NONPARAMETRIC COVARIANCE ESTIMATION
FOR MIXED LONGITUDINAL STUDIES,
WITH APPLICATIONS IN MIDLIFE WOMEN'S HEALTH

Anru R. Zhang[1,2] and Kehui Chen[3]

[1]*University of Wisconsin-Madison,* [2]*Duke University*
and [3]*University of Pittsburgh*

Abstract: In mixed longitudinal studies, a group of subjects enter the study at different ages (cross-sectional) and are followed for successive years (longitudinal). In the context of such studies, we consider nonparametric covariance estimation with samples of noisy and partially observed functional trajectories. The proposed algorithm is based on a noniterative sequential-aggregation scheme with only basic matrix operations and closed-form solutions in each step. The good performance of the proposed method is supported by both theory and numerical experiments. We also apply the proposed procedure to a study on the working memory of midlife women, based on data from the Study of Women's Health Across the Nation (SWAN).

Key words and phrases: Consistency, covariance estimation, cross-sectional, functional data, longitudinal studies, partial trajectories.

1. Introduction

A mixed longitudinal study is a mixture of a longitudinal and a cross-sectional study (Berger (1986); Helms (1992)). Suppose the researchers intend to study the social and cognitive development of children aged four to twelve. In an ideal longitudinal design, a group of four-year-old children will be recruited and followed over eight successive years. In a mixed longitudinal design, one can recruit a group of children between the ages of four and eight, and then follow them for four years (within a typical funding period). Because the age requirement is more flexible at recruitment, this type of mixed longitudinal design results in shorter completion times and potentially larger group sizes. However, this type of mixed longitudinal design also brings new challenges for statistical analysis, because the trajectory is only partially observed for each subject.

Specifically, we consider a data example from the Study of Women's Health Across the Nation (SWAN). The SWAN is a community-based, longitudinal study

---

Corresponding author: Anru R. Zhang, Department of Statistics, University of Wisconsin-Madison, Madison, WI 53706, USA. E-mail: anruzhang@stat.wisc.edu.
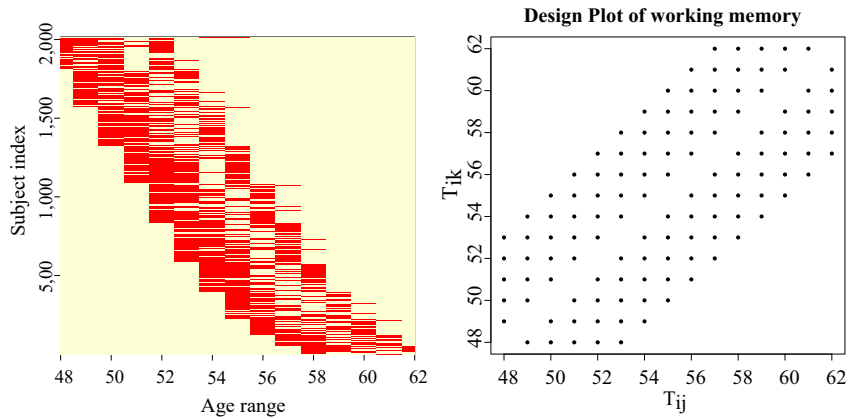
Figure 1. Left: For each of the 2,016 subjects, measurements were between age x and $x + 5$, for some $x \in [48, 58]$. Right: The design plots for covariance $G(s, t)$, that is, the assembled pairs of (t_{ij}, t_{il}), for $1 \leq i \leq n, 1 \leq j < l \leq n_i$. The pooled pairs do not fill the entire domain $\mathcal{T}^2$ because there are no measurements available for pairs of (t_{ij}, t_{il}) whenever $|j - l| > 5$.

of midlife women. Women aged between 42 and 52 years were enrolled around 1996/97, and followed annually thereafter. Currently, SWAN data up to the 10th follow up visit are available in a publicly accessible repository managed by the ICPSR, at `http://www.icpsr.umich.edu/icpsrweb/ICPSR/series/00253`. Although enormous studies have examined cognitive functioning in midlife, few are longitudinal, and most are based on three or fewer cognition assessments (Karlamangla et al. (2017)). As a result, there are insufficient studies on within-person longitudinal decline in cognitive performance in those under 60 years of age (Hedden and Gabrieli (2004); Rönnlund et al. (2005)). In contrast, the SWAN data contain more follow-ups and a wider age range, providing a good opportunity for a longer-term study of women's midlife health. In particular, we focus on working memory measurements available from Visits six to ten. By pooling the subjects, the age range under consideration is a span of 15 years: $\mathcal{T} = [48, 62]$. However, given its mixed longitudinal design, the longitudinal follow-ups of each subject in the SWAN only capture a piece of the chronological aging trajectory, and the shape might have a complex interaction with age (Rönnlund et al. (2005); Fuh et al. (2006)). As shown in the left panel of Figure 1, the measurements for each subject are only a subset within a period of at most five years. Traditional parametric models, such as linear mixed models (with age as a between-subject effect, and the time of follow-up as a within-subject effect), often assume a linear trend over time. However, the individual chronological aging trajectories of working

memory might have a complex shape. For example, working memory might improve first and then decline, and the age when working memory starts to decline varies among subjects. Therefore, we believe that nonparametric models, such as a functional principal component analysis, may reveal interesting features.

We consider a mixed longitudinal design for n subjects, where for each subject k, measurements are obtained at times t_{kj}, for $k = 1, \ldots, n$ and $j = 1, \ldots, n_k$. We use the notation

$$X_k(t_{kj}) = Z_k(t_{kj}) + \epsilon_{kj}, \quad t_{kj} \in \mathcal{T}, \tag{1.1}$$

where ϵ_{kj} are zero mean independent and identically distributed (i.i.d.) measurement errors that are uncorrelated with all other random components and satisfy $\text{var}(\epsilon_{kj}) = \sigma^2$. Here, $Z(t)$, for $t \in \mathcal{T}$, is assumed to be a square-integrable random process with mean and covariance functions $\mu(t)$ and $G(s, t) = \text{Cov}(Z(s), Z(t))$. In a mixed longitudinal design, the observed time points $\{t_{kj}\}_{j=1,\ldots,n_k}$ for each subject k are restricted to a subject-specific partial domain. As shown in the right panel of Figure 1, we do not have within-subject correlation information for any two points that are more than five years apart in the SWAN data example. To apply a functional data approach for mixed longitudinal studies, the main methodological challenge is to nonparametrically estimate the covariance structure G of the underlying process.

Estimating the mean and covariance functions plays a fundamentally important role in a functional data analysis. Useful tools, such as a functional principal component analysis, often rely on a consistent covariance function estimation (Yao et al. (2005); Hall and Hosseini-Nasab (2006); Li and Hsing (2010)). For conventional functional data, where the pooled design (right panel of Figure 1) for the covariance is complete, various methods based on kernel smoothing and splines have been proposed (e.g., Rice and Silverman (1991); Yao et al. (2005); Peng and Paul (2009); Xiao et al. (2013)). In a study in which the covariance information is incomplete, Fan et al. (2007) considered a semiparametric covariance estimation, where the variance function $G(t, t) = \sigma^2(t)$ is modeled non-parametrically under smoothness conditions, while the off-diagonal correlation structures are assumed to have a parametric form $\rho(s, t, \theta)$. However, this problem differs from the banded covariance estimation considered in studies such as Bickel and Levina (2008), Cai et al. (2010), Cai and Yuan (2012), Cai, Ren and Zhou (2016), and the references therein, because there is no bandable covariance structure in our scenario, and the design pairs are only within a banded area.

We propose estimating the covariance suing a sequential-aggregation scheme

(see Section 2). The proposed algorithm is noniterative, with closed-form solutions and only basic matrix operations (such as matrix multiplication and singular value decomposition (SVD)) in each step. We prove that under moderate conditions (see Section 3), the proposed method consistently recovers the nonparametric covariance structure using data within a banded area. A key step of the proposed procedure is solving the orthogonal Procrustes or Wahba problem (Wahba (1965)), that is, finding a rotation matrix to best align two sets of points in two different Euclidean coordinate systems. This problem was first motivated by satellite attitude determination, then later applied to many other applications. To theoretically analyze the procedure, we introduce a new error bound for the solution to Wahba problem (Lemma 1). In the theoretical analysis, we introduce a series of technical tools on perturbation inequalities of singular subspaces, including Lemmas 3, 5, 7, and 8, which may be of independent interest.

Fragmentary functional observations have been studied under other modeling assumptions; see, for example, Delaigle and Hall (2013) and Delaigle and Hall (2016). Descary and Panaretos (2018) and Kneip and Liebl (2017) consider covariance estimation and reconstruction from fragmentary functional observations using an optimization framework. The implementations of both works involve iterations. In particular, Descary and Panaretos (2018) formulates the problem as a nonconvex optimization that aims to minimize the error within the observable diagonal band under a rank constraint. In contrast, we introduce a novel sequential-aggregation approach that provides explicit solutions and new insights into the covariance estimation problem. We also include numerical comparisons with the method of Descary and Panaretos (2018) in the simulation section. In addition, this problem is related to several recent works on high-dimensional covariance estimation with missing values. For example, Loh and Wainwright (2012) and Lounici (2014) consider a linear regression or covariance matrix estimation, where the observations are missing randomly with a fixed rate. In contrast, Kolar and Xing (2012) and Cai and Zhang (2016) consider a more general setting that allows a nonrandom missing pattern, but still requires that each pair of covariates simultaneously appear in a sufficient number of samples. The problem discussed in this paper is distinct from these existing settings, because a large portion of the covariate pairs will never appear in the same sample (such as the pairs between the earlier and latest observations in the longitudinal studies), by the nature of the design. Bishop and Byron (2014) studied a similar sequential-aggregation scheme for matrix completion. However, they mainly consider the completion of high-dimensional low-rank positive semidefinite matrices in a deterministic setting, whereas we provide a statistical guarantee for

covariance estimation from partially observed noisy functional data.

The rest of this paper is organized as follows. The methodology and algorithm are described in Section 2, followed by theoretical analyses in Section 3. In Section 4, we present a series of numerical experiments, including the application to the SWAN data. Section 5 concludes the paper. The proofs are collected in the Supplementary Materials.

2. Covariance Estimation for Mixed Longitudinal Design

We briefly introduce the notation that will be used throughout the paper. For a matrix $A \in \mathbb{R}^{p_1 \times p_2}$ or bivariate function G, let $\{\sigma_1(A), \sigma_2(A), \ldots\}$ and $\{\sigma_1(G), \sigma_2(G), \ldots\}$ be singular values in nonincreasing order. We adapt the R syntax to indicate matrices/functions restricted to the subsets of indices/domains: if $A \in \mathbb{R}^{p_1 \times p_2}$, and $a \leq b$ and $c \leq d$ are four positive integers, we use $A_{[a:b,c:d]}$ to denote the submatrix of A formed by its ath to bth rows and cth to dth columns. Here, "$:$" alone represents the entire index set, so $A_{[:,1:r]}$ and $A_{[a:b,:]}$ represent the first r columns of A and the $\{a, \ldots, b\}$th rows of A, respectively; similarly, $G_{[\mathcal{T}_1, \mathcal{T}_2]}$ represents a function G with domain $\mathcal{T}_1 \times \mathcal{T}_2$. Let $\mathcal{L}(\mathcal{T})$ be the Lebesgue measure of any domain $\mathcal{T}$. Let $\|A\|_F$ and $\|A\|$ be the matrix Frobenius norm and operator norm, respectively: $\|A\|_F = (\sum_{i,j} A_{ij}^2)^{1/2} = (\sum_i \sigma_i^2(A))^{1/2}$, $\|A\| = \sigma_{\max}(A)$. Denote $I_{r \times r}$ as the r-by-r identity matrix, and $\mathbb{O}_{p,r} = \{V : V^\top V = I_{r \times r}\}$ as the set of all p-by-r matrices with orthonormal columns. In particular, the set of all r-by-r orthogonal matrices can be denoted as $\mathbb{O}_r = \mathbb{O}_{r,r}$. Denote $\|G\|_{HS} = \left(\iint |G(s_1, s_2)|^2 ds_1 s_2\right)^{1/2}$ as the Hilbert–Schmidt norm of the bivariate function G. Finally, we use $C, C_0, C_1, c, c_0, \ldots$ to represent generic constants, the exact values of which may vary from line to line.

Suppose $\mathcal{T}$ is the entire period of interest. Consider an equally spaced grid of time points $T = \{t_1, \ldots, t_p\}$ on the time domain $\mathcal{T}$. In a mixed-longitudinal design, suppose $\mathcal{T}_k$ is the observational period for subject k, and we observe $X_k(T_k)$ in the contiguous band of the domain $\mathcal{T}_k$:

$$\mathcal{T}_k \subseteq \mathcal{T}, \quad \frac{\mathcal{L}(\mathcal{T}_k)}{\mathcal{L}(\mathcal{T})} = \delta, \quad T_k \subseteq T \cap \mathcal{T}_k = \{t_1, \ldots, t_p\} \cap \mathcal{T}_k, \quad k = 1, \ldots, n.$$

Here, the fraction of observation δ is assumed to be a constant between zero and one and T_k might not be consecutive, owing to missing values. If T_k is complete with no missing values, the number of observations is d, with $\delta = d/p$. Suppose the signal-noise decomposition (1.1) holds for each observation: $X_k(t_{kj}) = Z_k(t_{kj}) + \epsilon_{kj}$. Let Σ_0 denote the $p \times p$ discretized version of covariance

G, that is, the (i, j)th entry of Σ_0 is equal to $\text{Cov}(Z(t_i), Z(t_j)) = G(t_i, t_j)$. We estimate G using the discretized version Σ_0. Suppose G has approximate rank r. Then, we also have $\Sigma_0 \approx AA^\top$, where $A \in \mathbb{R}^{p \times r}$ can be regarded as the factors of Σ_0.

We consider a sequential-aggregation-based algorithm. We first divide $\mathcal{T}$ into a series of overlapping sub-domains, then obtain estimates of A on each sub-domain. Next, we aggregate all estimates on the sub-intervals into a full estimate of A. Here, a crucial rotation operation is involved in the aggregation step to ensure that the estimates of A on each sub-domain are aligned. Finally, we obtain an estimate of Σ_0 from $\tilde{A}\tilde{A}^\top$, where $\tilde{A}$ is an estimate of A up to a rotation. Then, G is recovered using a standard interpolation technique. The steps are as follows; see Figure 2. For any sub-index set $I \subseteq \{1, \ldots, p\}$, we use the notation $T(I) = \{t_i : i \in I\}$ and $(X_k)_I = X_k(T(I))$.

Step 1 For a chosen band parameter b and an increment parameter a satisfying $1 \leq a \leq b - r \leq b \leq d$, we construct the following sub-index set:

$$I_l = \{(l-1)a + 1, \ldots, \{(l-1)a + b\} \wedge p\}, \quad l = 1, \ldots, l_{\max}. \qquad (2.1)$$

Here, $l_{\max} = 1 + \lceil (p - b)/a \rceil$ is the total number of sub-index sets. Each I_l except the last one contains b indices, and the last one contains at most b indices.

Step 2 For $l = 1, \ldots, l_{\max}$, we search for all samples that have full observations in I_l, and denote the set of such samples as J_l:

$$J_l = \{1 \leq k \leq n : T(I_l) \subseteq T_k\}.$$

Then, the sample covariance matrix for the indices in I_l is calculated as

$$\hat{\Sigma}_l \in \mathbb{R}^{|I_l| \times |I_l|}, \quad \hat{\Sigma}_l = \frac{1}{n_l^*} \sum_{k \in J_l} \left((X_k)_{I_l} - \bar{X}_{I_l}\right) \left((X_k)_{I_l} - \bar{X}_{I_l}\right)^\top,$$

$$n_l^* = |J_l|, \quad \bar{X}_{I_l} = \frac{1}{n_l^*} \sum_{k \in J_l} (X_k)_{I_l}. \qquad (2.2)$$

Step 2′ As an alternative to using only subjects that have full observations in I_l, we can use all data available for the pair $(I_l(i), I_l(j))$ when computing $\hat{\Sigma}_{l,[ij]}$. This scheme is preferred to Step 2 when large portions of subjects have missing values; that is, $X_k(T_k)$ are not complete consecutive observations (see Theorem 1 and Remark 4): $\hat{\Sigma}_l \in \mathbb{R}^{|I_l| \times |I_l|}, \bar{X}_{I_l} \in \mathbb{R}^{|I_l|}$,
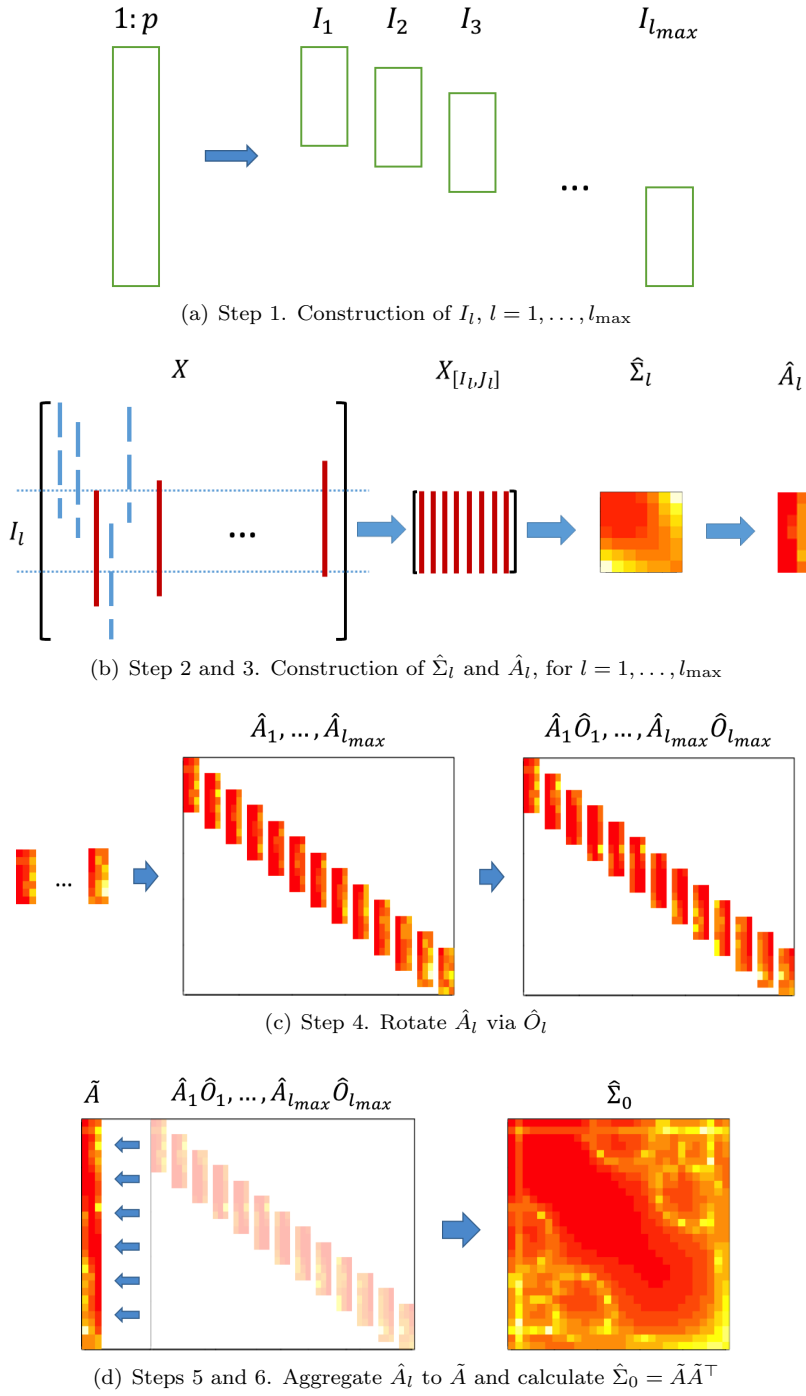
(a) Step 1. Construction of I_l, $l = 1, \ldots, l_{\max}$

(b) Step 2 and 3. Construction of $\hat{\Sigma}_l$ and $\hat{A}_l$, for $l = 1, \ldots, l_{\max}$

(c) Step 4. Rotate $\hat{A}_l$ via $\hat{O}_l$

(d) Steps 5 and 6. Aggregate $\hat{A}_l$ to $\tilde{A}$ and calculate $\hat{\Sigma}_0 = \tilde{A}\tilde{A}^\top$

Figure 2. Illustration of the procedure.

$$\hat{\Sigma}_{l,[ij]} = \frac{\sum_{k:T(I_l(i)),T(I_l(j))\in T_k}\left((X_k)_{I_l(i)} - \bar{X}_{I_l(i)}\right)\left((X_k)_{I_l(j)} - \bar{X}_{I_l(j)}\right)}{(n^*)_{i,j,l}},$$

$$n^*_{i,j,l} = |\{k : T(I_l(i)), T(I_l(j)) \in T_k\}|, \tag{2.3}$$

$$\bar{X}_{I_l(i)} = \frac{\sum_{k:T(I_l(i)),T(I_l(j))\in T_k}(X_k)_{I_l(i)}}{n^*_{i,l}}, \quad n^*_{i,l} = |\{k : T(I_l(i)) \in T_k\}|.$$

Step 3 Evaluate the eigenvalue decomposition and the rank-r truncation of $\hat{\Sigma}_l$ as

$$\hat{\Sigma}_l = \hat{U}_l \hat{D}_l \hat{U}_l^\top, \quad \hat{\Sigma}_l^{(r)} = \hat{U}_{l,[:,1:r]} \hat{D}_{l,[1:r,1:r]} \hat{U}_{l,[:,1:r]}^\top. \tag{2.4}$$

Then, for $l = 1, \ldots, l_{\max}$, we evaluate $\hat{\sigma}_l^2 = ((1/|I_l| - r)\sum_{i=r+1}^{|I_l|} \hat{D}_{l,[i,i]}) \vee 0$ as the sample variance of the noise and

$$\hat{A}_l = U_{l,[:,1:r]}\left\{(D_{l,[1:r,1:r]} - \hat{\sigma}_l^2 \cdot I_{r\times r}) \vee 0\right\}^{1/2} \in \mathbb{R}^{|I_l|\times r} \tag{2.5}$$

as the estimate of A on the sub-domain I_l. Here, $I_{r\times r}$ is the r-by-r identity matrix. By these calculations, we expect that $\hat{A}_l \hat{A}_l^\top \approx \Sigma_{0,l} = (\Sigma_0)_{[I_l,I_l]}$.

Step 4 We construct a suitable right rotation on $\hat{A}_l$ so that all the pieces can be aligned. Specifically, we first let $\hat{O}_1 = I_{r\times r}$, and then calculate $\hat{O}_{l+1}$ sequentially as

$$\hat{O}_{l+1} = \operatorname*{argmin}_{O\in\mathbb{O}_r} \left\|(\hat{A}_l)_{[(a+1):b,:]}\hat{O}_l - (\hat{A}_{l+1})_{[1:(b-a),:]}O\right\|_F^2, \quad l = 1, \ldots, l_{\max} - 1. \tag{2.6}$$

Here, the row indices of $(\hat{A}_l)_{[(a+1):b,:]}$ and $(\hat{A}_{l+1})_{[1:(b-a),:]}$ both correspond to $[la + 1, (l-1)a + b] \subseteq \{1, \ldots, p\}$. Note that (2.6) is actually the orthogonal Procrustes or Wahba problem (Wahba (1965)), which can be solved using

$$\hat{O}_{l+1} = \tilde{U}\tilde{V}^\top, \text{ where } \tilde{U}\tilde{\Sigma}\tilde{V}^\top = (\hat{A}_{l+1})_{[1:(b-a),:]}^\top(\hat{A}_l)_{[(a+1):b,:]}\hat{O}_l \text{ is the SVD.} \tag{2.7}$$

Step 5 In this step, we aggregate all pieces $\hat{A}_l\hat{O}_l$ into one complete factor $\tilde{A} \in \mathbb{R}^{p\times r}$. For convenience of notation, we "frame" the $|I_l|$-by-r matrix $\hat{A}_l$ to its original p-by-r factor scale, $\hat{A}_l^* \in \mathbb{R}^{p\times r}$, $\hat{A}_{l,[I_l,:]}^* = \hat{A}_l\hat{O}_l$, and $\hat{A}_{l,[I_l^c,:]}^* = 0$. For $1 \le i \le p$ and $1 \le j \le r$, we calculate

$$\tilde{A}_{[i,j]} = \frac{\sum_{l:i\in I_l}\hat{A}_{l,[i,j]}^*}{|\{l : i \in I_l\}|}. \tag{2.8}$$

Step 6 After the sequential aggregation, we estimate Σ_0 using

$$\hat{\Sigma}_0 = \tilde{A}\tilde{A}^\top \in \mathbb{R}^{p \times p}, \tag{2.9}$$

then linear interpolate between grid points to obtain $\hat{G}$ (Press et al. (1992, Chap. 3.6)). Some smoothing instead of linear interpolation might be useful in data applications for smoother results and better visualization.

Computation and Tuning Parameters: In summary, the proposed algorithm is noniterative, and uses only basic matrix calculations, such as matrix multiplications and SVD, which can be implemented efficiently. The algorithm takes input as a, b, and the rank r. According to our simulation studies in Section 4, the performance of the method is not sensitive to the selection of a and b. In our numerical implementation, we suggest selecting b to be slightly smaller than bandwidth d, and selecting a to be a small increment (in practice $a = 0.1 \times d$ usually provides a good enough result). In the following, we describe the random sub-sampling cross-validation method (Picard and Cook (1984)) used to select the rank r.

We first randomly split n observations $\{X_k(T_k)\}_{k=1}^n$ into training and testing groups of sizes $n_1 \approx (K-1)n/K$ and $n_2 \approx n/K$, respectively, T times. For the tth split, let $J_{\text{train}}^{(t)}$ and $J_{\text{test}}^{(t)}$ be the index sets for the training and testing groups, respectively. For each $r \in \{1, \ldots, b-a\}$, we apply the proposed procedure to the training dataset $\{X_k(T_k)\}_{k \in J_{\text{train}}^{(t)}}$, and denote the outcome as $\hat{\Sigma}^{(t)}(r)$. Then, we calculate the sample covariance matrix $\hat{\Sigma}_{\text{test}}^{(t)} \in \mathbb{R}^{p \times p}$ based on the samples from the testing group,

$$(\hat{\Sigma}_{\text{test}}^{(t)})_{[i,j]} =$$

$$\begin{cases} \sum_{\substack{k \in J_{\text{test}}^{(t)} \\ T_k \ni T(i), T(j)}} (X_{ik} - \bar{X}_i)(X_{jk} - \bar{X}_j) \Big/ \sum_{\substack{k \in J_{\text{test}}^{(t)} \\ T_k \ni T(i), T(j)}} 1, & \text{if } \sum_{\substack{k \in J_{\text{test}}^{(t)} \\ T_k \ni T(i), T(j)}} 1 \ge n_0, \\ \text{NA}, & \text{otherwise}, \end{cases}$$

where n_0 is the lower threshold when evaluating the testing sample covariance matrix. Then, we evaluate the prediction error as

$$E(r) = \sum_{t=1}^T \sum_{(\hat{\Sigma}_{\text{test}}^{(t)})_{[i,j]} \neq \text{NA}} \left((\hat{\Sigma}^{(t)}(r))_{[i,j]} - (\hat{\Sigma}_{\text{test}})_{[i,j]} \right)^2.$$

Here, to improve accuracy, we only evaluate the prediction errors on those (i, j)

pairs where $(\hat{\Sigma}_{\text{test}})_{[i,j]}$ is evaluated based on at least n_0 samples. Finally, we choose $\hat{r} = \arg\min_{1 \leq r \leq b-a} E(r)$, and apply the proposed procedure with $\hat{r}$ to obtain the final estimator $\hat{\Sigma}_0$. In our simulations, we use $K = 5$, $T = 10$, and $n_0 = 4$; other cross-validation methods are expected to yield similar results.

In practice, we propose using the cross-validation method, because this usually prevents under-selection. We observed a slight over-selection of r in our simulations, but this is not a problem in a covariance estimation because the components (eigenvalues) beyond r are all assumed to be very small. In Section 4, we examine the numerical performance of the proposed procedure based on cross-validation and the effect of the tuning parameters.

3. Theoretical Analysis

Before presenting the main theoretical results, we first introduce the following assumptions.

Assumption 1. *There is a positive integer r such that the eigenvalues of G satisfy $\lambda_1(G) \geq \cdots \geq \lambda_r(G) > \lambda_{r+1}(G) \geq \cdots \geq 0$. Let $G^{(r)}$ be the best rank-r approximation for G and $G^{-(r)} = G - G^{(r)}$. We also assume $\|G\|_{HS} < \infty$, $\|G^{(-r)}\|_{HS} \leq C/\sqrt{n^*}$, where n^* is the effective sample size, defined in Theorem 1.*

The rank r is allowed to increase slowly as n and p grow. The (approximate) reduced-rank covariance structure is explored by James et al. (2000) and Peng and Paul (2009) for sparse functional data, where only a few irregularly (randomly) spaced observations are available on each subject. They view the rank restriction as a form of regularization to avoid over-parametrization. The same reasoning applies to our scenario, because only a fraction of the trajectories are observed for each subject.

Assumption 2. *For any contiguous subdomain $\tilde{\mathcal{T}} \subseteq \mathcal{T}$, we define $G_{[\tilde{\mathcal{T}},\tilde{\mathcal{T}}]}^{(r)} = G^{(r)}($ $s,t)_{s\in\tilde{\mathcal{T}},t\in\tilde{\mathcal{T}}}$. There exists a constant $0 < \kappa < \delta$, such that $\gamma = \max_{\mathcal{L}(\tilde{\mathcal{T}})/\mathcal{L}(\mathcal{T})\geq\kappa}$ $\{\mathrm{tr}(G)\mathcal{L}(\tilde{\mathcal{T}})/\mathcal{L}(\mathcal{T})/\lambda_r(G_{[\tilde{\mathcal{T}},\tilde{\mathcal{T}}]}^{(r)})\}$ satisfies $\gamma = o((n^*)^{1/2})$.*

Intuitively speaking, this assumption imposes a lower bound of C/γ on the rth eigenvalues of $G_{[\tilde{\mathcal{T}},\tilde{\mathcal{T}}]}$. It essentially ensures that G restricted on different contiguous subdomain $[\tilde{\mathcal{T}},\tilde{\mathcal{T}}]$ is nonsingular, so that being able to estimate G using only segments of the functional observations is possible. As a counter-example, if G has two "spikes" in the sense that only $G_{[0:0.2,0:0.2]}$ and $G_{[0.8:1,0.8:1]}$ have significant amplitudes, while $G_{[0.2:0.8,0.2:0.8]}$ is zero, then the estimation of

the cross-covariance parts $G_{[0:0.2,0.8:1]}$ and $G_{[0.8:1,0:0.2]}$ is impossible when one can only observe functional segments of length no more than 0.6. In addition, γ is allowed to increase moderately as n and p grow. Note that $\gamma \geq r$, and in the scenarios in which γ/r is big, the method using complete observations only (step 2) is better than step 2'.

Assumption 3. *Assume X satisfies the moment condition $\sup_t \mathbb{E}|X(t)|^4 \leq C$.*

Assumption 4. *There exists $L > 0$, such that $|G(s,t) - G(s',t')| \leq L \max(|s - s'|, |t - t'|)$, for all $s, s', t, t' \in \mathcal{T}$.*

Because we use sample covariance approach and interpolate between observed grid points, the Lipschitz condition is almost necessary. It is easy to satisfy because we work with a finite domain $\mathcal{T}$, and it is weaker than the second differentiable conditions usually used in smoothing methods.

We can now state the main results of this study.

Theorem 1. *Suppose Assumptions 1–4 hold. We take $b = \beta p, a = \alpha p$ for some constants $0 < \alpha < \beta \leq \delta < 1$. Assume $\beta - \alpha \geq \kappa \geq 2r/p$ (κ and r were defined in the assumptions), $n \geq Cp$, and $p \geq C\gamma$. Then, the proposed procedure yields*

$$\mathbb{E}\|\hat{G} - G\|_{HS} = O\left(\sqrt{\frac{\gamma^2}{n^*}} + p^{-1}\right). \tag{3.1}$$

Here, $n^* = \min_l n_l^*$ and n_l^* is defined in (2.2). If we use complete samples to calculate $\hat{\Sigma}_l$ using (2.2) of Step 2; $n^* = \min_{i,j,l} n_{i,j,l}^*$ and $n_{i,j,l}^*$ is defined in (2.3) if we use both complete and incomplete samples to calculate $\hat{\Sigma}_l$ using (2.3) of Step 2'.

Remark 1. The first error term in (3.1) is due to estimating errors of the discretized covariance Σ_0. The second error term p^{-1} is from the linear interpolation of the discretized Σ_0.

Remark 2. Theorem 1 provides theoretical guarantees for the proposed procedure under general mixed longitudinal designs (conditional on T_k), where the effective sample size, n^*, is driven by the minimum number of samples that cover each sub-interval I_l. In a balanced design, where $T_k \subseteq T(\{w_k, \ldots, w_k + d - 1\})$, with w_k evenly chosen from $\{1, \ldots, p - d + 1\}$, for $k = 1, \ldots, n$, the boundary sub-intervals I_1 and $I_{l_{\max}}$ have less effective sample sizes than those of middle ones, which yield a higher estimation error for the boundary part of G. To overcome this bottleneck, we recommend a boundary-enriched design: beyond the balanced design, as mentioned above, we include $n_a = cn$ additional ones with

$T_k = T(\{1, \ldots, d\})$ or $T(\{p - d + 1, \ldots, p\})$ for a small constant $0 \leq c \leq 1$. Alternatively, one can apply an extended-domain design: for each $k = 1, \ldots, n$, $T_k = T(\{w_k, \ldots, w_k + d - 1\}) \cup T(\{1, \ldots, p\})$, with w_k uniformly chosen from $\{(2 - d), \ldots, p\}$. Under both the boundary-enriched and the extended-domain designs, the result of Theorem 1 yields (3.1).

Remark 3. (Proof sketch of Theorem 1). After introducing some notation, we develop error bounds for $\hat{A}_l$ (the outcome of Step 3), $\hat{O}_l$ (the outcome of Step 4), $\tilde{A}$ (the outcome of Step 5), $\hat{\Sigma}_0$, and the final estimator $\hat{G}$ (the outcome of Step 6). In particular, Step 4 of the proposed procedure involves solving the orthogonal Procrustes problem (or Wahba problem) (2.6). To derive the error bound of $\hat{O}_l$ from the error bound of $\hat{A}_l$, we introduce 1, which provides a theoretical guarantee for the solution of (2.6). In addition, Lemma 1 is stronger than previous results (cf., (Bishop and Byron, 2014, Lemma 16)), which may be of independent interest.

Lemma 1. *(Perturbation bound for Wahba problem). Suppose $A_1, A_2, A \in \mathbb{R}^{m \times r}$, $O_1, O_2 \in \mathbb{O}_r$, $\|A_1 - AO_1\|_F \leq a_1$, $\|A_2 - AO_2\|_F \leq a_2$, and $\sigma_r(A) \geq \lambda$. Suppose $\hat{O}$ is the solution to Wahba problem,*

$$\hat{O} = \operatorname*{argmin}_{O \in \mathbb{O}_r} \|A_2 O - A_1\|_F,$$

or equivalently, $\hat{O} = UV^\top$ *if* $A_2^\top A_1 = U\Sigma V^\top$ *is the SVD.*

Then, $\hat{O}$ satisfies

$$\left\| \hat{O} - O_2^\top O_1 \right\|_F \leq \frac{2(a_1 + a_2)}{\lambda}. \tag{3.2}$$

Proposition 1 provides a sharper convergence rate when Step 2 is applied (with only complete pieces) and the random scores are sub-Gaussian distributed.

Proposition 1. *Suppose $Z(t) = \mu(t) + \sum_{k \geq 1} \xi_k \phi_k(t)$ is the Karhunen–Loève decomposition, where $\{\phi_k(t)\}_{k \geq 1}$ is the fixed eigenfunction and $\{\xi_k\}_{\geq 1}$ are random scores. In addition to the assumptions of Theorem 1, we further assume the normalized leading r scores, $\tilde{\xi} = \{\xi_k / \lambda_k^{1/2}(G)\}_{k=1}^r$, are sub-Gaussian distributed, such that $\mathbb{E} \exp(t \tilde{\xi}^\top u) \leq \exp(C\|u\|_2^2)$, for any $u \in \mathbb{R}^r$, the tail part $Z^{(-r)}(t) = \sum_{k \geq r+1} \xi_k \phi_k(t)$ satisfies $\sup_t \mathbb{E}|Z^{(-r)}(t)|^4 \leq Cr/(n^*\gamma)$, and the noise satisfies $(\mathbb{E}|\epsilon|^4)^{1/2} \leq Cr/\gamma$. Then, the proposed procedure with Step 2 yields the following rate of convergence:*

$$\mathbb{E}\|\hat{G} - G\|_{HS} = O\left(\sqrt{\frac{r\gamma}{n^*}} + p^{-1}\right). \tag{3.3}$$

Here, $n^* = \min_l n_l^*$ and n_l^* is defined in (2.2).

Remark 4. We briefly compare the convergence rates of step 2 and step 2′. First, n^* when using the complete sample is no greater than that when using both complete and incomplete subjects. On the other hand, the factor γ^2 in (3.1) is greater than $r\gamma$ in the counterpart of (3.3). This is because $\hat{A}_l$ calculated using the standard sample covariance matrix, as in Step 2, possesses a sharper convergence rate than that calculated using the extended sample covariance matrix, as in Step 2′, as demonstrated by Lemma 7. Therefore, there is a trade-off between using Steps 2 or 2′. In general, we recommend using Step 2′ when most subjects have noncontiguous observations (missing values); otherwise Step 2 is preferred.

4. Numerical Experiments

Simulations: In this section, we investigate the numerical performance for the proposed procedure using a series of simulation studies. For each setting, we generate $X_{ij} = \sum_{k=1}^{K} \xi_{ik}\phi_k(t_{ij}) + \epsilon_{ij}$, where $i = 1, \ldots, n$, $j = 1, \ldots, p$, and t_{ij} are equally spaced p values on $[0, 1]$. We observe a contiguous ($\delta = d/p$) portion of the trajectory for each subject. All simulation results are based on 100 repetitions.

The first simulation setting is designed to assess the basic performance of the proposed method, and to explore the choices of the tuning parameters. In particular, we set $p = 30$, the true rank $K = 3$, and the eigenfunctions $\{\phi_k(t)\}$ as linear combinations of $M = 10$ cubic B-splines with equally spaced knots, as shown in Figure 3. The random scores $\{\xi_{ik}\}$ are i.i.d. normal with variances $(\lambda_1, \lambda_2, \lambda_3) = (4^2, 3^2, 2^2)$. The errors ϵ_{ij} are i.i.d. normal with variance one. We let the length of the observation band $d = 10$, so that each observation band covers one-third (d/p) of the total domain. We further let each contiguous subset of length d be observable by $n_{rep} = \{10, 20, 50\}$ subjects, which means the total sample size $n = n_{rep} \times (p - d + 1) = 210, 420, 1{,}050$. We apply the proposed method in Section 2, with the rank r selected by cross-validation, as described in Section 2, and report the relative estimation errors for different choices of tuning parameters b and a in Table 1. Here, the relative estimation error in all simulation settings is defined as $\|\hat{G} - G\|_{HS}/\|G\|_{HS}$. We can see that the estimation error decreases as the sample size increases, and the performance is not sensitive to the values of (b, a), as long as b is slightly smaller than the bandwidth d and a is small. The cross-validation of the proposed method tends to slightly over-select r, but over-selection does not affect the RMSE of the covariance estimation significantly in our simulation settings. In the following simulations, we always use the bandwidth $b = \lceil 0.7 \cdot d \rceil$ and the incremental parameter $a = \lceil 0.1 \cdot d \rceil$

The second simulation setting further explores the performance under dif-
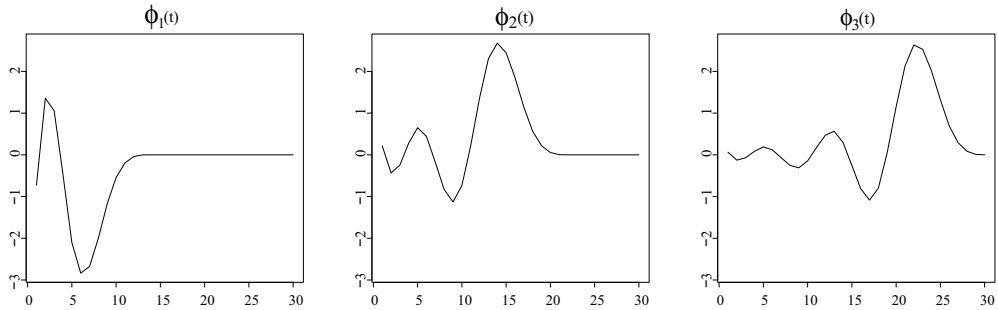
Figure 3. The first three eigenfunctions used in the simulations to generate the data.

Table 1. Results for simulation 1: the average relative error over 100 simulations are shown, with the standard error given in parentheses. Here, a and b are different choices of tuning parameters, and the results are stable.

	$n_{rep} = 10$	$n_{rep} = 20$	$n_{rep} = 50$
$b = 7, a = 1$	0.324 (0.17)	0.224 (0.14)	0.123 (0.06)
$b = 7, a = 2$	0.325 (0.16)	0.221 (0.13)	0.132 (0.1)
$b = 8, a = 1$	0.314 (0.17)	0.23 (0.14)	0.13 (0.09)
$b = 8, a = 2$	0.364 (0.17)	0.292 (0.19)	0.126 (0.08)
$b = 9, a = 1$	0.326 (0.16)	0.227 (0.15)	0.119 (0.07)
$b = 9, a = 2$	0.347 (0.15)	0.214 (0.11)	0.145 (0.11)

ferent settings. In particular, let $p = 30$ and the fraction of observable domain $\delta = \{1/5, 1/3, 1/2\}$. In addition to the previous setting with $K = 3$, we also consider $K = 10$, the score variances $(\lambda_1, \ldots, \lambda_{10}) = (4^2, 3^2, 2^2, 2^{-4}, \ldots, 2^{-10})$, ϕ_1, ϕ_2, ϕ_3 are the same as in the previous settings, and $\phi_k(t) = \sqrt{2} \sin(k\pi t)$, for $k = 4, \ldots, 10$ (all 10 functions are orthonormalized). Similarly to the first simulation setting, we implement the proposed procedure with r selected using cross-validation, and let $b = \lceil 0.7 \cdot d \rceil$ and $a = \lceil 0.1 \cdot d \rceil$; see Table 2. We can see that the proposed procedure still performs well when there are moderate deviations to the reduced-rank structure. The estimation error decreases as the observed partial trajectory covers a larger fraction of the entire trajectory. Note that the selected rank r for the cases $K = 10$ increases as the sample size increases, with an average value $r = 4.25$ for $d/p = 1/3$ and $n_{rep} = 50$.

The third simulation explores the performance when there are further missing values within the observable fraction of the domain. The setting is the same as that in the first simulation, except that the data have a 5%, 10%, or 15% missing rate. As in the previous two simulations, we implement the proposed procedure

Table 2. Results for simulation 2: the average relative error over 100 simulations are shown, with the standard error given in parentheses. Here, K is the total number of eigenfunctions used to generate the covariance, and δ denotes the fraction of domains observed.

	$K = 3$			$K = 10$		
	$n_{rep} = 10$	$n_{rep} = 20$	$n_{rep} = 50$	$n_{rep} = 10$	$n_{rep} = 20$	$n_{rep} = 50$
$\delta = 1/5$	0.43 (0.17)	0.397 (0.2)	0.294 (0.21)	0.461 (0.16)	0.403 (0.18)	0.304 (0.19)
$\delta = 1/3$	0.341 (0.17)	0.237 (0.16)	0.135 (0.1)	0.322 (0.16)	0.248 (0.14)	0.143 (0.06)
$\delta = 1/2$	0.243 (0.11)	0.17 (0.07)	0.113 (0.05)	0.248 (0.1)	0.165 (0.05)	0.114 (0.04)

Table 3. Results for simulation 3: the average relative error over 100 simulations are shown, with the standard error given in parentheses. Here, "missing" is the percentage of missing values within the observed domain.

missing	$n_{rep} = 10$	$n_{rep} = 20$	$n_{rep} = 50$	$n_{rep} = 100$
5%	0.36 (0.14)	0.24 (0.12)	0.16 (0.07)	0.12 (0.06)
10%	0.39 (0.16)	0.29 (0.13)	0.19 (0.08)	0.13 (0.05)
15%	0.42 (0.13)	0.32 (0.13)	0.22 (0.1)	0.16 (0.06)

with r selected by cross-validation, and let $b = \lceil 0.7 \cdot d \rceil$, and $a = \lceil 0.1 \cdot d \rceil$; see Table 3. We can see that the proposed procedure performs reasonably well when there is a moderate number of missing values, and the performance improves when the sample size becomes large.

The fourth simulation compares the performance of the proposed method with the matrix completion method proposed in Descary and Panaretos (2018). The data-generating procedure is the same as those of the previous simulations. The matrix completion method is implemented using the Matlab code downloaded from the authors' website. The method requires an input of rank r, and they propose using a scree-plot to manually determine the rank (looking for an "elbow" in the plot). Because this approach is not feasible in simulation settings, we use the true rank r for both methods. The results are reported in Table 4. The relative performance depends on the fraction of the domain observed. For $\delta = 1/2$, both methods work fine, and the matrix completion method is slightly better for a small sample size ($n_{rep} = 10$). For $\delta = 1/3$, both methods work fine, and the proposed method is slightly better for larger sample sizes. For $\delta = 1/5$, neither of the methods work well for a small sample size ($n_{rep} = 10$), although the error for the matrix completion method is not as large as that of the proposed method. When n increases, the error of the proposed method decreases to a reasonably small level; the matrix completion method is less satisfactory in

Table 4. Results for simulation 4: the average relative error over 200 simulations are shown, with the standard error given in parentheses. Here, "MatComp" is the matrix completion method proposed in Descary and Panaretos (2018), and δ denotes the fraction of domains observed.

		$n_{rep} = 10$	$n_{rep} = 20$	$n_{rep} = 50$	$n_{rep} = 100$
$\delta = 1/5$	proposed	0.37 (0.11)	0.27 (0.09)	0.17 (0.06)	0.12 (0.04)
	MatComp	0.32 (0.06)	0.27 (0.04)	0.23 (0.02)	0.22 (0.02)
$\delta = 1/3$	proposed	0.26 (0.10)	0.2 (0.06)	0.12 (0.04)	0.08 (0.03)
	MatComp	0.29 (0.08)	0.2 (0.05)	0.14 (0.03)	0.11 (0.02)
$\delta = 1/2$	proposed	0.26 (0.11)	0.18 (0.07)	0.12 (0.05)	0.08 (0.03)
	MatComp	0.24 (0.08)	0.17 (0.05)	0.11 (0.03)	0.08 (0.02)

this case.

Application to a study on the working memory of midlife women: We downloaded the data from the SWAN database (link: `http://www.icpsr.umich.edu/icpsrweb/ICPSR/series/00253`). The study examines the physical, biological, psychological, and social health of women during their middle years. In this section, we focus on the measurement of working memory, that is, the ability to manipulate information held in memory. In this study, working memory was assessed using digit span backwards (DSB) (Corporation (1997)): participants repeat strings of single-digit numbers backwards, with two trials at each string length, increasing from two to seven, stop after errors in both trials at a string length; score as the number of correct trials (range, 0–12). The testing was first administered at the fourth follow-up to 2,709 women, and then repeated in the sixth and subsequent visits. The data up to the tenth visit are publicly available. We exclude those subjects who dropped out before the tenth follow-up visit, leaving a sample size of $n = 2,016$. Following previous literature, we did not use the first measurement in order to alleviate the practice effect on the testing results (Karlamangla et al. (2017)). Instead, we focused on the age range $\mathcal{T} = [48, 62]$. Each subject has up to five years of consecutive data, and the average number of follow-ups is 3.3. We applied the proposed method described in Section 2 to estimate the covariance function, using a rank $r = 3$ selected by cross-validation, a band parameter $b = 4$, and an increment parameter $a = 1$. The estimated covariance surface is shown in the left panel of Figure 4. We can see that the variance is bigger at the middle part around age 55.

The nonparametric covariance estimation serves as a stepping stone for further functional data analysis. In the following, we perform a functional principal component analysis for the working memory trajectories, and examine how
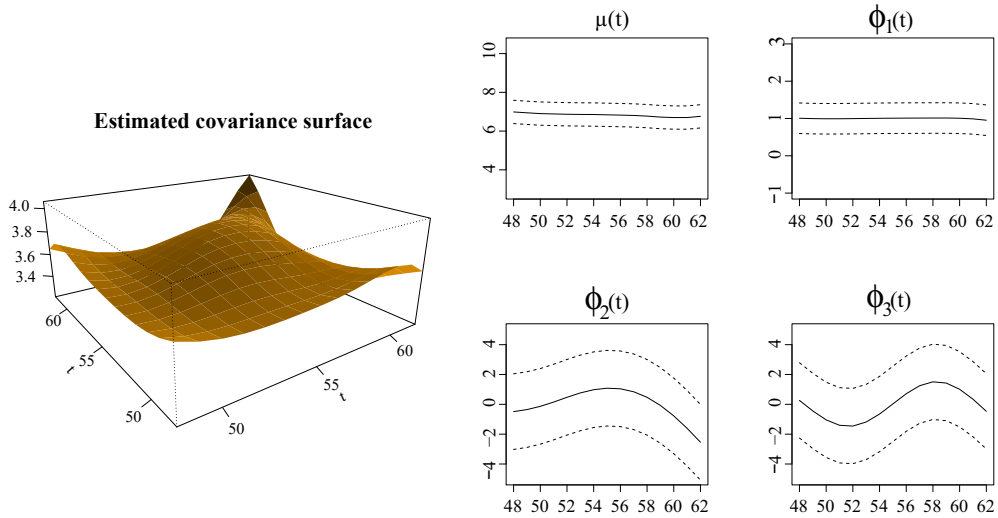
Figure 4. Left: The estimated covariance surface of the working memory data for women aged between 48 and 62. Right: The estimated mean function and estimated eigenfunctions corresponding to the largest three modes of variation, where the dashed lines are 95% bootstrap simultaneous confidence bands.

the shapes of the trajectories depend on education (less than high school, high school, some college/technical school, college graduate, postgraduate), controlling for race (Black, Chinese, Japanese, Caucasian/White, Non-Hispanic, Hispanic) and difficulty paying for basics (no hardship, somewhat hard, very hard). These are just for illustration of the functional data methods; a thorough analysis for this complex data set is beyond the scope of this study.

Given the estimated covariance, we conducted a functional principal component analysis based on the Karhunen–Loève expansion $Z(t) = \mu(t) + \sum_j \xi_j \phi_j(t)$. Here, $\{\phi_j(t), j \geq 1\}$ is an orthonormal basis that consists of eigenfunctions of G, and $\{\xi_j = \int (Z(t) - \mu(t))\phi_j(t)dt : j \geq 1\}$ are (random) scores. Intuitively, the first K terms expansion, $\mu(t) + \sum_{j=1}^{K} \xi_j \phi_j(t)$, forms a K-dimensional representation of $Z(t)$ with the smallest unexplained variance. The smoothed mean function and the first three estimated eigenfunctions $\{\phi_j(t), j = 1, 2, 3\}$ are visualized in the right panel of Figure 4. We also constructed 95% confidence bands for these quantities using the nonparametric bootstrap method, as outlined in Hall and Hosseini-Nasab (2006). The best linear prediction methods, as used in Yao et al. (2005), were applied to obtain estimates of ξ_j.

The mean function shows that the working memory function for a middle-age woman is, on average, decreases as she gets older. With longitudinal declines,

on average, there are individual differences in working memory aging and possible improvements in performance over multiple years. The first eigenfunction $\phi_1(t)$ is close to a horizontal line. Therefore, $\phi_1(t)$ can be interpreted as a size component: subjects with a positive score in the direction of this eigenfunction have better working memory function than that of an average woman for all ages between 48 and 62. The regression analysis show that this component is significantly and positively correlated with education level, which means that people with higher education tend to have higher working memory scores over the entire period. The other two covariates, financial status and race, are also statistically significant. The second eigenfunction $\phi_2(t)$ has a reversed U-shape with a maximum at around age = 55. This can be interpreted as a changing pattern before and after age 55, which possibly relates to the menopausal transition, resilience, and compensatory mechanisms (Fuh et al. (2006); Greendale et al. (2009); Hahn and Lachman (2015)). Subjects with a positive score in the direction of this eigenfunction have an increase in working memory before age 55, and a fast decline after age 55. The regression analysis show that education is a significant factor, with the postgraduate education group having a more prominent reversed U-shape pattern. The other two covariates are not statistically significant. The third component $\phi_3(t)$ crosses the zero line around age 55, representing a complementary effect to the second component.

This functional data analysis perspective differs from that of traditional linear mixed effect models, because the modes of variation for individual chronological aging trajectories are extracted nonparametrically from the data (FPC components), and one can examine how the shape of the trajectories interact with other covariates. In comparison, traditional linear mixed effect models (Karlamangla et al. (2017)) often control these covariates as fixed main effects.

5. Conclusions

We have focused on data observed on a regular equally spaced grid. The proposed sequential aggregating method can be readily extended to the setting in which the observational times are irregular and random. However, adjustments need to be made to step 2. In particular, the sample covariance estimate for Σ_l in step 2 is not applicable if the data are irregularly observed. In this case, one can first adopt a bivariate local linear smoothing method (Yao et al. (2005)) to estimate the covariance on the observable part (the diagonal banded area), say $\tilde{G}(s,t)$, for $|s-t| < \delta$. Then, for each piece l, take the corresponding sub-piece from $\tilde{G}(s,t)$, evaluate that on a predefined regular grid I_l, and use that as $\hat{\Sigma}_l$.

All other steps remain the same.

Acknowledgments

References

Berger, M. P. (1986). A comparison of efficiencies of longitudinal, mixed longitudinal, and cross-sectional designs. *Journal of Educational Statistics* **11**, 171–181.

Bickel, P. J. and Levina, E. (2008). Regularized estimation of large covariance matrices. *The Annals of Statistics* **36**, 199–227.

Bishop, W. E. and Byron, M. Y. (2014). Deterministic symmetric positive semidefinite matrix completion. In *Advances in Neural Information Processing Systems*, 2762–2770.

Cai, T. T., Ren, Z. and Zhou, H. H. (2016). Estimating structured high-dimensional covariance and precision matrices: Optimal rates and adaptive estimation. *Electronic Journal of Statistics* **10**, 1–59.

Cai, T. T. and Yuan, M. (2012). Adaptive covariance matrix estimation through block thresholding. *The Annals of Statistics* **40**, 2014–2042.

Cai, T. T. and Zhang, A. (2016). Minimax rate-optimal estimation of high-dimensional covariance matrices with incomplete data. *Journal of Multivariate Analysis* **150**, 55–74.

Cai, T. T., Zhang, C.-H. and Zhou, H. H. (2010). Optimal rates of convergence for covariance matrix estimation. *The Annals of Statistics* **38**, 2118–2144.

Corporation, P. (1997). WAIS-III and WMS-III: Technical manual. *San Antonio, TX: Psychological Corporation/Harcourt Brace.*

Delaigle, A. and Hall, P. (2013). Classification using censored functional data. *Journal of the American Statistical Association* **108**, 1269–1283.

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

Descary, M.-H. and Panaretos, V. M. (2018). Recovering covariance from functional fragments. *Biometrika* **106**, 145–160.

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

Fuh, J.-L., Wang, S.-J., Lee, S.-J., Lu, S.-R. and Juang, K.-D. (2006). A longitudinal study of cognition change during early menopausal transition in a rural community. *Maturitas* **53**, 447–453.

Greendale, G., Huang, M., Wight, R., Seeman, T., Luetters, C., Avis, N. et al. (2009). Effects of the menopause transition and hormone use on cognitive performance in midlife women. *Neurology* **72**, 1850–1857.

Hahn, E. A. and Lachman, M. E. (2015). Everyday experiences of memory problems and control: The adaptive role of selective optimization with compensation in the context of memory decline. *Aging, Neuropsychology and Cognition* **22**, 25–41.

Hall, P. and Hosseini-Nasab, M. (2006). On properties of functional principal components analysis. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **68**, 109–126.

Hedden, T. and Gabrieli, J. D. (2004). Insights into the ageing mind: A view from cognitive

neuroscience. *Nature Reviews Neuroscience* **5**, 87–96.

Helms, R. W. (1992). Intentionally incomplete longitudinal designs: I. methodology and comparison of some full span designs. *Statistics in Medicine* **11**, 1889–1913.

James, G. M., Hastie, T. J. and Sugar, C. A. (2000). Principal component models for sparse functional data. *Biometrika* **87**, 587–602.

Karlamangla, A. S., Lachman, M. E., Han, W., Huang, M. and Greendale, G. A. (2017). Evidence for cognitive aging in midlife women: Study of women's health across the nation. *PloS One* **12**, e0169008.

Kneip, A. and Liebl, D. (2017). On the optimal reconstruction of partially observed functional data. *arXiv preprint arXiv:1710.10099*.

Kolar, M. and Xing, E. P. (2012). Consistent covariance selection from data with missing values. In *Proceedings of the 29th International Conference on Machine Learning (ICML-12)*, 551–558.

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.

Loh, P.-L. and Wainwright, M. J. (2012). High-dimensional regression with noisy and missing data: Provable guarantees with nonconvexity. *The Annals of Statistics* **40**, 1637–1664.

Lounici, K. (2014). High-dimensional covariance matrix estimation with missing observations. *Bernoulli* **20**, 1029–1058.

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

Picard, R. R. and Cook, R. D. (1984). Cross-validation of regression models. *Journal of the American Statistical Association* **79**, 575–583.

Press, W. H., Teukolsky, S. A., Vetterling, W. T. and Flannery, B. P. (1992). *Numerical Recipes in C : The Art of Scientific Computing*. 2nd Edition. Cambridge University Press, New York.

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 (Methodological)* **53**, 233–243.

Rönnlund, M., Nyberg, L., Bäckman, L. and Nilsson, L.-G. (2005). Stability, growth, and decline in adult life span development of declarative memory: Cross-sectional and longitudinal data from a population-based study. *Psychology and Aging* **20**, 3–18.

Wahba, G. (1965). A least squares estimate of satellite attitude. *SIAM Review* **7**, 409–409.

Xiao, L., Li, Y. and Ruppert, D. (2013). Fast bivariate p-splines: the sandwich smoother. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **75**, 577–599.

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.

Anru R. Zhang

Department of Statistics, University of Wisconsin-Madison, Madison, WI 53706, USA.

Department of Biostatistics & Bioinformatics, Duke University, Durham, NC 27708, USA.

E-mail: anruzhang@stat.wisc.edu

Kehui Chen
Department of Statistics, University of Pittsburgh, Pittsburgh, PA 15260, USA.
E-mail: khchen@pitt.edu