

ORIGINAL ARTICLE

Fast covariance estimation for multivariate sparse functional data

Cai Li¹  | Luo Xiao¹  | Sheng Luo² ¹ Department of Statistics, North Carolina State University, Raleigh 27607, NC² Department of Biostatistics and Bioinformatics, Duke University, Durham 27705, NC**Correspondence**

Cai Li, Department of Statistics, North Carolina State University, Raleigh, NC 27607.

Email: cli9@ncsu.edu

Funding information

National Institute of Neurological Disorders and Stroke, Grant/Award Number: R01NS091307; National Institutes of Health, Grant/Award Number: U01 AG024904; U.S. Department of Defense, Grant/Award Number: W81XWH-12-2-0012

Covariance estimation is essential yet underdeveloped for analysing multivariate functional data. We propose a fast covariance estimation method for multivariate sparse functional data using bivariate penalized splines. The tensor-product B-spline formulation of the proposed method enables a simple spectral decomposition of the associated covariance operator and explicit expressions of the resulting eigenfunctions as linear combinations of B-spline bases, thereby dramatically facilitating subsequent principal component analysis. We derive a fast algorithm for selecting the smoothing parameters in covariance smoothing using leave-one-subject-out cross-validation. The method is evaluated with extensive numerical studies and applied to an Alzheimer's disease study with multiple longitudinal outcomes.

KEYWORDS

bivariate smoothing, covariance function, functional principal component analysis, longitudinal data, multivariate functional data, prediction

1 | INTRODUCTION

Functional data analysis (FDA) has been enjoying great successes in many applied fields, for example, neuroimaging (Reiss & Ogden, 2010; Lindquist, 2012; Goldsmith, Crainiceanu, Caffo, & Reich, 2012; Zhu, Li, & Kong, 2012), genetics (Leng & Müller, 2006; Reimherr & Nicolae, 2014; 2016), and wearable computing (Morris et al., 2006; Xiao et al., 2015). Functional principal component analysis (FPCA) conducts dimension reduction on the inherently infinite-dimensional functional data and thus facilitates subsequent modelling and analysis. Traditionally, functional data are densely observed on a common grid and can be easily connected to multivariate data, although the notion of smoothness distinguishes the former from the latter. In recent years, covariance-based FPCA (Yao, Müller, & Wang, 2005) has become a standard approach and has greatly expanded the applicability of functional data methods to irregularly spaced data such as longitudinal data. Various nonparametric methods have now been proposed to estimate the smooth covariance function, for example, Peng and Paul (2009), Cai and Yuan (2010), Goldsmith et al. (2012), Xiao, Li, Checkley, and Crainiceanu (2018), and Wong and Zhang (2019).

There has been growing interest in multivariate functional data where multiple functions are observed for each subject. For dense functional data, Ramsay and Silverman (2005, Chapter 8.5) proposed to concatenate multivariate functional data as a single vector and conduct multivariate PCA on the long vectors, and Berrendero, Justel, and Svarc (2011) repeatedly applied point-wise univariate PCA. For sparse and paired functional data, Zhou, Huang, and Carroll (2008) extended the low-rank mixed effects model in James, Hastie, and Sugar (2000). Chiou, Chen, and Yang (2014) considered normalized multivariate FPCA through standardizing the covariance operator. Petersen and Müller (2016) proposed various metrics for studying cross-covariance between multivariate functional data. More recently, Happ and Greven (2018) introduced a FPCA framework for multivariate functional data defined on different domains.

This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2019 The Authors Stat Published by John Wiley & Sons Ltd

The interest of the paper is FPCA for multivariate sparse functional data, where multiple responses are observed at time points that vary from subject to subject and may even vary between responses within subjects. There are much fewer works to handle such data. The approach in Zhou et al. (2008) focuses on bivariate functional data and can be extended to more than two-dimensional functional data, although model selection (e.g., selection of smoothing parameters) can be computationally difficult and convergence of the expectation–maximization estimation algorithm could also be an issue. The local polynomial method in Chiou et al. (2014) can be applied to multivariate sparse functional data, although a major drawback is the selection of multiple bandwidths. Moreover, because the local polynomial method is a local approach, there is no guarantee that the resulting estimates of covariance functions will lead to a properly defined covariance operator. The approach in Happ and Greven (2018) (denoted by mFPCA hereafter) estimates cross-covariances via scores from univariate FPCA and hence can be applied to multivariate sparse functional data. Although mFPCA is theoretically sound for dense functional data, it may not capture cross-correlations between functions because scores from univariate FPCA for sparse functional data are shrunk towards zero.

We propose a novel and fast covariance-based FPCA method for multivariate sparse functional data. Note that multiple auto-covariance functions for within-function correlations and cross-covariance functions for between-function correlations have to be estimated. Tensor-product B-splines are employed to approximate the covariance functions, and a smoothness penalty as in bivariate penalized splines (Eilers & Marx, 2003) is adopted to avoid overfit. Then, the individual estimates of covariance functions will be pooled and refined. The advantages of the new method are multifold. First, the tensor-product B-spline formulation is computationally efficient to handle multivariate sparse functional data. Second, a fast fitting algorithm for selecting the smoothing parameters will be derived, which alleviates the computational burden of conducting leave-one-subject-out cross-validation. Third, the tensor-product B-spline representation of the covariance functions enables a straightforward spectral decomposition of the covariance operator for the multivariate functional data; see Proposition 1. In particular, the eigenfunctions associated with the covariance operator are explicit functions of the B-spline bases. Last but not the least, via a simple truncation step, the refined estimates of the covariance functions lead to a properly defined covariance operator.

Compared with mFPCA, the proposed method does not rely on scores from univariate FPCA, which could be a severe problem for sparse functional data and hence could better capture the correlations between functions. And an improved correlation estimation will lead to improved subsequent FPCA analysis and curve prediction. The proposed method also compares favourably with the local polynomial method in Chiou et al. (2014) because of the computationally efficient tensor-product spline formulation of the covariance functions and the derived fast algorithm for selecting the smoothing parameters. Moreover, as mentioned above, there exists an explicit and easy-to-calculate relationship between the tensor-product spline representation of covariance functions and the associated eigenfunctions/eigenvalues, which greatly facilitate subsequent FPCA analysis.

In addition to FPCA, there are also abundant literatures on models for multivariate functional data with most focusing on dense functional data. For clustering of multivariate functional data, see Zhu, Brown, and Morris (2012), Jacques and Preda (2014), Huang, Li, and Guan (2014), and Park and Ahn (2017). For regression with multivariate functional responses, see Zhu et al. (2012), Luo and Qi (2017), Li, Huang, and Zhu, (2017), Wong, Li, and Zhu (2019), Zhu, Morris, Wei, and Cox (2017), Kowal, Matteson, and Ruppert (2017), and Qi and Luo (2018). Graphical models for multivariate functional data are studied in (Zhu, Strawn, & Dunson, 2016) and Qiao, Guo, and James (2019). Works on multivariate functional data include also Chiou and Müller (2014, 2016).

The remainder of the paper proceeds as follows. In Section 2, we present our proposed method. We conduct extensive simulation studies in Section 3 and apply the proposed method to an Alzheimer's disease (AD) study in Section 4. A discussion is given in Section 5. All technical details are enclosed in the Appendix.

2 | METHODS

2.1 | Fundamentals of multivariate functional principal component analysis

Let p be a positive integer and denote by \mathcal{J} a continuous and bounded domain in the real line \mathbb{R} . Consider the Hilbert space $\mathcal{H}: \underbrace{L^2(\mathcal{J}) \times \dots \times L^2(\mathcal{J})}_p$ equipped with the inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ and norm $\|\cdot\|_{\mathcal{H}}$ such that for arbitrary functions $\mathbf{f} = (f^{(1)}, \dots, f^{(p)})^\top$ and $\mathbf{g} = (g^{(1)}, \dots, g^{(p)})^\top$ in \mathcal{H} with each element in $L^2(\mathcal{J})$, $\langle \mathbf{f}, \mathbf{g} \rangle_{\mathcal{H}} = \sum_{k=1}^p \int f^{(k)}(t)g^{(k)}(t)dt$ and $\|\mathbf{f}\|_{\mathcal{H}} = \langle \mathbf{f}, \mathbf{f} \rangle_{\mathcal{H}}^{1/2}$. Let $\{x^{(k)}\}_{k=1, \dots, p}$ be a set of p random functions with each function in $L^2(\mathcal{J})$. Assume that the p -dimensional vector $\mathbf{x}(t) = (x^{(1)}, \dots, x^{(p)})^\top \in \mathbb{R}^p$ has a p -dimensional smooth mean function, $\boldsymbol{\mu}(t) = \mathbb{E}\{\mathbf{x}(t)\} = (\mathbb{E}\{x^{(1)}(t)\}, \dots, \mathbb{E}\{x^{(p)}(t)\})^\top = (\boldsymbol{\mu}^{(1)}(t), \dots, \boldsymbol{\mu}^{(p)}(t))^\top$. Define the covariance function as $\mathbf{C}(s, t) = \mathbb{E}\{(\mathbf{x}(s) - \boldsymbol{\mu}(s))(\mathbf{x}(t) - \boldsymbol{\mu}(t))^\top\} = [\mathbf{C}_{kk'}(s, t)]_{1 \leq k, k' \leq p}$ and $\mathbf{C}_{kk'}(s, t) = \text{Cov}\{x^{(k)}(s), x^{(k')}(t)\}$. Then, the covariance operator $\Gamma: \mathcal{H} \rightarrow \mathcal{H}$ associated with the kernel $\mathbf{C}(s, t)$ can be defined such that for any $\mathbf{f} \in \mathcal{H}$, the k th element of $\Gamma \mathbf{f}$ is given by

$$(\Gamma \mathbf{f})^{(k)}(s) = \langle \mathbf{C}_k(s, \cdot), \mathbf{f} \rangle_{\mathcal{H}} = \sum_{k'=1}^p \int \mathbf{C}_{kk'}(s, t) f^{(k')}(t) dt,$$

where $\mathbf{C}_k(s, t) = (C_{k1}(s, t), \dots, C_{kp}(s, t))^T$. Note that Γ is a linear, self-adjoint, compact, and non-negative integral operator. By the Hilbert-Schmidt theorem, there exists a set of orthonormal bases $\{\Psi_\ell\}_{\ell \geq 1} \in \mathcal{H}$, $\Psi_\ell = (\Psi_\ell^{(1)}, \dots, \Psi_\ell^{(p)})^T$, and $\langle \Psi_\ell, \Psi_{\ell'} \rangle_{\mathcal{H}} = \sum_{k=1}^p \int \Psi_\ell^{(k)}(t) \Psi_{\ell'}^{(k)}(t) dt = \mathbf{1}_{\{\ell=\ell'\}}$, such that

$$(\Gamma \Psi_\ell)^{(k)}(s) = \sum_{k'=1}^p \int \mathbf{C}_{kk'}(s, t) \Psi_\ell^{(k')}(t) dt = d_\ell \Psi_\ell^{(k)}(s), \quad (1)$$

where d_ℓ is the ℓ th largest eigenvalue corresponding to Ψ_ℓ . Then, the multivariate Mercer's theorem gives

$$\mathbf{C}(s, t) = \sum_{\ell=1}^{\infty} d_\ell \Psi_\ell(s) \Psi_\ell^T(t), \quad (2)$$

where $\mathbf{C}_{kk'}(s, t) = \sum_{\ell=1}^{\infty} d_\ell \Psi_\ell^{(k)}(s) \Psi_\ell^{(k')}(t)$. As shown in Saporta (1981), $\mathbf{x}(t)$ has the multivariate Karhunen-Loève representation, $\mathbf{x}(t) = \boldsymbol{\mu}(t) + \sum_{\ell=1}^{\infty} \xi_\ell \Psi_\ell(t)$, where $\xi_\ell = \langle \mathbf{x} - \boldsymbol{\mu}, \Psi_\ell \rangle_{\mathcal{H}}$ are the scores with $\mathbb{E}(\xi_\ell) = 0$ and $\mathbb{E}(\xi_\ell \xi_{\ell'}) = d_\ell \mathbf{1}_{\{\ell=\ell'\}}$. The covariance operator Γ has the positive semidefiniteness property; that is, for any $\mathbf{a} = (a_1, \dots, a_p)^T \in \mathbb{R}^p$, the covariance function of $\mathbf{a}^T \mathbf{x}$, denoted by $C_a(s, t)$, satisfies that for any sets of time points $(t_1, \dots, t_q) \subset \mathcal{T}$ with an arbitrary positive integer q , the square matrix $[C_a(t_i, t_j)]_{\{1 \leq i, j \leq q\}} \in \mathbb{R}^{q \times q}$ is positive semidefinite.

2.2 | Covariance estimation by bivariate penalized splines

Suppose that the observed data take the form $\{(y_{ij}^{(k)}, t_{ij}^{(k)}) : i = 1, \dots, n; k = 1, \dots, p; j = 1, \dots, m_{ik}\}$, where $t_{ij}^{(k)} \in \mathcal{T}$ is the observed time point, $y_{ij}^{(k)}$ is the observed k th response, n is the number of subjects, and m_{ik} is the number of observations for subject i 's k th response. The model is

$$y_{ij}^{(k)} = \mathbf{x}_i^{(k)}(t_{ij}^{(k)}) + \epsilon_{ij}^{(k)} = \boldsymbol{\mu}^{(k)}(t_{ij}^{(k)}) + \sum_{\ell=1}^{\infty} \xi_{i\ell} \Psi_\ell^{(k)}(t_{ij}^{(k)}) + \epsilon_{ij}^{(k)}, \quad (3)$$

where $\mathbf{x}_i(t) = (x_i^{(1)}(t), \dots, x_i^{(p)}(t))^T \in \mathcal{H}$, $\epsilon_{ij}^{(k)}$ are random noises with zero means and variances σ_k^2 and are independent across i, j , and k .

The goal is to estimate the covariance functions $C_{kk'}$. We adopt a three-step procedure. In the first step, empirical estimates of the covariance functions are constructed. Let $r_{ij}^{(k)} = y_{ij}^{(k)} - \boldsymbol{\mu}^{(k)}(t_{ij}^{(k)})$ be the residuals and $C_{ij_1 j_2}^{(kk')} = r_{ij_1}^{(k)} r_{ij_2}^{(k')}$ be the auxiliary variables. Note that $\mathbb{E}(C_{ij_1 j_2}^{(kk')}) = C_{kk'}(t_{ij_1}^{(k)}, t_{ij_2}^{(k')}) + \sigma_k^2 \mathbf{1}_{\{k=k', j_1=j_2\}}$ for $1 \leq j_1 \leq m_{ik}, 1 \leq j_2 \leq m_{ik'}$. Thus, $C_{ij_1 j_2}^{(kk')}$ is an unbiased estimate of $C_{kk'}(t_{ij_1}^{(k)}, t_{ij_2}^{(k')})$ whenever $k \neq k'$ or $j_1 \neq j_2$. In the second step, the noisy auxiliary variables are smoothed to obtain smooth estimates of the covariance functions. For smoothing, we use bivariate P -splines (Eilers & Marx, 2003) because it is an automatic smoother and is computationally simple. In the final step, we pool all estimates of the individual covariance functions and use an extra step of eigendecomposition to obtain refined estimates of covariance functions. The refined estimates lead to a covariance operator that is properly defined, that is, positive semidefinite. In practice, the mean functions $\boldsymbol{\mu}^{(k)}$ are unknown, and we estimate them using P -splines (Eilers & Marx, 1996) with the smoothing parameters selected by leave-one-subject-out cross-validation; see the Supporting Information for details. Denote the estimates by $\hat{\boldsymbol{\mu}}^{(k)}$. Let $\hat{r}_{ij}^{(k)} = y_{ij}^{(k)} - \hat{\boldsymbol{\mu}}^{(k)}(t_{ij}^{(k)})$ and $\hat{C}_{ij_1 j_2}^{(kk')} = \hat{r}_{ij_1}^{(k)} \hat{r}_{ij_2}^{(k')}$, the actual auxiliary variables.

The bivariate P -splines model $C_{kk'}(s, t)$ uses tensor-product splines $G_{kk'}(s, t)$ for $1 \leq k, k' \leq p$. Specifically, $G_{kk'}(s, t) = \sum_{1 \leq \gamma_1, \gamma_2 \leq c} \boldsymbol{\Theta}_{\gamma_1 \gamma_2}^{(kk')} B_{\gamma_1}(s) B_{\gamma_2}(t)$, where $\boldsymbol{\Theta}_{kk'} = [\boldsymbol{\Theta}_{\gamma_1 \gamma_2}^{(kk')}]_{1 \leq \gamma_1, \gamma_2 \leq c} \in \mathbb{R}^{c \times c}$ is a coefficient matrix, $\{B_1(\cdot), \dots, B_c(\cdot)\}$ is the collection of B-spline basis functions in \mathcal{T} , and c is the number of equally spaced interior knots plus the order (degree plus 1) of the B-splines. Because $C_{kk'}(s, t) = C_{k'k}(t, s) = \text{Cov}\{x^{(k)}(s), x^{(k')}(t)\}$, it is reasonable to impose the assumption that

$$\boldsymbol{\Theta}_{kk'} = \boldsymbol{\Theta}_{k'k}^T,$$

so that $G_{kk'}(s, t) = G_{k'k}(t, s)$. Therefore, in the rest of the section, we consider only $k \leq k'$.

Let $\mathbf{D} \in \mathbb{R}^{(c-2) \times c}$ denote a second-order differencing matrix such that for a vector $\mathbf{a} = (a_1, \dots, a_c)^\top \in \mathbb{R}^c$, $\mathbf{D}\mathbf{a} = (a_3 - 2a_2 + a_1, a_4 - 2a_3 + a_2, \dots, a_c - 2a_{c-1} + a_{c-2})^\top \in \mathbb{R}^{c-2}$. Also let $\|\cdot\|_F$ be the Frobenius norm. For the cross-covariance function $C_{kk'}(s, t)$ with $k < k'$, the bivariate P -splines estimate the coefficient matrix $\Theta_{kk'}$ by $\widehat{\Theta}_{kk'}$, which minimizes the penalized least squares

$$\sum_{i=1}^n \sum_{1 \leq j_1 \leq m_{k'} \leq j_2} \sum_{1 \leq j_1 \leq m_k} \left\{ G_{kk'}(t_{j_1}^{(k)}, t_{j_2}^{(k')}) - \widehat{C}_{j_1 j_2}^{(kk')} \right\}^2 + \lambda_{kk'1} \|\mathbf{D}\Theta_{kk'}\|_F^2 + \lambda_{kk'2} \|\mathbf{D}\Theta_{kk'}^\top\|_F^2, \quad (4)$$

where $\lambda_{kk'1}$ and $\lambda_{kk'2}$ are two non-negative smoothing parameters that balance the model fit and smoothness of the estimate and will be determined later. Indeed, the column penalty $\|\mathbf{D}\Theta_{kk'}\|_F^2$ penalizes the second-order consecutive differences of the columns of $\Theta_{kk'}$, and similarly, the row penalty $\|\mathbf{D}\Theta_{kk'}^\top\|_F^2$ penalizes the second-order consecutive differences of the rows of $\Theta_{kk'}$. The two penalty terms are essentially penalizing the second-order partial derivatives of $G_{kk'}(s, t)$ along the s and t directions. The two smoothing parameters are allowed to differ to accommodate different levels of smoothing along the two directions.

For the auto-covariance functions $C_{kk}(s, t)$ with $k=1, \dots, p$, we conduct bivariate covariance smoothing by enforcing the following constraint on the coefficient matrix Θ_{kk} (Xiao et al., 2018):

$$\Theta_{kk} = \Theta_{kk}^\top. \quad (5)$$

It follows that $G_{kk}(s, t)$ is a symmetric function. Then, the coefficient matrix Θ_{kk} and the error variance σ_k^2 are jointly estimated by $\widehat{\Theta}_{kk}$ and $\widehat{\sigma}_k^2$, which minimize the penalized least squares

$$\sum_{i=1}^n \sum_{1 \leq j_1, j_2 \leq m_k} \left\{ G_{kk}(t_{j_1}^{(k)}, t_{j_2}^{(k)}) + \sigma_k^2 \mathbf{1}_{\{j_1=j_2\}} - \widehat{C}_{j_1 j_2}^{(kk)} \right\}^2 + \lambda_k \|\mathbf{D}\Theta_{kk}\|_F^2, \quad (6)$$

over all symmetric Θ_{kk} and λ_k is a smoothing parameter. Note that the two penalty terms in Equation (4) become the same when Θ_{kk} is symmetric, and thus, only one smoothing parameter is needed for auto-covariance estimation.

2.2.1 | Estimation

We first introduce the notation. Let $\text{vec}(\cdot)$ be an operator that stacks the columns of a matrix into a column vector and denote by \otimes the Kronecker product. Fix k and k' with $k \leq k'$. Let $\boldsymbol{\theta}_{kk'} = \text{vec}(\Theta_{kk'}) \in \mathbb{R}^{c^2}$ be a vector of the coefficients and $\mathbf{b}(t) = \{B_1(t), \dots, B_c(t)\}^\top \in \mathbb{R}^c$ denotes the B-spline base. Then,

$$G_{kk'}(s, t) = \mathbf{b}(s)^\top \Theta_{kk'} \mathbf{b}(t) = \{\mathbf{b}(t) \otimes \mathbf{b}(s)\}^\top \boldsymbol{\theta}_{kk'}.$$

We now organize the auxiliary responses $\widehat{C}_{j_1 j_2}^{(kk')}$ for each pair of k and k' . Let $\mathbf{r}_i^{(k)} = (r_{i1}^{(k)}, \dots, r_{im_k}^{(k)})^\top \in \mathbb{R}^{m_k}$, $\widehat{\mathbf{C}}_i^{(kk')} = \mathbf{r}_i^{(k)} \otimes \mathbf{r}_i^{(k')} \in \mathbb{R}^{m_k m_{k'}}$ and $\widehat{\mathbf{C}}^{(kk')} = (\widehat{\mathbf{C}}_1^{(kk'), \top}, \dots, \widehat{\mathbf{C}}_n^{(kk'), \top})^\top \in \mathbb{R}^{N_{kk'}}$, where $N_{kk'} = \sum_{i=1}^n m_{ik} m_{ik'}$ is the total number of auxiliary responses for the pair of k and k' . As for the B-splines, let $\mathbf{b}_i^{(k)} = [\mathbf{b}(t_{i1}^{(k)}), \dots, \mathbf{b}(t_{im_k}^{(k)})] \in \mathbb{R}^{c \times m_k}$, $\mathbf{B}_i^{(kk')} = (\mathbf{b}_i^{(k')} \otimes \mathbf{b}_i^{(k)})^\top \in \mathbb{R}^{(m_{ik} m_{k'}) \times c^2}$, and $\mathbf{B}^{(kk')} = [\mathbf{B}_1^{(kk'), \top}, \dots, \mathbf{B}_n^{(kk'), \top}]^\top \in \mathbb{R}^{N_{kk'} \times c^2}$.

For estimation of the cross-covariance functions $C_{kk'}$ with $k < k'$, the penalized least squares in Equation (4) can be rewritten as

$$\left(\widehat{\mathbf{C}}^{(kk')} - \mathbf{B}^{(kk')} \boldsymbol{\theta}_{kk'} \right)^\top \left(\widehat{\mathbf{C}}^{(kk')} - \mathbf{B}^{(kk')} \boldsymbol{\theta}_{kk'} \right) + \lambda_{kk'1} \boldsymbol{\theta}_{kk'}^\top \mathbf{P}_1 \boldsymbol{\theta}_{kk'} + \lambda_{kk'2} \boldsymbol{\theta}_{kk'}^\top \mathbf{P}_2 \boldsymbol{\theta}_{kk'}, \quad (7)$$

where $\mathbf{P}_1 = \mathbf{I}_c \otimes \mathbf{D}^\top \mathbf{D}$ and $\mathbf{P}_2 = \mathbf{D}^\top \mathbf{D} \otimes \mathbf{I}_c$. The expression in Equation (7) is a quadratic function of the coefficient vector $\boldsymbol{\theta}_{kk'}$. Therefore, we derive that

$$\widehat{\boldsymbol{\theta}}_{kk'} = \left(\mathbf{B}^{(kk'), \top} \mathbf{B}^{(kk')} + \lambda_{kk'1} \mathbf{P}_1 + \lambda_{kk'2} \mathbf{P}_2 \right)^{-1} \mathbf{B}^{(kk'), \top} \widehat{\mathbf{C}}^{(kk')}$$

and the estimate of the cross-covariance function $C_{kk'}(s, t)$ is $\widehat{C}_{kk'}(s, t) = \{\mathbf{b}(t) \otimes \mathbf{b}(s)\}^\top \widehat{\boldsymbol{\theta}}_{kk'}$.

For estimation of the auto-covariance functions, because of the constraint on the coefficient matrix in Equation (5), let $\boldsymbol{\chi}_k \in \mathbb{R}^{c(c+1)/2}$ be a vector obtained by stacking the columns of the lower triangle of Θ_{kk} , and let $\mathbf{G}_c \in \mathbb{R}^{c^2 \times c(c+1)/2}$ be a duplication matrix such that $\boldsymbol{\theta}_{kk} = \mathbf{G}_c \boldsymbol{\eta}_k$ (Seber, 2008, p. 246). Let $\mathbf{Z}_i^{(k)} = \text{vec}(\mathbf{I}_{m_k}) \in \mathbb{R}^{m_k^2}$ and $\mathbf{Z}^{(k)} = (\mathbf{Z}_1^{(k), \top}, \dots, \mathbf{Z}_n^{(k), \top})^\top \in \mathbb{R}^{N_{kk}}$. Finally, let $\boldsymbol{\beta}_k = (\boldsymbol{\chi}_k^\top, \sigma_k^2)^\top \in \mathbb{R}^{\widetilde{c}}$ with $\widetilde{c} = c(c+1)/2 + 1$. It follows that the

penalized least squares in Equation (6) can be rewritten as

$$\left(\widehat{\mathbf{C}}^{(kk)} - \mathbf{X}^{(k)} \widehat{\boldsymbol{\beta}}_k\right)^\top \left(\widehat{\mathbf{C}}^{(kk)} - \mathbf{X}^{(k)} \widehat{\boldsymbol{\beta}}_k\right) + \lambda_k \widehat{\boldsymbol{\beta}}_k^\top \mathbf{Q} \widehat{\boldsymbol{\beta}}_k,$$

where $\mathbf{X}^{(k)} = \left[\mathbf{B}^{(kk)}, \mathbf{Z}^{(k)}\right] \in \mathbb{R}^{N_{kk} \times \widetilde{c}}$ and $\mathbf{Q} = \text{blockdiag}\{\mathbf{G}_c^\top (\mathbf{I}_c \otimes \mathbf{D} \mathbf{D}^\top) \mathbf{G}_c^\top, \mathbf{0}\} \in \mathbb{R}^{\widetilde{c} \times \widetilde{c}}$. Therefore, we obtain

$$\widehat{\boldsymbol{\beta}}_k = \left(\widehat{\boldsymbol{\chi}}_k^\top, \widehat{\boldsymbol{\sigma}}_k^2\right)^\top = \left(\mathbf{X}^{(k)\top} \mathbf{X}^{(k)} + \lambda_k \mathbf{Q}\right)^{-1} \mathbf{X}^{(k)\top} \widehat{\mathbf{C}}^{(kk)}.$$

It follows that $\widehat{\boldsymbol{\theta}}_{kk} = \mathbf{G}_c \widehat{\boldsymbol{\chi}}_k$, and the estimate of the auto-covariance function $C_{kk}(s, t)$ is $\widehat{C}_{kk}(s, t) = \{\mathbf{b}(t) \otimes \mathbf{b}(s)\}^\top \widehat{\boldsymbol{\theta}}_{kk}$.

The above estimates of covariance functions may not lead to a positive semidefinite covariance operator and thus have to be refined. We pool all estimates together, and we shall use the following proposition.

Proposition 1. Assume that $C_{kk'}(s, t) = \mathbf{b}(s)^\top \boldsymbol{\Theta}_{kk'} \mathbf{b}(t)$. Let $\mathbf{G} = \int \mathbf{b}(t) \mathbf{b}(t)^\top dt \in \mathbb{R}^{c \times c}$ and assume that \mathbf{G} is positive definite (Zhou, Shen, & Wolfe, 1998). Then, $\left[\mathbf{G}^{\frac{1}{2}} \boldsymbol{\Theta}_{kk'} \mathbf{G}^{\frac{1}{2}}\right]_{1 \leq k, k' \leq p} \in \mathbb{R}^{p \times p}$ admits the spectral decomposition, $\sum_{\ell=1}^{\infty} d_\ell \mathbf{u}_\ell \mathbf{u}_\ell^\top$, where d_ℓ is the ℓ th largest eigenvalue of the covariance operator Γ , and $\mathbf{u}_\ell = \{\mathbf{u}_\ell^{(1)\top}, \dots, \mathbf{u}_\ell^{(p)\top}\}^\top \in \mathbb{R}^{pc}$ is the associated eigenvector with $\mathbf{u}_\ell^{(k)} \in \mathbb{R}^c$ and such that $\Psi_\ell^{(k)}(t) = \mathbf{b}(t)^\top \mathbf{G}^{-\frac{1}{2}} \mathbf{u}_\ell^{(k)}$.

The proof is provided in Appendix A. Proposition 1 implies that, with the tensor-product B-spline representation of the covariance functions, one spectral decomposition gives us the eigenvalues and eigenfunctions. In particular, the eigenfunctions $\Psi_\ell^{(k)}(t)$ are linear combinations of the B-spline basis functions, which means that they can be straightforwardly evaluated, an advantage of spline-based methods compared with other smoothing methods for which eigenfunctions are approximated by spectral decompositions of the covariance functions evaluated at a grid of time points.

Once we have $\widehat{\boldsymbol{\Theta}}_{kk'}$, the estimate of the coefficient matrix $\boldsymbol{\Theta}_{kk'}$, the spectral decomposition of $[\mathbf{G}^{\frac{1}{2}} \widehat{\boldsymbol{\Theta}}_{kk'} \mathbf{G}^{\frac{1}{2}}]_{kk'}$ gives us estimates \widehat{d}_ℓ and $\widehat{\mathbf{u}}_\ell = \{\widehat{\mathbf{u}}_\ell^{(1)\top}, \dots, \widehat{\mathbf{u}}_\ell^{(p)\top}\}^\top$. We discard negative \widehat{d}_ℓ to ensure that the multivariate covariance operator is positive semidefinite, and this leads to a refined estimate of the coefficient matrix $\boldsymbol{\Theta}_{kk'}$, $\widetilde{\boldsymbol{\Theta}}_{kk'} = \mathbf{G}^{-\frac{1}{2}} \left\{ \sum_{\ell: \widehat{d}_\ell > 0} \widehat{d}_\ell \widehat{\mathbf{u}}_\ell^{(k)} \widehat{\mathbf{u}}_\ell^{(k')\top} \right\} \mathbf{G}^{-\frac{1}{2}}$. Then, the refined estimate of the covariance functions is $\widetilde{C}_{kk'}(s, t) = \mathbf{b}(s)^\top \widetilde{\boldsymbol{\Theta}}_{kk'} \mathbf{b}(t)$. Proposition 1 also suggests that the eigenfunctions can be estimated by $\widetilde{\Psi}_\ell^{(k)}(t) = \mathbf{b}(t)^\top \mathbf{G}^{-\frac{1}{2}} \widehat{\mathbf{u}}_\ell^{(k)}$.

For principal component analysis or curve prediction in practice, one may select further the number of principal components by either the proportion of variance explained (PVE) (Greven, Crainiceanu, Caffo, & Reich, 2010) or an Akaike information criterion-type criterion (Li, Wang, & Carroll, 2013). Here, we follow Greven et al. (2010) using PVE with a value of 0.99.

2.2.2 | Selection of smoothing parameters

We select the smoothing parameters in each auto-covariance/cross-covariance estimation using leave-one-subject-out cross-validation; see, for example, Yao et al. (2005) and Xiao et al. (2018). A fast approximate algorithm for the auto-covariance has been derived in Xiao et al. (2018). So we focus on the cross-covariance and use the notation in Equation (7). Note that there are two smoothing parameters for each cross-covariance estimation.

For simplicity, we suppress the superscript and subscript kk' in Equation (7) for both $\widehat{\mathbf{C}}$ and \mathbf{B} . Let $\widetilde{\mathbf{C}}_i^{[j]}$ be the prediction of the auxiliary responses $\widehat{\mathbf{C}}_i$ from the estimate using data without the i th subject. Let $\|\cdot\|$ be the Euclidean norm, and the cross-validation error is

$$\text{iCV} = \sum_{i=1}^n \left\| \widehat{\mathbf{C}}_i - \widetilde{\mathbf{C}}_i^{[i]} \right\|^2. \quad (8)$$

We shall also now suppress the subscript k from m_{ik} and kk' from $N_{kk'}$. Let $\mathbf{S} = \mathbf{B}(\mathbf{B}^\top \mathbf{B} + \lambda_1 \mathbf{P}_1 + \lambda_2 \mathbf{P}_2)^{-1} \mathbf{B}^\top \in \mathbb{R}^{N \times N}$, $\mathbf{S}_i = \mathbf{B}_i(\mathbf{B}^\top \mathbf{B} + \lambda_1 \mathbf{P}_1 + \lambda_2 \mathbf{P}_2)^{-1} \mathbf{B}_i^\top \in \mathbb{R}^{m_i^2 \times N}$, and $\mathbf{S}_{ii} = \mathbf{B}_i(\mathbf{B}^\top \mathbf{B} + \lambda_1 \mathbf{P}_1 + \lambda_2 \mathbf{P}_2)^{-1} \mathbf{B}_i^\top \in \mathbb{R}^{m_i^2 \times m_i^2}$.

Then, a shortcut formula for Equation (8) is

$$\text{iCV} = \sum_{i=1}^n \left\| \left(\mathbf{I}_{m_i^2} - \mathbf{S}_{ii} \right)^{-1} \left(\mathbf{S}_i \widehat{\mathbf{C}}_i - \widehat{\mathbf{C}}_i \right) \right\|^2.$$

Similar to Xu and Huang (2012) and Xiao et al. (2018), the iCV can be further simplified by adopting the approximation $(\mathbf{I}_{m_i^2} - \mathbf{S}_{ii})^{-2} = \mathbf{I}_{m_i^2} + 2\mathbf{S}_{ii}$, which results in the generalized cross-validation, denoted by iGCV,

$$\text{iGCV} = \|\widehat{\mathbf{C}} - \mathbf{S}\widehat{\mathbf{C}}\|^2 + 2 \sum_{i=1}^n (\mathbf{S}_i \widehat{\mathbf{C}} - \widehat{\mathbf{C}}_i)^\top \mathbf{S}_{ii} (\mathbf{S}_i \widehat{\mathbf{C}} - \widehat{\mathbf{C}}_i). \quad (9)$$

Although iGCV is much easier to compute than iCV, the formula in Equation (9) is still computationally expensive to compute. Indeed, the smoother matrix \mathbf{S} is of dimension $2,500 \times 2,500$ if $n=100$ and $m_i=m=5$ for all i . Thus, we need to further simplify the formula.

Let $\mathbf{G}_n = \mathbf{B}^\top \mathbf{B}$, $\widetilde{\mathbf{B}} = \mathbf{B} \mathbf{G}_n^{-1/2} \in \mathbb{R}^{N \times c^2}$, $\widetilde{\mathbf{B}}_i = \mathbf{B}_i \mathbf{G}_n^{-1/2} \in \mathbb{R}^{m_i^2 \times c^2}$, $\mathbf{f} = \widetilde{\mathbf{B}}^\top \widehat{\mathbf{C}} \in \mathbb{R}^{c^2}$, $\mathbf{f}_i = \widetilde{\mathbf{B}}_i^\top \widehat{\mathbf{C}}_i \in \mathbb{R}^{c^2}$, and $\mathbf{L}_i = \widetilde{\mathbf{B}}_i^\top \widetilde{\mathbf{B}}_i \in \mathbb{R}^{c^2 \times c^2}$. Also let $\widetilde{\mathbf{P}}_1 = \mathbf{G}_n^{-1/2} \mathbf{P}_1 \mathbf{G}_n^{-1/2} \in \mathbb{R}^{c^2 \times c^2}$, $\widetilde{\mathbf{P}}_2 = \mathbf{G}_n^{-1/2} \mathbf{P}_2 \mathbf{G}_n^{-1/2} \in \mathbb{R}^{c^2 \times c^2}$, and $\Sigma = \mathbf{I}_{c^2} + \lambda_1 \widetilde{\mathbf{P}}_1 + \lambda_2 \widetilde{\mathbf{P}}_2$. Then, Equation (9) can be simplified as

$$\text{iGCV} = \|\widehat{\mathbf{C}}\|^2 - 2\mathbf{f}^\top \Sigma^{-1} \mathbf{f} + \mathbf{f}^\top \Sigma^{-2} \mathbf{f} + 2 \sum_{i=1}^n (\mathbf{L}_i \Sigma^{-1} \mathbf{f} - \mathbf{f}_i)^\top \Sigma^{-1} (\mathbf{L}_i \Sigma^{-1} \mathbf{f} - \mathbf{f}_i). \quad (10)$$

Note that Σ has two smoothing parameters. Following Wood (2000), we use an equivalent parameterization $\Sigma = \mathbf{I} + \rho\{\mathbf{w}\widetilde{\mathbf{P}}_1 + (1-w)\widetilde{\mathbf{P}}_2\}$, where $\rho = \lambda_1 + \lambda_2$ represents the overall smoothing level and $w = \lambda_1 \rho^{-1} \in [0, 1]$ is the relative weight of λ_1 . We conduct a two-dimensional grid search of (ρ, w) , as follows. For a given w , let $\mathbf{U} \text{diag}(\mathbf{s}) \mathbf{U}^\top$ be the eigendecomposition of $\mathbf{w}\widetilde{\mathbf{P}}_1 + (1-w)\widetilde{\mathbf{P}}_2$, where $\mathbf{U} \in \mathbb{R}^{c^2 \times c^2}$ is an orthonormal matrix and $\mathbf{s} = (s_1, \dots, s_{c^2}) \in \mathbb{R}^{c^2}$ is the vector of eigenvalues. Then, $\Sigma^{-1} = \mathbf{U} \text{diag}(\widetilde{\mathbf{d}}) \mathbf{U}^\top$ with $\widetilde{\mathbf{d}} = 1/(1 + \rho \mathbf{s}) \in \mathbb{R}^{c^2}$.

Proposition 2. Let \odot stand for the point-wise multiplication. Then,

$$\text{iGCV} = \|\widehat{\mathbf{C}}\|^2 + (\widetilde{\mathbf{f}} \odot \widetilde{\mathbf{d}})^\top (\widetilde{\mathbf{f}} \odot \widetilde{\mathbf{d}}) - 2\widetilde{\mathbf{d}}^\top \mathbf{g} - 4\widetilde{\mathbf{d}}^\top \mathbf{F} \widetilde{\mathbf{d}} + 2\widetilde{\mathbf{d}}^\top \left[\sum_{i=1}^n \{ \widetilde{\mathbf{L}}_i (\widetilde{\mathbf{f}} \odot \widetilde{\mathbf{d}}) \} \odot \{ \widetilde{\mathbf{L}}_i (\widetilde{\mathbf{f}} \odot \widetilde{\mathbf{d}}) \} \right],$$

where $\widetilde{\mathbf{f}}_i = \mathbf{U}^\top \mathbf{f}_i \in \mathbb{R}^{c^2}$, $\widetilde{\mathbf{f}} = \mathbf{U}^\top \mathbf{f} \in \mathbb{R}^{c^2}$, $\mathbf{g} = \widetilde{\mathbf{f}} \odot \widetilde{\mathbf{d}} - \sum_{i=1}^n \widetilde{\mathbf{f}}_i \odot \widetilde{\mathbf{d}}_i \in \mathbb{R}^{c^2}$, $\widetilde{\mathbf{L}}_i = \mathbf{U}^\top \mathbf{L}_i \mathbf{U} \in \mathbb{R}^{c^2 \times c^2}$, and $\mathbf{F} = \sum_{i=1}^n (\widetilde{\mathbf{f}}_i \widetilde{\mathbf{f}}_i^\top) \odot \widetilde{\mathbf{L}}_i \in \mathbb{R}^{c^2 \times c^2}$.

The proof is provided in Appendix A. For each w , note that only $\widetilde{\mathbf{d}}$ depends on ρ and needs to be calculated repeatedly, and all other terms need to be calculated only once. The entire algorithm is presented in Algorithm 1. We give an evaluation of the complexity of the proposed algorithm. Assume that $m_i=m$ for all i . The first initialization (Step 1) requires $O(nm^2c^2 + nc^4 + c^6)$ computations. For each w , the second initialization (Step 2) also requires $O(nc^4 \min(m^2, c^2) + c^6)$ computations. For each ρ , Steps 3–8 require $O(nc^4)$ computations. Therefore, the formula in Proposition 2 is most efficient to calculate for sparse data with small numbers of observations per subject; that is, m_i s are small.

Algorithm 1 Selection of smoothing parameters

Input: $\mathbf{B}, \widehat{\mathbf{C}}, \mathbf{P}_1, \mathbf{P}_2, \rho = \{\rho_1, \dots, \rho_K\}^\top, \mathbf{w} = \{w_1, \dots, w_K\}^\top$

Output: $\{\rho^*, w^*\}$

(1) Initialize $\|\widehat{\mathbf{C}}\|^2, \widetilde{\mathbf{P}}_1, \widetilde{\mathbf{P}}_2, \mathbf{G}_n^{-1/2}, \widetilde{\mathbf{B}}, \widetilde{\mathbf{B}}_i, \mathbf{f}, \mathbf{f}_i, \mathbf{L}_i$ for $i = 1, \dots, n$;

for each w in \mathbf{w} do

(2) Initialize $\mathbf{s}, \widetilde{\mathbf{f}}, \widetilde{\mathbf{f}}_i, \mathbf{g}, \mathbf{F}, \widetilde{\mathbf{L}}_i, i = 1, \dots, n$;

for each ρ in ρ do

(3) $\widetilde{\mathbf{d}} \leftarrow 1/(1 + \rho \mathbf{s})$;

(4) $\mathbf{I} \leftarrow (\widetilde{\mathbf{f}} \odot \widetilde{\mathbf{d}})^\top (\widetilde{\mathbf{f}} \odot \widetilde{\mathbf{d}})$;

(5) $\text{II} \leftarrow -2\widetilde{\mathbf{d}}^\top \mathbf{g}$;

(6) $\text{III} \leftarrow -4\widetilde{\mathbf{d}}^\top \mathbf{F} \widetilde{\mathbf{d}}$;

(7) $\text{IV} \leftarrow 2\widetilde{\mathbf{d}}^\top \left[\sum_{i=1}^n \{ \widetilde{\mathbf{L}}_i (\widetilde{\mathbf{f}} \odot \widetilde{\mathbf{d}}) \} \odot \{ \widetilde{\mathbf{L}}_i (\widetilde{\mathbf{f}} \odot \widetilde{\mathbf{d}}) \} \right]$;

(8) $\text{iGCV} \leftarrow \|\widehat{\mathbf{C}}\|^2 + \text{I} + \text{II} + \text{III} + \text{IV}$;

end for

end for

(9) $\{\rho^*, w^*\} \leftarrow \arg \min_{\rho, w} \text{iGCV}$;

2.3 | Prediction

For prediction, assume that the smooth curve $\mathbf{x}_i(t)$ is generated from a multivariate Gaussian process. Suppose that we want to predict the i th multivariate response $\mathbf{x}_i(t)$ at $\{s_{i1}, \dots, s_{im}\}$ for $m \geq 1$. Let $\mathbf{y}_i^{(k)} = (y_{i1}^{(k)}, \dots, y_{im_k}^{(k)})^\top$ be the vector of observations at $\{t_{i1}^{(k)}, \dots, t_{im_k}^{(k)}\}$ for the k th response. Let

$\boldsymbol{\mu}_i^{(k),o} = (\boldsymbol{\mu}^{(k)}(t_{i1}^{(k)}), \dots, \boldsymbol{\mu}^{(k)}(t_{im_k}^{(k)}))^T$ be the vector of the k th mean function at the observed time points. Let $\mathbf{y}_i = (\mathbf{y}_i^{(1),\top}, \dots, \mathbf{y}_i^{(p),\top})^T$ and $\boldsymbol{\mu}_i^o = (\boldsymbol{\mu}_i^{(1),o,\top}, \dots, \boldsymbol{\mu}_i^{(p),o,\top})^T$. Let $\boldsymbol{\mu}_i^n = (\boldsymbol{\mu}^{(1)}(s_{i1}), \dots, \boldsymbol{\mu}^{(1)}(s_{im}), \dots, \boldsymbol{\mu}^{(p)}(s_{i1}), \dots, \boldsymbol{\mu}^{(p)}(s_{im}))^T$ be the vector of mean functions at the time points for prediction.

It follows that

$$\begin{pmatrix} \mathbf{y}_i \\ \mathbf{x}_i \end{pmatrix} \sim \mathcal{N} \left\{ \begin{pmatrix} \boldsymbol{\mu}_i^o \\ \boldsymbol{\mu}_i^n \end{pmatrix}, \begin{pmatrix} \text{Cov}(\mathbf{y}_i) & \text{Cov}(\mathbf{x}_i, \mathbf{y}_i)^T \\ \text{Cov}(\mathbf{x}_i, \mathbf{y}_i) & \text{Cov}(\mathbf{x}_i) \end{pmatrix} \right\}.$$

Thus, we obtain

$$\mathbb{E}(\mathbf{x}_i | \mathbf{y}_i) = \text{Cov}(\mathbf{x}_i, \mathbf{y}_i) \text{Cov}(\mathbf{y}_i)^{-1} (\mathbf{y}_i - \boldsymbol{\mu}_i^o) + \boldsymbol{\mu}_i^n,$$

$$\text{Cov}(\mathbf{x}_i | \mathbf{y}_i) = \text{Cov}(\mathbf{x}_i) - \text{Cov}(\mathbf{x}_i, \mathbf{y}_i) \text{Cov}(\mathbf{y}_i)^{-1} \text{Cov}(\mathbf{y}_i, \mathbf{x}_i)^T.$$

Let $\mathbf{b}_i^{(k),o} = [\mathbf{b}(t_{i1}^{(k)}), \dots, \mathbf{b}(t_{im_k}^{(k)})]^T$ and $\mathbf{b}_i^n = [\mathbf{b}(s_{i1}), \dots, \mathbf{b}(s_{im})]^T$. Next, let $\mathbf{B}_i^o = \text{blockdiag}(\mathbf{b}_i^{(1),o}, \dots, \mathbf{b}_i^{(p),o})$ and $\mathbf{B}_i^n = \mathbf{I}_p \otimes \mathbf{b}_i^n$. Then, $\text{Cov}(\mathbf{y}_i)$ given by $\mathbf{B}_i^o \boldsymbol{\Theta} \mathbf{B}_i^{o,\top} + \text{blockdiag}(\sigma_1^2 \mathbf{I}_{m_1}, \dots, \sigma_p^2 \mathbf{I}_{m_p})$, and $\text{Cov}(\mathbf{x}_i)$ and $\text{Cov}(\mathbf{y}_i, \mathbf{x}_i)$ are given by $\mathbf{B}_i^o \boldsymbol{\Theta} \mathbf{B}_i^{n,\top}$ and $\mathbf{B}_i^o \boldsymbol{\Theta} \mathbf{B}_i^{n,\top}$, respectively. Let $\tilde{\boldsymbol{\Theta}} = [\tilde{\Theta}_{kk}]_{1 \leq k, k' \leq p} \in \mathbb{R}^{p \times p \times p}$.

Plugging in the estimates, we predict \mathbf{x}_i by

$$\begin{aligned} \hat{\mathbf{x}}_i &= (\hat{\mathbf{x}}_i^{(1)}(s_{i1}), \dots, \hat{\mathbf{x}}_i^{(1)}(s_{im}), \dots, \hat{\mathbf{x}}_i^{(p)}(s_{i1}), \dots, \hat{\mathbf{x}}_i^{(p)}(s_{im}))^T \\ &= (\mathbf{B}_i^n \tilde{\boldsymbol{\Theta}} \mathbf{B}_i^{o,\top}) \hat{\mathbf{V}}_i^{-1} (\mathbf{y}_i - \hat{\boldsymbol{\mu}}_i^o) + \hat{\boldsymbol{\mu}}_i^n, \end{aligned}$$

where $\hat{\boldsymbol{\mu}}_i^o = (\hat{\boldsymbol{\mu}}^{(1)}(t_{i1}^{(1)}), \dots, \hat{\boldsymbol{\mu}}^{(1)}(t_{im_1}^{(1)}), \dots, \hat{\boldsymbol{\mu}}^{(p)}(t_{i1}^{(p)}), \dots, \hat{\boldsymbol{\mu}}^{(p)}(t_{im_p}^{(p)}))^T$ is the estimate of $\boldsymbol{\mu}_i^o$, $\hat{\boldsymbol{\mu}}_i^n = (\hat{\boldsymbol{\mu}}^{(1)}(s_{i1}), \dots, \hat{\boldsymbol{\mu}}^{(1)}(s_{im}), \dots, \hat{\boldsymbol{\mu}}^{(p)}(s_{i1}), \dots, \hat{\boldsymbol{\mu}}^{(p)}(s_{im}))^T$ is the estimate of $\boldsymbol{\mu}_i^n$, $\hat{\mathbf{V}}_i = \mathbf{B}_i^o \tilde{\boldsymbol{\Theta}} \mathbf{B}_i^{o,\top} + \text{blockdiag}(\hat{\sigma}_1^2 \mathbf{I}_{m_1}, \dots, \hat{\sigma}_p^2 \mathbf{I}_{m_p})$. An approximate covariance matrix for $\hat{\mathbf{x}}_i$ is

$$\widehat{\text{Cov}}(\mathbf{x}_i | \mathbf{y}_i) = \mathbf{B}_i^n \tilde{\boldsymbol{\Theta}} \mathbf{B}_i^{n,\top} - \mathbf{B}_i^n \tilde{\boldsymbol{\Theta}} \mathbf{B}_i^{o,\top} \hat{\mathbf{V}}_i^{-1} (\mathbf{B}_i^o \tilde{\boldsymbol{\Theta}} \mathbf{B}_i^{o,\top})^T.$$

Therefore, a 95% point-wise confidence interval for the k th response is given by

$$\hat{\mathbf{x}}_i^{(k)}(s_{ij}) \pm 1.96 \sqrt{\widehat{\text{Var}}(\mathbf{x}_i^{(k)}(s_{ij}) | \mathbf{y}_i)},$$

where $\widehat{\text{Var}}(\mathbf{x}_i^{(k)}(s_{ij}) | \mathbf{y}_i)$ can be extracted from the diagonal of $\widehat{\text{Cov}}(\mathbf{x}_i | \mathbf{y}_i)$.

Finally, we predict the first $L \geq 1$ scores $\boldsymbol{\xi}_i = (\xi_{i1}, \dots, \xi_{iL})^T$ for the i th subject. Note that $\xi_{i\ell} = \int \Psi_\ell(t)^T \{\mathbf{x}_i(t) - \boldsymbol{\mu}(t)\} dt$. With a similar derivation as above, $\mathbf{x}_i(t) - \boldsymbol{\mu}(t)$ can be predicted by $\{\mathbf{I}_p \otimes \mathbf{b}(t)\}^T \tilde{\boldsymbol{\Theta}} \mathbf{B}_i^{o,\top} \hat{\mathbf{V}}_i^{-1} (\mathbf{y}_i - \hat{\boldsymbol{\mu}}_i^o)$. By Proposition 1, the eigenfunctions $\Psi_\ell^{(k)}(t)$ are estimated by $\mathbf{b}(t)^T \mathbf{G}^{-\frac{1}{2}} \hat{\mathbf{U}}_\ell^{(k)}$, and thus, $\Psi_\ell(t)^T = \hat{\mathbf{U}}_\ell^T \{\mathbf{I}_p \otimes \mathbf{G}^{-\frac{1}{2}} \mathbf{b}(t)\}$. It follows that

$$\hat{\xi}_{i\ell} = \hat{\mathbf{U}}_\ell^T (\mathbf{I}_p \otimes \mathbf{G}^{-\frac{1}{2}}) \tilde{\boldsymbol{\Theta}} \mathbf{B}_i^{o,\top} \hat{\mathbf{V}}_i^{-1} (\mathbf{y}_i - \hat{\boldsymbol{\mu}}_i^o).$$

3 | SIMULATIONS

We evaluate the finite sample performance of the proposed method (denoted by mFACES) against mFPCA via a synthetic simulation study and a simulation study mimicking the Alzheimer's Disease Neuroimaging Initiative (ADNI) data in the real data example. Here, we report the details and results of the former as the conclusions remain the same for the latter, and details are provided in the Supporting Information.

3.1 | Simulation settings and evaluation criteria

We generate data by model (3) with $p=3$ responses. The mean functions are $\boldsymbol{\mu}(t) = [5\sin(2\pi t), 5\cos(2\pi t), 5(t-1)^2]^T$. We first specify the auto-covariance functions. Let $\boldsymbol{\Phi}_1(t) = [\sqrt{2}\sin(2\pi t), \sqrt{2}\cos(4\pi t), \sqrt{2}\sin(4\pi t)]^T$, $\boldsymbol{\Phi}_2(t) = [\sqrt{2}\cos(\pi t), \sqrt{2}\cos(2\pi t), \sqrt{2}\cos(3\pi t)]^T$, and $\boldsymbol{\Phi}_3(t) = [\sqrt{2}\sin(\pi t), \sqrt{2}\sin(2\pi t), \sqrt{2}\sin(3\pi t)]^T$. Also let

$$\Lambda_{11} = \begin{pmatrix} 3 & 0 & 0 \\ 0 & 1.5 & 0 \\ 0 & 0 & 0.75 \end{pmatrix}, \Lambda_{22} = \begin{pmatrix} 3.5 & 0 & 0 \\ 0 & 1.75 & 0 \\ 0 & 0 & 0.5 \end{pmatrix}, \Lambda_{33} = \begin{pmatrix} 2.5 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

Then, the auto-covariance functions are $C_{kk}(s, t) = \Phi_k(s)^\top \Lambda_{kk} \Phi_k(t)$, $k = 1, 2, 3$. For the cross-covariance functions, let $C_{kk'}(s, t) = \rho \Phi_{k'}(s)^\top \Lambda_{kk'}^{\frac{\rho}{2}} \Lambda_{kk'}^{\frac{\rho}{2}} \Phi_{k'}(t)$ for $k \neq k'$, where $\rho \in [0, 1]$ is a parameter to be specified. The induced covariance operator from the above specifications is proper; see Lemma 1 in Appendix B. It is easy to derive that the absolute value of cross-correlation $\rho_{kk'}(s, t) = C_{kk'}(s, t) / \sqrt{C_{kk}(s, s)C_{k'k'}(t, t)}$ is bounded by ρ . Hence, ρ controls the overall level of correlation between responses: if $\rho = 0$, then the responses are uncorrelated from each other.

The eigendecomposition of the multivariate covariance function gives nine nonzero eigenvalues with associated multivariate eigenfunctions; hence, for $\ell = 1, \dots, 9$, we simulate the scores $\xi_{i\ell}$ from $\mathcal{N}(0, d_\ell)$, where d_ℓ are the induced eigenvalues. Next, we simulate the white noises $\epsilon_{ij}^{(k)}$ from $\mathcal{N}(0, \sigma_\epsilon^2)$, where σ_ϵ^2 is determined according to the signal-to-noise ratio $\text{SNR} = \sum_\ell d_\ell / (\rho \sigma_\epsilon^2)$. Here, we let $\text{SNR} = 2$. For each response, the sampling time points are drawn from a uniform distribution in the unit interval, and the number of observations for each subject, m_{ik} , is generated from a uniform discrete distribution on $\{3, 4, 5, 6, 7\}$. Thus, the sampling points vary not only from subject to subject but also across responses within each subject.

We use a factorial design with two factors: the number of subjects n and the correlation parameter ρ . We let $n = 100, 200$, or 400 . We let $\rho = 0.5$, which corresponds to a weak correlation between responses as the average absolute correlation between responses is only 0.36 . Another value of ρ is 0.9 , which corresponds to a moderate correlation between responses as the average absolute correlation between responses is about 0.50 .

In total, we have six model conditions, and for each model condition, we generate 200 datasets. To evaluate the prediction accuracy of the various methods, we draw 200 additional subjects as testing data. The true correlation functions and a sample of the simulated data are shown in the Supporting Information.

We compare mFACs and mFPCA in terms of estimation accuracy of the covariance functions, the eigenfunctions and eigenvalues, and prediction of new subjects. For covariance function estimation, we use the relative integrated square errors (RISE). Let $\hat{C}_{kk'}(s, t)$ be an estimate of $C_{kk'}(s, t)$, and then RISE are given by

$$\frac{\sum_{k=1}^p \sum_{k'=1}^p \int_0^1 \int_0^1 \{C_{kk'}(s, t) - \hat{C}_{kk'}(s, t)\}^2 ds dt}{\sum_{k=1}^p \sum_{k'=1}^p \int_0^1 \int_0^1 \{C_{kk'}(s, t)\}^2 ds dt}.$$

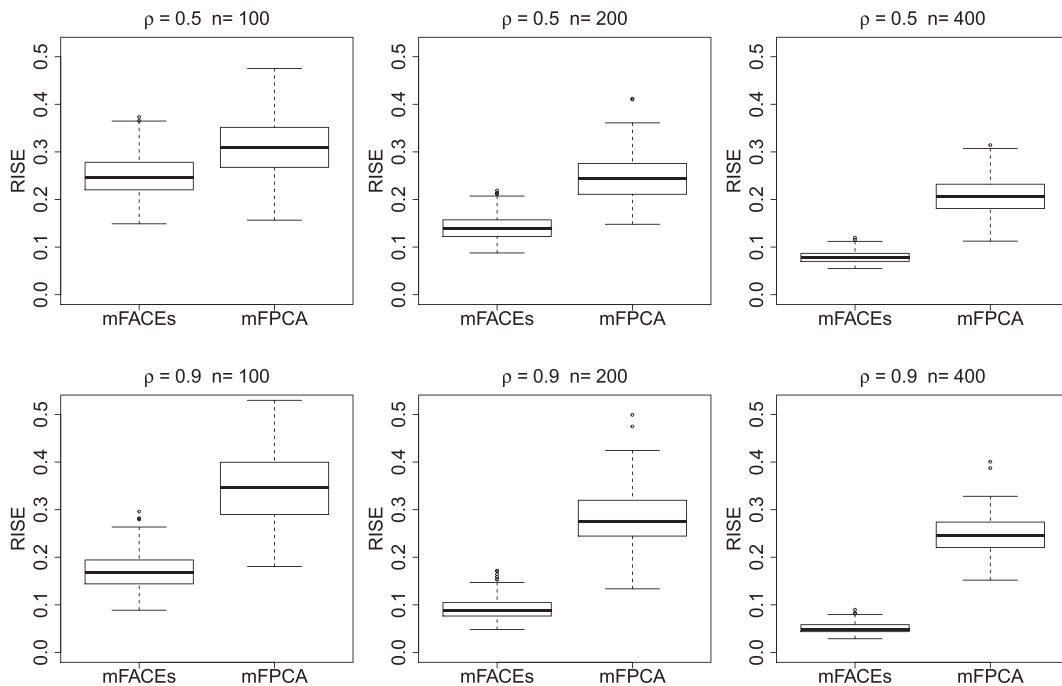


FIGURE 1 Boxplots of relative integrated square errors (RISEs) of mFACs and mFPCA for estimating the covariance function

For estimating the ℓ th eigenfunction, we use the integrated square errors (ISE), which are defined as

$$\min \left[\sum_{k=1}^p \int_0^1 \left\{ \Psi_\ell^{(k)}(t) - \widehat{\Psi}_\ell^{(k)}(t) \right\}^2 dt, \sum_{k=1}^p \int_0^1 \left\{ \Psi_\ell^{(k)}(t) + \widehat{\Psi}_\ell^{(k)}(t) \right\}^2 dt \right].$$

Note that the range of ISE is $[0,2]$. For estimating the eigenvalues, we use the ratio of the estimate against the truth; that is, \widehat{d}_ℓ/d_ℓ .

For predicting new curves, we use the mean integrated square errors (MISE), which are given by

$$\frac{1}{200p} \sum_{k=1}^p \sum_{i=1}^{200} \left[\int_0^1 \left\{ x_i^{(k)}(t) - \widehat{x}_i^{(k)}(t) \right\}^2 dt \right].$$

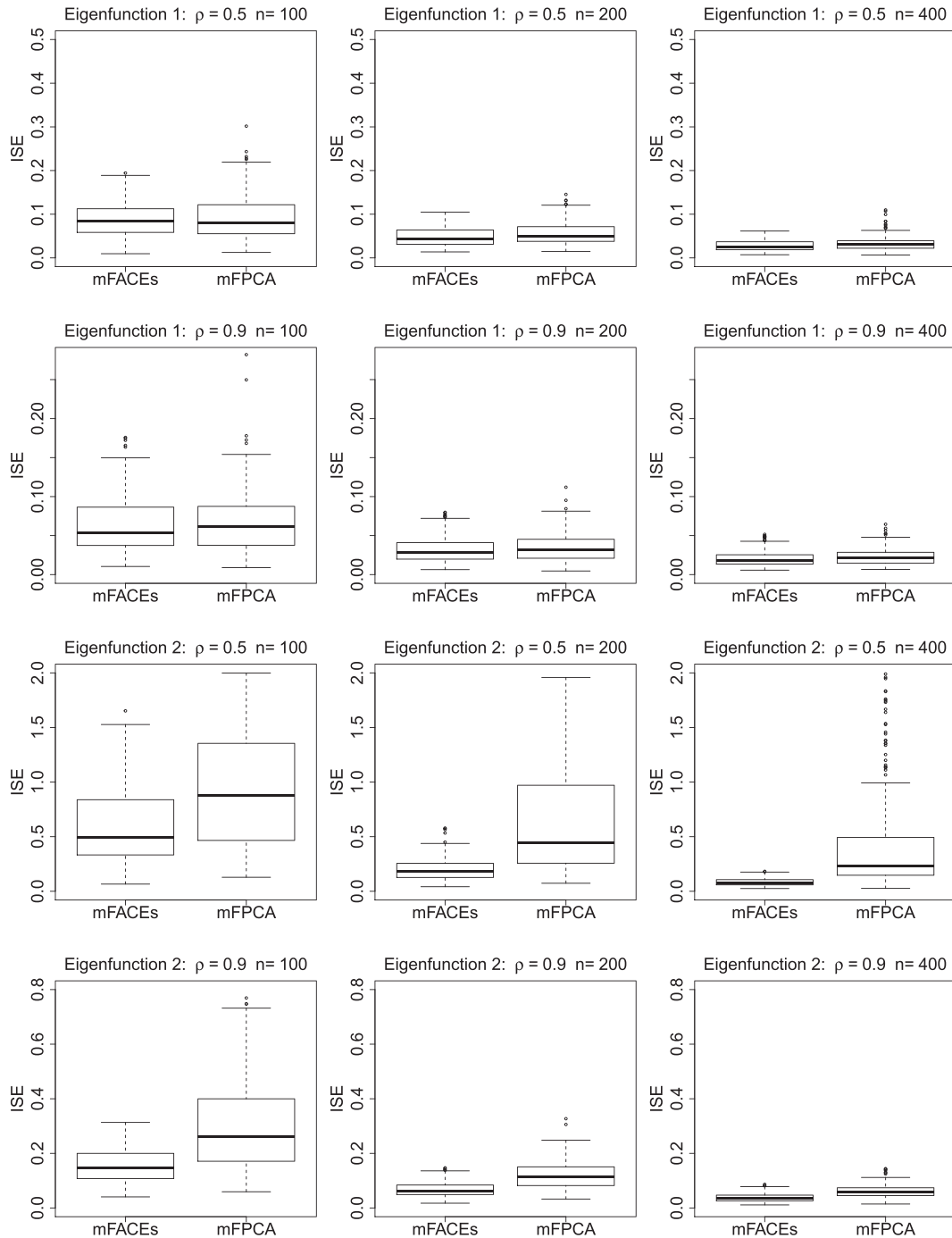


FIGURE 2 Boxplots of integrated square errors (ISEs) of mFACES and mFPCA for estimating the top two eigenfunctions

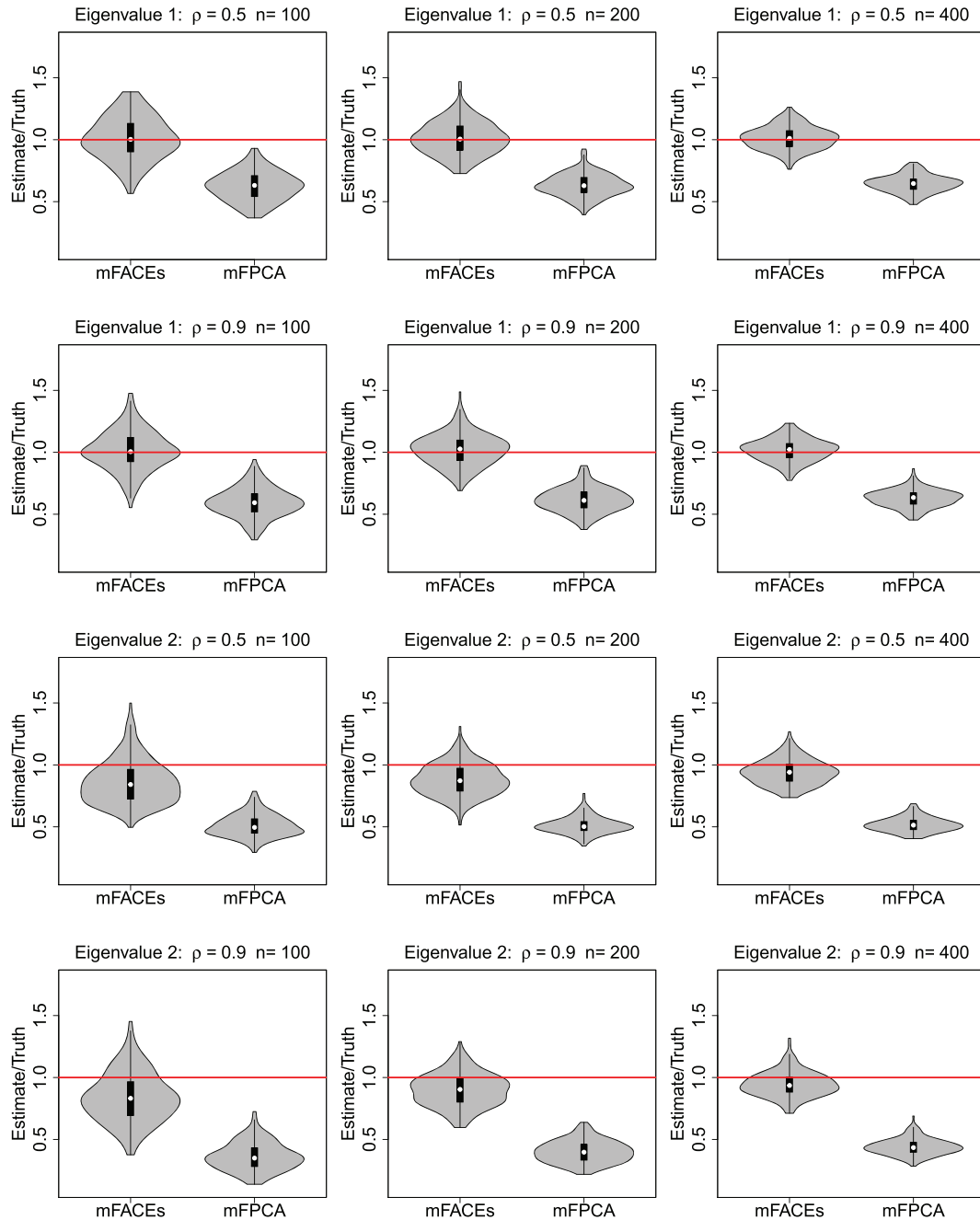


FIGURE 3 Violin plots of mFACES and mFPCA for estimating the top two eigenvalues. The red horizontal lines indicate that the estimates are equal to the truth

For the curve prediction using mFPCA, we truncate the number of principal components using a PVE of 0.99. It is worth noting that if no truncation is adopted, then the curve prediction using mFPCA reduces to curve prediction using univariate FPCA. We shall also consider the conditional expectation method based on the estimates of covariance functions from mFPCA. The method is denoted by mFPCA(CE), and its difference with mFACES is that different estimates of covariance functions are used.

3.2 | Simulation results

Figure 1 gives boxplots of RISEs of mFACES and mFPCA for estimating covariance functions. Under all model conditions, mFACES outperforms mFPCA, and the improvement in RISEs as the sample size increases is much more pronounced for mFACES. Under the model conditions with moderate correlations ($\rho=0.9$), the advantage of mFACES is substantial even for the small sample size $n=100$.

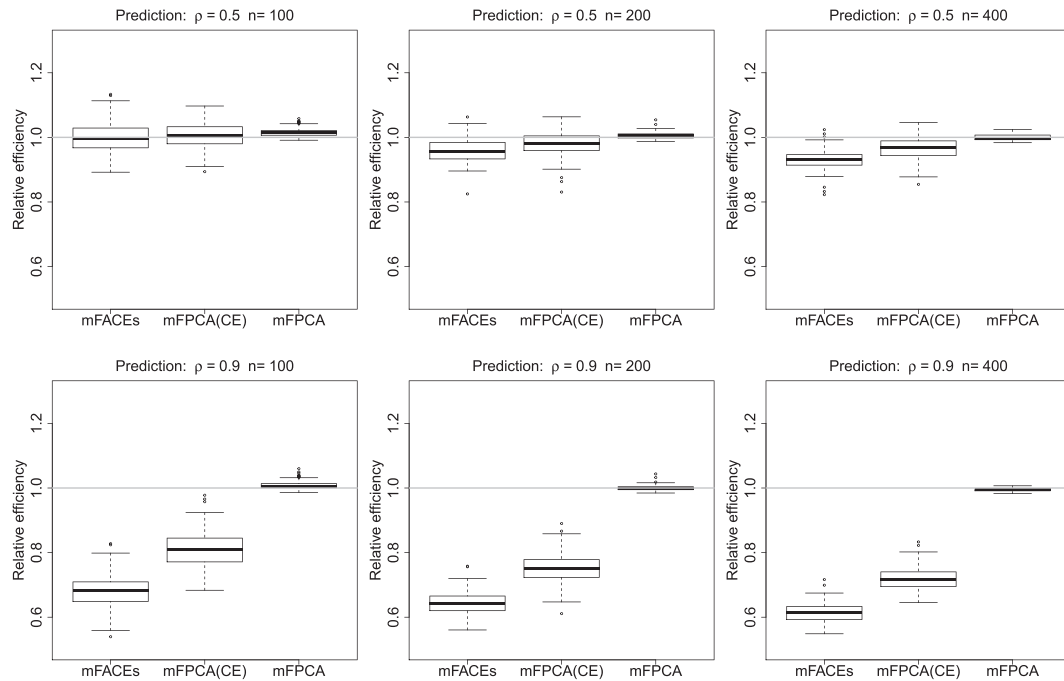


FIGURE 4 Boxplots of relative efficiency of three methods for curve prediction. The grey horizontal lines indicate the mean integrated square errors (MISEs) for univariate functional principal component analysis (FPCA)

Figures 2 and 3 give boxplots of ISEs and violin plots of mFACES and mFPCA for estimating the top two eigenfunctions and eigenvalues, respectively. The top two eigenvalues account for about 60% of the total variation in the functional data for $\rho=0.5$, and it is 80% for $\rho=0.9$. Figure 2 shows that although the two methods are overall comparable for estimating the first eigenfunction, mFACES has a much better accuracy for estimating the second eigenfunction than mFPCA. The violin plots in Figure 3 show that mFACES outperforms mFPCA substantially for estimating both eigenvalues under all model conditions. The mFPCA always underestimates the eigenvalues as the variation of scores from univariate FPCA is smaller than the true variation and hence leads to underestimates of eigenvalues.

Finally, we consider the prediction of new subjects by mFACES, mFPCA, and mFPCA(CE). We define the relative efficiencies of different methods as the ratios of MISEs with respect to those of univariate FPCA; see Figure 4. Univariate FPCA is implemented in the R package *face* (Xiao, Li, Checkley, & Crainiceanu, 2018). We have the following findings. Under all model conditions, mFACES has the smallest MISE, mFPCA(CE) has the second best performance, and mFPCA is close to univariate FPCA. Thus, on average, mFACES provides the most accurate curve prediction. These results indicate that (a) mFACES has better covariance estimation than mFPCA(CE) and so is the prediction based on it; (b) compared with mFPCA/univariate FPCA, mFPCA(CE) exploits the correlation information and hence results in better predictions.

In summary, mFACES shows competing performance against alternative methods.

4 | APPLICATION TO ALZHEIMER'S DISEASE STUDY

The ADNI is a two-stage longitudinal observational study launched in year 2003 with the primary goal of investigating whether serial neuroimages, biological markers, clinical and neuropsychological assessments can be combined to measure the progression of AD (Weiner et al., 2017). The ADNI-1 data from the first stage contain 379 patients with amnesic mild cognitive impairment (MCI, a risk state for AD) at baseline who had at least one follow-up visit. Participants were assessed at baseline, 6, 12, 18, 24, and 36 months with additional annual follow-ups included in the second stage of the study. At each visit, various neuropsychological assessments, clinical measures, and brain images were collected. The ADNI-2 data include 424 additional patients suffering from MCI and significant memory concern, with at least one follow-up visit and longitudinal data collected over 4 years. Thus, for the combined data, the total number of subjects is 803, and the average number of visits is 4.72. The data are publicly available at <http://ida.loni.ucla.edu/>.

We consider five longitudinal markers commonly measured in studies of AD with strong comparative predictive value (Li, Chan, Doody, Quinn, & Luo, 2017). Among the five markers, Disease Assessment Scale-Cognitive 13 items (ADAS-Cog 13), Rey Auditory Verbal Learning Test immediate recall (RAVLT.imme), Rey Auditory Verbal Learning Test learning curve (RAVLT.learn), and Mini-Mental State Examination (MMSE) are neuropsychological assessments. Functional Assessment Questionnaire (FAQ) is a functional and behavioural assessment. High values of ADAS-Cog 13 and FAQ indicate a high-risk state for AD, whereas low values of RAVLT.imme, RAVLT.learn, and MMSE reflect severe cognitive impairment. The longitudinal trajectories in ADNI-1 and ADNI-2 are defined on the same time domain with the largest follow-up time of 96 months from the start of ADNI-1 (Time 0).

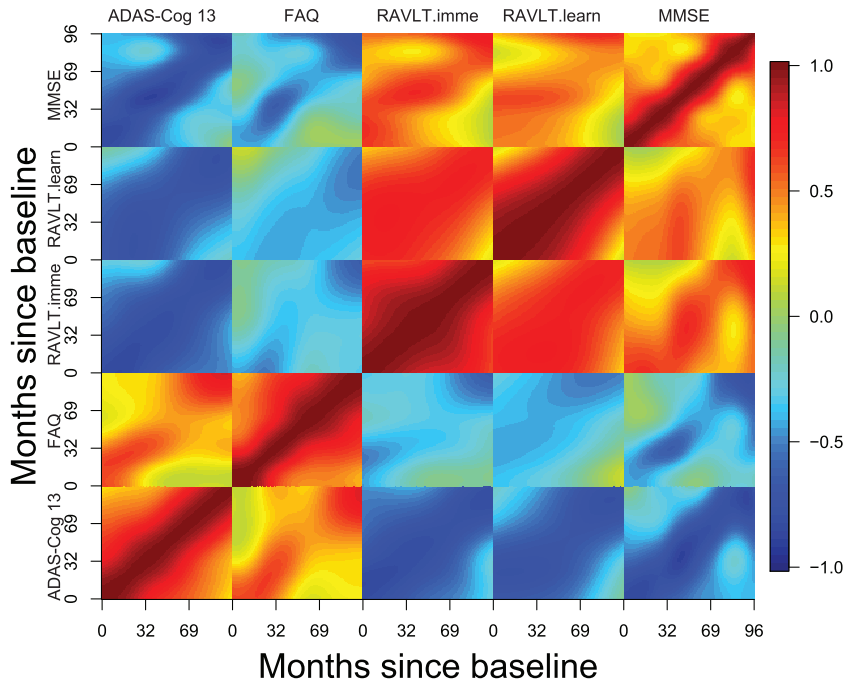


FIGURE 5 Estimated correlation functions for the longitudinal markers. ADAS-Cog 13, Disease Assessment Scale-Cognitive 13 items; FAQ, Functional Assessment Questionnaire; MMSE, Mini-Mental State Examination; RAVLT.imme, Rey Auditory Verbal Learning Test immediate recall; RAVLT.learn, Rey Auditory Verbal Learning Test learning curve

4.1 | Multivariate FPCA via mFACES

We analyse the five longitudinal biomarkers using mFACES. For better visualization, we plot in Figure 5 the estimated correlation functions $\rho_{kk'}(s, t) = C_{kk'}(s, t) / \sqrt{C_{kk}(s, s)C_{k'k'}(t, t)}$. The plot indicates two groups of biomarkers, ADAS-Cog 13 and FAQ, in one group and RAVLT.imme, RAVLT.learn, and MMSE in another group. The biomarkers within the groups are positively correlated and negatively correlated between groups, which makes sense as high values of ADAS-Cog 13 and FAQ and low values for the other biomarkers suggest of AD. Next, we display in Figure 6

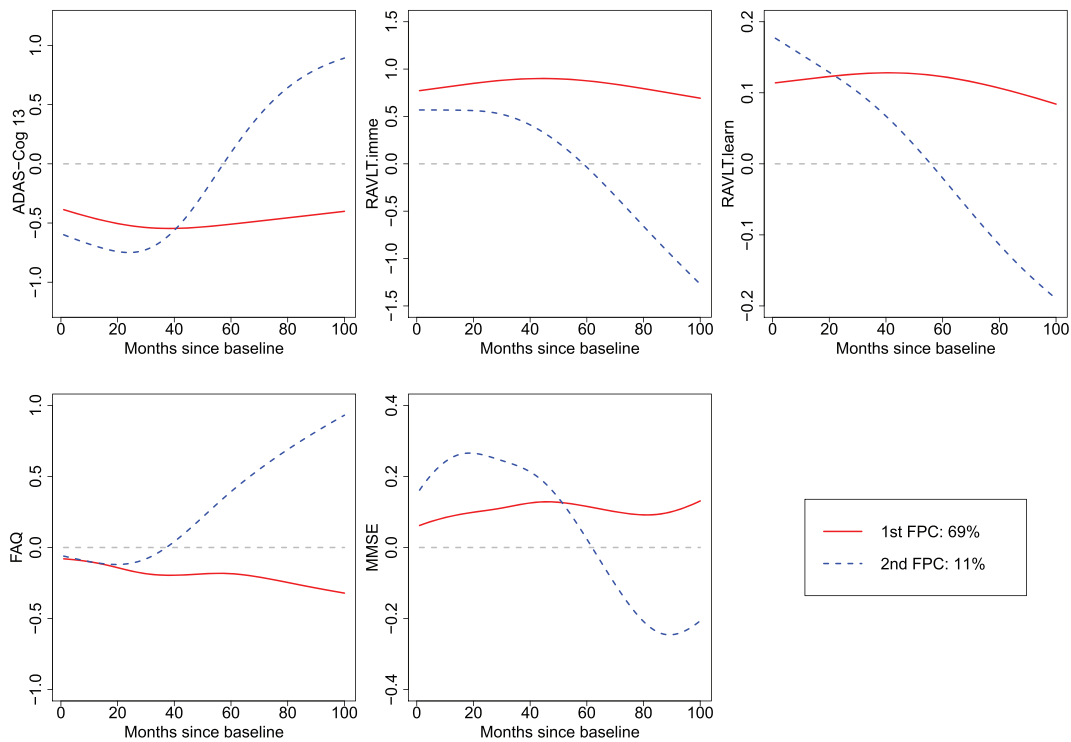


FIGURE 6 Estimated top two eigenfunctions for the longitudinal markers. ADAS-Cog 13, Disease Assessment Scale-Cognitive 13 items; FAQ, Functional Assessment Questionnaire; MMSE, Mini-Mental State Examination; RAVLT.imme, Rey Auditory Verbal Learning Test immediate recall; RAVLT.learn, Rey Auditory Verbal Learning Test learning curve

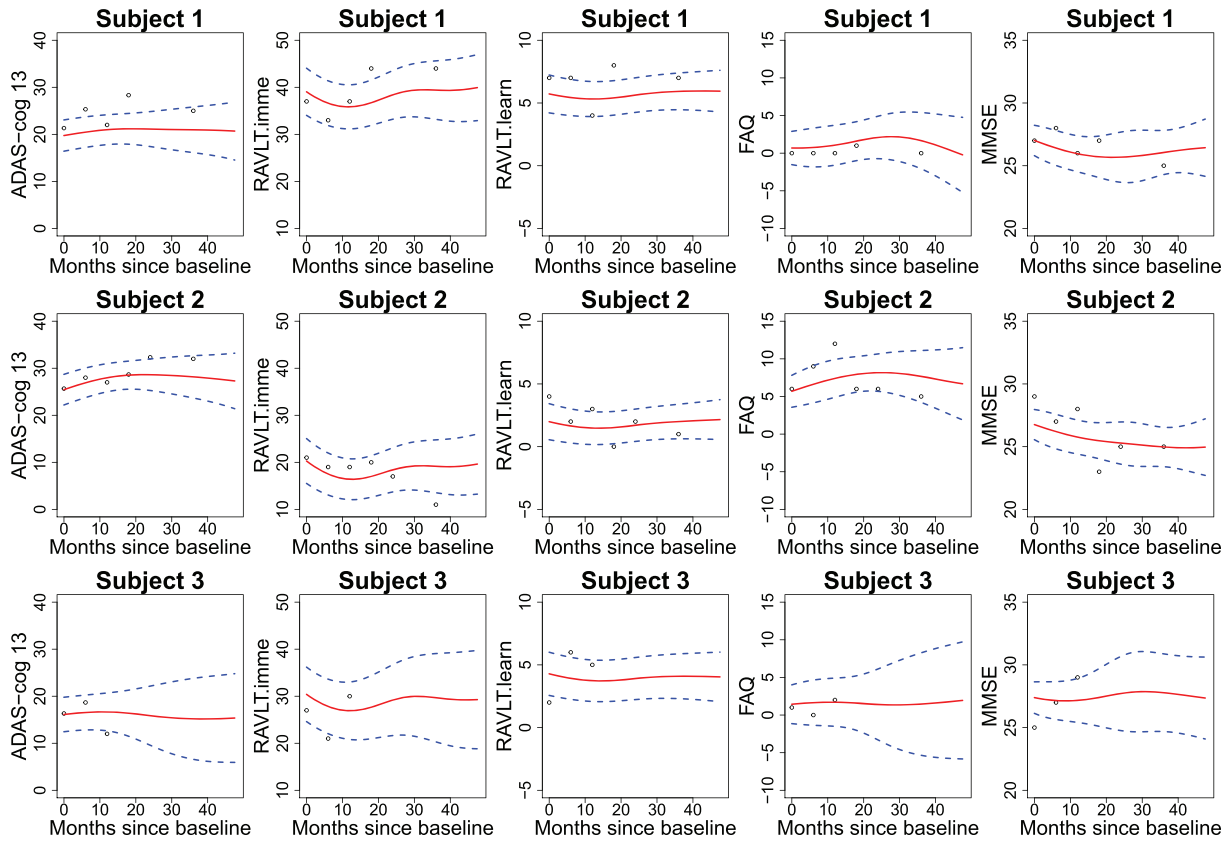


FIGURE 7 Predicted subject-specific curves (red solid line) of the longitudinal markers and associated 95% point-wise confidence bands (blue dashed line) for three subjects. ADAS-Cog 13, Disease Assessment Scale-Cognitive 13 items; FAQ, Functional Assessment Questionnaire; MMSE, Mini-Mental State Examination; RAVLT.imme, Rey Auditory Verbal Learning Test immediate recall; RAVLT.learn, Rey Auditory Verbal Learning Test learning curve

the two estimated (multivariate) eigenfunctions associated with the top two estimated eigenvalues, which account for 69% and 11% of the total variance in the functional part of the data. The eigenfunctions reveal how the five biomarkers covary and how a subject's trajectories of biomarkers deviate from the population mean. Indeed, we see from Figure 6 that the first eigenfunction (solid curves) is below the zero line for ADAS-Cog 13 and FAQ and above the zero line for the other three biomarkers. This means that the score corresponding to the first eigenfunction might be used as an indicator of AD. Indeed, a negative score for the first eigenfunction means higher-than-population-mean values of the former whereas lower-than-population-mean values of the latter, indicating more severe AD status. The second eigenfunction (dashed curves) for the five biomarkers is below the zero line at first and then above it or the other way around, potentially suggesting of a longitudinal pattern of the AD

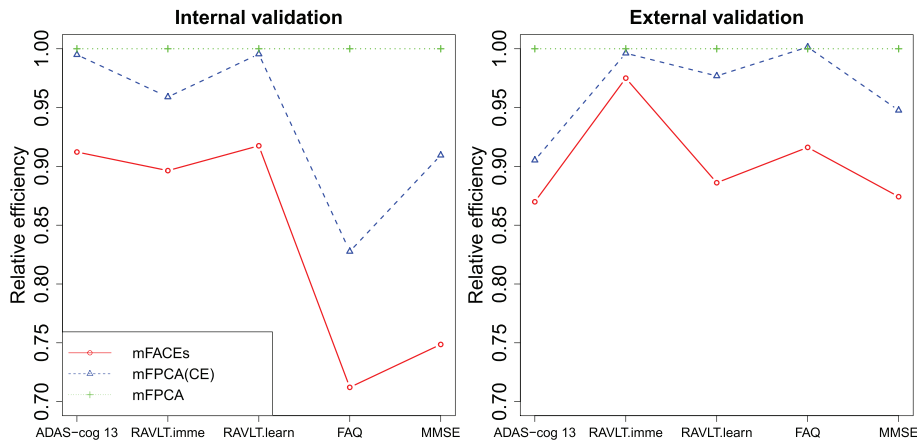


FIGURE 8 The internal and external prediction validations for the Alzheimer's Disease Neuroimaging Initiative longitudinal markers. ADAS-Cog 13, Disease Assessment Scale-Cognitive 13 items; FAQ, Functional Assessment Questionnaire; MMSE, Mini-Mental State Examination; RAVLT.imme, Rey Auditory Verbal Learning Test immediate recall; RAVLT.learn, Rey Auditory Verbal Learning Test learning curve

progression. Specifically, these subjects with a positive score for the second eigenfunction will have higher ADAS-Cog 13/FAQ and lower RAVLT and MMSE over the months, suggesting of AD progression. Finally, we illustrate in Figure 7 the predicted curves along with the associated 95% point-wise confidence bands for three subjects. We focus on predicting the trajectories over the first 4 years as there are more observations. We can see that the confidence bands are getting wider at the later time points because of fewer observations.

4.2 | Comparison of prediction performance of different methods

We compare the proposed mFACEs with mFPCA and mFPCA(CE) for predicting the five longitudinal biomarkers. The prediction performance is evaluated by the average squared prediction errors (APE),

$$APE_k = \frac{1}{n} \sum_{i=1}^n \left[\frac{1}{m_{ik}} \sum_{j=1}^{m_{ik}} \{Y_{ij}^{(k)} - \hat{Y}_{ij}^{(k)}\}^2 \right],$$

where $\hat{Y}_{ij}^{(k)}$ is the predicted value of the k th biomarker for the i th subject at time $t_{ij}^{(k)}$. We conduct two types of validation: an internal validation and an external validation. For the internal validation, we perform a 10-fold cross-validation to the combined data of ADNI-1 and ADNI-2. For the external validation, we fit the model using only the ADNI-1 data and then predict ADNI-2 data. Figure 8 summarizes the results. For simplicity, we present the relative efficiency of APE, which is the ratio of APEs of one method against the mFPCA. In both cases, mFACEs achieves better prediction accuracy than competing methods. Note that mFPCA(CE) outperforms mFPCA for predicting almost all biomarkers. The results suggest that (a) mFACEs is better than competing methods for analysing the longitudinal biomarkers. and (b) exploiting the correlations between the biomarkers improves prediction.

5 | DISCUSSION

The prevalence of multivariate functional data has sparked much research interests in recent years. However, covariance estimation for multivariate sparse functional data remains underdeveloped. We proposed a new method, mFACEs, and its features include the following: (a) A covariance smoothing framework is proposed to tackle multivariate sparse functional data; (b) an automatic and fast fitting algorithm is adopted to ensure the scalability of the method; (c) eigenfunctions and eigenvalues can be obtained through a one-time spectral decomposition, and eigenfunctions can be easily evaluated at any sampling points; and (d) a multivariate extension of the conditional expectation approach (Yao et al., 2005) is derived to exploit correlations between outcomes. The simulation study and the data example showed that mFACEs could better capture between-function correlations and thus gives improved principal component analysis and curve prediction.

When the magnitude of functional data are quite different, one may first normalize the functional data, as recommended by Chiou et al. (2014). One method of normalization is to rescale the functional data using the estimated variance function $\hat{C}_{kk}(t, t)^{-1/2}$ as in Chiou et al. (2014) and Jacques and Preda (2014). An alternative method is to use a global rescaling factor like $\left(\int \hat{C}_{kk}(t, t) dt\right)^{-1/2}$ as in Happ and Greven (2018). Both methods can be easily incorporated into our proposed method. In our data analysis, we find that the results with normalization are very close to those without normalization; thus, we present the results without normalization.

Because multivariate FPCA is more complex than univariate FPCA, weak correlations between the functions and small sample size may offset the benefit of conducting multivariate FPCA; see Section 7.3 in Wong et al. (2019). Thus, it is of future interest to develop practical tests to determine if correlations between multivariate functional data are different from 0.

The mFACEs method has been implemented in an R package `mfaces` and will be submitted to CRAN for public access.

DATA AVAILABILITY STATEMENT

Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (<http://adni.loni.usc.edu>). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this paper. A complete listing of ADNI investigators can be found at: http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf.

ACKNOWLEDGEMENTS

This work was supported partly by National Institute of Neurological Disorders and StrokeR01NS091307. Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (U.S. Department of Defense Award Number W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging,

and the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: AbbVie, Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Cogstate; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Lumosity; Lundbeck; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Therapeutic Research Institute at the University of Southern California. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California.

ORCID

Cai Li  <https://orcid.org/0000-0002-5624-3031>

Luo Xiao  <https://orcid.org/0000-0001-8707-0914>

Sheng Luo  <https://orcid.org/0000-0003-4214-5809>

REFERENCES

- Berrendero, J., Justel, A., & Svarc, M. (2011). Principal components for multivariate functional data. *Computational Statistics Data Analysis*, *55*(9), 2619–2634.
- Cai, T., & Yuan, M. (2010). Nonparametric covariance function estimation for functional and longitudinal data. University of Pennsylvania and Georgia Institute of Technology.
- Chiou, J. M., Chen, Y. T., & Yang, Y. F. (2014). Multivariate functional principal component analysis: A normalization approach. *Statistica Sinica*, *24*(4), 1571–1596.
- Chiou, J. M., & Müller, H. G. (2014). Linear manifold modelling of multivariate functional data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *76*(3), 605–626.
- Chiou, J. M., & Müller, H. G. (2016). A pairwise interaction model for multivariate functional and longitudinal data. *Biometrika*, *103*(2), 377–396.
- Eilers, P., & Marx, B. (1996). Flexible smoothing with B-splines and penalties (with discussion). *Statistical Science*, *11*(2), 89–121.
- Eilers, P., & Marx, B. (2003). Multivariate calibration with temperature interaction using two-dimensional penalized signal regression. *Chemometrics and Intelligent Laboratory Systems*, *66*, 159–174.
- Goldsmith, J., Crainiceanu, C. M., Caffo, B., & Reich, D. (2012). Longitudinal penalized functional regression for cognitive outcomes on neuronal tract measurements. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *61*(3), 453–469.
- Greven, S., Crainiceanu, C. M., Caffo, B., & Reich, D. (2010). Longitudinal functional principal component. *Electronic Journal of Statistical*, *4*, 1022–1054.
- Happ, C., & Greven, S. (2018). Multivariate functional principal component analysis for data observed on different (dimensional) domains. *Journal of the American Statistical Association*, *113*(522), 649–659.
- Huang, H., Li, Y., & Guan, Y. (2014). Joint modeling and clustering paired generalized longitudinal trajectories with application to cocaine abuse treatment data. *Journal of the American Statistical Association*, *109*(508), 1412–1424.
- Jacques, J., & Preda, C. (2014). Model-based clustering for multivariate functional data. *Computational Statistics & Data Analysis*, *71*, 92–106.
- James, G., Hastie, T., & Sugar, C. (2000). Principal component models for sparse functional data. *Biometrika*, *87*(3), 587–602.
- Kowal, D. R., Matteson, D. S., & Ruppert, D. (2017). A Bayesian multivariate functional dynamic linear model. *Journal of the American Statistical Association*, *112*(518), 733–744.
- Leng, X., & Müller, H. (2006). Classification using functional data analysis for temporal gene expression data. *Bioinformatics*, *22*(1), 68–76.
- Li, K., Chan, W., Doody, R. S., Quinn, J., & Luo, S. (2017). Prediction of conversion to Alzheimer's disease with longitudinal measures and time-to-event data. *Journal of Alzheimer's Disease*, *58*(2), 361–371.
- Li, J., Huang, C., & Zhu, H. (2017). A functional varying-coefficient single-index model for functional response data. *Journal of the American Statistical Association*, *112*(519), 1169–1181.
- 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.
- Lindquist, M. (2012). Functional causal mediation analysis with an application to brain connectivity. *Journal of the American Statistical Association*, *107*(500), 1297–1309.
- Luo, R., & Qi, X. (2017). Function-on-function linear regression by signal compression. *Journal of the American Statistical Association*, *112*(518), 690–705.
- Morris, J., Arroyo, C., Coull, B., Ryan, L., Herrick, R., & Gortmaker, S. (2006). Using wavelet-based functional mixed models to characterize population heterogeneity in accelerometer profiles: A case study. *Journal of the American Statistical Association*, *101*(476), 1352–1364.
- Park, J., & Ahn, J. (2017). Clustering multivariate functional data with phase variation. *Biometrics*, *73*(1), 324–333.

- Peng, J., & Paul, D. (2009). A geometric approach to maximum likelihood estimation of functional principal components from sparse longitudinal data. *Journal of Computational and Graphical Statistics*, 18(4), 995–1015.
- Petersen, A., & Müller, H. G. (2016). Fréchet integration and adaptive metric selection for interpretable covariances of multivariate functional data. *Biometrika*, 103(1), 103–120.
- Qi, X., & Luo, R. (2018). Function-on-function regression with thousands of predictive curves. *Journal of Multivariate Analysis*, 163, 51–66.
- Qiao, X., Guo, S., & James, G. M. (2019). Functional graphical models. *Journal of the American Statistical Association*, 114(525), 211–222.
- Ramsay, J., & Silverman, B. (2005). *Functional data analysis*. New York: Springer.
- Reimherr, M., & Nicolae, D. (2014). A functional data analysis approach for genetic association studies. *The Annals of Applied Statistics*, 8(1), 406–429.
- Reimherr, M., & Nicolae, D. (2016). Estimating variance components in functional linear models with applications to genetic heritability. *Journal of the American Statistical Association*, 111(513), 407–422.
- Reiss, P., & Ogden, R. (2010). Functional generalized linear models with images as predictors. *Biometrics*, 66(1), 61–69.
- Saporta, G. (1981). Méthodes exploratoires d'analyse de données temporelles. Cahiers du Bureau universitaire de recherche opérationnelle Série Recherche, 37, 7–194.
- Seber, G. (2008). *A matrix handbook for statisticians*. Hoboken, New Jersey: John Wiley & Sons.
- Weiner, M. W., Veitch, D. P., Aisen, P. S., Beckett, L. A., Cairns, N. J., Green, R. C., & Petersen, R. C. (2017). Recent publications from the Alzheimer's disease neuroimaging initiative: Reviewing progress toward improved ad clinical trials. *Alzheimer's & Dementia: The Journal of the Alzheimer's Association*, 13(4), e1–e85.
- Wong, R. K., Li, Y., & Zhu, Z. (2019). Partially linear functional additive models for multivariate functional data. *Journal of the American Statistical Association*, 114(525), 406–418.
- Wong, R. K., & Zhang, X. (2019). Nonparametric operator-regularized covariance function estimation for functional data. *Computational Statistics & Data Analysis*, 131, 131–144.
- Wood, S. N. (2000). Modelling and smoothing parameter estimation with multiple quadratic penalties. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 62(2), 413–428.
- Xiao, L., Huang, L., Schrack, J., Ferrucci, L., Zipunnikov, V., & Crainiceanu, C. M. (2015). Quantifying the life-time circadian rhythm of physical activity: A covariate-dependent functional approach. *Biostatistics*, 16(2), 352–367.
- Xiao, L., Li, C., Checkley, W., & Crainiceanu, C. M. (2018). R package face: Fast covariance estimation for sparse functional data (version 0.1-4). <http://cran.r-project.org/web/packages/face/index.html>.
- Xiao, L., Li, C., Checkley, W., & Crainiceanu, C. M. (2018). Fast covariance estimation for sparse functional data. *Statistics and Computing*, 28(3), 511–522.
- Xu, G., & Huang, J. (2012). Asymptotic optimality and efficient computation of the leave-subject-out cross-validation. *Annals of Statistics*, 40(6), 3003–3030.
- Yao, F., Müller, H., & Wang, J. (2005). Functional data analysis for sparse longitudinal data. *Journal of the American Statistical Association*, 100(470), 577–590.
- Zhou, L., Huang, J. Z., & Carroll, R. J. (2008). Joint modelling of paired sparse functional data using principal components. *Biometrika*, 95(3), 601–619.
- Zhou, S., Shen, X., & Wolfe, D. A. (1998). Local asymptotics for regression splines and confidence regions. *Annals of Statistics*, 26(5), 1760–1782.
- Zhu, H., Brown, P., & Morris, J. (2012). Robust classification of functional and quantitative image data using functional mixed models. *Biometrics*, 68(4), 1260–1268.
- Zhu, H., Li, R., & Kong, L. (2012). Multivariate varying coefficient model for functional responses. *Annals of Statistics*, 40(5), 2634–2666.
- Zhu, H., Morris, J. S., Wei, F., & Cox, D. D. (2017). Multivariate functional response regression, with application to fluorescence spectroscopy in a cervical pre-cancer study. *Computational Statistics & Data Analysis*, 111, 88–101.
- Zhu, H., Strawn, N., & Dunson, D. B. (2016). Bayesian graphical models for multivariate functional data. *Journal of Machine Learning Research*, 17(1), 7157–7183.

SUPPORTING INFORMATION

The Supporting Information contains additional details and results of a simulation that mimics the ADNI data. R package `mfaces` (<https://github.com/cli9/mfaces>) contains an open-source implementation of the proposed method described in the article. Demo zip file contains codes to demonstrate the proposed method with a simulated dataset in Section 3.

Additional supporting information may be found online in the Supporting Information section at the end of the article.

How to cite this article: Li C, Xiao L, Luo S. Fast covariance estimation for multivariate sparse functional data. *Stat*. 2020;9:e245. <https://doi.org/10.1002/sta4.245>

APPENDIX A: PROOFS OF PROPOSITIONS 1 AND 2

Proof of Proposition 1 Define $\tilde{\mathbf{b}}(t) = \mathbf{G}^{-\frac{1}{2}}\mathbf{b}(t)$, and then $\int \tilde{\mathbf{b}}(t)\tilde{\mathbf{b}}(t)^\top dt = \mathbf{I}$. Define $\check{\Theta}_{kk'}(t) = \mathbf{G}^{\frac{1}{2}}\Theta_{kk'}\mathbf{G}^{\frac{1}{2}}$. According to Equation (1),

$$\begin{aligned} d_e \Psi_\ell^{(k)}(s) &= \sum_{k=1}^p \int \mathbf{b}(s)^\top \Theta_{kk'} \mathbf{b}(t) \Psi_\ell^{(k')}(t) dt \\ &= \tilde{\mathbf{b}}(s)^\top \sum_{k=1}^p \check{\Theta}_{kk'} \int \tilde{\mathbf{b}}(t) \Psi_\ell^{(k')}(t) dt. \end{aligned}$$

Thus, $\Psi_\ell^{(k)}(s) = \tilde{\mathbf{b}}(s)^\top \mathbf{u}_\ell^{(k)}$ with $\mathbf{u}_\ell^{(k)} = d_\ell^{-1} \left\{ \sum_{k'=1}^p \check{\Theta}_{kk'} \int \tilde{\mathbf{b}}(t) \Psi_\ell^{(k')}(t) dt \right\} \in \mathbb{R}^c$ and $\mathbf{u}_\ell = (\mathbf{u}_\ell^{(1)\top}, \dots, \mathbf{u}_\ell^{(p)\top})^\top \in \mathbb{R}^{pc}$. As $\sum_{k=1}^p \int \Psi_\ell^{(k)}(t) \Psi_\ell^{(k)}(t) dt = \mathbf{1}_{\{e=e'\}}$, we derive that

$$\mathbf{u}_\ell^\top \mathbf{u}_{\ell'} = \sum_{k=1}^p \mathbf{u}_\ell^{(k)\top} \mathbf{u}_{\ell'}^{(k)} = \mathbf{1}_{\{e=e'\}}. \quad (\text{B1})$$

By Equation (2),

$$\mathbf{C}_{kk'}(s, t) = \tilde{\mathbf{b}}(s)^\top \check{\Theta}_{kk'} \tilde{\mathbf{b}}(t) = \tilde{\mathbf{b}}(s)^\top \left\{ \sum_{e \geq 1} d_e \mathbf{u}_\ell^{(k)} \mathbf{u}_\ell^{(k')\top} \right\} \tilde{\mathbf{b}}(t),$$

which gives

$$\check{\Theta}_{kk'} = \sum_{e \geq 1} d_e \mathbf{u}_\ell^{(k)} \mathbf{u}_\ell^{(k')\top}.$$

As $1 \leq k, k' \leq p$, the above is equivalent to

$$\underbrace{\begin{pmatrix} \check{\Theta}_{11} & \dots & \check{\Theta}_{1p} \\ \vdots & \ddots & \vdots \\ \check{\Theta}_{p1} & \dots & \check{\Theta}_{pp} \end{pmatrix}}_{:= \check{\Theta}} = \sum_{e \geq 1} d_e \mathbf{u}_\ell \mathbf{u}_\ell^\top.$$

Because of Equation (B1), \mathbf{u}_ℓ s are orthonormal eigenvectors of $\check{\Theta}$ with d_e s the corresponding eigenvalues. The proof is now complete.

Proof of Proposition 2. By Equation (10),

$$\text{IGCV} = \|\hat{\mathbf{C}}\|^2 - 2\mathbf{f}^\top \Sigma^{-1} \mathbf{f} + \mathbf{f}^\top \Sigma^{-2} \mathbf{f} + 2 \sum_{i=1}^n (\mathbf{L}_i \Sigma^{-1} \mathbf{f} - \mathbf{f}_i)^\top \Sigma^{-1} (\mathbf{L}_i \Sigma^{-1} \mathbf{f} - \mathbf{f}_i). \quad (\text{B2})$$

Since $\Sigma^{-1} = \mathbf{U} \text{diag}(\tilde{\mathbf{d}}) \mathbf{U}^\top$, we have

$$\mathbf{f}^\top \Sigma^{-1} \mathbf{f} = \tilde{\mathbf{f}}^\top \text{diag}(\tilde{\mathbf{d}}) \tilde{\mathbf{f}} = \tilde{\mathbf{d}}^\top (\tilde{\mathbf{f}} \odot \tilde{\mathbf{f}}). \quad (\text{B3})$$

Similarly,

$$\mathbf{f}^\top \Sigma^{-2} \mathbf{f} = \tilde{\mathbf{f}}^\top \text{diag}(\tilde{\mathbf{d}}^2) \tilde{\mathbf{f}} = (\tilde{\mathbf{f}} \odot \tilde{\mathbf{d}})^\top (\tilde{\mathbf{f}} \odot \tilde{\mathbf{d}}). \quad (\text{B4})$$

Next we derive that

$$(\mathbf{L}_i \Sigma^{-1} \mathbf{f} - \mathbf{f}_i)^\top \Sigma^{-1} (\mathbf{L}_i \Sigma^{-1} \mathbf{f} - \mathbf{f}_i) = (\mathbf{U}^\top \mathbf{L}_i \Sigma^{-1} \mathbf{f} - \mathbf{U}^\top \mathbf{f}_i)^\top \text{diag}(\tilde{\mathbf{d}}) (\mathbf{U}^\top \mathbf{L}_i \Sigma^{-1} \mathbf{f} - \mathbf{U}^\top \mathbf{f}_i) = (\tilde{\mathbf{L}}_i \text{diag}(\tilde{\mathbf{d}}) \tilde{\mathbf{f}} - \tilde{\mathbf{f}}_i)^\top \text{diag}(\tilde{\mathbf{d}}) (\tilde{\mathbf{L}}_i \text{diag}(\tilde{\mathbf{d}}) \tilde{\mathbf{f}} - \tilde{\mathbf{f}}_i).$$

It follows that

$$(\mathbf{L}_i \Sigma^{-1} \mathbf{f} - \mathbf{f}_i)^\top \Sigma^{-1} (\mathbf{L}_i \Sigma^{-1} \mathbf{f} - \mathbf{f}_i) = \tilde{\mathbf{d}}^\top \left[\left\{ \tilde{\mathbf{L}}_i (\tilde{\mathbf{f}} \odot \tilde{\mathbf{d}}) \right\} \odot \left\{ \tilde{\mathbf{L}}_i (\tilde{\mathbf{f}} \odot \tilde{\mathbf{d}}) \right\} + (\tilde{\mathbf{f}}_i \odot \tilde{\mathbf{f}}_i) \right] - 2\tilde{\mathbf{d}}^\top \left\{ (\tilde{\mathbf{f}}_i \tilde{\mathbf{f}}_i^\top) \odot \tilde{\mathbf{L}}_i \right\} \tilde{\mathbf{d}}. \quad (\text{B5})$$

By combining Equations (B2), (B3), (B4), and (B5), the proof is complete.

APPENDIX B: LEMMA

Lemma 1. *The covariance operator with the covariance functions defined in Section 4.1 is positive semidefinite.*

Proof. Let $\mathbf{a} = (a_1, \dots, a_p)^\top \in \mathbb{R}^p$ and $\tilde{X} = \mathbf{a}^\top \mathbf{x}$, and then \tilde{X} is a stochastic process with covariance function

$$\text{Cov}\{\tilde{X}(s), \tilde{X}(t)\} = \sum_{kk'} a_k a_{k'} C_{kk'}(s, t) = \sum_{kk'} a_k a_{k'} (\rho + (1 - \rho) \mathbf{1}_{\{k=k'\}}) \tilde{\Phi}_k(s)^\top \tilde{\Phi}_{k'}(t),$$

where $\tilde{\Phi}_k(s) = \Phi_k(s)^\top \Lambda_{kk}^{\frac{1}{2}}$. Let $\Psi(s) = \sum_{k=1}^p a_k \tilde{\Phi}_k(s)$. Then

$$\text{Cov}\{\tilde{X}(s), \tilde{X}(t)\} = \rho \Psi(s)^\top \Psi(t) + (1 - \rho) \sum_k a_k^2 \tilde{\Phi}_k(s)^\top \tilde{\Phi}_k(t),$$

which is always positive semidefinite, and the proof is complete.