Low-Rank Covariance Function Estimation for Multidimensional Functional Data

Jiayi Wang, Raymond K. W. Wong, and Xiaoke Zhang

Abstract
Multidimensional functional data arise from many fields nowadays. The covariance function plays an important role in the analysis of such increasingly common data. In this article, we propose a novel nonparametric covariance function estimation approach under the framework of reproducing kernel Hilbert spaces (RKHS) that can handle both sparse and dense functional data. We extend multilinear rank structures for (finite-dimensional) tensors to functions, which allow for flexible modeling of both covariance operators and marginal structures. The proposed framework can guarantee that the resulting estimator is automatically semipositive definite, and can incorporate various spectral regularizations. The trace-norm regularization in particular can promote low ranks for both covariance operator and marginal structures. Despite the lack of a closed form, under mild assumptions, the proposed estimator can achieve unified theoretical results that hold for any relative magnitudes between the sample size and the number of observations per sample field, and the rate of convergence reveals the phase-transition phenomenon from sparse to dense functional data. Based on a new representer theorem, an ADMM algorithm is developed for the trace-norm regularization. The appealing numerical performance of the proposed estimator is demonstrated by a simulation study and the analysis of a dataset from the Argo project. Supplementary materials for this article are available online.

1. Introduction

In recent decades, functional data analysis (FDA) has become a popular branch of statistical research. General introductions to FDA can be found in a few monographs (e.g., Ramsay and Silverman 2005; Ferraty and Vieu 2006; Horváth and Kokoszka 2012; Hsing and Eubank 2015; Kokoszka and Reimherr 2017). While traditional FDA deals with a sample of time-varying trajectories, many new forms of functional data have emerged due to improved capabilities of data recording and storage, as well as advances in scientific computing. One particular new form of functional data is multidimensional functional data, which becomes increasingly common in various fields such as climate science, neuroscience, and chemometrics. Multidimensional functional data are generated from random fields, that is, random functions of several input variables. We note that this is different from those settings with multiple output variables, which are sometimes referred to as multivariate functional data (e.g., Ramsay and Silverman 2005; Ferraty and Vieu 2006). One example is multi-subject magnetic resonance imaging (MRI) scans, such as those collected by the Alzheimer’s Disease Neuroimaging Initiative. A human brain is virtually divided into three-dimensional boxes called “voxels” and brain signals obtained from these voxels form a three-dimensional functional sample indexed by spatial locations of the voxels. Despite the growing popularity of multidimensional functional data, statistical methods for such data are limited apart from very few existing works (e.g., Huang, Shen, and Buja 2009; Allen 2013; Zhang, Shen, and Huang 2013; Zhou and Pan 2014; Wang and Huang 2017).

In FDA, covariance function estimation plays an important role. Many methods have been proposed for unidimensional functional data (e.g., Rice and Silverman 1991; James, Hastie, and Sugar 2000; Yao, Müller, and Wang 2005; Paul and Peng 2009; Li and Hsing 2010; Goldsmith et al. 2011; Xiao, Li, and Ruppert 2013), and a few were particularly developed for two-dimensional functional data (e.g., Zhou and Pan 2014; Wang and Huang 2017). In general when the input domain is of dimension $p$, one needs to estimate a $2p$-dimensional covariance function. Since covariance function estimation in FDA is typically nonparametric, the curse of dimensionality emerges soon when $p$ is moderate or large. For general $p$, most work are restricted to regular and fixed designs (e.g., Zipunnikov et al. 2011; Allen 2013), where all random fields are observed over a regular grid like MRI scans. Such sampling plan leads to a tensor dataset, so one may apply tensor/matrix decompositions to estimate the covariance function. When random fields are observed at irregular locations, the dataset is no longer a completely observed tensor so tensor/matrix decompositions are not directly applicable. If observations are densely collected for each random field, a two-step approach is a natural solution, which involves pre-smoothing every random field followed by tensor/matrix decompositions at a fine discretized grid. However, this solution is infeasible for sparse data where there are a limited number of observations per random field. One example is the data collected by the...
international Argo project (http://www.argo.net). See Section 8 for more details. In such sparse data setting, one may apply the local smoothing method of Chen and Jiang (2017), but it suffers from the curse of dimensionality when the dimension $p$ is moderate due to a $2p$-dimensional nonparametric regression.

We notice that there is a related class of literature on longitudinal functional data (e.g., Chen and Müller 2012; Park and Staicu 2015; Chen, Delcado, and Müller 2017), a special type of multidimensional functional data where a function is repeatedly measured over longitudinal times. Typically multi-step methods are needed to model the functional and longitudinal dimensions either separately (one dimension at a time) or sequentially (one dimension given the other), as opposed to the joint estimation procedure proposed in this article. We also notice a recent work on longitudinal functional data under the Bayesian framework (Shamshoian et al. 2019).

The contribution of this article is 3-fold. First, we propose a new and flexible nonparametric method for low-rank covariance function estimation for multidimensional functional data, via the introduction of (infinite-dimensional) unfolding operators (see Section 4). This method can handle both sparse and dense functional data, and can achieve joint structural reductions in all dimensions, in addition to rank reduction of the covariance operator. The proposed estimator is a one-step estimator that achieves positivity and low-rankness, unlike those multi-stage estimators (e.g., Hall and Vial 2006; Poskitt and Sengarapillai 2013) which require a truncation and reconstruction step. The one-step nature reduces the theoretical complexities in the development of asymptotic properties.

Second, we generalize the representer theorem for unidimensional functional data by Wong and Zhang (2019) to the multidimensional case with more complex spectral regularizations. The new representer theorem makes the estimation procedure practically computable by generating a finite-dimensional parameterization to the solution of the underlying infinite-dimensional optimization.

Finally, a unified asymptotic theory is developed for the proposed estimator. It automatically incorporates the settings of dense and sparse functional data, and reveals a phase transition in the rate of convergence. Different from existing theoretical work heavily based on closed-form representations of estimators (Cai and Yuan 2010; Li and Hsing 2010; Zhang and Wang 2016; Liebl 2019), this article provides the first unified theory for penalized global M-estimators of covariance functions which does not require a closed-form solution. Furthermore, a near-optimal (i.e., optimal up to a logarithmic order) one-dimensional nonparametric rate of convergence is attainable for the $2p$-dimensional covariance function estimator for Sobolev–Hilbert spaces.

The rest of the article is organized as follows. Section 2 provides some background on reproducing kernel Hilbert space (RKHS) frameworks for functional data. Section 3 introduces marginal low-rank structure of the covariance function. Section 4 presents Tucker decomposition for finite-dimensional tensors and the generalization to tensor product RKHS operators, which is the foundation for our estimation procedure. The proposed estimation method is given in Section 5, together with an algorithm. The unified theoretical results are presented in Section 6. The numerical performance of the proposed method is evaluated by a simulation study in Section 7 and a real data application in Section 8. Additional simulation results and technical details are collected in a separate online supplementary materials (SM).

2. RKHS Framework for Functional Data

In recent years there is a surge of RKHS methods in FDA (e.g., Yuan and Cai 2010; Zhu, Yao, and Zhang 2014; Li and Song 2017; Reimherr, Sriperumbudur, and Taoufik 2018; Sun et al. 2018; Wong, Li, and Zhu 2019). However, covariance function estimation, a seemingly well-studied problem, does not receive the same amount of attention in the development of RKHS methods, even for unidimensional functional data. Interestingly, we find that the RKHS modeling provides a versatile framework for both unidimensional and multidimensional functional data.

Let $X$ be a random field defined on an index set $\mathcal{T} \subset \mathbb{R}^p$, with a mean function $\mu_0(\cdot) = E[X(\cdot)]$ and a covariance function $\gamma_0(s, \cdot) = \text{cov}(X(s), X(\cdot))$, and let $\{X_i : i = 1, \ldots, n\}$ be i.i.d. copies of $X$. Typically, a functional dataset is represented by $\{(T_{ij}, Y_{ij}) : j = 1, \ldots, m_i; i = 1, \ldots, n\}$, where

$$Y_{ij} = X_i(T_{ij}) + \epsilon_{ij} \in \mathbb{R} \quad (1)$$

is the noisy measurement of the $i$th random field $X_i$ taken at the corresponding index $T_{ij} \in \mathcal{T}$, $m_i$ is the number of measurements observed from the $i$th random field, and $\{\epsilon_{ij} : i = 1, \ldots, n; j = 1, \ldots, m_i\}$ are independent errors with mean zero and finite variance. For simplicity and without loss of generality, we assume $m_i = m$ for all $i$.

As in many nonparametric regression setups such as penalized regression splines (e.g., Pearl and Wand 2006) and smoothing splines (e.g., Wahba 1990; Gu 2013), the sample field of $X$, that is, the realized $X$ (as opposed to the sample path of a unidimensional random function), is assumed to reside in an RKHS $\mathcal{H}$ of functions defined on $\mathcal{T}$ with a continuous and square integrable reproducing kernel $K$. Let $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ and $\|\cdot\|_{\mathcal{H}}$ denote the inner product and norm of $\mathcal{H}$, respectively. With the technical condition $E\|X\|_{\mathcal{H}}^2 < \infty$, the covariance function $\gamma_0$ resides in the tensor product RKHS $\mathcal{H} \otimes \mathcal{H}$. It can be shown that $\mathcal{H} \otimes \mathcal{H}$ is an RKHS, equipped with the reproducing kernel $K \otimes K$ defined as $(K \otimes K)(s_1, t_1)(s_2, t_2) = K(s_1, s_2)K(t_1, t_2)$, for any $s_1, s_2, t_1, t_2 \in \mathcal{T}$. This result has been exploited by Cai and Yuan (2010) and Wong and Zhang (2019) for covariance estimation in the unidimensional setting.

For any function $f \in \mathcal{H} \otimes \mathcal{H}$, there exists an operator mapping $\mathcal{H}$ to $\mathcal{H}$ defined by $g \in \mathcal{H} \mapsto \langle f(\cdot, \cdot), g(\cdot) \rangle_{\mathcal{H}} \in \mathcal{H}$. When $f$ is a covariance function, we call the induced operator an $\mathcal{H}$-covariance operator, or simply a covariance operator as below. To avoid clutter, the induced operator will share the same notation with the generating function. Similar to $L^2$-covariance operators, the definition of an induced operator is obtained by replacing the $L^2$ inner product by the RKHS inner product.

The benefits of considering this operator have been discussed in Wong and Zhang (2019). We also note that a singular value
decomposition (e.g., Hsing and Eubank 2015) of the induced operator exists whenever the corresponding function \( f \) belongs to the tensor product \( \mathcal{H} \otimes \mathcal{H} \). The idea of induced operator can be similarly extended to general tensor product space \( \mathcal{F}_1 \otimes \mathcal{F}_2 \) where \( \mathcal{F}_1 \) and \( \mathcal{F}_2 \) are two generic RKHSs of functions.

For any \( \gamma \in \mathcal{H} \otimes \mathcal{H} \), let \( \gamma^\top \) be the transpose of \( \gamma \), that is, \( \gamma^\top(s, t) = \langle \gamma(t, s), s, t \rangle \in T \). Define \( M = \{ \gamma \in \mathcal{H} \otimes \mathcal{H} : \gamma \equiv \gamma^\top \} \). To guarantee symmetry and positive semidefiniteness of the estimators, Wong and Zhang (2019) adopted \( M^+ = \{ \gamma \in M : \langle \gamma f, f \rangle \mathcal{H} \geq 0, \forall f \in \mathcal{H} \} \) as the hypothesis class of \( \gamma_0 \) and considered the following regularized estimator:

\[
\arg \min_{\gamma \in M^+} \{ \ell(\gamma) + \tau \Psi(\gamma) \},
\]

where \( \ell \) is a convex and smooth loss function characterizing the fidelity to the data, \( \Psi(\gamma) \) is a spectral penalty function (see Definition 2), and \( \tau \) is a tuning parameter. Due to the constraints specified in \( M^+ \), the resulting covariance estimator is always positive semidefinite.

In particular, if the spectral penalty function \( \Psi(\gamma) \) imposes the trace-norm regularization, an \( \ell_1 \)-type shrinkage penalty on the respective singular values, the estimator is usually of low rank. Cai and Yuan (2010) adopted a similar objective function as in (2) but with the hypothesis class \( \mathcal{H} \otimes \mathcal{H} \) and an \( \ell_2 \)-type penalty \( \Psi(\gamma) = \|\gamma\|^2_{\mathcal{H} \otimes \mathcal{H}} \), so the resulting estimator may neither be positive semidefinite nor low-rank.

Although Cai and Yuan (2010) and Wong and Zhang (2019) focused on unidimensional functional data, their frameworks can be directly extended to the multidimensional setting. Explicitly, similar to (2), as long as a proper \( \mathcal{H} \) for the random fields with dimension \( p > 1 \) is selected, an efficient “one-step” covariance function estimation with the hypothesis class \( M^+ \) can be obtained immediately, which results in a positive semidefinite and possibly low-rank estimator. Since an RKHS is identified by its reproducing kernel, we simply need to pick a multivariate reproducing kernel \( K \) for multidimensional functional data. However, even when the low-rank approximation/estimation is adopted (e.g., by trace-norm regularization), we still need to estimate several \( p \)-dimensional eigenfunctions nonparametrically. This curse of dimensionality calls for a more efficient modeling.

### 3. Marginal Low-Rank Structure

The proposed method we will present in Section 5 relies on the promotion of sharing structure among eigenfunctions. In this section, we provide an illustration to explain it.

Suppose that \( X \in \mathcal{H} = \bigotimes_{k=1}^p \mathcal{H}_k \), where each \( \mathcal{H}_k \) is an RKHS of functions equipped with an inner product \( \langle \cdot, \cdot \rangle_k \) and corresponding norm \( \| \cdot \|_k \) \( k = 1, \ldots, p \). Then the covariance function \( \gamma_0 \) resides in \( \mathcal{H} \otimes \mathcal{H} = \bigotimes_{j=1}^p \mathcal{H}_j \otimes \bigotimes_{k=1}^p \mathcal{H}_k \). For a general \( \mathcal{H} = \bigotimes_{k=1}^p \mathcal{H}_k \), let \( \{ e_{k,k} : k = 1, \ldots, q_k \} \) be a set of orthonormal basis functions of \( \mathcal{H}_k \) for \( k = 1, \ldots, p \), where \( q_k \) is allowed to be infinite, depending on the dimensionality of \( \mathcal{H}_k \). Then \( \{ \bigotimes_{k=1}^p e_{k,k} : k = 1, \ldots, q \} \) forms a set of orthonormal basis functions for \( \mathcal{H} \). To have an easy illustration, we consider \( p = 2 \), and then the covariance function \( \gamma_0 \) can be expressed by

\[
\gamma_0((s_1, s_2), (t_1, t_2)) = \sum_{k=1}^q \sum_{l=1}^{q_1} \sum_{m=1}^{q_2} \sum_{n=1}^{q_3} b_{k, l, m, n}^3 e_{k, l}(s_1) e_{m, n}(s_2) e_{l, m}(t_1) e_{n, k}(t_2),
\]

(3)

where \( B \in \mathbb{R}^{q_1 \times q_2 \times q_3} \).

Let \( \{ u_j : l = 1, \ldots, r_1 \} \) and \( \{ v_j : l = 1, \ldots, r_2 \} \) be two sets whose elements are orthonormal linear combinations of \( \{ e_{1,l} : l = 1, \ldots, q_1 \} \) and \( \{ e_{2,l} : l = 1, \ldots, q_2 \} \), respectively, such that

\[
\gamma_0((s_1, s_2), (t_1, t_2)) = \sum_{j=1}^{r_1} \sum_{k=1}^{r_2} E_{j,k} r_1 r_2 u_j(s_1) v_k(s_2) u_j(t_1) v_k(t_2),
\]

(4)

where \( E \in \mathbb{R}^{q_1 \times q_2 \times q_3} \). To explain the sharing structure, we consider those associated with the minimal \( r_1 \), \( r_2 \), which correspond to the “best” dimension reduction along each dimension. Consider eigen-decomposition of \( E = \text{PDP} \) where \( D = \text{diag}(\lambda_1, \lambda_2, \ldots, \lambda_{r_2}) \) and \( \text{P} \in \mathbb{R}^{r_1 \times r_2} \) has orthonormal columns. Then we obtain the eigen-decomposition of the covariance function \( \gamma_0 \):

\[
\gamma_0((s_1, s_2), (t_1, t_2)) = \sum_{k=1}^{R} \lambda_k \gamma_k(s_1, s_2) \gamma_k^T(t_1, t_2),
\]

where the eigenfunction is

\[
f_k(s_1, s_2) = \sum_{j=1}^{r_1} \sum_{k=1}^{r_2} P_{j,k} f_j(s_1) f_k(s_2),
\]

with \( a_{j,k} \) is the common basis for the variation in \( \mathcal{H}_1 \), hence describing the marginal structure in \( \mathcal{H}_1 \). Similarly \( \{ v_j : j = 1, \ldots, r_2 \} \) is the common basis that characterizes the marginal variation in \( \mathcal{H}_2 \). We call them the marginal basis along the respective dimension.

Similarly, for \( p \)-dimensional functional data, each eigenfunction can be represented by a linear combination of \( p \)-products of univariate functions. Compared to typical low-rank covariance modelings only in terms of \( R \), we also intend to find the efficient marginal bases that have small \( r_k \) required to characterize the \( p \)-dimensional eigenfunctions of the covariance operator. Intuitively, the above illustration shows the small \( r_k \) encourages a “sharing” structure of one-dimensional variations among different eigenfunctions. Promoting small \( r_k \) shrinks the number of these one-dimensional functions to be estimated and further alleviates the curse of dimensionality. Moreover, one-dimensional marginal structures can potentially help with a better understanding of \( p \)-dimensional eigenfunctions, as illustrated in Section 8.
Below, we explore low-rank modeling of marginal structures through the lens of tensor decomposition in finite-dimensional vector spaces and its extension to infinite-dimensional function spaces.

4. Unfolding Operations

In this section, we first review the well-known Tucker decomposition for finite-dimensional tensors, and then introduce the concept of functional unfolding for low-rank modeling.

4.1. Review of Tucker Decomposition

Let \( G = \bigotimes_{k=1}^{d} G_k \) denote a generic tensor product space. To begin with, we consider that \( G_k, k = 1, \ldots, d \), are all finite-dimensional. Let the dimension of \( G_k \) be \( q_k, k = 1, \ldots, d \). Then each element in \( G = \bigotimes_{k=1}^{d} G_k \) can be identified by an array \( A \) in \( \mathbb{R}^{\prod_{k=1}^{d} q_k} \), which contains the coefficients through an orthonormal basis. By Tucker decomposition (Tucker 1966), we can decompose an array \( A \) in the following way:

\[
A = G \times_1 U_1 \times_2 \cdots \times_d U_d, \tag{5}
\]

where \( \times_n \) represents the \( n \)-mode product (see Definition S1 in the SM), \( U_i \in \mathbb{R}^{d_i \times \ell_i} \) where \( \ell_i \leq q_i \), and \( G \in \mathbb{R}^{d_1 \times \cdots \times d_d} \) is called the core tensor. By Tucker decomposition (Tucker 1966), we can decompose an array \( A \) in the following way:

\[
A(n) = U_n G(n) (U_d \otimes \cdots \otimes U_{n+1} \otimes U_{n-1} \otimes \cdots \otimes U_1)^T, \tag{6}
\]

where, with a slightly notational abuse, \( \otimes \) also represents the Kronecker product between matrices.

Figure 1 provides a pictorial illustration of a Tucker decomposition. Tucker decomposition naturally leads to a particular form of rank, called "multilinear rank," which is defined as the vector \( \text{rank}(A_{(1)}), \ldots, \text{rank}(A_{(d)}) \). If one chooses a Tucker decomposition such that \( |U_k| : k = 1, \ldots, d \) are orthonormal matrices and \( \text{rank}(U_k) = r_k \), then \( \text{rank}(A_{(d)}) = \text{rank}(G_{(k)}) \).

To encourage low-rank structures in covariance function estimation, the matricization operation will be generalized for finite-dimensional arrays to infinite-dimensional tensors (Hackbusch 2012). Their relationship to the marginal structures explained in Section 3 will become clear in Section 4.3.

4.2. Functional Unfolding

Now we take \( G_k \) as an RKHS of functions with an inner product \( \langle \cdot, \cdot \rangle_{G_k} \). Notice that the tensor product space \( G = \bigotimes_{k=1}^{d} G_k \) can be generated by some elementary tensors of the form \( \bigotimes_{k=1}^{d} f_k(x_k) \) where \( f_k \in G_k, k = 1, \ldots, d \). More specifically, \( G \) is the completion of the linear span of all elementary tensors under the inner product \( \langle \bigotimes_{k=1}^{d} f_k, \bigotimes_{k=1}^{d} f'_k \rangle = \prod_{k=1}^{d} \langle f_k, f'_k \rangle_{G_k} \), for any \( f_k, f'_k \in G_k \).

In Definition 1, we generalize matricization/unfolding for finite-dimensional arrays to infinite-dimensional elementary tensors. We also define a square unfolding for infinite-dimensional tensors that will be used to describe the spectrum of covariance operators.

**Definition 1 (Functional unfolding operators).** The one-way unfolding operator and square unfolding operators are defined as follows for any elementary tensor of the form \( \bigotimes_{k=1}^{d} f_k \).

1. One-way unfolding operator \( U_j \) for \( j = 1, \ldots, d \): The \( j \)-mode one-way unfolding operator \( U_j : \bigotimes_{k=1}^{d} G_k \rightarrow G_j \otimes (\bigotimes_{k \neq j} G_k) \) is defined by \( U_j(\bigotimes_{k=1}^{d} f_k) = f_j \otimes (\bigotimes_{k \neq j} f_k) \).
2. Square unfolding operator \( S \): When \( d \) is even, the square unfolding operator \( S : \bigotimes_{j=1}^{d} G_j \rightarrow (\bigotimes_{j=1}^{d/2} G_j) \otimes (\bigotimes_{k=d/2+1}^{d} G_k) \) is defined by \( S(\bigotimes_{j=1}^{d/2} f_j) = (\bigotimes_{j=1}^{d/2} f_j) \otimes (\bigotimes_{k=d/2+1}^{d} f_k) \).

These definitions extend to any function \( f \in G \) by linearity. For notational simplicity, we denote \( U_j(f) \) by \( f_{(j)} \).

Note that the image of each functional unfolding operator, either \( U_j, j = 1, \ldots, d \) or \( S \), is a tensor product of two RKHSs, so its output can be interpreted as an (induced) operator. Given a function \( f \in G \), the multilinear rank can be defined as \( \text{rank}(f_{(1)}), \ldots, \text{rank}(f_{(d)}) \), where \( f_{(j)} \)'s are interpreted as an operator here and \( \text{rank}(A) \) is the rank of any operator \( A \). If all \( G_k, k = 1, \ldots, d \) are finite-dimensional, the singular values of the output of any functional unfolding operator match with those of the \( j \)-mode matricization (of the corresponding array representation).

**Remark 1.** For an array \( A \in \mathbb{R}^{\prod_{k=1}^{d} q_k} \), the one-way unfolding \( U_j(A) \) is the same as matricization, if we further impose the same ordering of the columns in the output of \( U_j(A) \), \( j = 1, \ldots, d \). This ordering is just related to how we represent the array, and is not crucial in the general definition of \( U_j \). Since the description of the computational strategy depends on the explicit representation, we will always assume this ordering. Similarly, we also define a specific ordering of rows and columns for \( A \in \mathbb{R}^{\prod_{m=1}^{M} q_m} \times (\prod_{k=1}^{d} q_k) \) when \( d \) is even, such that its \((j_1, j_2)\)th entry is \( A_{j_1, \ldots, j_d, k_1, \ldots, k_d} \) where \( j_1 = 1 + \sum_{m=1}^{d/2} (k_{2m-1} - 1)(\prod_{m=i+1}^{d/2} q_m) \) and \( j_2 = 1 + \sum_{i=d/2+1}^{d} (k_i - 1) (\prod_{m=i+1}^{d} q_m) \).
4.3. One-Way and Two-Way Ranks in Covariance Functions

Recall that we aim to estimate $\gamma_0 \in \mathcal{H} \otimes \mathcal{H}$. We could consider a special case of $\mathcal{G} = \bigotimes_{j=1}^d \mathcal{G}_j$ by letting $d = 2p$, $\mathcal{G}_j = \mathcal{H}_j$ for $j = 1, \ldots, p$, $\mathcal{G}_j = \mathcal{H}_{j-p}$ for $j = p+1, \ldots, d$, and $(\cdot, \cdot)_{\mathcal{G}_j} = (\cdot, \cdot)$ for $j = 1, \ldots, d$. Clearly, the elements of $\mathcal{H} \otimes \mathcal{H}$ are identified by those in $\mathcal{G} = \bigotimes_{j=1}^d \mathcal{G}_j$. In the folding structure, $\mathcal{H} \otimes \mathcal{H}$ has a squarely unfolded structure. Since a low-multilinear-rank representation is different from unfolded forms, it would be easier to study the completely unfolded space $\bigotimes_{k=1}^m \mathcal{G}_k$ instead of the squarely unfolded space $\mathcal{H} \otimes \mathcal{H}$. We use $\Gamma_0$ to represent the folded covariance function, the corresponding element of $\gamma_0$ in $\mathcal{G}$. In other words, $\Gamma_0 = \gamma_0$.

For any $\Gamma \in \mathcal{G}$, rank$(\Gamma)$ is defined as the two-way rank of $\Gamma$ while rank$(\Gamma(1))$, $\ldots$, rank$(\Gamma(p))$ are defined as the one-way ranks of $\Gamma$. Now, if we link the unfolding to the illustration in Section 3. In Section 3, let us define tensors $B^{\text{fold}} \in \mathbb{R}^{q_1 \times q_2 \times q_1 \times q_2}$ and $E^{\text{fold}} \in \mathbb{R}^{r_1 \times r_2 \times r_1 \times r_2}$ such that $B^{\text{fold}} = B$ and $E^{\text{fold}} = E$. Then the fully folded covariance function $\Gamma_0$ can be expressed as $\Gamma_0 = \sum_{k_1, \ldots, k_p} B^{\text{fold}}_{k_1, k_2; k_q} \bigotimes_{i=1}^d \mathcal{G}_i$, as compared to the square unfolded covariance function in (3).

Similar to the construction of (4), there exist $U_k \in \mathbb{R}^{q_k \times r_k}, k = 1, 2$, with orthonormal columns, such that $B^{\text{fold}} = E^{\text{fold}} \times_1 U_1 \times_2 U_2 \times_3 U_1 \times_4 U_2$. Now, one can see that $R$ and $(r_1, r_2)$ defined in Section 3 have the following interpretations: $R = \text{rank}(B^{\text{fold}}) = \text{rank}(E^{\text{fold}})$ is the two-way rank and (minimal) $r_1, r_2 = (\text{rank}(B^{\text{fold}}_{(1)}), \text{rank}(B^{\text{fold}}_{(2)})) = (\text{rank}(E^{\text{fold}}_1), \text{rank}(E^{\text{fold}}_2))$ are the one-way ranks. Therefore, to promote marginal low-rank structure as discussed in Section 3, we should encourage a low-multilinear-rank structure.

Remark 2. Obviously, rank$(\Gamma_0) \leq \prod_{k=1}^p \text{rank}(\Gamma(k))$ for $p$-dimensional functional data. If the random field $X$ has the property of “weak separability” as defined by Lynch and Chen (2018), then max(rank$(\Gamma(1))$, rank$(\Gamma(p))$) $\leq R$ so the low-rank structure in terms of $R$ will be automatically translated to low one-way ranks. Note that the construction of our estimator and corresponding theoretical analysis do not require separability conditions.

Next, we will utilize both one-way and two-way structures and propose an estimation procedure that regularizes one-way and two-way ranks jointly and flexibly, with the aim of seeking the “sharing” of marginal structures while controlling the number of eigen-components simultaneously.

5. Covariance Function Estimation

In this section, we propose a low-rank covariance function estimation framework based on functional unfolding operators and spectral regularizations. Spectral penalty functions (Abernethy et al. 2009; Wong and Zhang 2019) are defined as follows.

Definition 2 (Spectral penalty function). Given a compact operator $A$, a spectral penalty function takes the form $\Psi(A) = \sum_{k=1}^{\infty} \psi(\lambda_k(A))$ with the singular values of the operator $A$, $\lambda_1(A), \lambda_2(A), \ldots$ in a descending order of magnitude and a nondecreasing penalty function $\psi$ such that $\psi(0) = 0$.

Recall $\mathcal{H} = \bigotimes_{j=1}^p \mathcal{H}_j$ and $\mathcal{G} = \bigotimes_{j=1}^d \mathcal{G}_j$ where $d = 2p$, $\mathcal{G}_j = \mathcal{H}_j$ for $j = 1, \ldots, p$, and $\mathcal{G}_j = \mathcal{H}_{j-p}$ for $j = p+1, \ldots, d$. Clearly, a covariance operator is self-adjoint and positive semidefinite. Therefore, we consider the hypothesis space $\mathcal{M}^+ = \{\Gamma \in \mathcal{M} : (\Gamma(f)^2)_{\mathcal{H}} \geq 0, \forall f \in \mathcal{H}\}$, where $\mathcal{M} = \{\Gamma \in \mathcal{G} : \Gamma$ is self-adjoint$\}$, and propose a general class of covariance function estimators as follows:

$$\arg\min_{\Gamma \in \mathcal{M}^+} \left\{ \epsilon(\Gamma) + \lambda \left[ \beta \Psi_0(\Gamma) + \frac{1-\beta}{p} \sum_{j=1}^p \Psi_j(\Gamma(j)) \right] \right\},$$

where $\epsilon$ is a convex and smooth loss function, $\{\Psi_j : j = 1, \ldots, p\}$ are spectral penalty functions, and $\lambda \geq 0, \beta \in [0, 1]$ are tuning parameters. Here $\Psi_0$ penalizes the squarely unfolded operator $\Gamma_0$ while $\Psi_j$ regularizes one-way unfolded operator $\Gamma(j)$, respectively, for $j = 1, \ldots, p$. The tuning parameter $\beta$ controls the relative degree of regularization between one-way and two-way singular values. The larger the $\beta$ is, the more penalty is imposed on the two-way singular values relative to the one-way singular values. When $\beta = 1$, the penalization is only on the eigenvalues of the covariance operator (i.e., the two-way singular values), similarly as Wong and Zhang (2019).

To achieve low-rank estimation, we adopt a special form of (7):

$$\hat{\Gamma} = \arg\min_{\Gamma \in \mathcal{M}^+} \left\{ \epsilon_{\text{square}}(\Gamma) + \lambda \left[ \beta \|\Gamma\|_* + \frac{1-\beta}{p} \sum_{j=1}^p \|\Gamma(j)\|_* \right] \right\},$$

where $\|\cdot\|_*$ is the sum of singular values, also called trace norm, and $\epsilon_{\text{square}}$ is the squared error loss:

$$\epsilon_{\text{square}}(\Gamma) = \frac{1}{nm(m-1)} \sum_{i=1}^n \sum_{1 \leq j \neq f \leq m} (\Gamma(T_{ij1}, \ldots, T_{ijp}, T_{gf1}, \ldots, T_{gfP}) - Z_{ijf})^2,$$

with $Z_{ijf} = (Y_{ij} - \hat{\mu}(T_{ij1}, \ldots, T_{ijp}))(Y_{gf} - \hat{\mu}(T_{gf1}, \ldots, T_{gfP}))$, $\hat{\mu}$ as an estimate of the mean function, and $T_{ijk}$ is the $k$th element of location vector $T_i$. Notice that trace-norm regularizations promote low-rankness of the underlying operators, hence leading to a low-rank estimation in terms of both the one-way and two-way (covariance) ranks.

The main reason for penalizing one-way ranks is that only regularizing two-way ranks is often insufficient for dimension reduction for $p$-dimensional functional data. To nonparametrically estimate the covariance function for $p$-dimensional functional data, even if the rank of the covariance operator, that is, the two-way rank, is $R$, one needs to estimate $R$ eigenfunctions of which each is $p$-dimensional. Thus, unless $p$ is small, the covariance function estimation will suffer from the curse of dimensionality.

5.1. Representer Theorem and Parameterization

Before deriving a computational algorithm, we notice that the optimization (8) is an infinite-dimensional optimization
which is generally unsolvable. To overcome this challenge, we show that the solution to (8) always lies in a known finite-dimensional sub-space given data, hence allowing a finite-dimensional parameterization. Indeed, we are able to achieve a stronger result in Theorem 1 which holds for the general class of estimators obtained by (7). In below, we use $\hat{\Gamma}$ to represent a generic solution to (7). Let $L_{n,m} = \{T_{ij} : i = 1,\ldots,n, j = 1,\ldots,m, k = 1,\ldots,p\}$.

**Theorem 1 (Representer theorem).** If the solution set of (7) is not empty, then the solution $\hat{\Gamma}$ lies in the space $G(L_{n,m}) := \bigotimes_{k=1}^{p} K_{k}$, where $K_{p+k} = K_{k}$ and

$$K_{k} = \text{span} \{ K_{k}(T_{ij}, \cdot) : i = 1,\ldots,n, j = 1,\ldots,m \},$$

for $k = 1,\ldots,p$. It also takes the form:

$$\hat{\Gamma}(s_{1},\ldots,s_{p},t_{1},\ldots,t_{p}) = A \times_{1} z_{1}^{T}(s_{1}) \times_{2} z_{2}^{T}(s_{2}) \cdots \times_{p} z_{p}^{T}(s_{p}) \times_{p+1} z_{1}^{T}(t_{1}) \cdots \times_{2p} z_{2p}^{T}(t_{2p}),$$

(10)

where the $l$th element of $z_{k}(\cdot) \in \mathbb{R}^{m \times n}$ is $K(T_{ij}, \cdot)$ with $l = (i-1)n + j$, $A$ is a $2p$th order tensor where the dimension of each mode is $nm$ and $A_{k}$ is a symmetric matrix.

The proof of Theorem 1 is given in Section S2 of the SM. By Theorem 1, we can now only focus on covariance function estimators of the form (10). Let $B = A \times_{1} M_{1}^{T} \cdots \times_{p} M_{p}^{T} \times_{p+1} M_{p+1}^{T} \cdots \times_{2p} M_{2p}^{T}$, where $M_{k}$ is an $nm \times q_{k}$ matrix such that

$$M_{k} M_{k}^{T} = K = \left[ K(T_{ij}, T_{i'j'}) \right]_{1 \leq i,j \leq n, 1 \leq i', j' \leq m}.$$

(11)

With $B$, we can express

$$\hat{\Gamma}(s_{1},\ldots,s_{p},t_{1},\ldots,t_{p}) = B \times_{1} \{ M_{1}^{T} z_{1}(s_{1}) \} \cdots \times_{p} \{ M_{p}^{T} z_{p}(s_{p}) \}^{T} \times_{p+1} \{ M_{p+1}^{T} z_{1}(t_{1}) \} \cdots \times_{2p} \{ M_{2p}^{T} z_{p}(t_{p}) \}^{T},$$

(12)

where $z_{k}(\cdot)$ is defined in Theorem 1 and $M_{k}^{T}$ is the Moore–Penrose inverse of matrix $M_{k}$.

For smooth modeling of $X_{t}$, the selected reproducing kernel is often smooth, which leads to an approximately low-rank kernel matrix $K_{k}$. Thus, computationally, one can find a matrix $M_{k}$ with the number of columns $q_{k}$ much smaller than the rank of $K_{k}$ and $nm$, such that the first equation in (11) approximately holds. This greatly reduces the dimensions of $B$, and thus benefits the computation.

Ideally we can obtain $M_{k}$ by the “best” low-rank approximation with respect to the Frobenius norm by eigen-decomposition, but a full eigen-decomposition is computationally expensive. Instead, randomized algorithms can be used to obtain low-rank approximations in an efficient manner (Halko, Martinsson, and Tropp 2009).

One can easily show that the eigenvalues of the operator $\hat{\Gamma}$ are the same as those of the matrix $B_{0}$ and that the singular values of the operator $\tilde{\Gamma}$ are the same as those of the matrix $B_{(j)}$. Therefore, solving (8) is equivalent to solving the following optimization:

$$\tilde{B} = \text{argmin}_{B} \left\{ \ell_{\text{square}}(B) + \lambda \beta \text{h}(B_{0}) + \frac{1 - \beta}{p} \sum_{k=1}^{p} \text{\| } B_{(j)} \text{\| }^{2}_{\text{F}} \right\},$$

(13)

where $\lambda \| \cdot \|_{\text{F}}$ also represents the trace norm of matrices, $h(H) = \| H_{k} \|_{\text{f}}$ if matrix $H$ is positive semidefinite, and $h(H) = \infty$ otherwise, and $\ell_{\text{square}}(B) = \ell_{\text{square}}(\Gamma)$, where $\tilde{\Gamma}$ is constructed from (12). Then $\tilde{\Gamma}$ in (8) can be obtained by substituting $\tilde{B}$ into (12).

Beyond estimating the covariance function, one may be further interested in the eigen-decomposition of $\hat{\Gamma}$ via the $L^{2}$ inner product, for example, to perform functional principal component analysis in the usual sense. Due to the finite-dimensional parameterization, a closed-form expression of $L^{2}$ eigen-decomposition can be derived from our estimator without further discretization or approximation. In addition, we can obtain a similar one-way analysis in terms of the $L_{2}$ inner product. We can define a $L^{2}$ singular value decomposition via the Tucker form and obtain the $L^{2}$ marginal basis. Details are given in Appendix A.

### 5.2. Computational Algorithm

We solve (13) by the accelerated alternating direction method of multipliers (ADMM) algorithm (Kadkhodaie et al. 2015). We begin with an alternative form of (13):

$$\min_{B \in \mathbb{R}^{q_{1} \times \cdots \times q_{2p}}} \left\{ \ell_{\text{square}}(B) + \lambda \beta \text{h}(D_{0}) + \frac{1 - \beta}{p} \sum_{k=1}^{p} \| D_{(j)} \|_{\text{F}}^{2} \right\},$$

subject to $B = D_{0} = D_{1} = \cdots = D_{p}$,

(14)

(15)

where $q_{p+k} = q_{k}$ for $k = 1,\ldots,p$. Then a standard ADMM algorithm solves the optimization problem (14) by minimizing the augmented Lagrangian with respect to different variables alternatively. More explicitly, at the $t$th iteration, the following updates are implemented:

$$B^{(t+1)} = \text{argmin}_{B} \left\{ \ell_{\text{square}}(B) + \frac{\eta}{2} \| B_{0} - D^{(t)} + V_{(k)}^{(t)} \|_{\text{F}}^{2} \right\},$$

(16a)

$$D^{(t+1)} = \text{argmin}_{D_{k}} \left\{ \lambda \beta \text{h}(D_{0}) + \frac{\eta}{2} \| D_{(k)}^{(t+1)} - D_{(k)} + V_{(k)}^{(t)} \|_{\text{F}}^{2} \right\},$$

(16b)

$$D^{(t+1)} = \text{argmin}_{D_{k}} \left\{ \lambda \frac{1 - \beta}{p} \| D_{(k)}^{(t+1)} - D_{(k)} + V_{(k)}^{(t)} \|_{\text{F}}^{2} \right\},$$

(16c)

where $V_{k} \in \mathbb{R}^{q_{1} \times \cdots \times q_{2p}}$, for $k = 0,\ldots,p$, are scaled Lagrange multipliers and $\eta > 0$ is an algorithmic parameter. An adaptive strategy to tune $\eta$ is provided in Boyd et al. (2010). One can see that Steps (16a), (16b), and (16c) involve additional optimizations. Now we discuss how to solve them.

The objective function of (16a) is a quadratic function, and so we can easily solve this with a closed-form solution, given in line 2 of Algorithm 1. To solve (16b) and (16c), we use proximal operator $\text{prox}^{\ell}_{\psi}$, $k = 1,\ldots,p$ and $\text{prox}^{\ell}_{\psi} : \mathbb{R}^{q_{1} \times \cdots \times q_{2p}} \rightarrow \mathbb{R}^{q_{1} \times \cdots \times q_{2p}}$. 

for $v \geq 0$. By Lemma 1 in Mazumder, Hastie, and Tibshirani (2010), the solutions to (17) have closed forms. For (17a), write the singular value decomposition of $A(k)$ as $U\text{diag}((\tilde{a}_1, \ldots, \tilde{a}_q))V^\top$, then $[\text{prox}^v_y(A)](k) = U\text{diag}(\tilde{v})V^\top$ where $\tilde{v} = ((\tilde{a}_1 - v)_+, (\tilde{a}_2 - v)_+, \ldots, (\tilde{a}_q - v)_+)$. As for (17b), the solution is restricted to be a symmetric matrix since the penalty $h$ equals infinity otherwise. Thus, (17b) is equivalent to minimizing $\left\{ \frac{1}{2}\|W_\ast - (A_\ast + A_\ast^\top)/2\|_F^2 + vh(W_\ast) \right\}$ when $\langle W_\ast, (A_\ast - A_\ast^\top)/2 \rangle = \langle W_\ast + W_\ast^\top, (A_\ast - A_\ast^\top)/2 \rangle = 0$. Suppose that $(A_\ast + A_\ast^\top)/2$ yields eigen-decomposition $P\text{diag}(\tilde{c}_1, \ldots, \tilde{c}_q)P^\top$. Then $[\text{prox}^v_y(A)] = P\text{diag}(\tilde{v})P^\top$, where $\tilde{v} = ((\tilde{c}_1 - v)_+, (\tilde{c}_2 - v)_+, \ldots, (\tilde{c}_q - v)_+)$. Unlike singular values, the eigenvalues may be negative. Hence, as opposed to $\text{prox}^v_y$, this procedure $\text{prox}^+_{\ast y}$ also removes eigen-components with negative eigenvalues.

The details of computational algorithm are given in Algorithm 1, an accelerated version of ADMM which involves additional steps for a faster algorithmic convergence.

### 6. Asymptotic Properties

In this section, we conduct an asymptotic analysis for the proposed estimator $\hat{\Gamma}$ as defined in (8). Our analysis has a unified flavor such that the derived convergence rate of the proposed estimator automatically adapts to sparse and dense settings. Throughout this section, we neglect the mean function estimation error by setting $\mu_0(t) = \tilde{\mu}(t) = 0$ for any $t \in \mathcal{T}$, which leads to a cleaner and more focused analysis. The additional error from the mean function estimation can be incorporated into our proofs without any fundamental difficulty.

#### 6.1. Assumptions

Without loss of generality let $\mathcal{T} = [0, 1]^p$. The assumptions needed in the asymptotic results are listed as follows.

**Assumption 1.** Sample fields $\{X_i : i = 1, \ldots, n\}$ reside in $\mathcal{H} = \bigotimes_{k=1}^p \mathcal{H}_k$ where $\mathcal{H}_k$ is an RKHS of functions on $[0, 1]$ with a continuous and square integrable reproducing kernel $K_k$.

**Assumption 2.** The true (folded) covariance function $\Gamma_0 \neq 0$ and $\Gamma_0 \in \mathcal{G} = \bigotimes_{j=1}^d \mathcal{G}_j$, where $d = 2p$, $\mathcal{G}_j = \mathcal{H}_j$ for $j = 1, \ldots, p$ and $\mathcal{G}_j = \mathcal{H}_{j-p}$ for $j = p + 1, \ldots, d$.

**Assumption 3.** The locations $\{T_{ij} : i = 1, \ldots, n; j = 1, \ldots, m\}$ are independent random vectors from Uniform,$[0, 1]^p$, and they are independent of $\{X_i : i = 1, \ldots, n\}$. The errors $\{e_{ij} : i = 1, \ldots, n; j = 1, \ldots, m\}$ are independent of both locations and sample fields.

**Algorithm 1:** Accelerated ADMM for solving (13)

**Input:** $V_k^{(0)} \in \mathbb{R}^{q_k \times \cdots q_p}$, $k = 0, 1, \ldots, p$, and $B(0) \in \mathbb{R}^{q_k \times \cdots q_p}$ such that $V_k^{(0)}$ and $B_k^{(0)}$ are symmetric matrices; $M_k = [M_{1k}, \ldots, M_{nk}]^\top$, $k = 1, \ldots, p; G_t = (Z_{ij}/1\leq j \leq m_i = 1, \ldots, n_i; I = [1 \leq i \leq m; n_i = 0] > 0; T$

**Initialization:** $\alpha_k^{(0)} \leftarrow 1, D_k^{(-1)} \leftarrow B_k^{(-1)}$, $V_k^{(-1)} = V_k^{(0)}$, $k = 1, \ldots, p$

$L_i \leftarrow [M_{1i}^\top \odot \cdots \odot M_{ki}^\top]^\top$ for $i = 1, \ldots, n$, where $\odot$ is the Khatri–Rao product defined as

$A \odot B = [a_1 b_1; \ldots a_n b_n] \in \mathbb{R}^{n \times k}$ for $A \in \mathbb{R}^{n \times k}$, $B \in \mathbb{R}^{m \times k}$ and $a_i, b_i$ $i$th column of matrices $A$ and $B$.

$G \leftarrow \frac{1}{n(m-n)} \sum_{i=1}^n (L_i \odot L_i)^\top \text{diag}((vec(I))(I) \odot L_i)\odot L_i$ $h \leftarrow \frac{1}{n(m-n)} \sum_{i=1}^n (L_i \odot L_i)^\top \text{diag}((vec(I))\text{vec}(Z_i))\ainv Q \leftarrow (2(G + \frac{p+1}{2} \ast I)^{-1}$

for $t = 0, 1, \ldots, T$

vec $(B_k(t+1)) \leftarrow Q(h + \eta \sum_{k=0}^p \text{vec}[(D_k(t) - \hat{V}_k(t)] \ast a)$

for $k = 0, 1, \ldots, p$

if $k = 0$ then

$D_0(t) \leftarrow \text{prox}^\pm_{\beta(0)}(B_k^{(t+1)} + \hat{V}_k^{(t)})$

else

$D_k^{(t)} \leftarrow \text{prox}^\pm_{\lambda(t)}(B_k^{(t+1)} + \hat{V}_k^{(t)})$

end

$V_k(t) \leftarrow \hat{V}_k^{(t)} + B_k^{(t+1)} - D_k^{(t)}$

$\alpha_k^{(t+1)} \leftarrow \frac{1 + \sqrt{1 + 4(\alpha_k^{(t)})^2}}{2}$

$D_k^{(t+1)} \leftarrow D_k^{(t)} + \frac{\alpha_k^{(t+1)} - 1}{\alpha_k^{(t+1)}}(D_k^{(t)} - D_k^{(t-1)})$

$\hat{V}_k^{(t+1)} \leftarrow V_k(t) + \frac{\alpha_k^{(t+1)} - 1}{\alpha_k^{(t+1)}}(V_k(t) - V_k^{(t-1)})$

end

Stop if objective value change less than tolerance.

Output: $D_0^{(T)}$
distribution of which density function $\pi$ satisfies $c_\pi \leq \pi(t) \leq c_\pi'$ for all $t \in [0, 1]^p$, for some constants $0 < c_\pi < c_\pi' < 1$, to ensure both Theorems 2 and 3 still hold. Assumptions 4 and 5 involve sub-Gaussian conditions of the stochastic process and measurement error, which are standard tail conditions.

### 6.2. Reproducing Kernels

In Assumption 1, the “smoothness” of the function in the underlying RKHS is not explicitly specified. It is well known that such smoothness conditions are directly related to the eigen-decay of the respective reproducing kernel. By Mercer’s theorem (Mercer 1909), the reproducing kernel $K_H((t_1, \ldots, t_p), (t'_1, \ldots, t'_p))$ possesses the eigen-decomposition

$$
K_H((t_1, \ldots, t_p), (t'_1, \ldots, t'_p)) = \sum_{l=1}^{\infty} \mu_l \phi_l(t_1, \ldots, t_p) \phi_l(t'_1, \ldots, t'_p),
$$

where $\{\mu_l : l \geq 1\}$ are nonnegative eigenvalues and $\{\phi_l : l \geq 1\}$ are $L^2$ eigenfunctions on $[0,1]^p$. Then for the space $H \otimes H$, which is also identified by $G = \bigotimes_{k=1}^{d} G_k$, its corresponding reproducing kernel $K_G$ has the following eigen-decomposition

$$
K_G((x_1, \ldots, x_{2p}), (x'_1, \ldots, x'_{2p})) = \sum_{l,h=1}^{\infty} \mu_l \mu_h \phi_l(x_1, \ldots, x_p) \phi_h(x'_1, \ldots, x'_p) \phi_l(x_{p+1}, \ldots, x_{2p}) \phi_h(x'_{p+1}, \ldots, x'_{2p}),
$$

where $\{\mu_l \mu_h : l, h \geq 1\}$ are the eigenvalues of $K_G$. Due to continuity assumption (Assumption 1) of the univariate kernels, there exists a constant $b$ such that

$$
\sup_{(x_1, \ldots, x_{2p}) \in [0,1]^{2p}} K_G((x_1, \ldots, x_{2p}), (x_1, \ldots, x_{2p})) \leq b.
$$

The decay rate of the eigenvalues $\{\mu_l \mu_h : l, h \geq 1\}$ is involved in our analysis through two quantities $\kappa_{n,m}$ and $\eta_{n,m}$, which have relatively complex forms defined in Appendix B. Similar quantities can be found in other analyses of RKHS-based estimators (e.g., Raskutti, Wainwright, and Yu 2012) that accommodate general choices of RKHS. Generally, $\kappa_{n,m}$ and $\eta_{n,m}$ are expected to diminish in certain orders of $n$ and $m$, characterized by the decay rate of the eigenvalues $\{\mu_l \mu_h\}$. The smoother the functions in the RKHS, the faster these two quantities diminish. Our general results in Theorems 2 and 3 are specified in terms of these quantities. To provide a solid example, we derive the orders of $\kappa_{n,m}$ and $\eta_{n,m}$ under a Sobolev–Hilbert space setting and provide the convergence rate of the proposed estimator in Corollary 1.

### 6.3. Unified Rates of Convergence

We write the penalty in (8) as $I(\Gamma) = \beta \| \Gamma \|_s + (1 - \beta)p^{-1} \sum_{k=1}^{p} \| \Gamma(h_k) \|_s$. For arbitrary functions $g_1, g_2 \in G$, define their empirical inner product and the corresponding (squared) empirical norm as

$$
\langle g_1, g_2 \rangle_{n,m} = \frac{1}{nm(m-1)} \sum_{i=1}^{n} \sum_{1 \leq j' \leq m} g_1(T_{ij_1}, \ldots, T_{ij_p}),
$$

$$
\| g_1 \|_{n,m}^2 = \langle g_1, g_1 \rangle_{n,m}.
$$

Additionally, the $L^2$ norm of a function $g$ is defined as $\|g\|_2 = (\int_T g^2(t) dt)^{1/2}$.

Define $\xi_{n,m} = \max\{\eta_{n,m}, \kappa_{n,m}, (n^{-1} \log n)^{1/2}\}$. We first provide the empirical $L^2$ rate of convergence for $\hat{\Gamma}$.

**Theorem 2.** Suppose that Assumptions 1–5 hold. If $\xi_{n,m}$ satisfies $(\log n)/n \leq \xi_{n,m}^2/(\log \log \xi_{n,m})$, and $\lambda \geq L_1 \xi_{n,m}$ for some constant $L_1 > 0$ depending on $b_X, b_e$ and $b$, we have

$$
\| \hat{\Gamma} - \Gamma_0 \|_{n,m} \leq \sqrt{2I(\Gamma_0)\lambda} + L_2 \xi_{n,m},
$$

with probability at least $1 - \exp(-cn\xi_{n,m}^2/\log n)$ for some positive universal constant $c$.

Next, we provide the $L^2$ rate of convergence for $\hat{\Gamma}$.

**Theorem 3.** Under the same conditions as Theorem 2, there exists a positive constant $L_2$ depending on $b_X, b_e$ and $I(\Gamma_0)$, such that

$$
\| \hat{\Gamma} - \Gamma_0 \|_2 \leq 2\sqrt{I(\Gamma_0)\lambda} + L_2 \xi_{n,m},
$$

with probability at least $1 - \exp(-c_P n \xi_{n,m}^2/\log n)$ for some constant $c_P$ depending on $b$.

The proofs of Theorems 2 and 3 can be found in Section S2 in the SM. Theorems 2 and 3 are applicable to general RKHS $H$ which satisfies Assumption 1. The convergence rate depends on the eigen-decay rates of the reproducing kernel. A special case of polynomial decay rates for univariate RKHS will be given in Corollary 1. Moreover, our analysis has a unified flavor in the sense that the resulting convergence rates automatically adapt to the orders of both $n$ and $m$. After Corollary 1, we will provide a discussion of a “phase transition” between dense and sparse functional data revealed by our theory.

**Remark 3.** With a properly chosen $\lambda$, Theorems 2 and 3 bound the convergence rates (in terms of both the empirical and theoretical $L^2$ norm) by $\xi_{n,m}$, which cannot be faster than $(n^{-1} \log n)^{1/2}$. The logarithmic order is due to the use of Adamczak bound in Lemma S2 in the SM. If one further assumes boundedness for the sample fields $X_i$’s (in terms of the sup-norm) and the noise variables $e_i$’s, we can instead use Talagrand concentration inequality (Bousquet bound in Koltchinskii (2011)) and the results in Theorems 2 and 3 can be improved to max$\{\| \hat{\Gamma} - \Gamma_0 \|_{n,m}^2, \| \hat{\Gamma} - \Gamma_0 \|_2^2\} = O_P(\xi_{n,m}^2)$, where $\xi_{n,m} = \max\{\eta_{n,m}, \kappa_{n,m}, n^{-1/2}\}$.
Next we focus on a special case where the reproducing kernels of the univariate RKHS $\mathcal{H}_k$’s exhibit polynomial eigen-decay rates, which holds for a range of commonly used RKHS. A canonical example is orthonormal Sobolev–Hilbert space: $$\mathcal{H}_k = \{ f : f^{(r)} \in L^2([0, 1]), r = 0, \ldots, \alpha, \text{are absolutely continuous};
$$ $$f^{(\alpha)} \in L^2([0, 1]) \},$$ where $k = 1, \ldots, p$. Here $\alpha$ is the same as $\alpha$ in Corollary 1. To derive the convergence rates, we relate the eigenvalues $\nu_l$ in (18) to the univariate RKHS $\mathcal{H}_k, k = 1, \ldots, p$. Due to Mercer’s theorem, the reproducing kernel $K_k$ of $\mathcal{H}_k$ yields an eigen-decomposition with nonnegative eigenvalues $\{\mu_{\ell}^{(k)}: l \geq 1\}$ and an $L^2$ eigenfunction $\{\phi_{\ell}^{(k)}: l \geq 1\}$, that is, $K_k(t, t') = \sum_{l=1}^{\infty} \mu_{\ell}^{(k)} \phi_{\ell}^{(k)}(t) \phi_{\ell}^{(k)}(t')$. Therefore, the set of eigenvalues $\{\mu_{\ell}: l \geq 1\}$ in (18) is the same as the set $\{\prod_{k=1}^{p} \mu_{\ell}^{(k)}: l_1, \ldots, l_p \geq 1\}$. Given the eigen-decay of $\mu_{\ell}^{(k)}$, one can obtain the order of $\xi_{n,m}$ and hence the convergence rates from Theorems 2 and 3. Here are the results under the setting of a polynomial eigen-decay.

**Corollary 1.** Suppose that the same conditions in Theorem 3 hold. If the eigenvalues of $K_k$ for $\mathcal{H}_k, k = 1, \ldots, p$, have polynomial decaying rates, that is, there exists $\alpha > 1/2$ such that $\mu_{\ell}^{(k)} \asymp \ell^{-2\alpha}$ for all $k = 1, \ldots, p$, then

$$\max \left\{ \| \hat{\Gamma} - \Gamma_0 \|_{n,m}, \| \hat{\Gamma} - \Gamma_0 \|_2 \right\} = O_{p}\left( \max \left\{ (nm)^{-\frac{1}{2}}, (\log(nm))^{2\alpha - 1}, \log \frac{n}{m} \right\} \right).$$

The proof of Corollary 1 can be found in Section S2 in the SM. All Theorems 2 and 3 and Corollary 1 reveal a “phase-transition” of the convergence rate depending on the relative magnitudes between $n$, the sample size, and $m$, the number of observations per field. When $\kappa_{n,m}^2 \ll (\log n/n)$, which is equivalent to $m \gg n^{1/(2\alpha)} (\log n)^{2p-2-1/(2\alpha)}$ in Corollary 1, both empirical and theoretical $L^2$ rates of convergence can achieve the near-optimal rate $\sqrt{\log n/n}$. Under the stronger assumptions in Remark 3, the convergence rate will achieve the optimal order $\sqrt{1/n}$ when $\kappa_{n,m}^2 \ll 1/n$ (or $m \gg n^{1/(2\alpha)} (\log n)^{2p-2-1}$ in Corollary 1). In this case, the observations are so densely sampled that we can estimate the covariance function as precisely as if the entire sample fields are observable. On the contrary, when $\kappa_{n,m}^2 \gg (\log n/n)$ (or $m \ll n^{1/(2\alpha)} (\log n)^{2p-2-1/(2\alpha)}$ in Corollary 1), the convergence rate is determined by the total number of observations $nm$. When $p = 1$, the asymptotic result in Corollary 1, up to some log $m$ and log $n$ terms, is the same as the minimax optimal rate obtained by Cai and Yuan (2010), and is comparable to the $L^2$ rate obtained by Paul and Peng (2009) for $\alpha = 2$.

Our estimator is closely related to Wong and Zhang (2019), but the corresponding theories are substantially distinct. First, the theoretical results of Wong and Zhang (2019) hold only for one-dimensional functional data and Sobolev–Hilbert spaces, while our results apply to multidimensional functional data and general RKHS. Moreover, unlike Theorems 2, 3 or Corollary 1, Wong and Zhang (2019) did not provide unified theories. Their theories can only achieve nonparametric rates which do not change over $m$, so they are not optimal for dense functional data as $m$ goes to infinity. In contrast, as shown in Corollary 1 by taking $p = 1$, our result gives a significantly better rate when $m$ diverges. If $m$ is bounded, by Theorem 3 of Wong and Zhang (2019), their best rate $n^{-2\alpha/(1+2\alpha)} \log(n)$ is obtained if the covariance function estimator is only searched among periodic functions. In comparison, Corollary 1 indicates that our rate is $n^{-2\alpha/(1+2\alpha)} \log(n)^{(2\alpha)/(1+2\alpha)}$ for one-dimensional functional data, that is, $p = 1$, and bounded $m$, which is slightly better than theirs, even if we do not assume periodic covariance functions in our theory.

For covariance function estimation for unidimensional functional data, that is, $p = 1$, a limited number of approaches, including Cai and Yuan (2010), Li and Hsing (2010), Zhang and Wang (2016), and Liebl (2019), can achieve unified theoretical results in the sense that they hold for all relative magnitudes of $n$ and $m$. The similarity of these approaches is the availability of a closed form for each covariance function estimator. In contrast, our estimator obtained from (8) does not have a closed form due to the non-differentiability of the penalty, but it can still achieve unified theoretical results which hold for both unidimensional and multidimensional functional data. Due to the lack of a closed form of our covariance estimator, we used the empirical process techniques (e.g., Bartlett, Bousquet, and Mendelson 2005; Koltchinskii 2011) in the theoretical development. In particular, we have developed a novel grouping lemma (Lemma S4 in the SM) to deterministically decouple the dependence within a U-statistics of order 2. We believe this lemma is of independent interest. In our analysis, the corresponding U-statistics is indexed by a function class, and this grouping lemma provides a tool to obtain uniform results (see Lemma S3 in the SM). In particular, this allows us to relate the empirical and theoretical $L^2$ norm of the underlying function class, in precise enough order dependence on $n$ and $m$ to derive the unified theory. See Lemma S3 for more details. To the best of our knowledge, this article is one of the first in the FDA literature that derives a unified result in terms of empirical process theories, and the proof technique is potentially useful for some other estimators without a closed form.

### 7. Simulation

To evaluate the practical performance of the proposed method, we conducted a simulation study. We in particular focused on two-dimensional functional data. Let $\mathcal{H}_1$ and $\mathcal{H}_2$ both be the RKHS with kernel $K(t_1, t_2) = \sum_{k=1}^{\infty} (k\pi)^{-4} \phi_k(t_1) \phi_k(t_2)$, where $e_k(t) = \sqrt{2} \cos(k\pi t)$, $k \geq 1$. This RKHS has been used in various studies in FDA, for example, the simulation study of Cai and Yuan (2012). Each $X_t$ is generated from a mean-zero Gaussian random field with a covariance function

$$\gamma_0((s_1, s_2), (t_1, t_2)) = \Gamma_0(s_1, s_2, t_1, t_2) = \sum_{k=1}^{R} k^{-\alpha} \psi_k(s_1, s_2) \psi_k(t_1, t_2),$$

where the eigenfunctions $\psi_k(t_1, t_2) \in P_{r_1, r_2} := \{ e_k(t_1) e_j(t_2) : i = 1, \ldots, r_1; j = 1, \ldots, r_2 \}$, and $\alpha$ controls the decay rate of eigenvalues. We considered three different choices of decay rates ($\alpha = 1.1, 2, 4$). Due to space limitation, we only present the
results for $\sigma = 2$ here. Corresponding results of the other two decay rates can be found in Section 3.1 of the SM.

Three combinations of one-way ranks $(r_1, r_2)$ and two-way rank $R$ were studied for $\Gamma_2$:  

Setting 1: $R = 6$, $r_1 = 3$, $r_2 = 2$;  
Setting 2: $R = 6$, $r_1 = 2$;  
Setting 3: $R = r_1 = r_2 = 4$.

For each setting, we chose $R$ functions out of $\mathcal{P}_{r_1, r_2}$ to be $\{\psi_k\}$ such that smoother functions are associated with larger eigenvalues. The details are described in Section S3.1.1 of the SM.

In terms of sampling plans, we considered both sparse and dense designs. Here, we only show and discuss the results for the sparse design, while defer those for the dense design to Section S3.1.1 of the SM.

For the sparse design, the random locations $T_{ij} = 1, \ldots, m$, were independently generated from the continuous uniform distribution on $[0, 1]^2$ within each field and across different fields, and the random errors $\{e_{ij} : i = 1, \ldots, n; j = 1, \ldots, m\}$ were independently generated from $N(0, \sigma^2)$. In each of the 200 simulation runs, the observed data were obtained following (1) with various combinations of $m = 10, 20$, $n = 100, 200$ and noise level $\sigma = 0.1, 0.4$.

We compared the proposed method, denoted by mOpCov, with three existing methods: (1) OpCov: the estimator based on Wong and Zhang (2019) with adaption to two dimensional case (see Section 2); (2) II-smooth: local linear smoothing with Epanechnikov kernel; (3) II-smooth+: the two-step estimator constructed by retaining eigen-components of II-smooth selected by 99% fraction of variation explained (FVE), and then removing the eigen-components with negative eigenvalues. For both OpCov and mOpCov, 5-fold cross-validation was adopted to select the corresponding tuning parameters.

Table 1 shows the average integrated squared error (AISE), average of estimated two-way rank $(\bar{R})$, as well as average of estimated one-way ranks $(\bar{r}_1, \bar{r}_2)$ of the above covariance estimators over 200 simulated datasets in respective settings when sample size is $n = 200$. Corresponding results for $n = 100$ can be

| Setting | $m$ | $\sigma$ | AISE | OpCov | II-smooth | II-smooth+ |
|---------|-----|----------|------|--------|-----------|-----------|
| 1       | 10  | 0.1      | 0.053 (1.96e−03) | 0.0632 (3.22e−03) | 0.652 (1.92e−01) | 0.337 (5.35e−02) |
|         |     |          | $\bar{R}$ | 8.35 | 2.94 | – | 172.70 |
|         |     |          | $\bar{r}_1, \bar{r}_2$ | 5.22, 5.19 | – | – | – |
| 20      | 0.4 |          | 0.0527 (1.39e−03) | 0.0656 (2.72e−03) | 0.714 (2.11e−01) | 0.366 (5.96e−02) |
|         |     |          | $\bar{R}$ | 9.16 | 2.84 | – | 177.3 |
|         |     |          | $\bar{r}_1, \bar{r}_2$ | 5.29, 5.26 | – | – | – |
| 2       | 0.4 |          | 0.0340 (1.35e−03) | 0.0421 (1.97e−03) | 0.297 (1.39e−02) | 0.206 (4.62e−03) |
|         |     |          | $\bar{R}$ | 8.265 | 3.78 | – | 314.47 |
|         |     |          | $\bar{r}_1, \bar{r}_2$ | 5.81, 5.8 | – | – | – |
| 2       | 0.4 |          | 0.0349 (1.38e−03) | 0.044 (2.21e−03) | 0.325 (1.58e−02) | 0.223 (4.94e−03) |
|         |     |          | $\bar{R}$ | 8.86 | 3.76 | – | 326.31 |
|         |     |          | $\bar{r}_1, \bar{r}_2$ | 5.8, 5.8 | – | – | – |
| 2       | 0.4 |          | 0.0516 (1.96e−03) | 0.0636 (3.12e−03) | 2.33 (1.13e+00) | 0.795 (2.98e−01) |
|         |     |          | $\bar{R}$ | 8.48 | 3.02 | – | 191.75 |
|         |     |          | $\bar{r}_1, \bar{r}_2$ | 5.47, 5.47 | – | – | – |
| 20      | 0.4 |          | 0.0532 (1.96e−03) | 0.0686 (3.53e−03) | 2.44 (1.17e+00) | 0.828 (3.04e−01) |
|         |     |          | $\bar{R}$ | 9.04 | 3.04 | – | 196.34 |
|         |     |          | $\bar{r}_1, \bar{r}_2$ | 5.45, 5.43 | – | – | – |
| 2       | 0.4 |          | 0.0339 (1.39e−03) | 0.0419 (2.02e−03) | 0.301 (1.58e−02) | 0.208 (4.50e−03) |
|         |     |          | $\bar{R}$ | 8.745 | 3.74 | – | 318.645 |
|         |     |          | $\bar{r}_1, \bar{r}_2$ | 5.87, 5.88 | – | – | – |
| 2       | 0.4 |          | 0.0349 (1.43e−03) | 0.043 (2.22e−03) | 0.328 (1.78e−02) | 0.225 (4.74e−03) |
|         |     |          | $\bar{R}$ | 7.785 | 3.6 | – | 327.395 |
|         |     |          | $\bar{r}_1, \bar{r}_2$ | 5.75, 5.75 | – | – | – |
| 3       | 0.4 |          | 0.0581 (2.68e−03) | 0.0692 (5.33e−03) | 0.454 (7.28e−02) | 0.286 (2.89e−02) |
|         |     |          | $\bar{R}$ | 6.76 | 3.12 | – | 182.74 |
|         |     |          | $\bar{r}_1, \bar{r}_2$ | 4.07, 4.165 | – | – | – |
| 20      | 0.4 |          | 0.0602 (2.76e−03) | 0.0733 (6.14e−03) | 0.531 (1.07e−01) | 0.323 (4.23e−02) |
|         |     |          | $\bar{R}$ | 6.93 | 3.2 | – | 185.82 |
|         |     |          | $\bar{r}_1, \bar{r}_2$ | 4.12, 4.17 | – | – | – |
| 2       | 0.4 |          | 0.0399 (1.32e−03) | 0.0535 (2.64e−03) | 0.267 (5.04e−03) | 0.196 (3.59e−03) |
|         |     |          | $\bar{R}$ | 6.17 | 4.49 | – | 332.09 |
|         |     |          | $\bar{r}_1, \bar{r}_2$ | 3.50, 3.49 | – | – | – |
| 2       | 0.4 |          | 0.0405 (1.33e−03) | 0.0494 (2.42e−03) | 0.292 (5.30e−03) | 0.212 (3.72e−03) |
|         |     |          | $\bar{R}$ | 6.12 | 3.36 | – | 338.725 |
|         |     |          | $\bar{r}_1, \bar{r}_2$ | 3.53, 3.55 | – | – | – |

NOTE: The AISE values with standard errors (SE) in parentheses are provided for the four covariance estimators in comparison, together with average two-way ranks ($\bar{R}$) for those estimators which can lead to rank reduction (i.e., mOpCov, OpCov, and ll-smooth+) and average one-way ranks ($\bar{r}_1, \bar{r}_2$) for mOpCov.
found in Table S5 of the SM, and they lead to similar conclusions. Obviously ll-smooth and ll-smooth+, especially ll-smooth, perform significantly worse than the other two methods in both estimation accuracy and rank reduction (if applicable). Below we only compare mOpCov and OpCov.

Regarding estimation accuracy, the proposed mOpCov has uniformly smaller AISE values than OpCov, with around 10%–20% improvement of AISE over OpCov in most cases under Settings 1 and 2. If the standard error (SE) of AISE is taken into account, the improvements of AISE by mOpCov are more distinguishable in Settings 1 and 2 than those in Setting 3 since the SEs of OpCov in Setting 3 are relatively high. This is due to the fact that in Setting 3, the marginal basis is not shared by different two-dimensional eigenfunctions, and hence mOpCov cannot benefit from the structure sharing among eigenfunctions. Setting 3 is in fact an extreme setting we designed to challenge the proposed method.

For rank estimation, OpCov almost always underestimates two-way ranks, while mOpCov typically overestimates both one-way and two-way ranks. For mOpCov, the average one-way rank estimates are always smaller than the average two-way rank estimates, and their differences are substantial in Settings 1 and 2. This demonstrates the benefit of mOpCov of detecting structure sharing of one-dimensional basis among two-dimensional eigenfunctions.

We also tested the performance of mOpCov in the dense and regular designs, and compared it with the existing methods mentioned above together with the one by Wang and Huang (2017), which is not applicable to the sparse design. Details are given in Section S3.1.3 of the SM. Overall, all methods achieve similar AISE values, but mOpCov performs slightly better in estimation accuracy in smoother cases (i.e., those with larger $\alpha$), and when the noise level is high.

We also investigated the performance of mOpCov when the true covariance function is of high rank and obtained a similar conclusion as above. See details in Section S3.2 of the SM.

8. Real Data Application

We applied the proposed method to an Argo profile dataset, obtained from https://argo.ucsd.edu. The Argo project has a global array of approximately 3800 free-drifting profiling floats, which measure temperature and salinity of the ocean. These floats drift freely in the depths of the ocean most of the time, and ascend regularly to the sea surface, where they transmit the collected data to the satellites. Every day only a small subset of floats show up on the sea surface. Due to the drifting process, these floats measure temperature and salinity at irregular locations over the ocean. See Figure 2 for examples.

In this analysis, we focus on the different changes of sea surface temperature between the tropical western and eastern Indian Ocean, which is called the Indian Ocean Dipole (IOD). The IOD is known to be associated with droughts in Australia (Unmehua et al. 2009) and has a significant effect on rainfall patterns in southeast Australia (Behera and Yamagata 2003). According to Shinoda, Hendon, and Alexander (2004), the IOD phenomenon is a predominant inter-annual variation of sea surface temperature during late boreal summer and autumn (Shinoda, Hendon, and Alexander 2004), so in this application we focused on the sea surface temperature in the Indian Ocean region of longitude 40 to 120 and latitude −20 to 20 between September and November every year from 2003 to 2018.

Based on a simple autocorrelation analysis on the gridded data, we decided to use measurements for every 10 days to reduce the temporal dependence among the data. At each location of a float on a particular day, the average temperature between 0 and 5 hPa from the float is regarded as a measurement. The Argo float dataset provides multiple versions of data, and we adopted the quality controlled (QC) version. Eventually we have a two-dimensional functional data collected of $n = 144$ days, where the number of observed locations $T_{ij} = (\text{longitude, latitude})$ per day varies from 7 to 47, that is, $7 \leq m_i \leq 47$, $i = 1, \ldots, n$, with an average of 21.83. The locations are rescaled to $[0, 1] \times [0, 1]$. As shown in Figure 2, the data have a random sparse design.

First we used kernel ridge regression with the corresponding kernel for the tensor product of two second-order Sobolev spaces (e.g., Wong and Zhang 2019) to obtain a mean function estimate for every month. Then, we applied the proposed covariance function estimator with the same kernel.

The estimates of the top two-dimensional $L^2$ eigenfunctions are illustrated in Figure 3. The first eigenfunction shows the east-west dipole mode, which aligns with existing scientific findings (e.g., Shinoda, Hendon, and Alexander 2004; Deser et al. 2010; Chu et al. 2014). The second eigenfunction can be interpreted as the basin-wide mode, which is a dominant mode all around the year (e.g., Deser et al. 2010; Chu et al. 2014).

![Figure 2](image_url). Observations on 2013/09/04 (left), and all observations in the dataset (right). Points on the map indicate locations (longitude, latitude) of observations and the color scale of every point shows the corresponding Celsius temperature.
In this section, we present a transformation procedure to produce $L^2$ basis functions and corresponding eigenvalues from our estimator $\hat{B}$ obtained by (13).

Let $Q_k = \int_0^1 K(s, T_{ik})K(s, T_{fj})ds \leq i, j \leq m$, $k = 1, \ldots, p$. Then $Q_k = M_k R_k M_k^\top$, where $R_k = \int_0^1 \varphi_h(s) \varphi_h(s) ds \leq \varphi_k$ and $[\varphi_l : l = 1, \ldots, q_k]$ form a basis of $L^2$, so $R_k = M_k^+ Q_k (M_k^+)^\top$. The $L^2$ eigenvalues of $\hat{\Gamma}$ coincide with the eigenvalues of matrix $\hat{B}_{\text{square}}^L := (R_1 \otimes \cdots \otimes R_p)^{1/2} \hat{B}_{\text{square}} [(R_1 \otimes \cdots \otimes R_p)^{1/2}]^\top$, and the number of nonzero eigenvalues is the same as the rank of $\hat{B}_{\text{square}}$. The $L^2$ eigenfunction $\hat{\phi}_i$ that corresponds to the $i$th eigenvalue of $\hat{\Gamma}$ can be expressed as $\hat{\phi}_i(s_1, \ldots, s_p) = u_{i}^L [z_1(s_1) \otimes \cdots \otimes z_p(s_p)]$, where $z_k(\cdot)$, $k = 1, \ldots, p$ are defined in Theorem 1, and $u_i = (M_1^+ \otimes \cdots \otimes M_p^+) \otimes (R_1 \otimes \cdots \otimes R_p)^{-1/2} \varphi_i$ being the $i$th eigenvector of matrix $\hat{B}_{\text{square}}^L$. Using the property of Kronecker products, we have $\hat{\phi}_i(s_1, \ldots, s_p) = v_i^T [(R_1^{-1/2} M_1^+ z_1(s_1)) \otimes \cdots \otimes (R_p^{-1/2} M_p^+ z_p(s_p))].$

By simple verification, we can see that $R_k^{-1/2} M_k^+ z_k(\cdot)$ are $q_k$ one-dimensional orthonormal $L^2$ functions for dimension $k$, $k = 1, \ldots, p$. Therefore, we can also express $\hat{\Gamma}$ with these $L^2$ one-dimensional basis and the coefficients will form a $2p$th order tensor of dimension $q_1 \times \cdots \times q_p \times q_1 \times \cdots \times q_p$. We use $\hat{B}_{\text{square}}^L$ to represent this new coefficient tensor and extend our unfolding operators to $L^2$ space. It is easy to see that $\hat{B}_{\text{square}} = \hat{B}_{\text{square}}^L$.

Since $\hat{\Gamma}(k)$ is a compact operator in the $L^2$ space, this yields a singular value decomposition which leads to a $L^2$ basis characterizing the marginal variation along the $k$th dimension. We call it a $L^2$ marginal basis for the $k$th dimension. Obviously, the marginal basis function $\hat{\psi}_k$ corresponding to the $k$th singular value for dimension $k$ can be

To provide a clearer understanding of the covariance function structure, we derived a marginal $L^2$ basis along longitude and latitude, respectively. The details are given in Appendix A. The left panel of Figure 4 demonstrates that the first longitudinal marginal basis reflects a large variation in the western region while the second one corresponds to the variation in the eastern region. Due to different linear combinations, the variation along longitude may contribute to not only opposite changes between the eastern and western sides of the Indian Ocean as shown in the first two-dimensional eigenfunction, but also an overall warming or cooling tendency as shown in the second two-dimensional eigenfunction. The second longitudinal marginal basis reveals that the closer to the east boundary, the greater the variation is, which suggests that the IOD may be related to the Pacific Ocean. This aligns with the evidence that the IOD has a link with El Niño Southern Oscillation (ENSO) (Stuecker et al. 2017), an irregularly periodic variation in sea surface temperature over the tropical eastern Pacific Ocean. As shown in the right panel of Figure 4, the overall trend for the first latitude marginal basis is almost a constant function. This provides evidence that the IOD is primarily associated with the variation along longitude.

**Appendix A: $L^2$ Eigensystem and $L^2$ Marginal Basis**

In this section, we present a transformation procedure to produce $L^2$ eigenfunctions and corresponding eigenvalues from our estimator $\hat{B}$ obtained by (13).
expressed as $\hat{\psi}^k_1(\cdot) = u^k_1 z_k(\cdot)$, where $u^k_1 = (M^L_k)^{1/2} v^k_1$, and $v^k_1$ is the $i$th singular vector of $\hat{B}^L_1$. And the $L^2$ marginal singular values of $\hat{\Gamma}^{L}_1(k)$ coincide with the singular values of matrix $\hat{B}^L_k$.

**Appendix B: Definitions of $\kappa_{n,m}$ and $\eta_{n,m}$**

Here, we provide the specific forms of $\kappa_{n,m}$ and $\eta_{n,m}$, which are closely related to the decay of $|\mu^h \mu^h_1|$, $h = 1, \ldots, n$. Specifically, $\kappa_{n,m}$ is defined as the smallest positive $\kappa$ such that

$$
\chi^3 \left[ \frac{1}{m(m-1)} \sum_{h=1}^{\infty} \min \left\{ \kappa^2, \mu^h \right\} \right]^{1/2} \leq \kappa^2, \tag{B.1}
$$

$$
32b \left[ \frac{1}{m(m-1)} \sum_{h=1}^{\infty} \min \left\{ \kappa^2 / b^2, \mu^h \right\} \right]^{1/2} \leq \kappa^2, \tag{B.2}
$$

where $c$ is a universal constant, and $\eta_{n,m}$ is defined as the smallest positive $\eta$ such that

$$
\left( \frac{c_n}{nm} \sum_{h=1}^{\infty} \min \left\{ \eta^2, \mu^h \right\} + \frac{\eta^2}{n} \right)^{1/2} \leq \eta^2, \tag{B.3}
$$

where $c_n$ is a constant depending on $b, b_{\chi}, b_{\eta}$. The existences of $\kappa_{n,m}$ and $\eta_{n,m}$ are shown in the proof of Theorem 2.

**Supplementary Materials**

In the supplementary materials related to this article, we provide formal definitions related to Tucker decomposition for finite-dimensional tensors, proofs of our theoretical findings and additional simulation results.

**Acknowledgments**

The authors would like to thank Professor Yehua Li for insightful discussions, and the editor, an associate editor and two referees for their constructive comments and suggestions.

**Funding**

Raymond K. W. Wong’s research is partially supported by the National Science Foundation under grants DMS-1806063, DMS-1711952 (subcontract), and CCF-1934904. Xiaole Zhang’s research is partially supported by the National Science Foundation under grant DMS-1832046. Portions of this research were conducted with high performance research computing resources provided by Texas A&M University (https://hprc.tamu.edu).

**References**

Abernethy, J., Bach, F., Evgeniou, T., and Vert, J.-P. (2009), “A New Approach to Collaborative Filtering: Operator Estimation With Spectral Regularization,” *Journal of Machine Learning Research*, 10, 803–826. [813]

Allen, G. I. (2013), “Multi-Way Functional Principal Components Analysis,” in 2013 5th IEEE International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP), IEEE, pp. 220–223. [809]

Bartlett, P. L., Bouvier, O., and Mendelson, S. (2005), “Local Rademacher Complexities,” *The Annals of Statistics*, 33, 1497–1537. [817]

Behera, S. K., and Yamagata, T. (2003), “Influence of the Indian Ocean Dipole on the Southern Oscillation,” *Journal of the Meteorological Society of Japan*, Ser. II, 81, 169–177. [819]

Boyd, S., Parikh, N., Chu, E., Peleato, B., and Eckstein, J. (2010), “Distributed Optimization and Statistical Learning via the Alternating Direction Method of Multipliers,” *Foundations and Trends in Machine Learning*, 3, 1–122. [814]

Cai, T. T., and Yuan, M. (2010), “Nonparametric Covariance Function Estimation for Functional and Longitudinal Data,” Technical Report, Georgia Institute of Technology, Atlanta, GA. [810,811,817]

Chen, K., Delgado, P., and Müller, H.-G. (2017), “Modelling Function-Valued Stochastic Processes, With Applications to Fertility Dynamics,” *Journal of the Royal Statistical Society*, Series B, 79, 177–196. [810]

Chen, K., and Müller, H.-G. (2012), “Modeling Repeated Functional Observations,” *Journal of the American Statistical Association*, 107, 1599–1609. [810]

Chen, L.-H., and Jiang, C.-R. (2017), “Multi-Dimensional Functional Principal Component Analysis,” *Statistics and Computing*, 27, 1181–1192. [810]

Chu, J.-E., Ha, K.-J., Lee, J.-Y., Wang, B., Kim, B.-H., and Chung, C. E. (2014), “Future Change of the Indian Ocean Basin-Wide and Dipole Modes in the CMIP5,” *Climate Dynamics*, 43, 535–551. [819]

Deser, C., Alexander, M. A., Xie, S.-P., and Phillips, A. S. (2010), “Sea Surface Temperature Variability: Patterns and Mechanisms,” *Annual Review of Marine Science*, 2, 115–143. [819]

Ferraty, F., and Vieu, P. (2006), *Nonparametric Functional Data Analysis: Theory and Practice*, New York: Springer. [809]

Goldsmith, J., Bobb, J., Cina, M., Caffo, B., and Reich, D. (2011), “Penalized Functional Regression,” *Journal of Computational and Graphical Statistics*, 20, 830–851. [809]

Gu, C. (2013), *Smoothing Spline ANOVA Models* (2nd ed.), New York: Springer. [810,815]

Hackbusch, W. (2012), *Tensor Spaces and Numerical Tensor Calculus* (Vol. 42), Cham: Springer. [812]

Halko, N., Martinsson, P.-G., and Tropp, J. A. (2009), “Finding Structure With Randomness: Stochastic Algorithms for Constructing Approximate Matrix Decompositions,” arXiv no. 0909.4061. [814]

Hall, P., and Vial, C. (2006), “Assessing the Finite Dimensionality of Functional Data,” *Journal of the Royal Statistical Society*, Series B, 68, 689–705. [810]

Horváth, L., and Kokoszka, P. (2012), *Inference for Functional Data With Applications* (Vol. 200), New York: Springer. [809]

Hsing, T., and Eubank, R. (2015), *Theoretical Foundations of Functional Data Analysis, With an Introduction to Linear Operators*, New York: Wiley. [809,811]

Huang, J. Z., Shen, H., and Buja, A. (2009), “The Analysis of Two-Way Functional Data Using Two-Way Regularized Singular Value Decompositions,” *Journal of the American Statistical Association*, 104, 1609–1620. [809]

James, G. M., Hastie, T. J., and Sugar, C. A. (2000), “Principal Component Models for Sparse Functional Data,” *Biometrika*, 87, 587–602. [809]

Kadkhodaie, M., Christakopoulou, K., Sanjabi, M., and Banerjee, A. (2015), “Accelerated Alternating Direction Method of Multipliers,” in *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ACM, pp. 497–506. [814]

Kokoszka, P., and Reimherr, M. (2017), *Introduction to Functional Data Analysis*, Boca Raton, FL: CRC Press. [809]

Koltchinskii, V. (2011), *Oracle Inequalities in Empirical Risk Minimization and Sparse Recovery Problems: Eole d’Ete de Probabilites de Saint-Flour XXXVIII-2008* (Vol. 2033), Berlin, Heidelberg: Springer. [816,817]

Li, B., and Song, J. (2017), “Nonlinear Sufficient Dimension Reduction for Functional Data,” *The Annals of Statistics*, 45, 1059–1095. [810]

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. [809,810,817]

Liebl, D. (2019), “Inference for Sparse and Dense Functional Data With Covariate Adjustments,” *Journal of Multivariate Analysis*, 170, 315–335. [810,817]
Lynch, B., and Chen, K. (2018), "A Test of Weak Separability for Multi-Way Functional Data, With Application to Brain Connectivity Studies," *Biometrika*, 105, 815–831. [813]

Mazumder, R., Hastie, T., and Tibshirani, R. (2010), "Spectral Regularization Algorithms for Learning Large Incomplete Matrices," *Journal of Machine Learning Research*, 11, 2287–2322. [815]

Mercer, J. (1909), "XVI. Functions of Positive and Negative Type, and Their Connection the Theory of Integral Equations," *Philosophical Transactions of the Royal Society of London*, Series A, 209, 415–446. [816]

Park, S. Y., and Staicu, A.-M. (2015), "Longitudinal Functional Data Analysis," *Stat*, 4, 212–226. [810]

Paul, D., and Peng, J. (2009), "Consistency of Restricted Maximum Likelihood Estimators of Principal Components," *The Annals of Statistics*, 37, 1229–1271. [809,817]

Pearce, N. D., and Wand, M. P. (2006), " Penalized Splines and Reproducing Kernel Methods," *The American Statistician*, 60, 233–240. [810]

Poskitt, D. S., and Sengarapillai, A. (2013), "Description Length and Dimensionality Reduction in Functional Data Analysis," *Computational Statistics & Data Analysis*, 58, 98–113. [810]

Ramsay, J., and Silverman, B. (2005), *Functional Data Analysis*, New York: Springer. [809]

Raskutti, G., Wainwright, M. J., and Yu, B. (2012), " Minimax-Optimal Rates for Sparse Additive Models Over Kernel Classes via Convex Programming," *The Journal of Machine Learning Research*, 13, 389–427. [816]

Reimherr, M., Sriperumbudur, B., and Taoufik, B. (2018), "Optimal Prediction for Additive Function-on-Function Regression," *Electronic Journal of Statistics*, 12, 4571–4601. [810]

Rice, J. A., and Silverman, B. W. (1991), " Estimating the Mean and Covariance Structure Nonparametrically When the Data Are Curves," *Journal of the Royal Statistical Society*, Series B, 53, 233–243. [809]

Shahshahani, J., Senturk, D., Jeste, S., and Telesca, D. (2019), "Bayesian Analysis of Multidimensional Functional Data," arXiv no. 1909.08763. [810]

Shinoda, T., Hendon, H. H., and Alexander, M. A. (2004), "Surface and Subsurface Dipole Variability in the Indian Ocean and Its Relation With ENSO," *Deep Sea Research Part I: Oceanographic Research Papers*, 51, 619–635. [819]

Stuecker, M. F., Timmermann, A., Jin, E.-F., Chikamoto, Y., Zhang, W., Wittenberg, A. T., Widiasih, E., and Zhao, S. (2017), "Revisiting ENSO/Indian Ocean Dipole Phase Relationships," *Geophysical Research Letters*, 44, 2481–2492. [820]

Sun, X., Du, P., Wang, X., and Ma, P. (2018), "Optimal Penalized Function-on-Function Regression Under a Reproducing Kernel Hilbert Space Framework," *Journal of the American Statistical Association*, 113, 1601–1611. [810]

Tucker, L. R. (1966), "Some Mathematical Notes on Three-Mode Factor Analysis," *Psychometrika*, 31, 279–311. [812]

Ummenhofer, C. C., England, M. H., McIntosh, P. C., Meyers, G. A., Pook, M. J., Risbey, J. S., Gupta, A. S., and Taschetto, A. S. (2009), "What Causes Southeast Australia’s Worst Droughts?," *Geophysical Research Letters*, 36, L04706. [819]

Wahba, G. (1990), *Spline Models for Observational Data*, Philadelphia, PA: SIAM. [810,815]

Wang, W.-T., and Huang, H.-C. (2017), "Regularized Principal Component Analysis for Spatial Data," *Journal of Computational and Graphical Statistics*, 26, 14–25. [809,819]

Wong, R. K. W., Li, Y., and Zhu, Z. (2019), "Partially Linear Functional Additive Models for Multivariate Functional Data," *Journal of the American Statistical Association*, 114, 406–418. [810]

Wong, R. K. W., and Zhang, X. (2019), "Nonparametric Operator-Regularized Covariance Function Estimation for Functional Data," *Computational Statistics & Data Analysis*, 131, 131–144. [810,811,813,817,818,819]

Xiao, L., Li, Y., and Ruppert, D. (2013), "Fast Bivariate P-Splines: The Sandwich Smoother," *Journal of the Royal Statistical Society*, Series B, 75, 577–599. [809]

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. [809]

Yuan, M., and Cai, T. T. (2010), "A Reproducing Kernel Hilbert Space Approach to Functional Linear Regression," *The Annals of Statistics*, 38, 3412–3444. [810]

Zhang, L., Shen, H., and Huang, J. Z. (2013), "Robust Regularized Singular Value Decomposition With Application to Mortality Data," *The Annals of Applied Statistics*, 7, 1540–1561. [809]

Zhang, X., and Wang, J.-L. (2016), "From Sparse to Dense Functional Data and Beyond," *The Annals of Statistics*, 44, 2281–2321. [810,817]

Zhou, L., and Pan, H. (2014), "Principal Component Analysis of Two-Dimensional Functional Data," *Journal of Computational and Graphical Statistics*, 23, 779–801. [809]

Zhu, H., Yao, F., and Zhang, H. H. (2014), "Structured Functional Additive Regression in Reproducing Kernel Hilbert Spaces," *Journal of the Royal Statistical Society*, Series B, 76, 581–603. [810]

Zipunnikov, V., Caffo, B., Yousem, D. M., Davatzikos, C., Schwartz, B. S., and Crainiceanu, C. (2011), "Multilevel Functional Principal Component Analysis for High-Dimensional Data," *Journal of Computational and Graphical Statistics*, 20, 852–873. [809]