1}^{L} \zeta_k \phi_k(s)\phi_k(t)$, with $\zeta_k = 1/(k+1)^2$, $\phi_{2l-1}(t) = 2^{1/2}\cos(2\pi lt)$, and $\phi_{2l}(t) = 2^{1/2}\sin(2\pi lt)$ for any positive integers k, l. We also sampled $\{T_{ij} : i = 1, \ldots, n; j = 1, \ldots m\}$ independently from the uniform distribution on $[0,1]$ and $\{\varepsilon_{ij} : i = 1, \ldots, n; j = 1, \ldots m\}$ independently from $N(0, \sigma^2)$ with $\sigma^2 = 0.01$ to produce $Y_{ij} = X_i(T_{ij}) + \varepsilon_{ij}$. We studied 18 settings in total, where $L = 2$, 4 or 10, $n = 50$, 100 or 200, and $m = 10$ or 20. In each setting, we simulated 300 datasets where we compared various covariance function estimators. Other than our proposed estimators, we also studied popular alternatives, including the covariance smoothing estimators by local polynomial regression (Yao et al., 2005a) and bivariate P-splines (Goldsmith et al., 2011; Xiao et al., 2018) respectively, and the restricted maximum likelihood (REML) estimator by Peng and Paul (2009). For the first two alternative methods, a raw smoothed estimate is first computed. Then a truncation step via FPCA is applied to reconstruct a covariance function that is both positive semi-definite and of low rank. That means, the reconstructed covariance estimator, which we refer to as a *truncated* estimator below, takes the form: $\sum_{k=1}^{J} \hat{\zeta}_k \hat{\phi}_k(s)\hat{\phi}_k(t)$ where $\hat{\zeta}_k$'s and $\hat{\phi}_k$'s are the largest J positive estimated eigenvalues and corresponding estimated eigenfunctions based on the raw smoothed estimate.

Altogether we compared the following ten estimators, of which the first five are based on our proposed framework while the rest are popular alternatives: 1) $\hat{C}^{+}_{\mathsf{trace}}$: obtained from (3) with the trace-norm regularization; 2) $\hat{C}_{\mathsf{trace}}$: obtained from (3) with the trace-norm regularization but without the semi-positivity constraint, i.e., $\mathcal{S}^{+}(K)$ replaced by $\mathcal{S}(K)$; 3) $\hat{C}^{+}_{\mathsf{HS}}$: obtained from (3) with the Hilbert-Schmidt norm regularization; 4) $\hat{C}_{\mathsf{HS}}$: obtained from (3) with the Hilbert-Schmidt norm regularization but without the semi-positivity constraint. 5) $\hat{C}_{\mathsf{CY}}$: the estimator proposed by Cai and Yuan (2010) and implemented in their R package (`http://stat.wharton.upenn.edu/~tcai/paper/html/Covariance-Function.html`); 6) $\hat{C}_{\mathsf{PACE}}$: the raw smoothed covariance estimator using local polynomial regression (Yao et al., 2005a), implemented in the R package fdapace; 7) $\hat{C}^{+}_{\mathsf{PACE,BIC}}$: the truncated estimator based on $\hat{C}_{\mathsf{PACE}}$ with J selected by Baysian Information

Criterion (BIC), implemented in fdapace; 8) $\hat{C}^{+}_{\mathsf{FACE}}$: the truncated estimator based on a generalized version of bivariate P-spline smoothing (Xiao et al., 2018) with J selected by fraction of variation explained FVE= 0.99 by default, implemented in the R package face; 9) $\hat{C}^{+}_{\mathsf{SC}}$: the truncated estimator based on tensor product bivariate P-spline smoothing (Goldsmith et al., 2011) with J selected by FVE= 0.99 by default, implemented in the R package refund; 10) $\hat{C}^{+}_{\mathsf{PP}}$: the REML estimator based on a Newton-Raphson procedure on the Stiefel manifold (Peng and Paul, 2009), implemented in the R package fpca, with the number of B-spline basis functions and rank ranging in $[4, 20]$ and $[2, 7]$ respectively. We remark that, for the estimators based on local polynomial regression, fdapace provides other methods for choosing J such as Akaike Information Criterion and FVE in addition to BIC. Since these three methods give similar results, here we only report $\hat{C}^{+}_{\mathsf{PACE,BIC}}$ which achieves the best numerical performance among them. Besides, both refund and face do not output raw smoothed estimates so we did not compare them.

To obtain $\hat{C}^{+}_{\mathsf{trace}}, \hat{C}_{\mathsf{trace}}, \hat{C}^{+}_{\mathsf{HS}}, \hat{C}_{\mathsf{HS}}$ and $\hat{C}_{\mathsf{CY}}$, a smoothing spline was first applied to estimate μ, where its smoothing parameter was selected by generalized cross-validation (GCV), ignoring the functional data structure. To obtain each of these five covariance estimators, we always used the loss function ℓ as in Section 3.2, and $\mathcal{H}(K) = \mathcal{W}^2$ with the squared norm $\|g\|^2 = \sum_{v=0}^{1}\{\int_0^1 g^{(v)}(t)dt\}^2 + \int_0^1\{g^{(2)}(t)\}^2 dt$. The tuning parameters λ of the first four methods were chosen by five-fold cross-validation (CV), with $n/5$ curves in each fold. The computations of $\hat{C}_{\mathsf{trace}}$, $\hat{C}^{+}_{\mathsf{HS}}$ andd $\hat{C}_{\mathsf{HS}}$ were achieved by Algorithm 1 with different proximal operators (line 7 of Algorithm 1) due to the change of penalty and semi-positivity constraint. For the remaining five methods, μ is estimated by the corresponding computational packages. See their documentation for further implementation details.

4.1. Comparisons between variations of (3)

First, we restrict our attention to the first five methods which can all be regarded as variations of (3). Table 1 shows the average integrated squared errors (AISE) and average ranks of these covariance estimators over 300 simulated data sets, which reflect their performances in estimation accuracy and rank reduction respectively. Due to space limitations, we only present those settings with $n = 50$ and $m = 20$, which have an n-to-m ratio most similar to the real data in Section

5. We can obtain similar conclusions in other settings, which are reported in the supplementary material. Although $\hat{C}_{\mathsf{CY}}$ and $\hat{C}_{\mathsf{HS}}$ share the same definition, they differ in various implementation details and hence the practical performance. However, their differences in AISE are too small to affect subsequent comparisons, so hereafter it suffices to simply focus only on $\hat{C}_{\mathsf{HS}}$, rather than both of them, to study rank reduction and the effect of the semi-positivity constraint.

When we compare the two pairs, $\hat{C}^{+}_{\mathsf{trace}}$ versus $\hat{C}_{\mathsf{trace}}$, and $\hat{C}^{+}_{\mathsf{HS}}$ versus $\hat{C}_{\mathsf{HS}}$, obviously the covariance estimators with the positivity constraint always achieve smaller AISE values than their counterparts. This suggests that not only can the semi-positivity constraint produce a valid estimator, but also improve estimation accuracy. Notice in Table 1 that rank reduction can also be observed for $\hat{C}^{+}_{\mathsf{HS}}$ because the semi-positivity constraint often results in the truncation of eigenvalues at zero. When $\hat{C}^{+}_{\mathsf{trace}}$ is compared with $\hat{C}^{+}_{\mathsf{HS}}$, the former performs slightly worse in AISE, but significantly better in rank reduction. For settings with $L = 2$ or 4, the average ranks of $\hat{C}^{+}_{\mathsf{trace}}$ are in fact the closest to the true rank among the three rank-reduced estimators. This highlights the benefits of trace-norm regularizations in computation, storage and subsequent uses as mentioned in Section 1. Next we only compare $\hat{C}^{+}_{\mathsf{trace}}$ and $\hat{C}^{+}_{\mathsf{HS}}$ with popular alternatives.

Table 1: AISE ($\times 10^3$) values with standard errors ($\times 10^3$) in parentheses for the five variations of (3), and average ranks with standard errors in parentheses for those estimators with rank reduction. Only settings with $n = 50, m = 20$ are presented. See full results in the supplementary material.

L	n	m		$\hat{C}^{+}_{\mathsf{trace}}$	$\hat{C}_{\mathsf{trace}}$	$\hat{C}^{+}_{\mathsf{HS}}$	$\hat{C}_{\mathsf{HS}}$	$\hat{C}_{\mathsf{CY}}$
2	50	20	AISE	6.88 (0.290)	7.66 (0.266)	6.70 (0.260)	8.07 (0.263)	7.77 (0.263)
			rank	2.6 (0.051)	7.5 (0.158)	13.0 (0.051)	-	-
4	50	20	AISE	11.64 (0.338)	14.97 (0.311)	11.27 (0.317)	14.98 (0.285)	12.36 (0.316)
			rank	3.8 (0.045)	13.2 (0.614)	13.5 (0.048)	-	-
10	50	20	AISE	15.99 (0.398)	19.12 (0.367)	15.20 (0.370)	19.37 (0.354)	15.74 (0.371)
			rank	3.7 (0.061)	9.7 (0.515)	14.5 (0.045)	-	-

4.2. Comparisons with popular alternatives

Due to space constraints, here we only present the results for the settings with $n = 50$ or 200, and defer the remaining ones (with $n = 100$) to the supplementary material. In Table 2, we report AISE values and average ranks for $\hat{C}^{+}_{\mathsf{trace}}$, $\hat{C}^{+}_{\mathsf{HS}}$, $\hat{C}^{+}_{\mathsf{PACE,BIC}}$, $\hat{C}^{+}_{\mathsf{FACE}}$, and $\hat{C}^{+}_{\mathsf{SC}}$. We exclude $\hat{C}_{\mathsf{PACE}}$ here since it is uniformly worse than $\hat{C}^{+}_{\mathsf{PACE,BIC}}$ in both estimation accuracy and rank reduction, which illustrates the benefit of the reconstruction step. We also omit $\hat{C}^{+}_{\mathsf{PP}}$ in Table 2 since it suffers the

most from computational instability, probably due to the algorithmic convergence issue as reported in Peng and Paul (2009). The fpca package for computing $\hat{C}^+_{\mathsf{PP}}$ failed to provide an output in some simulation runs, and the failure rate may be very high, e.g., 38% when $(L, n, m) = (4, 200, 20)$. See the supplementary material for full results with both $\hat{C}_{\mathsf{PACE}}$ and $\hat{C}^+_{\mathsf{PP}}$ included.

We first focus on the settings with $L = 2$ or 4 in Table 2. As also observed in Table 1, $\hat{C}^+_{\mathsf{trace}}$ has slightly larger AISE values than $\hat{C}^+_{\mathsf{HS}}$ but is considerably superior in rank reduction. The average ranks and AISE values of $\hat{C}^+_{\mathsf{trace}}$ are both much lower than those of $\hat{C}^+_{\mathsf{PACE,BIC}}$ in most settings. When compared with $\hat{C}^+_{\mathsf{FACE}}$, the performance of $\hat{C}^+_{\mathsf{trace}}$ is better throughout all settings in Table 2 except for the estimation accuracy when $(L, n, m) = (4, 50, 10)$. Compared with $\hat{C}^+_{\mathsf{SC}}$, $\hat{C}^+_{\mathsf{trace}}$ achieves similar AISE values and performs slightly but uniformly better in rank reduction. Table 2 also shows that $\hat{C}^+_{\mathsf{trace}}$ is numerically more stable than $\hat{C}^+_{\mathsf{SC}}$, since no computational error appeared for $\hat{C}^+_{\mathsf{trace}}$, but some occurred in a fraction of simulation runs where no output was returned to obtain $\hat{C}^+_{\mathsf{SC}}$. Here the results for $\hat{C}^+_{\mathsf{SC}}$ were computed based on successful runs which has no computational error.

Next we turn to the settings with a high rank $L = 10$, where the covariance function estimation is more difficult. All estimators, except for $\hat{C}^+_{\mathsf{HS}}$, generally shrink the rank to some extent. Regarding estimation accuracy, the performance $\hat{C}^+_{\mathsf{trace}}$ is not as strong as $\hat{C}^+_{\mathsf{HS}}$, which is expected due to the penalization of the trace-norm regularization on high-rank solutions. However, $\hat{C}^+_{\mathsf{trace}}$ surprisingly remains very competitive compared to the other three: It is significantly better than $\hat{C}^+_{\mathsf{PACE,BIC}}$ and $\hat{C}^+_{\mathsf{FACE}}$ for $n = 200$, and similar to, if not slightly better than, $\hat{C}^+_{\mathsf{SC}}$ for $n = 50$.

At last we compare the performances of the five covariance estimators in estimating the principal eigenvalue ζ_1 and principal eigenfunction ϕ_1 in Table 3, where the bias and mean squared error (MSE) for ζ_1, and the AISE for ϕ_1 are given for each method. Here we only report those settings with $n = 50$ and $m = 20$, since similar patterns can be seen in other settings. The full results are reported in the supplementary material, together with those for the second eigen-component. Both $\hat{C}^+_{\mathsf{PACE,BIC}}$ and $\hat{C}^+_{\mathsf{FACE}}$ perform well in estimating ϕ_1, but their eigenvalue estimations are significantly worse than $\hat{C}^+_{\mathsf{trace}}$, $\hat{C}^+_{\mathsf{HS}}$ and $\hat{C}^+_{\mathsf{SC}}$. The biases for ζ_1 are negative for most methods, which indicates that their eigenvalue estimates are on average smaller than the true value. The performances of $\hat{C}^+_{\mathsf{trace}}$ and $\hat{C}^+_{\mathsf{HS}}$ are very similar in both eigenvalue and eigenfunction estimations. Despite not

always being the smallest, the magnitude of their biases for ζ_1 is sufficiently small ($< 1.5 \times 10^{-2}$) compared with the true principal eigenvalue $\zeta_1 = 0.25$. Their MSEs for ζ_1 and AISEs for ϕ_1 are usually smaller than those of $\hat{C}^+_{\mathsf{SC}}$, and always among the smallest.

In summary, the overall performance of $\hat{C}^+_{\mathsf{trace}}$ is the best regarding rank reduction, estimation accuracy, and computational stability. This conclusion is also confirmed by an additional simulation study in the supplementary material with a higher error variance $\sigma^2 = 0.1$. This motivates us to use $\hat{C}^+_{\mathsf{trace}}$ in the following real data application.

Table 2: AISE ($\times 10^3$) values with standard errors ($\times 10^3$) in parentheses, and average ranks with standard errors in parentheses, for $\hat{C}^+_{\mathsf{trace}}$, $\hat{C}^+_{\mathsf{HS}}$, $\hat{C}^+_{\mathsf{PACE,BIC}}$, $\hat{C}^+_{\mathsf{FACE}}$, and $\hat{C}^+_{\mathsf{SC}}$. For $\hat{C}^+_{\mathsf{SC}}$, its statistics in each setting are computed only based on successful runs, that is, those simulation runs where its corresponding package does not return an output due to computational errors are not counted, with the proportion of successful runs additionally shown in square brackets.

L	n	m		$\hat{C}^+_{\mathsf{trace}}$	$\hat{C}^+_{\mathsf{HS}}$	$\hat{C}^+_{\mathsf{PACE,BIC}}$	$\hat{C}^+_{\mathsf{FACE}}$	$\hat{C}^+_{\mathsf{SC}}$
2	50	10	AISE	9.25 (0.327)	9.00 (0.310)	11.94 (0.314)	9.37 (0.616)	10.64 (0.364)
			RANK	2.6 (0.031)	13.4 (0.049)	5.2 (0.047)	4.0 (0.036)	4.2 (0.039)
2	50	20	AISE	6.88 (0.290)	6.70 (0.260)	9.54 (0.253)	9.73 (0.967)	7.59 (0.296)
			RANK	2.6 (0.051)	13.0 (0.051)	4.5 (0.044)	3.7 (0.033)	3.9 (0.037)
2	200	10	AISE	2.85 (0.092)	2.81 (0.089)	6.37 (0.126)	3.40 (0.294)	3.18 (0.098)
			RANK	2.7 (0.033)	14.5 (0.043)	4.9 (0.045)	4.1 (0.033)	4.1 (0.037)
2	200	20	AISE	2.07 (0.078)	2.04 (0.077)	5.56 (0.112)	3.58 (0.393)	2.23 (0.080)
			RANK	2.7 (0.036)	14.3 (0.045)	4.4 (0.044)	4.0 (0.025)	3.9 (0.034)
4	50	10	AISE	17.04 (0.416)	15.94 (0.395)	15.56 (0.335)	14.69 (0.555)	16.09 (0.516) [99.7%]
			RANK	3.1 (0.055)	13.8 (0.047)	5.7 (0.049)	5.0 (0.036)	5.1 (0.039) [99.7%]
4	50	20	AISE	11.64 (0.338)	11.27 (0.317)	12.34 (0.294)	12.87 (0.869)	11.42 (0.360)
			RANK	3.8 (0.045)	13.5 (0.048)	5.2 (0.047)	5.2 (0.029)	5.0 (0.032)
4	200	10	AISE	4.94 (0.107)	4.74 (0.097)	8.61 (0.136)	6.50 (0.495)	4.64 (0.104)
			RANK	4.4 (0.032)	14.9 (0.046)	5.5 (0.047)	5.6 (0.030)	5.4 (0.032)
4	200	20	AISE	3.27 (0.081)	3.20 (0.080)	7.58 (0.116)	4.70 (0.295)	3.14 (0.081)
			RANK	4.5 (0.029)	15.0 (0.044)	5.0 (0.042)	5.7 (0.030)	5.1 (0.030)
10	50	10	AISE	19.99 (0.491)	18.45 (0.420)	17.74 (0.409)	18.70 (0.847)	20.83 (0.718) [99.7%]
			RANK	3.1 (0.054)	14.3 (0.045)	6.1 (0.048)	5.2 (0.041)	6.0 (0.036) [99.7%]
10	50	20	AISE	15.99 (0.398)	15.20 (0.370)	14.84 (0.308)	20.08 (1.754)	16.00 (0.445)
			RANK	3.7 (0.061)	14.5 (0.045)	6.1 (0.045)	5.6 (0.040)	6.7 (0.033)
10	200	10	AISE	8.08 (0.158)	7.54 (0.144)	10.41 (0.160)	10.35 (0.694)	7.14 (0.166) [96.0%]
			RANK	4.7 (0.052)	15.9 (0.048)	6.4 (0.045)	6.4 (0.036)	6.9 (0.035) [96.0%]
10	200	20	AISE	5.42 (0.099)	5.09 (0.087)	9.23 (0.124)	10.52 (0.611)	4.57 (0.096)
			RANK	5.8 (0.068)	16.5 (0.044)	6.5 (0.040)	7.2 (0.031)	7.7 (0.030)

4.3. Computation times

In this section we performed a simple experiment to evaluate the practicality of our positive semi-definite estimators and popular alternatives in terms of computational times. The time in seconds was recorded for each method to fit the real data described in Section 5, with $n = 78$ and $m = 31$, on a laptop computer (Macbook Pro with a 2.8 GHz Intel Core i7 processor). We repeated

Table 3: Bias ($\times 10^2$) and MSE ($\times 10^4$) values with their standard errors (multiplied by 10^2 and 10^4 respectively) in parentheses for the principal eigenvalue ζ_1, and AISE ($\times 10^2$) values with standard errors ($\times 10^2$) in parentheses for the principal eigenfunction ϕ_1.

L	n	m		$\hat{C}^+_{\mathsf{trace}}$	$\hat{C}^+_{\mathsf{HS}}$	$\hat{C}^+_{\mathsf{PACE,BIC}}$	$\hat{C}^+_{\mathsf{FACE}}$	$\hat{C}^+_{\mathsf{SC}}$
2	50	20	ζ_1(BIAS)	-1.166 (0.29)	-1.464 (0.29)	-4.971 (0.25)	-0.188 (0.41)	-0.675 (0.30)
			ζ_1(MSE)	27.16 (2.32)	26.71 (2.23)	43.13 (2.43)	49.96 (6.64)	26.89 (2.34)
			ϕ_1(AISE)	5.53 (0.410)	5.34 (0.378)	4.85 (0.263)	5.89 (0.509)	6.09 (0.448)
4	50	20	ζ_1(BIAS)	-0.485 (0.29)	-0.761 (0.29)	-4.865 (0.25)	0.864 (0.39)	-0.021 (0.30)
			ζ_1(MSE)	25.90 (2.16)	25.51 (2.06)	43.03 (2.50)	45.34 (6.26)	27.52 (2.39)
			ϕ_1(AISE)	7.66 (0.481)	7.41 (0.477)	6.34 (0.317)	7.84 (0.637)	8.39 (0.484)
10	50	20	ζ_1(BIAS)	-0.176 (0.32)	-0.509 (0.32)	-4.430 (0.28)	1.506 (0.52)	0.635 (0.33)
			ζ_1(MSE)	31.51 (2.48)	31.50 (2.44)	42.67 (2.61)	84.57 (14.41)	32.39 (2.66)
			ϕ_1(AISE)	9.53 (0.740)	9.41 (0.765)	7.85 (0.520)	8.97 (0.726)	11.58 (0.874)

the experiment five times with the same seed of randomness, and report their average computing times shown in Table 4, so as to remove the random effect in the computing environment. Our proposed method takes advantage of parallel computing with five threads, each for an individual fold in the five-fold CV. Table 4 shows that among $\hat{C}^+_{\mathsf{trace}}$, $\hat{C}^+_{\mathsf{HS}}$, $\hat{C}^+_{\mathsf{PACE,BIC}}$, $\hat{C}^+_{\mathsf{FACE}}$, $\hat{C}^+_{\mathsf{SC}}$ and $\hat{C}^+_{\mathsf{PP}}$, $\hat{C}^+_{\mathsf{SC}}$ is the fastest while $\hat{C}^+_{\mathsf{PP}}$ and $\hat{C}^+_{\mathsf{trace}}$ are the slowest, but all methods are practical since their computing times are up to 3.5 minutes on the tested laptop computer.

Different from those faster estimators $\hat{C}^+_{\mathsf{PACE,BIC}}$, $\hat{C}^+_{\mathsf{FACE}}$ and $\hat{C}^+_{\mathsf{SC}}$ that adopt BIC or FVE to select tuning parameters, both $\hat{C}^+_{\mathsf{trace}}$ and $\hat{C}^+_{\mathsf{HS}}$ use five-fold CV for tuning parameter selection, which is the primary cause of their slowness. For illustration, we also report the average computing time of both estimators (without parallel computing) after the value of λ is selected by five-fold CV, which are only 4.82 seconds and 4.75 seconds. See $\hat{C}^+_{\mathsf{trace}}(\lambda\,\mathsf{fixed})$ and $\hat{C}^+_{\mathsf{HS}}(\lambda\,\mathsf{fixed})$ in Table 4 respectively. Therefore, a computationally efficient approach to tuning parameter selection is an important future direction for our method.

For situations when computation is an issue to apply our proposed method, an ad-hoc remedy is to tune down the maximum number of allowable iterations of Algorithm 1, which is set as 10,000 by default (together with a strict stopping criterion.) If we allow up to 1,000 iterations, for instance, the average times for computing $\hat{C}^+_{\mathsf{trace}}$ and $\hat{C}^+_{\mathsf{HS}}$, including five-fold CV, are reduced by about 2/3. See $\hat{C}^+_{\mathsf{trace,fast}}$ and $\hat{C}^+_{\mathsf{HS,fast}}$ in Table 4 respectively. However, the impact of this change on the resulted estimators is not significant. The relative differences $\|\hat{C}^+_{\mathsf{trace}} - \hat{C}^+_{\mathsf{trace,fast}}\|_F / \|\hat{C}^+_{\mathsf{trace}}\|_F$ and $\|\hat{C}^+_{\mathsf{HS}} - \hat{C}^+_{\mathsf{HS,fast}}\|_F / \|\hat{C}^+_{\mathsf{HS}}\|_F$ are only 0.020 and 0.007 respectively.

Table 4: Means and standard deviations (SD) of computing times (in seconds) with respect to various methods when applied to the real data in Section 5.

	$\hat{C}^{+}_{\mathsf{trace}}$	$\hat{C}^{+}_{\mathsf{HS}}$	$\hat{C}^{+}_{\mathsf{PACE,BIC}}$	$\hat{C}^{+}_{\mathsf{FACE}}$	$\hat{C}^{+}_{\mathsf{SC}}$	$\hat{C}^{+}_{\mathsf{PP}}$	$\hat{C}^{+}_{\mathsf{trace}}(\lambda\,\mathsf{fixed})$	$\hat{C}^{+}_{\mathsf{HS}}(\lambda\,\mathsf{fixed})$	$\hat{C}^{+}_{\mathsf{trace,fast}}$	$\hat{C}^{+}_{\mathsf{HS,fast}}$
mean	205.49	133.21	1.31	3.34	0.16	210.66	4.82	4.75	77.32	42.69
SD	0.78	0.47	0.03	0.12	0.01	2.97	0.06	0.03	3.42	2.14

5. Real data application

We apply the proposed method to a loop sensor dataset which contains vehicle counts recorded every five minutes at an on-ramp on the 101 North freeway in Los Angeles, U.S.A.. This on-ramp is located near Dodger Stadium, the home field of the Los Angeles Dodgers baseball team, so unusual traffic is expected after a Dodgers home game. These measurements were collected by the Freeway Performance Measurement System (PeMS, `http://pems.dot.ca.gov`) and can be obtained from the UCI Machine Learning Repository (`https://archive.ics.uci.edu/ml/datasets/Dodgers+Loop+Sensor`). We focus on the after-game traffic measurements of 78 games between April 2005 and October 2005 available in this dataset. For each game, we have 31 measurements that cover the time interval from 30 minutes before the end of the game, to 120 minutes after the end of the game. This time interval is presented as $[-30, 120]$, where zero marks the end of a game.

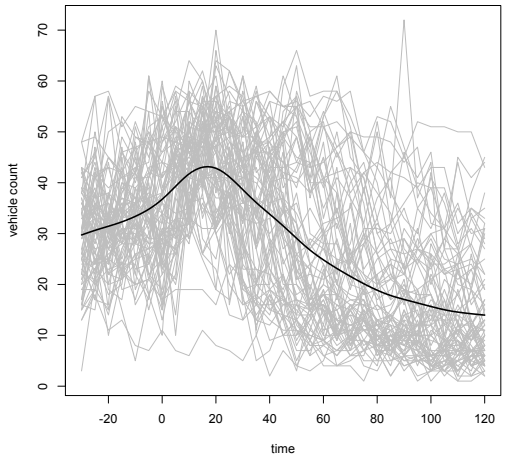

Figure 1: Vehicle counts over a time interval from 30 minutes before the end of a game, to 120 minutes after the end of the game. The black line represents a smoothing spline estimate of the mean function.

The vehicle counts of the 78 games are displayed in Figure 1, where the mean function was estimated by smoothing splines with its tuning parameter determined by GCV. The estimated mean curve demonstrates a traffic peak that emerges at around 20 minutes after the end of a game. This characteristic is consistent with the finding of Zhang and Wang (2015) and conforms to common sense.

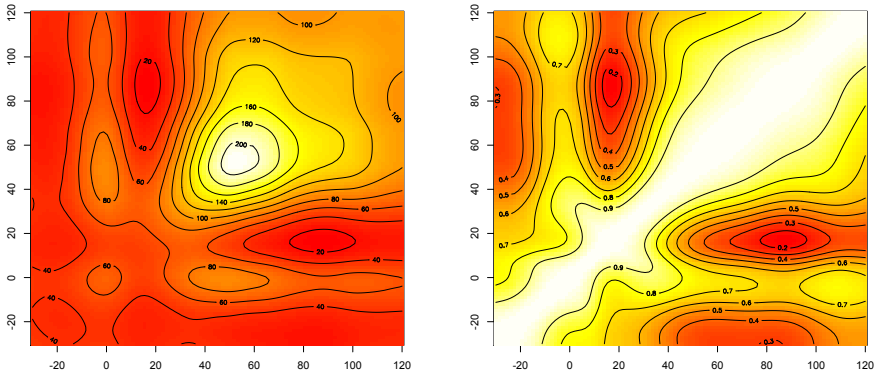

Figure 2: Contour plots of $\hat{C}^{+}_{\mathsf{trace}}$ (left) and its corresponding correlation function (right).

We provided the covariance estimator $\hat{C}^{+}_{\mathsf{trace}}$ of the vehicle counts, as described in Section 4, and constructed the corresponding correlation function estimate by the simple transformation: $\hat{C}(s,t)/\{\hat{C}(s,s)\hat{C}(t,t)\}^{1/2}$ for any covariance estimate $\hat{C}$ with $\hat{C}(s,s) > 0$ for all s. Note that positive semi-definiteness guarantees the validity of the correlation function estimate obtained by the above simple transformation. Namely, it has value between -1 and 1. However, this property could be violated for non-positive semi-definite estimators such as $\hat{C}_{\mathsf{CY}}$. The covariance and correlation estimates for $\hat{C}^{+}_{\mathsf{trace}}$ are depicted in Figure 2. One intriguing feature with respect to the temporal dependency of the vehicle counts is the high correlations of traffic between time 0 and time points after around time 30. When compared with adjacent time points such as -20 and 20, this feature is so distinctive that a ridge is formed at time 0.

To provide further insights of such phenomenon, we investigate the L^2 eigen-decomposition of $\hat{C}^{+}_{\mathsf{trace}}$. Due to the built-in low-rank estimation, $\hat{C}^{+}_{\mathsf{trace}}$ is automatically of rank 5 without further truncation of eigenvalues. Its corresponding five L^2 eigenfunctions, as described in Section 2.5,

are shown in Figure 3 (Left). The first eigenfunction explains over 80% of the total variance, i.e., the first eigenvalue is greater than 80% of the sum of all five eigenvalues. Therefore, the first eigenfunction plays a major role in the variation of the traffic profile. Of interest is that this eigenfunction possesses two peaks located near times 0 and 50, where the second peak is spanning over the time interval roughly between 30 and 120. This eigenfunction characterizes the high correlation we have observed between time 0 and the time interval between 30 and 120. Since a positive variation along this eigenfunction will add traffic to these two peaks, this implies that some audiences may choose to leave shortly after the game or even earlier, while some others take longer than usual to leave. As suggested by Zhang and Wang (2015), one possible explanation for this phenomenon is high game attendance. For games with high attendance, one may choose to leave earlier than usual to avoid traffic. Meanwhile, heavy traffic would also last longer due to high attendance. To further verify this explanation, we produced the functional principal component (FPC) scores by pre-smoothing individual vehicle count curves and then projecting them onto the first eigenfunction. Smoothing spline with GCV was used to implement the pre-smoothing. The scatter plot between FPC scores and game attendance as shown in Figure 3 (Right), together with the fact that their Pearson correlation is 0.57, indicates a positive association.

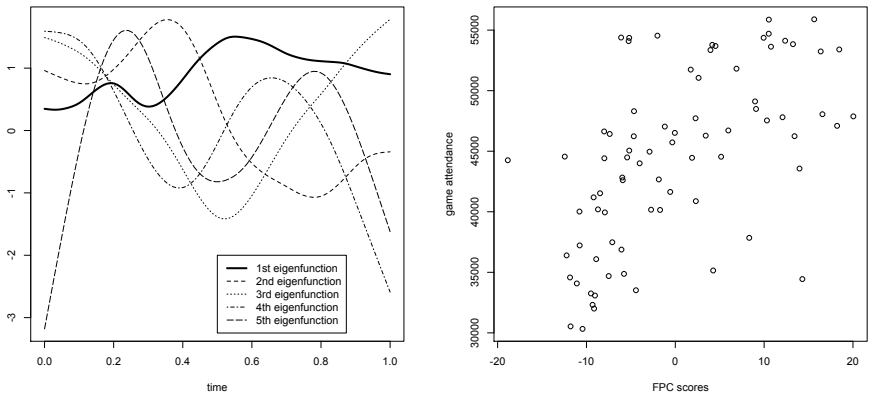

Figure 3: Left: L^2 eigenfunctions of $\hat{C}^{+}_{\mathsf{trace}}$. Right: Scatterplot of game attendance versus functional principal component scores (with respect to the first eigenfunction).

6. Asymptotic properties

In this section, we develop the empirical L^2 rate of convergence for a variety of spectrally regularized covariance estimators in tensor product Sobolev-Hilbert spaces in Theorems 2 and 3 below. The results are broad in three aspects. First, they hold for both fixed and random designs. Second, they allow a variety of spectral regularizations, where the trace-norm and Hilbert-Schmidt regularizations are both special cases. Third, these results are not restricted to positive semi-definite estimators.

Without loss of generality $\mathcal{T} = [0, 1]$. Here we focus on the r-th order Sobolev-Hilbert space on $[0, 1]$ where $r \geq 2$, i.e.,

$$\mathcal{H}(K) = \{g : g^{(v)}, v = 0, \ldots, r-1, \text{ are absolutely continuous; } g^{(r)} \in L^2([0,1])\},$$

equipped with squared norm $\|g\|^2 = \sum_{v=0}^{r} \int_0^1 \{g^{(v)}(t)\}^2 \, dt$. The asymptotic results also hold for its equivalent norms, e.g., $\|g\|^2 = \int_0^1 \{g(t)\}^2 \, dt + \int_0^1 \{g^{(r)}(t)\}^2 \, dt$, $\|g\|^2 = ([\int_0^1 \{g(t)\}^2 \, dt]^{1/2} + [\int_0^1 \{g^{(r)}(t)\}^2 \, dt]^{1/2})^2$, and $\|g\|^2 = \sum_{v=0}^{r-1} \{\int_0^1 g^{(v)}(t) \, dt\}^2 + \int_0^1 g^{(r)}(t)^2 \, dt$.

We investigate the asymptotic property of a class of covariance estimators given by

$$\hat{C}_\lambda = \underset{C \in \mathcal{F}}{\arg\min} \{\ell(C) + \lambda \Psi(C)\}, \tag{9}$$

where $\Psi(C) = \sum_{k \geq 1} |\tau_k(C)|^p$ for $1 \leq p \leq 2$, the loss function ℓ is chosen as (8), and $\mathcal{F} \subseteq \mathcal{H}(K \otimes K)$ is the hypothesis space for estimation.

Apparently, the penalty term Ψ incorporates both trace-norm ($p = 1$) and Hilbert-Schmidt-norm ($p = 2$) regularizations, so the asymptotic results below are not restricted to low-rank estimators. Moreover, since the choice of $\mathcal{F}$ is flexible, the results hold for estimators that are positive semi-definite, e.g., when $\mathcal{F} = \mathcal{S}^+(K)$, as well as for those that are not. In particular, if $\mathcal{F} = \mathcal{H}(K \otimes K)$ and $p = 2$, $\hat{C}_\lambda$ becomes the estimator by Cai and Yuan (2010) .

6.1. Assumptions

We list the assumptions needed for the asymptotic properties as follows.

Assumption 1. $C_0 \neq 0$ and $C_0 \in \mathcal{F} \subseteq \mathcal{H}(K \otimes K)$.

Assumption 2. The time points $\{T_{ij} : i = 1, \ldots, n; j = 1, \ldots, m\}$ are either fixed or random, and are independent of $\{X_i : i = 1, \ldots, n\}$. The errors $\{\varepsilon_{ij} : i = 1, \ldots, n; j = 1, \ldots m\}$ are independent of both $\{T_{ij} : i = 1, \ldots, n; j = 1, \ldots, m\}$ and $\{X_i : i = 1, \ldots, n\}$.

Assumption 3. For each $t \in [0, 1]$, $X(t)$ is sub-Gaussian with a parameter $b_X > 0$ which does not depend on t, i.e., $\mathbb{E}(\exp\{\beta X(t)\}) \leq \exp\{b_X^2 \beta^2/2\}$ for all $\beta > 0$ and $t \in [0, 1]$.

Assumption 4. For each i, j, ε_{ij} is sub-Gaussian with a parameter b_ε independent of i and j.

Assumption 2 is standard in FDA modeling. Assumptions 3 and 4 are sub-gaussian conditions of the stochastic process and noise.

6.2. Rate of convergence

For simplicity, we assume known $\mu_0 = 0$ so we let $\hat{\mu} = 0$ and accordingly $Z_{ijk} = Y_{ij}Y_{ik}$. For arbitrary bivariate functions g_1 and g_2, define an empirical inner product and the corresponding empirical norm as follows:

$$\langle g_1, g_2 \rangle_n = \frac{1}{nm(m-1)} \sum_{i=1}^{n} \sum_{1 \leq j \neq k \leq m} g_1(T_{ij}, T_{ik}) g_2(T_{ij}, T_{ik}) \quad \text{and} \quad \|g_1\|_n^2 = \langle g_1, g_1 \rangle_n.$$

Recall that we say a random variable $S_n = \mathcal{O}_p(k_n)$ if

$$\lim_{L \to \infty} \limsup_{n \to \infty} \Pr(S_n \geq L k_n) = 0.$$

To accommodate the flexibility of the design $\mathbb{T} = \{T_{ij} : i = 1, \ldots, n; j = 1, \ldots, m\} \in \mathcal{T}^{nm}$, we denote $S_n = \mathcal{O}_p^T(k_n)$ if

$$\lim_{L \to \infty} \limsup_{n \to \infty} \sup_{\mathbb{T} \in \mathcal{T}^{nm}} \Pr(S_n \geq L k_n \mid \mathbb{T}) = 0.$$

We first provide the empirical L^2 rate of convergence for $\hat{C}_\lambda$ from (9).

Theorem 2. *Under Assumptions 1–4, if $\Psi(C_0) > 0$ and $\lambda^{-1} = \mathcal{O}_p\{n^{r/(1+r)}\}$, we have $\|\hat{C}_\lambda - C_0\|_n = \mathcal{O}_p(\lambda^{1/2})$. Further, if $\lambda^{-1} = \mathcal{O}_p^T\{n^{r/(1+r)}\}$, we have $\|\hat{C}_\lambda - C_0\|_n = \mathcal{O}_p^T(\lambda^{1/2})$.*

In Theorem 2, the asymptotic accuracy of $\hat{C}_\lambda$ is guaranteed for both fixed and random designs. In particular, both independent and dependent designs are allowed if the design is random. Furthermore, Theorem 2 provides a uniform result over all designs under a stronger condition of λ. For instance, such $\mathcal{O}_p^T$-condition degenerates to the weaker $\mathcal{O}_p$-condition if the choice of λ is nonrandom or independent of the design.

Theorem 2 incorporates a variety of regularizations as long as $1 \leq p \leq 2$, where the commonly used trace-norm ($p = 1$) and Hilbert-Schmidt-norm ($p = 2$) penalties are both special cases. It shows that the empirical L^2 rate of convergence of $\hat{C}_\lambda$ is comparable to that of standard two-dimensional nonparametric smoothers. For example, the rate of convergence is $n^{1/3}$ for the second order Sobolev-Hilbert space, i.e., $r = 2$. The conclusion in Theorem 2 is generally true for all two-dimensional Sobolev spaces, but the rate is sub-optimal within the scope of tensor product Sobolev-Hilbert spaces. For periodic functions, however, we are able to significantly improve this rate by utilizing pinpoint entropy results for tensor product Sobolev-Hilbert spaces.

Theorem 3. *Suppose that* $\mathcal{F} \subseteq \{C \in \mathcal{H}(K \otimes K) : C$ *is a periodic function*$\}$. *Under Assumptions 1–4, if* $\Psi(C_0) > 0$, *and* $\lambda^{-1} = \mathcal{O}_p\{n^{2r/(1+2r)}/\log n\}$, *we have* $\|\hat{C}_\lambda - C_0\|_n = \mathcal{O}_p(\lambda^{1/2})$. *Further, if* $\lambda^{-1} = \mathcal{O}_p^T\{n^{2r/(1+2r)}/\log n\}$, *we have* $\|\hat{C}_\lambda - C_0\|_n = \mathcal{O}_p^T(\lambda^{1/2})$.

Similar to Theorem 2, Theorem 3 also allows for both fixed and random designs. Theorem 3 demonstrates that $\hat{C}_\lambda$ can achieve the empirical L^2 rate of convergence for one-dimensional nonparametric estimation, up to some order of $\log n$, although the target function C_0 is two-dimensional. For instance, if we let $r = 2$, the rate of $\hat{C}_\lambda$ is $(\log n)^{-1/2} n^{2/5}$, which is much faster than the two-dimensional nonparametric rate $n^{1/3}$. For sparse functional data, i.e., $m < \infty$, up to some order of $\log n$, the rate of $\hat{C}_\lambda$ is comparable to the minimax rate obtained by Cai and Yuan (2010) and the L^2 rate achieved by Paul and Peng (2009) for $r = 4$. However, the rates in both theorems are sub-optimal for functional data that are not sparse (Zhang and Wang, 2016).

The covariance estimator $\hat{C}_\lambda$ defined in (9) does not have a closed form due to the possible non-differentiability of the penalty term (e.g., when $p = 1$), and the flexibility of $\mathcal{F}$. This explains the technical challenges and highlights the novelties of the proofs for Theorems 2 and 3. In Theorem 3,

the particular structure of the tensor product RKHS accounts for the appealing rate of convergence of $\hat{C}_\lambda$. The upper bound of the entropy for tensor product Sobolev-Hilbert spaces, as given in Lemma 1 of the supplementary material, is a crucial component for the technical success. To our best knowledge, this article is the first in the FDA literature that achieves this result.

7. Conclusion

In this article, we propose a new class of covariance function estimators under a tensor product RKHS framework in terms of a variety of spectral regularizations of an operator. All covariance estimators are automatically positive semi-definite via a one-step procedure. Low rank of the estimators can be additionally achieved if a proper penalty, e.g., the trace-norm penalty, is chosen. We establish an unconventional representer theorem for the entire class of covariance estimators, based on which we develop an efficient algorithm tailored for the trace-norm regularization. Through an asymptotic analysis, a simulation study and a real data application, the proposed estimators are shown to enjoy excellent theoretical and numerical performances.

The focus of this article is covariance function estimation. Since covariance function estimation is usually an initial step to perform advanced FDA methods, such as trajectory prediction and functional linear regression, one direction for future work is to study how the proposed covariance estimators may improve the performances of those methods. Although the rates of convergence obtained in Theorems 2 and 3 are competitive under sparse functional data setups, such rates are not optimal in non-sparse settings. Consequently, another future exploration is to establish the optimal rate of convergence for all types of functional data following the work by Cai and Yuan (2010), Li and Hsing (2010), Zhang and Wang (2016) and Wang et al. (2018). Finally, as suggested by the above numerical experiments, a computationally efficient approaches for tuning parameter selection is also an important direction for future research.

8. Supplementary Material

The algorithm for Hilbert-Schmidt-norm regularization, additional simulation results, and all technical proofs are in the supplemental material.

Acknowledgements

The research of Raymond K. W. Wong is partially supported by National Science Foundation grant DMS-1612985. The research of Xiaoke Zhang is partially supported by National Science Foundation grant DMS-1613018. Portions of this research were conducted with high performance research computing resources provided by Texas A&M University (`https://hprc.tamu.edu`). The authors thank a Guest Editor and two referees for insightful suggestions that have helped improve early versions of this article.

References

Abernethy, J., Bach, F., Evgeniou, T., Vert, J.-P., 2009. A new approach to collaborative filtering: operator estimation with spectral regularization. Journal of Machine Learning Research 10, 803–826.

Avery, M., Wu, Y., Helen Zhang, H., Zhang, J., 2014. RKHS-based functional nonparametric regression for sparse and irregular longitudinal data. Canadian Journal of Statistics 42 (2), 204–216.

Beck, A., Teboulle, M., 2009. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. SIAM Journal on Imaging Sciences 2 (1), 183–202.

Becker, S. R., Candès, E. J., Grant, M. C., 2011. Templates for convex cone problems with applications to sparse signal recovery. Mathematical Programming Computation 3 (3), 165–218.

Cai, T. T., Yuan, M., 2010. Nonparametric covariance function estimation for functional and longitudinal data. Tech. rep., Georgia Institute of Technology, Atlanta, GA.

Chiou, J.-M., 2012. Dynamical functional prediction and classification, with application to traffic flow prediction. The Annals of Applied Statistics 6 (4), 1588–1614.

Eggermont, P. P., LaRiccia, V. N., 2009. Maximum penalized likelihood estimation: volume II: regression. Springer, New York.

Ferraty, F., Vieu, P., 2006. Nonparametric functional data analysis: theory and practice. Springer, New York.

Goldsmith, J., Bobb, J., Crainiceanu, C. M., Caffo, B., Reich, D., 2011. Penalized functional regression. Journal of Computational and Graphical Statistics 20 (4), 830–851.

Gu, C., 2013. Smoothing spline ANOVA models, 2nd Edition. Springer, New York.

Hall, P., Vial, C., 2006. Assessing the finite dimensionality of functional data. Journal of the Royal Statistical Society: Series B 68 (4), 689–705.

Horváth, L., Kokoszka, P., 2012. Inference for functional data with applications. Vol. 200. Springer, New York.

Hsing, T., Eubank, R., 2015. Theoretical foundations of functional data analysis, with an introduction to linear operators. John Wiley & Sons.

James, G., Hastie, T., Sugar, C., 2000. Principal component models for sparse functional data. Biometrika 87 (3), 587–602.

Jiang, C.-R., Aston, J. A., Wang, J.-L., 2016. A functional approach to deconvolve dynamic neuroimaging data. Journal of the American Statistical Association 111 (513), 1–13.

Li, Y., Hsing, T., 2010. Uniform convergence rates for nonparametric regression and principal component analysis in functional/longitudinal data. The Annals of Statistics 38 (6), 3321–3351.

Li, Y., Wang, N., Carroll, R. J., 2013. Selecting the number of principal components in functional data. Journal of the American Statistical Association 108 (504), 1284–1294.

Mazumder, R., Hastie, T., Tibshirani, R., 2010. Spectral regularization algorithms for learning large incomplete matrices. Journal of Machine Learning Research 11, 2287–2322.

Paul, D., Peng, J., 2009. Consistency of restricted maximum likelihood estimators of principal components. The Annals of Statistics 37 (3), 1229–1271.

Pearce, N. D., Wand, M. P., 2006. Penalized splines and reproducing kernel methods. The American Statistician 60 (3), 233–240.

Peng, J., Paul, D., 2009. A geometric approach to maximum likelihood estimation of the functional principal components from sparse longitudinal data. Journal of Computational and Graphical Statistics 18 (4), 995–1015.

Poskitt, D. S., Sengarapillai, A., 2013. Description length and dimensionality reduction in functional data analysis. Computational Statistics & Data Analysis 58, 98–113.

584 Ramsay, J. O., Silverman, B. W., 2005. Functional data analysis, 2nd Edition. Springer, New York.
585 Rice, J. A., Silverman, B. W., 1991. Estimating the mean and covariance structure nonparametri-
586 cally when the data are curves. Journal of the Royal Statistical Society: Series B 55 (1), 233–243.
587 Rice, J. A., Wu, C. O., 2001. Nonparametric mixed effects models for unequally sampled noisy
588 curves. Biometrics 57 (1), 253–259.
589 Wahba, G., 1990. Spline Models for Observational Data. SIAM, Philadelphia.
590 Wang, H., Zhong, P.-S., Cui, Y., Li, Y., 2018. Unified empirical likelihood ratio tests for functional
591 concurrent linear models and the phase transition from sparse to dense functional data. Journal
592 of the Royal Statistical Society: Series B (Statistical Methodology) 80 (2), 343–364.
593 Wang, X., Ruppert, D., 2015. Optimal prediction in an additive functional model. Statistica Sinica
594 25 (2), 567–589.
595 Wong, R. K. W., Li, Y., Zhu, Z., 2017. Partially linear functional additive models for multivariate
596 functional data. Journal of the American Statistical Association, to appear.
597 Xiao, L., Li, C., Checkley, W., Crainiceanu, C., 2018. Fast covariance estimation for sparse func-
598 tional data. Statistics and Computing 28 (3), 511–522.
599 Xiao, L., Li, Y., Ruppert, D., 2013. Fast bivariate p-splines: the sandwich smoother. Journal of
600 the Royal Statistical Society: Series B (Statistical Methodology) 75 (3), 577–599.
601 Yao, F., Müller, H.-G., Wang, J.-L., 2005a. Functional data analysis for sparse longitudinal data.
602 Journal of the American Statistical Association 100 (470), 577–590.
603 Yao, F., Müller, H.-G., Wang, J.-L., 2005b. Functional linear regression analysis for longitudinal
604 data. The Annals of Statistics 33 (6), 2873–2903.
605 Zhang, X., Wang, J.-L., 2015. Varying-coefficient additive models for functional data. Biometrika
606 102 (1), 15–32.
607 Zhang, X., Wang, J.-L., 2016. From sparse to dense functional data and beyond. The Annals of
608 Statistics 44 (5), 2281–2321.
609 Zhu, H., Yao, F., Zhang, H. H., 2014. Structured functional additive regression in reproducing
610 kernel hilbert spaces. Journal of the Royal Statistical Society: Series B (Statistical Methodology)
611 76 (3), 581–603.