Adaptive Functional Thresholding for Sparse Covariance Function Estimation in High Dimensions

Qin Fang[1], Shaojun Guo[2], and Xinghao Qiao[1]

[1]*Department of Statistics, London School of Economics and Political Science, U.K.*
[2]*Institute of Statistics and Big Data, Renmin University of China, P.R. China*

Abstract

Covariance function estimation is a fundamental task in multivariate functional data analysis and arises in many applications. In this paper, we consider estimating sparse covariance functions for high-dimensional functional data, where the number of random functions p is comparable to, or even larger than the sample size n. Aided by the Hilbert–Schmidt norm of functions, we introduce a new class of functional thresholding operators that combine functional versions of thresholding and shrinkage, and propose the adaptive functional thresholding estimator by incorporating the variance effects of individual entries of the sample covariance function into functional thresholding. To handle the practical scenario where curves are partially observed with errors, we also develop a nonparametric smoothing approach to obtain the smoothed adaptive functional thresholding estimator and its binned implementation to accelerate the computation. We investigate the theoretical properties of our proposals when p grows exponentially with n under both fully and partially observed functional scenarios. Finally, we demonstrate that the proposed adaptive functional thresholding estimators significantly outperform the competitors through extensive simulations and the functional connectivity analysis of two neuroimaging datasets.

Keywords: Binning; High-dimensional functional data; Functional connectivity; Functional sparsity; Local linear smoothing; Partially observed functional data.

1 Introduction

The covariance function estimation plays an important role in functional data analysis, while existing methods are restricted to data with a single or small number of random functions. Recent advances in technology have made multivariate or even high-dimensional functional datasets increasingly common in various applications: e.g., time-course gene expression data in genomics (Storey et al., 2005), air pollution data in environmental studies (Kong et al., 2016) and different types of brain imaging data in neuroscience (Li and Solea, 2018; Qiao et al., 2019). Under such scenarios, suppose we observe n independent samples $\mathbf{X}_i(\cdot) = \{X_{i1}(\cdot), \ldots, X_{ip}(\cdot)\}^{\mathrm{T}}$ for $i = 1, \ldots, n$ defined on a compact interval $\mathcal{U}$ with covariance function

$$\boldsymbol{\Sigma}(u, v) = \{\Sigma_{jk}(u, v)\}_{p\times p} = \mathrm{cov}\{\mathbf{X}_i(u), \mathbf{X}_i(v)\}, \quad u, v \in \mathcal{U}.$$

From a heuristic interpretation, we can simply treat each curve $X_{ij}(\cdot)$ as an infinitely long vector and replace the (j, k)th entry of $\boldsymbol{\Sigma}$ by $\Sigma_{jk}(\cdot, \cdot) = \mathrm{cov}\{X_{ij}(\cdot), X_{ik}(\cdot)\}$, the cross-covariance matrix of two infinitely long vectors. Then $\boldsymbol{\Sigma}$ can be understood as a block matrix with infinite sizes and its (j, k)th block being $\Sigma_{jk}(\cdot, \cdot)$. Besides being of interest in itself, an estimator of $\boldsymbol{\Sigma}$ is useful for many applications including, e.g., multivariate functional principal components analysis (FPCA) (Happ and Greven, 2018), multivariate functional linear regression (Chiou et al., 2016), functional factor model (Guo et al., 2022) and functional classification (Park et al., 2021). See Section 2.3 for details.

Our paper focuses on estimating $\boldsymbol{\Sigma}$ under high-dimensional scaling, where p can be comparable to, or even larger than n. In this setting, the sample covariance function

$$\widehat{\boldsymbol{\Sigma}}(u, v) = \{\widehat{\Sigma}_{jk}(u, v)\}_{p\times p} = \frac{1}{n-1}\sum_{i=1}^{n}\{\mathbf{X}_i(u) - \bar{\mathbf{X}}(u)\}\{\mathbf{X}_i(v) - \bar{\mathbf{X}}(v)\}^{\mathrm{T}}, \quad u, v \in \mathcal{U},$$

where $\bar{\mathbf{X}}(\cdot) = n^{-1}\sum_{i=1}^{n}\mathbf{X}_i(\cdot)$, performs poorly, and some lower-dimensional structural assumptions need to be imposed to estimate $\boldsymbol{\Sigma}$ consistently. In contrast to extensive work on estimating high-dimensional sparse covariance matrices (Bickel and Levina, 2008; Rothman

et al., 2009; Cai and Liu, 2011; Chen and Leng, 2016; Avella-Medina et al., 2018; Wang et al., 2021), research on sparse covariance function estimation in high dimensions remains largely unaddressed in the literature.

In this paper, we consider estimating sparse covariance functions via adaptive functional thresholding in the sense of shrinking some blocks $\widehat{\Sigma}_{jk}(\cdot,\cdot)$'s in an adaptive way. To achieve this, we introduce a new class of functional thresholding operators that combine functional versions of thresholding and shrinkage based on the Hilbert-Schmidt norm of functions, and develop an adaptive functional thresholding procedure on $\widehat{\mathbf{\Sigma}}(\cdot,\cdot)$ using entry-dependent functional thresholds that automatically adapt to the variability of blocks $\widehat{\Sigma}_{jk}(\cdot,\cdot)$'s. To provide theoretical guarantees of our method under high-dimensional scaling, it is essential to develop standardized concentration results taking into account the variability adjustment. Compared with adaptive thresholding for non-functional data (Cai and Liu, 2011), the intrinsic infinite-dimensionality of each $X_{ij}(\cdot)$ leads to a substantial rise in the complexity of sparsity modeling and theoretical analysis, as one needs to rely on some functional norm of standardized $\widehat{\Sigma}_{jk}$'s, e.g., the Hilbert–Schmidt norm, to enforce the functional sparsity in $\widehat{\mathbf{\Sigma}}$ and tackle more technical challenges for standardized processes within an abstract Hilbert space. To handle the practical scenario where functions are partially observed with errors, it is desirable to apply nonparametric smoothers in conjunction with adaptive functional thresholding. This poses a computationally intensive task especially when p is large, thus calling for the development of fast implementation strategy.

There are many applications of the proposed sparse covariance function estimation method in neuroimaging analysis, where brain signals are measured over time at a large number of regions of interest (ROIs) for individuals. Examples include the brain-computer interface classification (Lotte et al., 2018) and the brain functional connectivity identification (Rogers et al., 2007). Traditional neuroimaging analysis models brain signals for each subject as multivariate random variables, where each ROI is represented by a random variable, and hence the covariance/correlation matrices of interest are estimated by

treating the time-course data of each ROI as repeated observations. However, due to the non-stationary and dynamic features of signals (Chang and Glover, 2010), the strategy of averaging over time fails to characterize the time-varying structure leading to the loss of information in the original space. To overcome these drawbacks, we follow recent proposals to model signals directly as multivariate random functions with each ROI represented by a random function (Li and Solea, 2018; Qiao et al., 2019; Zapata et al., 2022; Lee et al., 2021). The identified functional sparsity pattern in our estimate of $\mathbf{\Sigma}$ can be used to recover the functional connectivity network among different ROIs, which is illustrated using examples of functional magnetic resonance imaging (fMRI) datasets in Section 6 and Section E.3 of the Supplementary Material.

Our paper makes useful contributions at multiple fronts. On the method side, it generalizes the thresholding/sparsity concept in multivariate statistics to the functional setting and offers a novel adaptive functional thresholding proposal to handle the heteroscedastic problem of the sparse covariance function estimation motivated from neuroimaging analysis and many statistical applications, e.g., those in Section 2.3 and Section C.2 of the Supplementary Material. It also provides an alternative way of identifying correlation-based functional connectivity with no need to specify the correlation function, the estimation of which poses challenges as the inverses of $\Sigma_{jj}(u, v)$'s are unbounded. In practice when functions are observed with errors at either a dense grid of points or a small subset of points, we also develop a unified local linear smoothing approach to obtain the smoothed adaptive functional thresholding estimator and its fast implementation via binning (Fan and Marron, 1994) to speed up the computation without sacrificing the estimation accuracy. On the theory side, we show that the proposed estimators enjoy the convergence and support recovery properties under both fully and partially observed functional scenarios when p grows exponentially fast relative to n. The proof relies on tools from empirical process theory due to the infinite-dimensional nature of functional data and some novel standardized concentration bounds in the Hilbert–Schmidt norm to deal with issues of high-dimensionality

and variance adjustment. Our theoretical results and adopted techniques are general, and can be applied to other settings in high-dimensional functional data analysis.

The remainder of this paper is organized as follows. Section 2 introduces a class of functional thresholding operators, based on which we propose the adaptive functional thresholding of the sample covariance function. We then discuss a couple of applications of the sparse covariance function estimation. Section 3 presents convergence and support recovery analysis of our proposed estimator. In Section 4, we develop a nonparametric smoothing approach and its binned implementation to deal with partially observed functional data, and then investigate its theoretical properties. In Sections 5 and 6, we demonstrate the uniform superiority of the adaptive functional thresholding estimators over the universal counterparts through an extensive set of simulation studies and the functional connectivity analysis of a neuroimaging dataset, respectively. All technical proofs are relegated to the Supplementary Material. We also provide the codes to reproduce the results for simulations and real data analysis in supplementary materials.

2 Methodology

2.1 Functional thresholding

We begin by introducing some notation. Let $L_2(\mathcal{U})$ denotes a Hilbert space of square integrable functions defined on $\mathcal{U}$ and $\mathbb{S} = L_2(\mathcal{U}) \otimes L_2(\mathcal{U})$, where $\otimes$ is the Kronecker product. For any $Q \in \mathbb{S}$, we denote its Hilbert–Schmidt norm by $\|Q\|_{\mathcal{S}} = \{\int\int Q(u,v)^2 \mathrm{dudv}\}^{1/2}$. With the aid of Hilbert–Schmidt norm, for any regularization parameter $\lambda \geqslant 0$, we first define a class of functional thresholding operators $s_\lambda : \mathbb{S} \to \mathbb{S}$ that satisfy the following conditions:

(i) $\|s_\lambda(Z)\|_{\mathcal{S}} \leqslant c\|Y\|_{\mathcal{S}}$ for all Z and $Y \in \mathbb{S}$ that satisfy $\|Z - Y\|_{\mathcal{S}} \leqslant \lambda$ and some $c > 0$;

(ii) $\|s_\lambda(Z)\|_{\mathcal{S}} = 0$ for $\|Z\|_{\mathcal{S}} \leqslant \lambda$;

(iii) $\|s_\lambda(Z) - Z\|_{\mathcal{S}} \leqslant \lambda$ for all $Z \in \mathbb{S}$.

Our proposed functional thresholding operators can be viewed as the functional generalization of thresholding operators (Cai and Liu, 2011). Instead of a simple pointwise extension of such thresholding operators under functional domain, we advocate a global thresholding rule based on the Hilbert–Schmidt norm of functions that encourages the functional sparsity, in the sense that $s_\lambda(Z)(u,v) = 0$, for all $u, v \in \mathcal{U}$, if $\|Z\|_{\mathcal{S}} \leqslant \lambda$ under condition (ii). Condition (iii) limits the amount of (global) functional shrinkage in the Hilbert–Schmidt norm to be no more than λ.

Conditions (i)–(iii) are satisfied by functional versions of some commonly adopted thresholding rules, which are introduced as solutions to the following penalized quadratic loss problem with various penalties:

$$s_\lambda(Z) = \arg\min_{\theta \in \mathbb{S}} \left\{ \frac{1}{2}\|\theta - Z\|_{\mathcal{S}}^2 + p_\lambda(\theta) \right\} \tag{1}$$

with $p_\lambda(\theta) = \tilde{p}_\lambda(\|\theta\|_{\mathcal{S}})$ being a penalty function of $\|\theta\|_{\mathcal{S}}$ to enforce the functional sparsity.

The soft functional thresholding rule results from solving (1) with an ℓ_1/ℓ_2 type of penalty, $p_\lambda(\theta) = \lambda\|\theta\|_{\mathcal{S}}$, and takes the form of $s_\lambda^{\mathrm{S}}(Z) = Z(1 - \lambda/\|Z\|_{\mathcal{S}})_+$, where $(x)_+ = \max(x, 0)$ for $x \in \mathbb{R}$. This rule can be viewed as a functional generalization of the group lasso solution under the multivariate setting (Yuan and Lin, 2006). To solve (1) with an ℓ_0/ℓ_2 type of penalty, $p_\lambda(\theta) = 2^{-1}\lambda^2 I(\|\theta\|_{\mathcal{S}} \neq 0)$, we obtain hard functional threhsolding rule as $ZI(\|Z\|_{\mathcal{S}} \geqslant \lambda)$, where $I(\cdot)$ is an indicator function. As a comparison, soft functional thresholding corresponds to the maximum amount of functional shrinkage allowed by condition (iii), whereas no shrinkage results from hard functional thresholding. Taking the compromise between soft and hard functional thresholding, we next propose functional versions of SCAD (Fan and Li, 2001) and adaptive lasso (Zou, 2006) thresholding rules. With a SCAD penalty (Fan and Li, 2001) operating on $\|\cdot\|_{\mathcal{S}}$ instead of $|\cdot|$ for the univariate scalar case, SCAD functional thresholding $s_\lambda^{\mathrm{SC}}(Z)$ is the same as soft functional thresholding if $\|Z\|_{\mathcal{S}} < 2\lambda$, and equals $Z\{(a-1) - a\lambda/\|Z\|_{\mathcal{S}}\}/(a-2)$ for $\|Z\|_{\mathcal{S}} \in [2\lambda, a\lambda]$ and Z if $\|Z\|_{\mathcal{S}} > a\lambda$, where $a > 2$. Analogously, adaptive lasso functional thresholding rule is $s_\lambda^{\mathrm{AL}}(Z) = Z(1 - \lambda^{\eta+1}/\|Z\|_{\mathcal{S}}^{\eta+1})_+$ with $\eta \geqslant 0$.

Our proposed functional generalizations of soft, SCAD and adaptive lasso thresholding rules can be checked to satisfy conditions (i)–(iii), see Section B.1 of the Supplementary Material for details. To present a unified theoretical analysis, we focus on functional thresholding operators $s_\lambda(Z)$ satisfying conditions (i)–(iii). Note that, although the hard functional thresholding does not satisfy condition (i), theoretical results in Section 3 still hold for hard functional thresholding estimators under similar conditions with corresponding proofs differing slightly. For examples of functional data with some local spikes, one may possibly suggest supremum-norm-based class of functional thresholding operators. See the detailed discussion in Section C.1 of the Supplementary Material.

2.2 Estimation

We now discuss our estimation procedure based on $s_\lambda(Z)$. Note the variance of $\widehat{\Sigma}_{jk}(u,v)$ depends on the distribution of $\{X_{ij}(u), X_{ik}(v)\}$ through higher-order moments, which is intrinsically a heteroscedastic problem. Hence it is more desirable to use entry-dependent functional thresholds that automatically takes into account the variability of blocks $\widehat{\Sigma}_{jk}(\cdot,\cdot)$'s to shrink some blocks to zero adaptively. To achieve this, define the variance factors $\Theta_{jk}(u,v) = \mathrm{var}\big([X_{ij}(u) - \mathbb{E}\{X_{ij}(u)\}][X_{ik}(v) - \mathbb{E}\{X_{ik}(v)\}]\big)$ with corresponding estimators

$$\widehat{\Theta}_{jk}(u,v) = \frac{1}{n}\sum_{i=1}^{n}\Big[\big\{X_{ij}(u) - \bar{X}_j(u)\big\}\big\{X_{ik}(v) - \bar{X}_k(v)\big\} - \widehat{\Sigma}_{jk}(u,v)\Big]^2, \quad j,k = 1,\ldots,p.$$

Then the adaptive functional thresholding estimator $\widehat{\mathbf{\Sigma}}_{\mathrm{A}} = \{\widehat{\Sigma}^{\mathrm{A}}_{jk}(\cdot,\cdot)\}_{p\times p}$ is defined by

$$\widehat{\Sigma}^{\mathrm{A}}_{jk} = \widehat{\Theta}^{1/2}_{jk} \times s_\lambda\left(\frac{\widehat{\Sigma}_{jk}}{\widehat{\Theta}^{1/2}_{jk}}\right), \tag{2}$$

which uses a single threshold level to functionally threshold standardized entries, $\widehat{\Sigma}_{jk}/\widehat{\Theta}^{1/2}_{jk}$ for all j,k, resulting in entry-dependent functional thresholds for $\widehat{\Sigma}_{jk}$'s. The selection of the optimal regularization parameter $\hat{\lambda}$ is discussed in Section 5.

An alternative approach to estimate $\mathbf{\Sigma}$ is the universal functional thresholding estimator

$$\widehat{\mathbf{\Sigma}}_{\mathrm{U}} = \{\widehat{\Sigma}^{\mathrm{U}}_{jk}(\cdot,\cdot)\}_{p\times p} \quad \text{with} \quad \widehat{\Sigma}^{\mathrm{U}}_{jk} = s_\lambda\big(\widehat{\Sigma}_{jk}\big),$$

where a universal threshold level is used for all entries. In a similar spirit to Rothman et al. (2009), the consistency of $\widehat{\mathbf{\Sigma}}_{\mathrm{U}}$ requires the assumption that marginal-covariance functions are uniformly bounded in nuclear norm, i.e., $\max_j \|\Sigma_{jj}\|_{\mathcal{N}} \leqslant M$, where $\|\Sigma_{jj}\|_{\mathcal{N}} = \int_{\mathcal{U}} \Sigma_{jj}(u,u)\mathrm{d}u$. However, intuitively, such universal method does not perform well when nuclear norms vary over a wide range, or even fails when the uniform boundedness assumption is violated. Section 5 provides some empirical evidence to support this intuition.

2.3 Applications

Many statistical problems involving multivariate functional data $\{\mathbf{X}_i(\cdot)\}_{i=1}^n$ require estimating the covariance function $\mathbf{\Sigma}$. Under a high-dimensional regime, the functional sparsity assumption can be imposed on $\mathbf{\Sigma}$ to facilitate its consistent sparse estimates. Here we outline three applications of our proposals for the sparse covariance function estimation.

Our first application is *multivariate FPCA* serving as a natural dimension reduction approach for $\mathbf{X}_i(\cdot)$. With the aid of Karhunen-Loéve expansion for multivariate functional data (Happ and Greven, 2018), $\mathbf{X}_i(\cdot)$ admits the following expansion

$$\mathbf{X}_i(\cdot) = \mathbb{E}\{\mathbf{X}_i(\cdot)\} + \sum_{l=1}^{\infty} \xi_{il}\boldsymbol{\phi}_l(\cdot), \quad i = 1, \ldots, n, \tag{3}$$

where the principal component scores $\xi_{il} = \sum_{j=1}^{p} \int [X_{ij}(u) - \mathbb{E}\{X_{ij}(u)\}]\phi_{lj}(u)\mathrm{d}u$ and eigenfunctions $\boldsymbol{\phi}_l(\cdot) = \{\phi_{l1}(\cdot), \ldots, \phi_{lp}(\cdot)\}^{\mathrm{T}}$ are attainable by the eigenanalysis of $\mathbf{\Sigma}$. Under a large p scenario, we can adopt the proposed functional thresholding technique to obtain the sparse estimation of $\mathbf{\Sigma}$, which guarantees the consistencies of estimated eigenvalues/eigenfunctions pairs. In Section E.1 of the Supplementary Material, we follow the proposal of a normalized version of multivariate FPCA in Happ and Greven (2018) and use a simulated example to illustrate the superior sample performance of our functional thresholding approaches.

Our second application, *multivariate functional linear regression* (Chiou et al., 2016), takes the form of

$$Y_i = \beta_0 + \int_{\mathcal{U}} \mathbf{X}_i(u)^{\mathrm{T}}\boldsymbol{\beta}(u)\mathrm{d}u + \epsilon_i, \quad i = 1, \ldots, n, \tag{4}$$

where $\boldsymbol{\beta}(\cdot) = \{\beta_1(\cdot), \ldots, \beta_p(\cdot)\}^{\mathrm{T}}$ is p-vector of functional coefficients to be estimated. The standard three-step procedure involves performing (normalized) multivariate FPCA on $\mathbf{X}_i(\cdot)$'s based on $\widehat{\boldsymbol{\Sigma}}$, then estimating the basis coefficients vector of $\boldsymbol{\beta}(\cdot)$ and finally recovering the estimated functional coefficients, where details are presented in Section E.1 of the Supplementary Material and Chiou et al. (2016). When p is large, we can implement our functional thresholding proposals to obtain consistent estimators of $\boldsymbol{\Sigma}$ and hence $\boldsymbol{\beta}$. In Section E.1 of the Supplementary Material, we demonstrate via a simulated example the superiority of our adaptive-functional-thresholding-based estimator over its competitors.

Our third application considers another dimension reduction framework via *functional factor model* (Guo et al., 2022) in the form of $\mathbf{X}_i(\cdot) = \mathbf{A}\mathbf{f}_i(\cdot) + \boldsymbol{\varepsilon}_i(\cdot)$, where the common components are driven by r functional factors $\mathbf{f}_i(\cdot) = \{f_{i1}(\cdot), \ldots, f_{ir}(\cdot)\}^{\mathrm{T}}$, the idiosyncratic components are $\boldsymbol{\varepsilon}_i(\cdot)$ and $\mathbf{A} \in \mathbb{R}^{p\times r}$ is the factor loading matrix. Denote the covariance functions of $\mathbf{X}_i(\cdot)$, $\mathbf{f}_i(\cdot)$ and $\boldsymbol{\varepsilon}_i(\cdot)$ by $\boldsymbol{\Sigma}_X$, $\boldsymbol{\Sigma}_f$ and $\boldsymbol{\Sigma}_\varepsilon$, respectively. Under the orthogonality of $\mathbf{A}$, $\int\int \boldsymbol{\Sigma}_X(u,v)\boldsymbol{\Sigma}_X(u,v)^{\mathrm{T}}\mathrm{d}u\mathrm{d}v$ can be decomposed as the sum of $\mathbf{A}\int\int \boldsymbol{\Sigma}_f(u,v)\boldsymbol{\Sigma}_f(u,v)^{\mathrm{T}}\mathrm{d}u\mathrm{d}v\mathbf{A}^{\mathrm{T}}$ and the remaining smaller order terms. Intuitively, with certain identifiable conditions, $\mathbf{A}$ can be recovered by carrying out an eigenanalysis of $\int\int \boldsymbol{\Sigma}_X(u,v)\boldsymbol{\Sigma}_X(u,v)^{\mathrm{T}}\mathrm{d}u\mathrm{d}v$. To provide a parsimonious model and enhance interpretability for near-zero loadings, we can impose subspace sparsity conditions (Vu and Lei, 2013) on $\mathbf{A}$ that results in a functional sparse $\boldsymbol{\Sigma}_X$ and hence our functional thresholding estimators become applicable. See an application of our functional thresholding technique to improve the estimation quality when fitting sparse functional factor model in Guo et al. (2022). See also Section C.2 of the Supplementary Material for other applications including functional graphical model estimation (Qiao et al., 2019) and multivariate functional classification.

3 Theoretical properties

We begin with some notation. For a random variable W, define $\|W\|_\psi = \inf\{c > 0 : \mathbb{E}[\psi(|W|/c)] \leqslant 1\}$, where $\psi : [0,\infty) \to [0,\infty)$ is a nondecreasing, nonzero convex function

with $\psi(0) = 0$ and the norm takes the value ∞ if no finite c exists for which $\mathbb{E}[\psi(|W|/c)] \leqslant 1$. Denote $\psi_k(x) = \exp(x^k) - 1$ for $k \geqslant 1$. Let the packing number $D(\epsilon, d)$ be the maximal number of points that can fit in the compact interval $\mathcal{U}$ while maintaining a distance greater than ϵ between all points with respect to the semimetric d. We refer to Chapter 8 of Kosorok (2008) for further explanations. For $\{X_{ij}(u) : u \in \mathcal{U}, i = 1, \ldots, n, j = 1, \ldots, p\}$, define the standardized processes by $Y_{ij}(u) = [X_{ij}(u) - \mathbb{E}\{X_{ij}(u)\}]/\sigma_j(u)^{1/2}$, where $\sigma_j(u) = \Sigma_{jj}(u, u)$.

To present the main theorems, we need the following regularity conditions.

Condition 1 *(i) For each i and j, $Y_{ij}(\cdot)$ is a separable stochastic process with the semimetric $d_j(u, v) = \|Y_{1j}(u) - Y_{1j}(v)\|_{\psi_2}$ for $u, v \in \mathcal{U}$; (ii) For some $u_0 \in \mathcal{U}$, $\max_{1 \leqslant j \leqslant p} \|Y_{1j}(u_0)\|_{\psi_2}$ is bounded.*

Condition 2 *The packing numbers $D(\epsilon, d_j)$'s satisfy $\max_{1 \leqslant j \leqslant p} D(\epsilon, d_j) \leqslant C\epsilon^{-r}$ for some constants $C, r > 0$ and $\epsilon \in (0, 1]$.*

Condition 3 *There exists some constant $\tau > 0$ such that $\min_{j,k} \inf_{u,v \in \mathcal{U}} var\{Y_{1j}(u)Y_{1k}(v)\} \geqslant \tau$.*

Condition 4 *The pair (n, p) satisfies $\log p / n^{1/4} \to 0$ as n and $p \to \infty$.*

Conditions 1 and 2 are standard to characterize the modulus of continuity of sub-Gaussian processes $Y_{ij}(\cdot)$'s, see Chapter 8 of Kosorok (2008). These conditions also imply that there exist some positive constants C_0 and η such that $\mathbb{E}[\exp(t\|Y_{1j}\|^2)] \leqslant C_0$ for all $|t| \leqslant \eta$ and j with $\|Y_{1j}\| = \{\int_{\mathcal{U}} Y_{1j}(u)^2 \mathrm{du}\}^{1/2}$, which plays a crucial role in our proof when applying concentration inequalities within Hilbert space. Condition 3 restricts the variances of $Y_{ij}(u)Y_{ik}(v)$'s to be uniformly bounded away from zero so that they can be well estimated. It also facilitates the development of some standardized concentration results. This condition precludes the case of a Brownian motion $X_{ij}(\cdot)$ starting at 0 for some j. However, replacing $X_{ij}(\cdot)$ with a contaminated process $X_{ij}(\cdot) + \xi_{ij}$, where ξ_{ij}'s are independent from a normal distribution with zero mean and a small variance and are

independent of $X_{ij}(\cdot)$'s, Condition 3 is fulfilled while the cross-covariance structure in $\mathbf{\Sigma}$ remains the same in the sense of $\mathrm{cov}\{X_{ij}(u)+\xi_{ij}, X_{ik}(v)\} = \mathrm{cov}\{X_{ij}(u), X_{ik}(v)\}$ for $k \neq j$ and $u, v \in \mathcal{U}$. Condition 4 allows the high-dimensional case, where p can diverge at some exponential rate as n increases.

We next establish the convergence rate of the adaptive functional thresholding estimator $\widehat{\mathbf{\Sigma}}_{\mathrm{A}}$ over a large class of "approximately sparse" covariance functions defined by

$$\mathcal{C}(q, s_0(p), \epsilon_0; \mathcal{U}) = \Big\{\mathbf{\Sigma} : \mathbf{\Sigma} \succeq 0, \max_{1\leqslant j\leqslant p} \sum_{k=1}^{p} \|\sigma_j\|_\infty^{(1-q)/2} \|\sigma_k\|_\infty^{(1-q)/2} \|\Sigma_{jk}\|_{\mathcal{S}}^q \leqslant s_0(p), \max_j \|\sigma_j^{-1}\|_\infty \|\sigma_j\|_\infty \leqslant \epsilon_0^{-1} < \infty\Big\}$$

for some $0 \leqslant q < 1$, where $\|\sigma_j\|_\infty = \sup_{u\in\mathcal{U}} \sigma_j(u)$ and $\mathbf{\Sigma} \succeq 0$ means that $\mathbf{\Sigma} = \{\Sigma_{jk}(\cdot,\cdot)\}_{p\times p}$ is positive semidefinite, i.e., $\sum_{j,k} \int\int \Sigma_{jk}(u,v) a_j(u) a_k(v) \mathrm{d}u\mathrm{d}v \geqslant 0$ for any $a_j(\cdot) \in L^2(\mathcal{U})$ and $j = 1, \ldots, p$. See Cai and Liu (2011) for a similar class of covariance matrices for non-functional data. Compared with the class

$$\mathcal{C}^*(q, s_0(p), M; \mathcal{U}) = \Big\{\mathbf{\Sigma} : \mathbf{\Sigma} \succeq 0, \max_j \|\sigma_j\|_{\mathcal{N}} \leqslant M, \max_j \sum_{k=1}^{p} \|\Sigma_{jk}\|_{\mathcal{S}}^q \leqslant s_0(p)\Big\},$$

over which the universal functional thresholding estimator $\widehat{\mathbf{\Sigma}}_{\mathrm{U}}$ can be shown to be consistent, the columns of a covariance function in $\mathcal{C}(q, s_0(p), \epsilon_0; \mathcal{U})$ are required to be within a weighted ℓ_q/ℓ_2 ball instead of a standard ℓ_q/ℓ_2 ball, where the weights are determined by $\|\sigma_j\|_\infty$'s. Unlike $\mathcal{C}^*(q, s_0(p), M; \mathcal{U})$, $\mathcal{C}(q, s_0(p), \epsilon_0; \mathcal{U})$ no longer requires the uniform boundedness assumption on $\|\sigma_j\|_{\mathcal{N}}$'s and allows $\max_j \|\sigma_j\|_{\mathcal{N}} \to \infty$. In the special case $q = 0$, $\mathcal{C}(q, s_0(p), \epsilon_0; \mathcal{U})$ corresponds to a class of truly sparse covariance functions. Notably, $s_0(p)$ can depend on p and be regarded implicitly as the restriction on functional sparsity.

Theorem 1 *Suppose that Conditions 1-4 hold. Then there exists some constant $\delta > 0$ such that, uniformly on $\mathcal{C}(q, s_0(p), \epsilon_0; \mathcal{U})$, if $\lambda = \delta(\log p/n)^{1/2}$,*

$$\|\widehat{\mathbf{\Sigma}}_{\mathrm{A}} - \mathbf{\Sigma}\|_1 = \max_{1\leqslant k\leqslant p} \sum_{j=1}^{p} \|\widehat{\Sigma}_{jk}^{\mathrm{A}} - \Sigma_{jk}\|_{\mathcal{S}} = O_P\left\{s_0(p)\Big(\frac{\log p}{n}\Big)^{\frac{1-q}{2}}\right\}. \tag{5}$$

Theorem 1 presents the convergence result in the functional version of matrix ℓ_1 norm. The rate in (5) is consistent to those of sparse covariance matrix estimates in Rothman et al. (2009); Cai and Liu (2011).

We finally turn to investigate the support recovery consistency of $\widehat{\mathbf{\Sigma}}_{\text{A}}$ over the parameter space of truly sparse covariance functions defined by

$$\mathcal{C}_0(s_0(p);\mathcal{U}) = \Big\{\mathbf{\Sigma} : \mathbf{\Sigma} \succeq 0, \max_{1\leqslant j\leqslant p}\sum_{k=1}^{p} I(\|\Sigma_{jk}\|_{\mathcal{S}} \neq 0) \leqslant s_0(p)\Big\},$$

which assumes that $(\Sigma_{jk})_{p\times p}$ has at most $s_0(p)$ non-zero functional entries on each row. The following theorem shows that, with the choice of $\lambda = \delta(\log p/n)^{1/2}$ for some constant $\delta > 0$, $\widehat{\mathbf{\Sigma}}_{\text{A}}$ exactly recovers the support of $\mathbf{\Sigma}$, $\text{supp}(\mathbf{\Sigma}) = \{(j,k) : \|\Sigma_{jk}\|_{\mathcal{S}} \neq 0\}$, with probability approaching one.

Theorem 2 *Suppose that Conditions 1-4 hold and* $\big\|\Sigma_{jk}/\Theta_{jk}^{1/2}\big\|_{\mathcal{S}} > (2\delta+\gamma)(\log p/n)^{1/2}$ *for all* $(j,k) \in supp(\mathbf{\Sigma})$ *and some* $\gamma > 0$, *where* δ *is stated in Theorem 1. Then we have that*

$$\inf_{\Sigma\in\mathcal{C}_0} P\big\{supp(\widehat{\mathbf{\Sigma}}_{\text{A}}) = supp(\mathbf{\Sigma})\big\} \to 1 \text{ as } n\to\infty.$$

Theorem 2 ensures that $\widehat{\mathbf{\Sigma}}_{\text{A}}$ achieves the exact recovery of functional sparsity structure in $\mathbf{\Sigma}$, i.e., the graph support in functional connectivity analysis, with probability tending to 1. This theorem holds under the condition that the Hilbert-Schmidt norms of non-zero standardized functional entries exceed a certain threshold, which ensures that non-zero components are correctly retained. See an analogous minimum signal strength condition for sparse covariance matrices in Cai and Liu (2011).

4 Partially observed functional data

In this section we consider a practical scenario where each $X_{ij}(\cdot)$ is partially observed, with errors, at random measurement locations $U_{ij1},\ldots,U_{ijL_{ij}} \in \mathcal{U}$. Let Z_{ijl} be the observed value of $X_{ij}(U_{ijl})$. Then

$$Z_{ijl} = X_{ij}(U_{ijl}) + \varepsilon_{ijl}, \quad l = 1,\ldots,L_{ij}, \tag{6}$$

where ε_{ijl}'s are i.i.d. errors with $\mathbb{E}(\varepsilon_{ijl}) = 0$ and $\text{var}(\varepsilon_{ijl}) = \sigma^2$, independent of $X_{ij}(\cdot)$. For dense measurement designs all L_{ij}'s are larger than some order of n, while for sparse designs all L_{ij}'s are bounded (Zhang and Wang, 2016; Qiao et al., 2020).

4.1 Estimation procedure

Based on the observed data, $\{(U_{ijl}, Z_{ijl})\}_{1\leqslant i\leqslant n, 1\leqslant j\leqslant p, 1\leqslant l\leqslant L_{ij}}$, we next present a unified estimation procedure that handles both densely and sparsely sampled functional data.

We first develop a nonparametric smoothing approach to estimate $\Sigma_{jk}(u, v)$'s. Without loss of generality, we assume that $\mathbf{X}_i(\cdot)$ has been centered to have mean zero. Denote $K_h(\cdot) = h^{-1}K(\cdot/h)$ for a univariate kernel function K with a bandwidth $h > 0$. A local linear surface smoother (LLS) is employed to estimate cross-covariance functions $\Sigma_{jk}(u, v)$ $(j \neq k)$ by minimizing

$$\sum_{i=1}^{n}\sum_{l=1}^{L_{ij}}\sum_{m=1}^{L_{ik}}\Big\{Z_{ijl}Z_{ikm} - \alpha_0 - \alpha_1(U_{ijl} - u) - \alpha_2(U_{ikm} - v)\Big\}^2 K_{h_C}(U_{ijl} - u)K_{h_C}(U_{ikm} - v), \quad (7)$$

with respect to $(\alpha_0, \alpha_1, \alpha_2)$. Let the minimizer of (7) be $(\hat{\alpha}_0, \hat{\alpha}_1, \hat{\alpha}_2)$ and the resulting estimator is $\widetilde{\Sigma}_{jk}(u, v) = \hat{\alpha}_0$. To estimate marginal-covariance functions $\Sigma_{jj}(u, v)$'s, we observe that $\text{cov}(Z_{ijl}, Z_{ijm}) = \Sigma_{jj}(U_{ijl}, U_{ijm}) + \sigma^2 I(l = m)$, and hence apply a LLS to the off-diagonals of the raw covariances $(Z_{ijl}Z_{ijm})_{1\leqslant l\leqslant m\leqslant L_{ij}}$. We consider minimizing

$$\sum_{i=1}^{n}\sum_{1\leqslant l\neq m\leqslant L_{ij}}\Big\{Z_{ijl}Z_{ijm} - \beta_0 - \beta_1(U_{ijl} - u) - \beta_2(U_{ikm} - v)\Big\}^2 K_{h_M}(U_{ijl} - u)K_{h_M}(U_{ikm} - v)$$

with respect to $(\beta_0, \beta_1, \beta_2)$, thus obtaining the estimate $\widetilde{\Sigma}_{jj}(u, v) = \hat{\beta}_0$. Note that we drop subscripts j, k of $h_{C,jk}$ and j of $h_{M,j}$ to simplify our notation in this section. However, we select different bandwidths $h_{C,jk}$ and $h_{M,j}$ across $j, k = 1, \ldots, p$ in our empirical studies.

To construct the corresponding adaptive functional thresholding estimator, a standard approach is to incorporate the variance effect of each $\widetilde{\Sigma}_{jk}(u, v)$ into functional thresholding. However, the estimation of $\text{var}\{\widetilde{\Sigma}_{jk}(u, v)\}$'s involves estimating multiple complicated fourth moment terms (Zhang and Wang, 2016), which results in high computational burden especially for large p. Since our focus is on characterizing the main variability of $\widetilde{\Sigma}_{jk}(u, v)$

rather than estimating its variance precisely, we next develop a computationally simple yet effective approach to estimate the main terms in the asymptotic variance of $\widetilde{\Sigma}_{jk}(u,v)$. For $a,b=0,1,2$, let

$$T_{ab,ijk}(u,v)=\sum_{l=1}^{L_{ij}}\sum_{m=1}^{L_{ik}} g_{ab}\{h_C,(u,v),(U_{ijl},U_{ikm})\}Z_{ijl}Z_{ikm}, \tag{8}$$

where $g_{ab}\big\{h,(u,v),(U_{ijl},U_{ikm})\big\}=K_h(U_{ijl}-u)K_h(U_{ikm}-v)(U_{ijl}-u)^a(U_{ikm}-v)^b$. According to Section D.1 of the Supplementary Material, minimizing (7) yields the resulting estimator

$$\widetilde{\Sigma}_{jk}=\sum_{i=1}^{n}\big(W_{1,jk}T_{00,ijk}+W_{2,jk}T_{10,ijk}+W_{3,jk}T_{01,ijk}\big), \tag{9}$$

where $W_{1,jk},W_{2,jk},W_{3,jk}$ can be represented via (S.12) in terms of

$$S_{ab,jk}(u,v)=\sum_{i=1}^{n}\sum_{l=1}^{L_{ij}}\sum_{m=1}^{L_{ik}} g_{ab}\big\{h_C,(u,v),(U_{ijl},U_{ikm})\big\},\quad a,b=0,1,2. \tag{10}$$

It is notable that the estimator $\widetilde{\Sigma}_{jk}$ in (9) is expressed as the sum of n independent terms. Ignoring the cross-covariances among observations within the subject that are dominated by the corresponding variances, we propose a surrogate estimator for the asymptotic variance of $\widetilde{\Sigma}_{jk}$ by

$$\widetilde{\Psi}_{jk}=I_{jk}\sum_{i=1}^{n}\big(W_{1,jk}V_{00,ijk}+W_{2,jk}V_{10,ijk}+W_{3,jk}V_{01,ijk}\big)^2, \tag{11}$$

where

$$I_{jk}=\Big(\sum_{i=1}^{n}L_{ij}L_{ik}\Big)^2\Big\{\sum_{i=1}^{n}\big(L_{ij}L_{ik}h_C^{-2}+L_{ij}^2L_{ik}h_C^{-1}+L_{ij}L_{ik}^2h_C^{-1}+L_{ij}^2L_{ik}^2\big)\Big\}^{-1}, \tag{12}$$

$$V_{ab,ijk}(u,v)=\sum_{l=1}^{L_{ij}}\sum_{m=1}^{L_{ik}} g_{ab}\big\{h_C,(u,v),(U_{ijl},U_{ikm})\big\}\big\{Z_{ijl}Z_{ikm}-\widetilde{\Sigma}_{jk}(u,v)\big\}. \tag{13}$$

The rationale of multiplying the rate I_{jk} in (11) is to ensure that $\widetilde{\Psi}_{jk}(u,v)$ converges to some finite function when $n\to\infty$ and $h_C\to 0$ as justified in Section D.4 of the Supplementary Material. In particular, the rate I_{jk} can be simplified to $\sum_{i=1}^{n}L_{ij}L_{ik}h_C^2$ for the sparse or moderately dense case and to $(\sum_{i=1}^{n}L_{ij}L_{ik})^2(\sum_{i=1}^{n}L_{ij}^2L_{ik}^2)^{-1}$ for the very dense case. Note that I_{jk} is imposed in (11) mainly for the theoretical purpose and hence will not place a practical constraint on our method.

In a similar procedure as above, the estimated variance factor $\widetilde{\Psi}_{jj}$ of $\widetilde{\Sigma}_{jj}$ for each j can be obtained by operating on $\{Z_{ijl}Z_{ijm}\}_{1\leqslant i\leqslant n, 1\leqslant l\neq m\leqslant L_{ij}}$ instead of $\{Z_{ijl}Z_{ikm}\}_{1\leqslant i\leqslant n, 1\leqslant l\leqslant L_{ij}, 1\leqslant m\leqslant L_{ik}}$ for $j \neq k$. Substituting $\widehat{\Theta}_{jk}$ in (2) by $\widetilde{\Psi}_{jk}$, we obtain the smoothed adaptive functional thresholding estimator

$$\widetilde{\boldsymbol{\Sigma}}_{\mathrm{A}} = (\widetilde{\Sigma}^{\mathrm{A}}_{jk})_{p\times p} \quad \text{with} \quad \widetilde{\Sigma}^{\mathrm{A}}_{jk} = \widetilde{\Psi}^{1/2}_{jk} \times s_\lambda\left(\frac{\widetilde{\Sigma}_{jk}}{\widetilde{\Psi}^{1/2}_{jk}}\right). \tag{14}$$

For comparison, we also define the smoothed universal functional thresholding estimator as $\widetilde{\boldsymbol{\Sigma}}_{\mathrm{U}} = (\widetilde{\Sigma}^{\mathrm{U}}_{jk})_{p\times p}$ with $\widetilde{\Sigma}^{\mathrm{U}}_{jk} = s_\lambda\big(\widetilde{\Sigma}_{jk}\big)$.

A natural alternative to the proposed LLS-based smoothing procedure considers pre-smoothing each individual data. For densely sampled functional data, the observations $Z_{ij1}, \ldots, Z_{ijL_{ij}}$ for each i and j can be pre-smoothed through the local linear smoother to eliminate the contaminated noise, thus producing reconstructed random curves $\widehat{X}_{ij}(\cdot)$'s before subsequent analysis (Zhang and Chen, 2007). See detailed implementation of pre-smoothing in Section D.2 of the Supplementary Material. For sparsely sampled functional data, such pre-smoothing step is not viable, while our smoothing proposal builds strength across functions by incorporating information from all the observations, and hence is still applicable. See also Section 5.3 for the numerical comparison between pre-smoothing and our smoothing approach under different measurement designs.

4.2 Theoretical properties

In this section, we investigate the theoretical properties of $\widetilde{\boldsymbol{\Sigma}}_{\mathrm{A}}$ for partially observed functional data. We begin by introducing some notation. For two positive sequences $\{a_n\}$ and $\{b_n\}$, we write $a_n \lesssim b_n$ if there exits a positive constant c_0 such that $a_n/b_n \leqslant c_0$. We write $a_n \asymp b_n$ if and only if $a_n \lesssim b_n$ and $b_n \lesssim a_n$ hold simultaneously. Before presenting the theory, we impose the following regularity conditions.

Condition 5 *(i) Let* $\{U_{ijl} : i = 1, \ldots, n, j \in 1, \ldots, p, l = 1, \ldots, L_{ij}\}$ *be i.i.d. copies of a random variable* U *with density* $f_U(\cdot)$ *defined on the compact set* $\mathcal{U}$*, with the* L_{ij}*'s fixed.*

There exist some constants m_f *and* M_f *such that* $0 < m_f \leqslant \inf_{\mathcal{U}} f_U(u) \leqslant \sup_{\mathcal{U}} f_U(u) \leqslant M_f < \infty$; *(ii)* X_{ij}, ε_{ijl} *and* U_{ijl} *are independent for each* i, j, l.

Condition 6 *(i) Under the sparse measurement design,* $L_{ij} \leqslant L_0 < \infty$ *for all* i, j *and, under the dense design,* $L_{ij} = L \to \infty$ *as* $n \to \infty$ *with* U_{ijl}*'s independent of* i; *(ii) The bandwidth parameters* $h_C \asymp h_M \asymp h \to 0$ *as* $n \to \infty$.

Condition 5 is standard in functional data analysis literature (Zhang and Wang, 2016). Condition 6 (i) treats the number of measurement locations L_{ij} as bounded and diverging under sparse and dense measurement designs, respectively. To simplify notation, we assume that $L_{ij} = L$ for the dense case and h_C is of the same order as h_M in Condition 6 (ii).

Condition 7 *There exists some constant* $\gamma_1 \in (0, 1/2]$ *such that*

$$\max_{1 \leqslant j,k \leqslant p} \left\| \widetilde{\Sigma}_{jk} - \Sigma_{jk} \right\|_{\mathcal{S}} \lesssim \sqrt{\frac{\log p}{n^{2\gamma_1}}} + h^2 \quad \textit{with probability approaching one.} \tag{15}$$

Condition 8 *There exist some positive constants* c_1, $\gamma_2 \in (0, 1/2]$ *and some deterministic functions* $\Psi_{jk}(u, v)$*'s with* $\min_{j,k} \inf_{u,v \in \mathcal{U}} \Psi_{jk}(u, v) \geqslant c_1$ *such that*

$$\max_{1 \leqslant j,k \leqslant p} \sup_{u,v \in \mathcal{U}} \left| \widetilde{\Psi}_{jk}(u, v) - \Psi_{jk}(u, v) \right| \lesssim \sqrt{\frac{\log p}{n^{2\gamma_2}}} + h^2 \quad \textit{with probability approaching one.} \tag{16}$$

Condition 9 *The pair* (n, p) *satisfies* $\log p / n^{\min(\gamma_1, \gamma_2)} \to 0$ *and* $\log p \geqslant c_2 n^{2\gamma_1} h^4$ *for some positive constant* c_2 *as* n *and* $p \to \infty$.

We follow Qiao et al. (2020) to impose Condition 7, in which the parameter γ_1 depends on h and possibly L under the dense design. This condition is satisfied if there exist some positive constants c_3, c_4, c_5 such that for each $j, k = 1, \ldots, p$ and $t \in (0, 1]$,

$$P\big(\|\widetilde{\Sigma}_{jk} - \Sigma_{jk}\|_{\mathcal{S}} \geqslant t + c_5 h^2\big) \leqslant c_4 \exp(-c_3 n^{2\gamma_1} t^2). \tag{17}$$

The presence of h^2 comes from the standard results for bias terms under the boundedness condition for the second-order partial derivatives of $\Sigma_{jk}(u, v)$ over $\mathcal{U}^2$ (Yao et al., 2005; Zhang and Wang, 2016). This concentration result is fulfilled under different measurement

schedules ranging from sparse to dense designs as γ_1 increases. For sparsely sampled functional data, Lemma 4 of Qiao et al. (2020) established L_2 concentration inequality for $\widetilde{\Sigma}_{jk}$ for $j = k$, which not only results in the same L_2 rate as that in the sparse case (Zhang and Wang, 2016) but also ensures (17) with the choice of $\gamma_1 = 1/2 - a$ and $h \asymp n^{-a}$ for some positive constant $a < 1/2$. Following the same proof procedure, the same concentration inequality also applies for $j \neq k$ and hence Condition 7 is satisfied. This condition is also satisfied by densely sampled functional data, since it follows from Lemma 5 of Qiao et al. (2020) that (17) holds for $j = k$ and, with more efforts, also for $j \neq k$ by choosing $\gamma_1 = \min(1/2, 1/3 + b/6 - \epsilon'/2 - 2a/3)$ for some small constant $\epsilon' > 0$ when $h \asymp n^{-a}$ and $L \asymp n^b$ for some constants $a, b > 0$. As L grows sufficiently large, $\gamma_1 = 1/2$, thus leading to the same rate as that in the ultra-dense case (Zhang and Wang, 2016). Condition 8 gives the uniform convergence rate for $\widetilde{\Psi}_{jk}(u, v)$ in the same form as (15) but with different parameter γ_2. A denser measurement design corresponds to a larger value of γ_2 and a faster rate in (16). See the heuristic verification of Condition 8 in Section D.4 of the Supplementary Material. Condition 9 indicates that p can grow exponentially fast relative to n.

We next present the convergence rate of the smoothed adaptive functional thresholding estimator $\widetilde{\mathbf{\Sigma}}_{\mathrm{A}}$ over a class of "approximate sparse" covariance functions defined by

$$\begin{aligned}\widetilde{\mathcal{C}}(q, \tilde{s}_0(p), \epsilon_0; \mathcal{U}) \;=\; \Big\{\mathbf{\Sigma} : \mathbf{\Sigma} \succeq 0, \max_{1\leqslant j\leqslant p} \sum_{k=1}^{p} \|\Psi_{jk}\|_\infty^{(1-q)/2} \|\Sigma_{jk}\|_{\mathcal{S}}^q \leqslant \tilde{s}_0(p), \\ \max_{j,k} \|\Psi_{jk}^{-1}\|_\infty \|\Psi_{jk}\|_\infty \leqslant \epsilon_0^{-1} < \infty\Big\},\end{aligned}$$

for some $0 \leqslant q < 1$.

Theorem 3 *Suppose that Conditions 5–9 hold. Then there exists some constants $\tilde{\delta} > 0$ such that, uniformly on $\widetilde{\mathcal{C}}(q, \tilde{s}_0(p), \epsilon_0; \mathcal{U})$, if $\lambda = \tilde{\delta}(\log p / n^{2\gamma_1})^{1/2}$,*

$$\|\widetilde{\mathbf{\Sigma}}_{\mathrm{A}} - \mathbf{\Sigma}\|_1 = \max_{1\leqslant k\leqslant p} \sum_{j=1}^{p} \|\widetilde{\Sigma}_{jk}^{\mathrm{A}} - \Sigma_{jk}\|_{\mathcal{S}} = O_P\left\{\tilde{s}_0(p)\Big(\frac{\log p}{n^{2\gamma_1}}\Big)^{\frac{1-q}{2}}\right\}. \tag{18}$$

The convergence rate of $\widetilde{\mathbf{\Sigma}}_{\mathrm{A}}$ in (18) is governed by internal parameters (γ_1, q) and other dimensionality parameters. Larger values of γ_1 correspond to a more frequent measurement

schedule with larger L and result in a faster rate. The convergence result implicitly reveals interesting phase transition phenomena depending on the relative order of L to n. As L grows fast enough, $\gamma_1 = 1/2$ and the rate is consistent to that for fully observed functional data in (5), presenting that the theory for very densely sampled functional data falls in the parametric paradigm. As L grows moderately fast, $\gamma_1 < 1/2$ and the rate is faster than that for sparsely sampled functional data but slower than the parametric rate.

We finally present Theorem 4 that guarantees the support recovery consistency of $\widetilde{\mathbf{\Sigma}}_{\text{A}}$.

Theorem 4 *Suppose that Conditions 5–9 hold and* $\left\|\Sigma_{jk}/\Psi_{jk}^{1/2}\right\|_{\mathcal{S}} > (2\tilde{\delta} + \tilde{\gamma})(\log p/n^{2\gamma_1})^{1/2}$ *for all* $(j,k) \in supp(\mathbf{\Sigma})$ *and some* $\tilde{\gamma} > 0$, *where* $\tilde{\delta}$ *is stated in Theorem 3, then*

$$\inf_{\Sigma \in \mathcal{C}_0} P\big\{supp(\widetilde{\mathbf{\Sigma}}_A) = supp(\mathbf{\Sigma})\big\} \to 1 \text{ as } n \to \infty.$$

4.3 Fast computation

Consider a common situation in practice, where, for each $i = 1, \ldots, n$, we observe the noisy versions of $X_{i1}(\cdot), \ldots, X_{ip}(\cdot)$ at the same set of points, $U_{i1}, \ldots, U_{iL_i} \in \mathcal{U}$, across $j = 1, \ldots, p$. Then the original model in (6) is simplified to

$$Z_{ijl} = X_{ij}(U_{il}) + \varepsilon_{ijl}, \quad l = 1, \ldots, L_i, \tag{19}$$

under which the proposed estimation procedure in Section 4.1 can still be applied. Suppose that the estimated covariance function is evaluated at a grid of $R \times R$ locations, $\{(u_{r_1}, u_{r_2}) \in \mathcal{U}^2 : r_1, r_2 = 1, \ldots, R\}$. To serve the estimation of $p(p+1)/2$ marginal- and cross-covariance functions and the corresponding variance factors, LLSs under the simplified model in (19) reduce the number of kernel evaluations from $O(\sum_{i=1}^n \sum_{j=1}^p L_{ij} R)$ to $O(\sum_{i=1}^n L_i R)$, which substantially accelerate the computation under a high-dimensional regime.

Apparently, such nonparametric smoothing approach is conceptually simple but suffers from high computational cost in kernel evaluations. To further reduce the computational burden, we consider fast implementations of LLSs by adopting a simple approximation technique, known as linear binning (Fan and Marron, 1994), to the covariance function

estimation. The key idea of the binning method is to greatly reduce the number of kernel evaluations through the fact that many of these evaluations are nearly the same. We start by dividing $\mathcal{U}$ into an equally-spaced grid of R points, $u_1 < \cdots < u_R \in \mathcal{U}$, with binwidth $\Delta = u_2 - u_1$. Denote by $w_r(U_{il}) = \max(1 - \Delta^{-1}|U_{il} - u_r|, 0)$ the linear weight that U_{il} assigns to the grid point u_r for $r = 1, \ldots, R$. For the i-th subject, we define its "binned weighted counts" and "binned weighted averages" as

$$\varpi_{r,i} = \sum_{l=1}^{L_i} w_r(U_{il}) \quad \text{and} \quad \mathcal{D}_{r,ij} = \sum_{l=1}^{L_i} w_r(U_{il}) Z_{ijl},$$

respectively. The binned implementation of smoothed adaptive functional thresholding can then be done using this modified dataset $\{(\varpi_{r,i}, \mathcal{D}_{r,ij})\}_{1 \leqslant i \leqslant n, 1 \leqslant j \leqslant p, 1 \leqslant r \leqslant R}$ and related kernel functions $g_{ab}\{h, (u, v), (u_{r_1}, u_{r_2})\}$ for $r_1, r_2 = 1, \ldots, R$. It is notable that, with the help of such binned implementation, the number of kernel evaluations required in the covariance function estimation is further reduced from $O(\sum_{i=1}^n L_i R)$ to $O(R)$, while only $O(\sum_{i=1}^n L_i)$ additional operations are involved for each j in the binning step (Fan and Marron, 1994).

We next illustrate the binned implementation of LLS, denoted as BinLLS, using the example of smoothed estimates $\widetilde{\Sigma}_{jk}$ for $j \neq k$ in (9). Under Model (19), we drop subscripts j, k in $W_{1,jk}$, $W_{2,jk}$, $W_{3,jk}$ and $S_{ab,jk}$ due to the same set of points $\{U_{i1}, \ldots, U_{iL_i}\}$ across j, k. Denote the binned approximations of $T_{ab,ijk}$ and S_{ab} by $\breve{T}_{ab,ijk}$ and $\breve{S}_{ab}$, respectively. It follows from (8) and (10) that

$$\breve{T}_{ab,ijk}(u, v) = \sum_{r_1=1}^{R} \sum_{r_2=1}^{R} g_{ab}\{h_C, (u, v), (u_{r_1}, u_{r_2})\} \mathcal{D}_{r_1,ij} \mathcal{D}_{r_2,ik},$$

$$\breve{S}_{ab}(u, v) = \sum_{i=1}^{n} \sum_{r_1=1}^{R} \sum_{r_2=1}^{R} g_{ab}\{h_C, (u, v), (u_{r_1}, u_{r_2})\} \varpi_{r_1,i} \varpi_{r_2,i},$$

both of which together with (9) yield the binned approximation of $\widetilde{\Sigma}_{jk}$ as

$$\breve{\Sigma}_{jk} = \sum_{i=1}^{n} \big(\breve{W}_1 \breve{T}_{00,ijk} + \breve{W}_2 \breve{T}_{10,ijk} + \breve{W}_3 \breve{T}_{01,ijk} \big),$$

where $\breve{W}_1, \breve{W}_2$ and $\breve{W}_3$ are the binned approximations of W_1, W_2 and W_3, computed by replacing the related S_{ab}'s in (S.12) of the Supplementary Material with the $\breve{S}_{ab}$'s. It is worth

Table 1: The computational complexity analysis of LLS, BinLLS under Models (6), (19) when evaluating the corresponding smoothed covariance function estimates at a grid of $R \times R$ points.

Method	Model	Number of kernel evaluations	Number of operations (additions and multiplications)
LLS	(6)	$O(\sum_{i=1}^{n}\sum_{j=1}^{p} L_{ij}R)$	$O(R^2\sum_{i=1}^{n}\sum_{j,k=1}^{p} L_{ij}L_{ik})$
LLS	(19)	$O(\sum_{i=1}^{n} L_iR)$	$O(p^2R^2\sum_{i=1}^{n} L_i^2)$
BinLLS	(19)	$O(R)$	$O(np^2R^2 + p^2R^4 + p\sum_{i=1}^{n} L_i)$

noting that, for each pair (j,k), the above binned implementation reduces the number of operations (i.e., additions and multiplications) from $O(R^2\sum_{i=1}^{n} L_i^2)$ to $O(nR^2 + R^4)$, since the kernel evaluations in $g_{ab}\{h_C, (u,v), (u_{r_1}, u_{r_2})\}$ no longer depend on individual observations. Table 1 presents the computational complexity analysis of LLS and BinLLS under Models (6) and (19). It reveals that the binned implementation dramatically improves the computational speed for both densely and sparsely sampled functional data, which is also supported by the empirical evidence in Section 5.3.

To aid the binned implementation of the smoothed adaptive functional thresholding estimator, we then derive the binned approximation of the variance factor $\widetilde{\Psi}_{jk}$, denoted by $\breve{\Psi}_{jk}$. It follows from (13) that $V_{ab,ijk}$ can be approximated by

$$\breve{V}_{ab,ijk}(u,v) = \sum_{r_1=1}^{R}\sum_{r_2=1}^{R} g_{ab}\big(h_C, (u,v), (u_{r_1}, u_{r_2})\big)\big\{\mathcal{D}_{r_1,ij}\mathcal{D}_{r_2,ik} - \breve{\Sigma}_{jk}(u,v)\varpi_{r_1,i}\varpi_{r_2,i}\big\}.$$

Substituting each term in (11) with its binned approximation, we obtain that

$$\breve{\Psi}_{jk} = I_{jk}\sum_{i=1}^{n}\big(\breve{W}_1\breve{V}_{00,ijk} + \breve{W}_2\breve{V}_{10,ijk} + \breve{W}_3\breve{V}_{01,ijk}\big)^2.$$

It is worth mentioning that, when $j = k$, the binned approximations of $\widetilde{\Sigma}_{jj}$ and $\widetilde{\Psi}_{jj}$ can be computed in a similar fashion except that the terms corresponding to $r_1 = r_2$ should be excluded from all double summations over $\{1, \ldots, R\}^2$. Finally, we obtain the binned adaptive functional thresholding estimator $\breve{\mathbf{\Sigma}}_{\mathrm{A}} = (\breve{\Sigma}^{\mathrm{A}}_{jk})_{p\times p}$ with $\breve{\Sigma}^{\mathrm{A}}_{jk} = \breve{\Psi}^{1/2}_{jk} \times s_\lambda\big(\breve{\Sigma}_{jk}/\breve{\Psi}^{1/2}_{jk}\big)$ and the corresponding universal thresholding estimator $\breve{\mathbf{\Sigma}}_{\mathrm{U}} = (\breve{\Sigma}^{\mathrm{U}}_{jk})_{p\times p}$ with $\breve{\Sigma}^{\mathrm{U}}_{jk} = s_\lambda\big(\breve{\Sigma}_{jk}\big)$.

5 Simulations

5.1 Setup

We conduct a number of simulations to compare adaptive functional thresholding estimators to universal functional thresholding estimators. Sections 5.2 and 5.3 consider scenarios where random functions are fully and partially observed, respectively.

In each scenario, to mimic the infinite-dimensionality of random curves, we generate functional variables by $X_{ij}(u) = \mathbf{s}(u)^{\mathrm{T}}\boldsymbol{\theta}_{ij}$ for $i = 1, \ldots, n, j = 1, \ldots, p$ and $u \in \mathcal{U} = [0, 1]$, where $\mathbf{s}(u)$ is a 50-dimensional Fourier basis function and $\boldsymbol{\theta}_i = (\boldsymbol{\theta}_{i1}^{\mathrm{T}}, \ldots, \boldsymbol{\theta}_{ip}^{\mathrm{T}})^{\mathrm{T}} \in \mathbb{R}^{50p}$ is generated from a mean zero multivariate Gaussian distribution with block covariance matrix $\boldsymbol{\Omega} \in \mathbb{R}^{50p\times 50p}$, whose (j, k)-th block is $\boldsymbol{\Omega}_{jk} \in \mathbb{R}^{50\times 50}$ for $j, k = 1, \ldots, p$. The functional sparsity pattern in $\boldsymbol{\Sigma} = \{\Sigma_{jk}(\cdot, \cdot)\}_{p\times p}$ with its (j, k)th entry $\Sigma_{jk}(u, v) = \mathbf{s}(u)^{\mathrm{T}}\boldsymbol{\Omega}_{jk}\mathbf{s}(v)$ can be characterized by the block sparsity structure in $\boldsymbol{\Omega}$. Define $\boldsymbol{\Omega}_{jk} = \omega_{jk}\mathbf{D}$ with $\mathbf{D} = \mathrm{diag}(1^{-2}, \ldots, 50^{-2})$ and hence $\mathrm{cov}(\theta_{ijk}, \theta_{ijk'}) \sim k^{-2}I(k = k')$ for $k, k' = 1, \ldots, 50$. Then we generate $\boldsymbol{\Omega}$ with different block sparsity patterns as follows.

- Model 1 (block banded). For $j, k = 1, \ldots, p/2$, $\omega_{jk} = (1 - |j - k|/10)_+$. For $j, k = p/2 + 1, \ldots, p$, $\omega_{jk} = 4I(j = k)$.
- Model 2 (block sparse without any special structure). For $j, k = p/2+1, \ldots, p$, $\omega_{jk} = 4I(j = k)$. For $j, k = 1, \ldots, p/2$, we generate $\boldsymbol{\omega} = (\omega_{jk})_{p/2\times p/2} = \mathbf{B} + \delta'\mathbf{I}_{p/2}$, where elements of $\mathbf{B}$ are sampled independently from Uniform$[0.3, 0.8]$ with probability 0.2 or 0 with probability 0.8, and $\delta' = \max\{-\lambda_{\min}(\mathbf{B}), 0\} + 0.01$ to guarantee the positive definiteness of $\boldsymbol{\Omega}$.

We implement a cross-validation approach (Bickel and Levina, 2008) for choosing the optimal thresholding parameter $\hat{\lambda}$ in $\widehat{\boldsymbol{\Sigma}}_{\mathrm{A}}$. Specifically, we randomly divide the sample $\{\mathbf{X}_i : i = 1, \ldots, n\}$ into two subsamples of size n_1 and n_2, where $n_1 = n(1 - 1/\log n)$ and $n_2 = n/\log n$ and repeat this N times. Let $\widehat{\boldsymbol{\Sigma}}_{\mathrm{A},1}^{(\nu)}(\lambda)$ and $\widehat{\boldsymbol{\Sigma}}_{\mathrm{S},2}^{(\nu)}$ be the adaptive functional

thresholding estimator as a function of λ and the sample covariance function based on n_1 and n_2 observations, respectively, from the νth split. We select the optimal $\hat{\lambda}$ by minimizing

$$\widehat{\mathrm{err}}(\lambda) = N^{-1} \sum_{\nu=1}^{N} \|\widehat{\boldsymbol{\Sigma}}_{\mathrm{A},1}^{(\nu)}(\lambda) - \widehat{\boldsymbol{\Sigma}}_{\mathrm{S},2}^{(\nu)}\|_{\mathrm{F}}^2,$$

where $\|\cdot\|_{\mathrm{F}}$ denotes the functional version of Frobenius norm, i.e., for any $\boldsymbol{Q} = \{Q_{jk}(\cdot,\cdot)\}_{p\times p}$ with each $Q_{jk} \in \mathbb{S}$, $\|\boldsymbol{Q}\|_{\mathrm{F}} = (\sum_{j,k} \|Q_{jk}\|_{\mathcal{S}}^2)^{1/2}$. The optimal thresholding parameters in $\widehat{\boldsymbol{\Sigma}}_{\mathrm{U}}$, $\widetilde{\boldsymbol{\Sigma}}_{\mathrm{A}}, \widetilde{\boldsymbol{\Sigma}}_{\mathrm{U}}, \breve{\boldsymbol{\Sigma}}_{\mathrm{A}}, \breve{\boldsymbol{\Sigma}}_{\mathrm{U}}$ can be selected in a similar fashion.

5.2 Fully observed functional data

We compare the adaptive functional thresholding estimator $\widehat{\boldsymbol{\Sigma}}_{\mathrm{A}}$ to the universal functional thresholding estimator $\widehat{\boldsymbol{\Sigma}}_{\mathrm{U}}$ under hard, soft, SCAD (with $a = 3.7$) and adaptive lasso (with $\eta = 3$) functional thresholding rules, where the corresponding $\hat{\lambda}$'s are selected by the cross-validation with $N = 5$. We generate $n = 100$ observations for $p = 50, 100, 150$ and replicate each simulation 100 times. We examine the performance of all competing approaches by estimation and support recovery accuracies. In terms of the estimation accuracy, Table 2 reports numerical summaries of losses measured by functional versions of Frobenius and matrix ℓ_1 norms. To assess the support recovery consistency, we present in Table 3 the average of true positive rates (TPRs) and false positive rates (FPRs), defined as TPR $= \#\{(j,k) : \|\widehat{\Sigma}_{jk}\|_{\mathcal{S}} \neq 0 \text{ and } \|\Sigma_{jk}\|_{\mathcal{S}} \neq 0\}/\#\{(j,k) : \|\Sigma_{jk}\|_{\mathcal{S}} \neq 0\}$ and FPR $= \#\{(j,k) : \|\widehat{\Sigma}_{jk}\|_{\mathcal{S}} \neq 0 \text{ and } \|\Sigma_{jk}\|_{\mathcal{S}} = 0\}/\#\{(j,k) : \|\Sigma_{jk}\|_{\mathcal{S}} = 0\}$. Since the results under Models 1 and 2 have similar trends, we only present the numerical results under Model 2 here to save space. See Tables 9 and 10 of the Supplementary Material for results under Model 1.

Several conclusions can be drawn from Tables 2–3 and 9–10. First, in all scenarios, $\widehat{\boldsymbol{\Sigma}}_{\mathrm{A}}$ provides substantially improved accuracy over $\widehat{\boldsymbol{\Sigma}}_{\mathrm{U}}$ regardless of the thresholding rule or the loss used. We also obtain the sample covariance function $\widehat{\boldsymbol{\Sigma}}_{\mathrm{S}}$, the results of which deteriorate severely compared with $\widehat{\boldsymbol{\Sigma}}_{\mathrm{A}}$ and $\widehat{\boldsymbol{\Sigma}}_{\mathrm{U}}$. Second, for support recovery, again $\widehat{\boldsymbol{\Sigma}}_{\mathrm{A}}$ uniformly outperforms $\widehat{\boldsymbol{\Sigma}}_{\mathrm{U}}$, which fails to recover the functional sparsity pattern especially when p is large. Third, the adaptive functional thresholding approach using the hard

Table 2: The average (standard error) functional matrix losses over 100 simulation runs.

Model	Method	$p = 50$		$p = 100$		$p = 150$	
		$\widehat{\mathbf{\Sigma}}_{\text{A}}$	$\widehat{\mathbf{\Sigma}}_{\text{U}}$	$\widehat{\mathbf{\Sigma}}_{\text{A}}$	$\widehat{\mathbf{\Sigma}}_{\text{U}}$	$\widehat{\mathbf{\Sigma}}_{\text{A}}$	$\widehat{\mathbf{\Sigma}}_{\text{U}}$
		Functional Frobenius norm					
2	Hard	5.67(0.03)	9.39(0.02)	9.48(0.04)	15.79(0.01)	14.00(0.05)	22.26(0.01)
	Soft	6.14(0.03)	8.55(0.04)	10.28(0.05)	15.00(0.05)	14.8(0.05)	21.89(0.04)
	SCAD	5.94(0.03)	8.59(0.04)	9.96(0.05)	15.02(0.05)	14.49(0.06)	21.91(0.04)
	Adap. lasso	5.44(0.03)	9.10(0.04)	8.99(0.04)	15.73(0.02)	13.02(0.05)	22.25(0.01)
	Sample	21.80(0.04)		43.51(0.06)		65.22(0.07)	
		Functional matrix ℓ_1 norm					
	Hard	2.85(0.03)	4.74(0.01)	4.77(0.05)	7.11(0.01)	7.65(0.07)	10.31(0.01)
	Soft	3.31(0.03)	4.51(0.04)	5.37(0.04)	6.90(0.02)	8.21(0.05)	10.21(0.01)
	SCAD	3.22(0.03)	4.48(0.03)	5.29(0.04)	6.91(0.02)	8.14(0.05)	10.21(0.01)
	Adap. lasso	2.75(0.03)	4.66(0.02)	4.62(0.05)	7.08(0.01)	7.35(0.07)	10.30(0.01)
	Sample	28.06(0.12)		56.01(0.19)		84.13(0.23)	

Table 3: The average TPRs/ FPRs over 100 simulation runs.

Model	Method	$p = 50$		$p = 100$		$p = 150$	
		$\widehat{\mathbf{\Sigma}}_{\text{A}}$	$\widehat{\mathbf{\Sigma}}_{\text{U}}$	$\widehat{\mathbf{\Sigma}}_{\text{A}}$	$\widehat{\mathbf{\Sigma}}_{\text{U}}$	$\widehat{\mathbf{\Sigma}}_{\text{A}}$	$\widehat{\mathbf{\Sigma}}_{\text{U}}$
2	Hard	0.77/0.00	0.00/0.00	0.68/0.00	0.00/0.00	0.63/0.00	0.00/0.00
	Soft	0.99/0.06	0.50/0.07	0.97/0.04	0.30/0.04	0.96/0.04	0.11/0.02
	SCAD	0.99/0.06	0.47/0.06	0.98/0.05	0.29/0.04	0.97/0.05	0.10/0.01
	Adap. lasso	0.91/0.00	0.10/0.01	0.86/0.00	0.01/0.00	0.83/0.00	0.00/0.00

and the adaptive lasso functional thresholding rules tends to have lower losses and lower TPRs/FPRs than that using the soft and the SCAD functional thresholding rules.

5.3 Partially observed functional data

In this section, we assess the finite-sample performance of LLS and BinLLS methods to handle partially observed functional data. We first generate random functions $X_{ij}(\cdot)$ for $i = 1, \ldots, n, j = 1, \ldots, p$ by the same procedure as in Section 5.1 with either non-sparse or sparse $\mathbf{\Sigma}$ depending on p. We then generate the observed values Z_{ijl} from equa-

tion (19), where the measurement locations U_{il} and errors ε_{ijl} are sampled independently from Uniform[0,1] and $\mathcal{N}(0, 0.5^2)$, respectively. We consider settings of $n = 100$ and $L_i = 11, 21, 51, 101$, changing from sparse to moderately dense to very dense measurement schedules. We use the Gaussian kernel with the optimal bandwidths proportional to $n^{-1/6}$, $(nL_i^2)^{-1/6}$ and $n^{-1/4}$, respectively, as suggested in Zhang and Wang (2016), so for the empirical work in this paper we choose the proportionality constants in the range $(0, 1]$, which gives good results in all settings we consider.

To compare BinLLS with LLS in terms of the computational speed and estimation accuracy, we first consider a low-dimensional example $p = 6$ with non-sparse $\mathbf{\Sigma}$ generated by modifying Model 1 with $\omega_{jk} = (1 - |j - k|/10)_+$ for $j, k = 1, \ldots, 6$. In addition to our proposed smoothing methods, we also implement local-linear-smoother-based pre-smoothing and its binned implementation, denoted as LLS-P and BinLLS-P, respectively. Table 4 reports numerical summaries of estimation errors evaluated at $R = 21$ equally-spaced points in $[0, 1]$ and the corresponding CPU time on the processor Intel(R) Xeon(R) CPU E5-2690 v3 @ 2.60GHz. The results for the sample covariance function $\widehat{\mathbf{\Sigma}}_{\mathrm{S}}$ based on fully observed $\mathbf{X}_1(\cdot), \ldots, \mathbf{X}_n(\cdot)$ are also provided as the baseline for comparison. Note that, LLS is too slow to implement for the case $L_i = 101$, so we do not report its result here.

A few trends are observable from Table 4. First, the binned implementations (BinLLS and BinLLS-P) attain similar or even lower estimation errors compared with their direct implementations (LLS and LLS-P) under all scenarios, while resulting in considerably faster computational speeds especially under dense designs. For example, BinLLS runs over 400 times faster than LLS when $L_i = 51$. Second, all methods provide higher estimation accuracies as L_i increases, and enjoy similar performance when functions are very densely observed, e.g., $L_i = 51$ and 101, compared with the fully observed functional case. However, the performance of LLS-P and BinLLS-P deteriorates severely under sparse designs, e.g., $L_i = 11$ and 21, since limited information is available from a small number of observations per subject. Among all competitors, we conclude that BinLLS is overall a unified approach

Table 4: The average (standard error) functional matrix losses and average CPU time for $p = 6$ over 100 simulation runs.

L_i	Method	Functional Frobenius norm	Functional matrix ℓ_1 norm	Elapsed time (sec)	Method	Functional Frobenius norm	Functional matrix ℓ_1 norm	Elapsed time (sec)
11	BinLLS	1.57(0.02)	1.72(0.03)	2.06	BinLLS-P	4.14(0.03)	4.36(0.04)	0.18
	LLS	1.62(0.02)	1.76(0.03)	50.52	LLS-P	4.23(0.04)	4.47(0.05)	0.22
21	BinLLS	1.28(0.02)	1.42(0.03)	2.07	BinLLS-P	2.66(0.02)	2.80(0.02)	0.19
	LLS	1.28(0.02)	1.42(0.03)	136.88	LLS-P	2.67(0.02)	2.82(0.03)	0.29
51	BinLLS	1.06(0.02)	1.20(0.03)	2.21	BinLLS-P	1.12(0.03)	1.26(0.03)	0.20
	LLS	1.04(0.02)	1.18(0.03)	967.75	LLS-P	1.12(0.03)	1.26(0.03)	0.39
101	BinLLS	1.00(0.02)	1.14(0.03)	2.23	BinLLS-P	0.99(0.02)	1.13(0.03)	0.21
	LLS	-	-	-	LLS-P	0.97(0.02)	1.11(0.03)	0.64

$\widehat{\mathbf{\Sigma}}_{\mathrm{S}}$	Functional Frobenius norm	Functional matrix ℓ_1 norm	Elapsed time (sec)
	1.04(0.03)	1.20(0.03)	0.11

that can handle both sparsely and densely sampled functional data well with increased computational efficiency and guaranteed estimation accuracy.

We next examine the performance of BinLLS-based adaptive and universal functional thresholding estimators in terms of estimation accuracy and support recovery consistency using the same performance measures as in Tables 2–3. Tables 5–6 and Tables 11–14 of the Supplementary Material report numerical results for settings of $p = 50$ and 100 satisfying Models 1 and 2 under different measurement schedules. We observe a few apparent patterns from Tables 5–6 and 11–14. First, $\check{\mathbf{\Sigma}}_{\mathrm{A}}$ substantially outperforms $\check{\mathbf{\Sigma}}_{\mathrm{U}}$ with significantly lower estimation errors in all settings. Second, $\check{\mathbf{\Sigma}}_{\mathrm{A}}$ works consistently well in recovering the functional sparsity structures especially under the soft and SCAD functional thresholding rules, while $\check{\mathbf{\Sigma}}_{\mathrm{U}}$ fails to identify such patterns. Third, the estimation and support recovery consistencies of $\check{\mathbf{\Sigma}}_{\mathrm{A}}$ and $\check{\mathbf{\Sigma}}_{\mathrm{U}}$ are improved as L_i increases. When curves are very densely observed, e.g., $L_i = 101$, we observe that both estimators enjoy similar performance with $\widehat{\mathbf{\Sigma}}_{\mathrm{A}}$ and $\widehat{\mathbf{\Sigma}}_{\mathrm{U}}$ in Tables 2–3 and Tables 9–10 of the Supplementary Material. Such observation provides empirical evidence to support our remark for Theorem 3 about the same convergence rate between very densely observed and fully observed functional scenarios.

Table 5: The average (standard error) functional matrix losses for partially observed functional scenarios and $p = 50$ over 100 simulation runs.

		$L_i = 11$		$L_i = 21$		$L_i = 51$		$L_i = 101$	
Model	Method	$\breve{\mathbf{\Sigma}}_{\text{A}}$	$\breve{\mathbf{\Sigma}}_{\text{U}}$	$\breve{\mathbf{\Sigma}}_{\text{A}}$	$\breve{\mathbf{\Sigma}}_{\text{U}}$	$\breve{\mathbf{\Sigma}}_{\text{A}}$	$\breve{\mathbf{\Sigma}}_{\text{U}}$	$\breve{\mathbf{\Sigma}}_{\text{A}}$	$\breve{\mathbf{\Sigma}}_{\text{U}}$
		Functional Frobenius norm							
2	Hard	8.12(0.03)	10.41(0.02)	6.85(0.04)	9.89(0.01)	6.06(0.04)	9.60(0.02)	5.75(0.04)	9.51(0.02)
	Soft	8.35(0.03)	10.37(0.01)	7.35(0.03)	9.60(0.03)	6.72(0.03)	8.86(0.04)	6.48(0.03)	8.56(0.04)
	SCAD	8.32(0.03)	10.37(0.01)	7.23(0.04)	9.60(0.03)	6.50(0.04)	8.89(0.04)	6.23(0.04)	8.61(0.04)
	Adap. lasso	7.83(0.03)	10.39(0.01)	6.69(0.04)	9.84(0.02)	5.97(0.04)	9.40(0.04)	5.71(0.04)	9.16(0.04)
		Functional matrix ℓ_1 norm							
	Hard	3.82(0.04)	4.91(0.01)	3.36(0.04)	4.82(0.01)	3.00(0.05)	4.78(0.01)	2.85(0.05)	4.77(0.01)
	Soft	3.96(0.02)	4.88(0.01)	3.71(0.03)	4.72(0.02)	3.50(0.03)	4.55(0.03)	3.44(0.03)	4.47(0.03)
	SCAD	3.96(0.02)	4.88(0.01)	3.67(0.03)	4.72(0.02)	3.41(0.03)	4.55(0.02)	3.32(0.03)	4.48(0.02)
	Adap. lasso	3.65(0.04)	4.90(0.01)	3.28(0.04)	4.80(0.01)	2.96(0.04)	4.73(0.01)	2.88(0.04)	4.69(0.02)

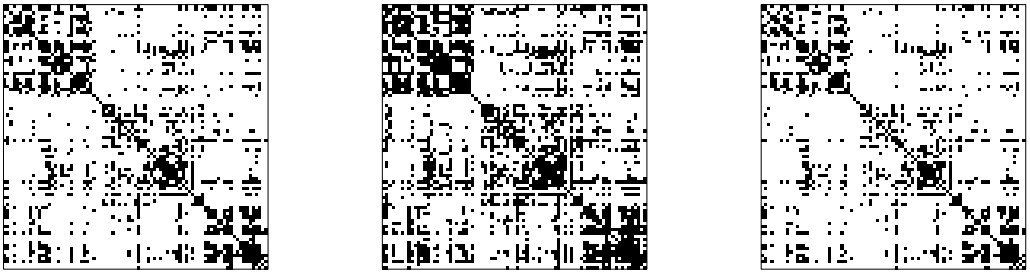
Table 6: The average TPRs/ FPRs for partially observed functional scenarios and $p = 50$ over 100 simulation runs.

		$L_i = 11$		$L_i = 21$		$L_i = 51$		$L_i = 101$	
Model	Method	$\breve{\mathbf{\Sigma}}_{\text{A}}$	$\breve{\mathbf{\Sigma}}_{\text{U}}$	$\breve{\mathbf{\Sigma}}_{\text{A}}$	$\breve{\mathbf{\Sigma}}_{\text{U}}$	$\breve{\mathbf{\Sigma}}_{\text{A}}$	$\breve{\mathbf{\Sigma}}_{\text{U}}$	$\breve{\mathbf{\Sigma}}_{\text{A}}$	$\breve{\mathbf{\Sigma}}_{\text{U}}$
2	Hard	0.58/0.00	0.00/0.00	0.69/0.00	0.00/0.00	0.75/0.00	0.01/0.00	0.79/0.00	0.01/0.00
	Soft	0.95/0.04	0.03/0.01	0.97/0.05	0.22/0.03	0.99/0.06	0.48/0.06	0.99/0.06	0.58/0.07
	SCAD	0.95/0.04	0.03/0.01	0.97/0.06	0.22/0.03	0.99/0.07	0.46/0.06	0.99/0.07	0.54/0.06
	Adap. lasso	0.80/0.00	0.00/0.00	0.86/0.00	0.02/0.00	0.90/0.00	0.08/0.00	0.91/0.00	0.15/0.01

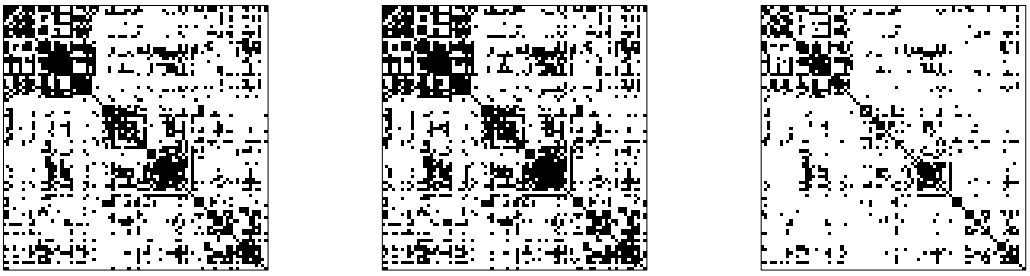
6 Real Data

In this section, we aim to investigate the association between the brain functional connectivity and fluid intelligence (gF), the capacity to solve problems independently of acquired knowledge (Cattell, 1987). The dataset contains subjects of resting-state fMRI scans and the corresponding gF scores, measured by the 24-item Raven's Progressive Matrices, from the Human Connectome Project (HCP). We follow many recent proposals based on HCP by modelling signals as multivariate random functions with each region of interest (ROI) representing one random function (Zapata et al., 2022; Lee et al., 2021; Miao et al., 2022). We focus our analysis on $n_{\text{low}} = 73$ subjects with intelligence scores $gF \leqslant 8$ and $n_{\text{high}} = 85$ subjects with $gF \geqslant 23$, and consider $p = 83$ ROIs of three generally acknowledged modules in neuroscience study (Finn et al., 2015): the medial frontal (29 ROIs), frontoparietal (34 ROIs) and default mode modules (20 ROIs). For each subject, the BOLD signals at each ROI are collected every 0.72 seconds for a total of $L = 1200$ measurement locations (14.4 minutes). We first implement the ICA-FIX preprocessed pipeline (Glasser et al., 2013) and a standard band-pass filter at $[0.01, 0.08]$ Hz to exclude frequency bands not implicated in resting state functional connectivity (Biswal et al., 1995). Figure 12 of the Supplementary Material displays examplified trajectories of pre-smoothed data. The adaptive functional thresholding method is then adopted to estimate the sparse covariance function and therefore the brain networks.

The sparsity structures in $\hat{\mathbf{\Sigma}}_{\text{A}}$ for both groups are displayed in Figure 1. With $\widehat{\lambda}$ selected by the cross-validation, the network associated with $\widehat{\mathbf{\Sigma}}_{\text{A}}$ for subjects with $gF \geqslant 23$ is more densely connected than that with $gF \leqslant 8$, as evident from Fig. 1(a)–(b). We further set the sparsity level to 70% and 85%, and present the corresponding sparsity patterns in Fig. 1(c)–(f). The results clearly indicate the existence of three diagonal blocks under all sparsity levels, complying with the identification of the medial frontal, frontoparietal and default mode modules in Finn et al. (2015). We also implement the universal functional thresholding method. However, compared with $\widehat{\mathbf{\Sigma}}_{\text{A}}$, the results of $\widehat{\mathbf{\Sigma}}_{\text{U}}$ suffer from

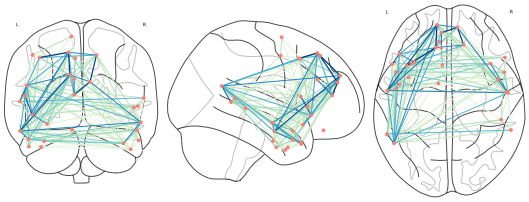


(a) $gF \leqslant 8$: $\widehat{\boldsymbol{\Sigma}}_{\text{A}}$ (80.42% zeros) (c) $gF \leqslant 8$: $\widehat{\boldsymbol{\Sigma}}_{\text{A}}$ (70% zeros) (e) $gF \leqslant 8$: $\widehat{\boldsymbol{\Sigma}}_{\text{A}}$ (85% zeros)

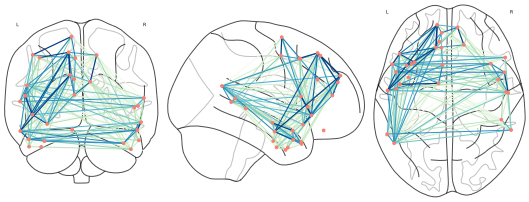


(b) $gF \geqslant 23$: $\widehat{\boldsymbol{\Sigma}}_{\text{A}}$ (72.93% zeros) (d) $gF \geqslant 23$: $\widehat{\boldsymbol{\Sigma}}_{\text{A}}$ (70% zeros) (f) $gF \geqslant 23$: $\widehat{\boldsymbol{\Sigma}}_{\text{A}}$ (85% zeros)

Figure 1: Estimated sparsity structures in $\widehat{\boldsymbol{\Sigma}}_{\text{A}}$ using soft functional thresholding rule at fluid intelligence $gF \leqslant 8$ and $gF \geqslant 23$: (a)–(b) with the corresponding $\hat{\lambda}$ selected by fivefold cross-validation; (c)–(f) with the estimated functional sparsity levels set at 70% and 85%.

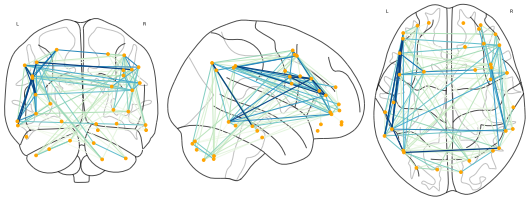


(a) $gF \leqslant 8$: the medial frontal module in Fig. 1(e)

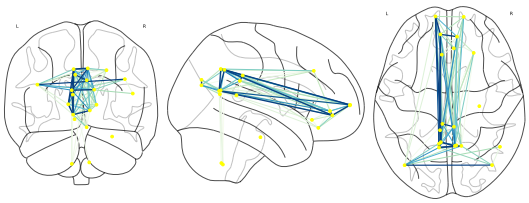


(d) $gF \geqslant 23$: the medial frontal module in Fig. 1(f)

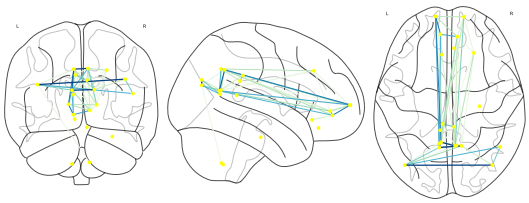
(b) $gF \leqslant 8$: the frontoparietal module in Fig. 1(e)



(e) $gF \geqslant 23$: the frontoparietal module in Fig. 1(f)



(c) $gF \leqslant 8$: the default mode module in Fig. 1(e)



(f) $gF \geqslant 23$: the default mode module in Fig. 1(f)

Figure 2: The connectivity strengths in Fig. 1(e)–(f) at fluid intelligence $gF \leqslant 8$ and $gF \geqslant 23$. Salmon, orange and yellow nodes represent the ROIs in the medial frontal, frontoparietal and default mode modules, respectively. The edge color from cyan to blue corresponds to the value of $\|\widehat{\Sigma}^{\mathrm{A}}_{jk}\|_{\mathcal{S}}/(\|\widehat{\Sigma}^{\mathrm{A}}_{jj}\|_{\mathcal{S}}\|\widehat{\Sigma}^{\mathrm{A}}_{kk}\|_{\mathcal{S}})^{1/2}$ from small to large.

the heteroscedasticity, as demonstrated in Section 5 and Section E.3 of the Supplementary Material, and fail to detect any noticeable block structure, hence we choose not to report them here. To explore the impact of gF on the functional connectivity, we compute the connectivity strength using the standardized form $\|\widehat{\Sigma}^{\mathrm{A}}_{jk}\|_{\mathcal{S}}/(\|\widehat{\Sigma}^{\mathrm{A}}_{jj}\|_{\mathcal{S}}\|\widehat{\Sigma}^{\mathrm{A}}_{kk}\|_{\mathcal{S}})^{1/2}$ for $j, k = 1 \ldots, p$. Interestingly, we observe from Figure 2 that subjects with $gF \geqslant 23$ tend to have enhanced brain connectivity in the medial frontal and frontoparietal modules, while the connectivity strength in the default mode module declines. This agrees with existing neuroscience literature reporting a strong positive association between intelligence score and the medial frontal/frontoparietal functional connectivity in the resting state (Van Den Heuvel et al., 2009; Finn et al., 2015), and lends support to the conclusion that lower default mode module activity is associated with better cognitive performance (Anticevic et al., 2012). See also Section E.3 of the Supplementary Material, in which we illustrate our adaptive functional thresholding estimation using another ADHD dataset.

Acknowledgements.

We are grateful to the editor, the associate editor and two referees for their insightful comments and suggestions, which have led to significant improvement of our paper. Shaojun Guo was partially supported by the National Natural Science Foundation of China (No. 11771447).

Disclosure Statement.

The authors report there are no competing interests to declare.

References

Anticevic, A., Cole, M. W., Murray, J. D., Corlett, P. R., Wang, X.-J. and Krystal, J. H. (2012). The role of default network deactivation in cognition and disease, *Trends in Cognitive Sciences* **16**: 584–592.

Avella-Medina, M., Battey, H. S., Fan, J. and Li, Q. (2018). Robust estimation of high-dimensional covariance and precision matrices, *Biometrika* **105**: 271–284.

Bickel, P. J. and Levina, E. (2008). Covariance regularization by thresholding, *The Annals of Statistics* **36**: 2577–2604.

Biswal, B., Zerrin Yetkin, F., Haughton, V. M. and Hyde, J. S. (1995). Functional connectivity in the motor cortex of resting human brain using echo-planar MRI, *Magnetic Resonance in Medicine* **34**: 537–541.

Cai, T. and Liu, W. (2011). Adaptive thresholding for sparse covariance matrix estimation, *Journal of the American Statistical Association* **106**: 672–684.

Cattell, R. B. (1987). *Intelligence: Its Structure, Growth and Action*, Elsevier.

Chang, C. and Glover, G. H. (2010). Time–frequency dynamics of resting-state brain connectivity measured with fMRI, *Neuroimage* **50**: 81–98.

Chen, Z. and Leng, C. (2016). Dynamic covariance models, *Journal of the American Statistical Association* **111**: 1196–1207.

Chiou, J.-M., Chen, Y.-T. and Yang, Y.-F. (2014). Multivariate functional principal component analysis: a normalization approach, *Statistica Sinica* **24**: 1571–1596.

Chiou, J.-M., Yang, Y.-F. and Chen, Y.-T. (2016). Multivariate functional linear regression and prediction, *Journal of Multivariate Analysis* **146**: 301–312.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties, *Journal of the American Statistical Association* **96**: 1348–1360.

Fan, J. and Marron, J. S. (1994). Fast implementations of nonparametric curve estimators, *Journal of Computational and Graphical Statistics* **3**: 35–56.

Finn, E. S., Shen, X., Scheinost, D., Rosenberg, M. D., Huang, J., Chun, M. M., Papademetris, X. and Constable, R. T. (2015). Functional connectome fingerprinting: identifying individuals using patterns of brain connectivity, *Nature Neuroscience* **18**: 1664–1671.

Glasser, M. F., Sotiropoulos, S. N., Wilson, J. A., Coalson, T. S., Fischl, B., Andersson, J. L., Xu, J., Jbabdi, S., Webster, M., Polimeni, J. R. et al. (2013). The minimal preprocessing pipelines for the human connectome project, *Neuroimage* **80**: 105–124.

Guo, S., Qiao, X. and Wang, Q. (2022). Factor modelling for high-dimensional functional time series, *arXiv:2112.13651v2* .

Happ, C. and Greven, S. (2018). Multivariate functional principal component analysis for data observed on different (dimensional) domains, *Journal of the American Statistical Association* **113**: 649–659.

Kong, D., Xue, K., Yao, F. and Zhang, H. H. (2016). Partially functional linear regression in high dimensions, *Biometrika* **103**: 147–159.

Kosorok, M. R. (2008). *Introduction to Empirical Processes and Semiparametric Inference*, Springer Series in Statistics, Springer, New York.

Lee, K.-Y., Ji, D., Li, L., Constable, T. and Zhao, H. (2021). Conditional functional graphical models, *Journal of the American Statistical Association, in press* .

Li, B. and Solea, E. (2018). A nonparametric graphical model for functional data with application to brain networks based on fMRI, *Journal of the American Statistical Association* **113**: 1637–1655.

Lotte, F., Bougrain, L., Cichocki, A., Clerc, M., Congedo, M., Rakotomamonjy, A. and Yger, F. (2018). A review of classification algorithms for EEG-based brain–computer interfaces: a 10 year update, *Journal of Neural Engineering* **15**: 031005.

Miao, R., Zhang, X. and Wong, R. K. (2022). A wavelet-based independence test for functional data with an application to MEG functional connectivity, *Journal of the American Statistical Association, in press* .

Park, J., Ahn, J. and Jeon, Y. (2021). Sparse functional linear discriminant analysis, *Biometrika* **109**: 209–226.

Qiao, X., Guo, S. and James, G. (2019). Functional graphical models, *Journal of the American Statistical Association* **114**: 211–222.

Qiao, X., Qian, C., James, G. M. and Guo, S. (2020). Doubly functional graphical models in high dimensions, *Biometrika* **107**: 415–431.

Rogers, B. P., Morgan, V. L., Newton, A. T. and Gore, J. C. (2007). Assessing functional connectivity in the human brain by fMRI, *Magnetic Resonance Imaging* **25**: 1347–1357.

Rothman, A. J., Levina, E. and Zhu, J. (2009). Generalized thresholding of large covariance matrices, *Journal of the American Statistical Association* **104**: 177–186.

Storey, J. D., Xiao, W., Leek, J. T., Tompkins, R. G. and Davis, R. W. (2005). Significance analysis of time course microarray experiments, *Proceedings of the National Academy of Sciences* **102**: 12837–12842.

Van Den Heuvel, M. P., Stam, C. J., Kahn, R. S. and Pol, H. E. H. (2009). Efficiency of functional brain networks and intellectual performance, *Journal of Neuroscience* **29**: 7619–7624.

Vu, V. Q. and Lei, J. (2013). Minimax sparse principal subspace estimation in high dimensions, *The Annals of Statistics* **41**: 2905–2947.

Wang, H., Peng, B., Li, D. and Leng, C. (2021). Nonparametric estimation of large covariance matrices with conditional sparsity, *Journal of Econometrics* **223**: 53–72.

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.

Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables, *Journal of the Royal Statistical Society: Series B* **68**: 49–67.

Zapata, J., Oh, S. Y. and Petersen, A. (2022). Partial separability and functional graphical models for multivariate Gaussian processes, *Biometrika* **109**: 665–681.

Zhang, J.-T. and Chen, J. (2007). Statistical inferences for functional data, *The Annals of Statistics* **35**: 1052–1079.

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

Zou, H. (2006). The adaptive lasso and its oracle properties, *Journal of the American Statistical Association* **101**: 1418–1429.

Supplementary material to "Adaptive functional thresholding for sparse covariance function estimation in high dimensions"

Qin Fang, Shaojun Guo and Xinghao Qiao

This supplementary material contains the technical proofs for the fully observed functional scenario in Section A, derivations of functional thresholding rules in Section B, further discussion in Section C, additional methodological details and technical proofs for the partially observed functional scenario in Section D and additional empirical results in Section E.

A Technical proofs

Before stating the regularity conditions, we make some notation. For a function $Z \in \mathbb{S}$, define $\|Z\|_\infty = \sup_{u,v\in\mathcal{U}}|Z(u,v)|$. For two sequences of real processes $\{a_n(u), u \in \mathcal{U}\}$ and $\{b_n(u), u \in \mathcal{U}\}$, we write $a_n(u) \lesssim b_n(u)$ if there exists some constant c such that $|a_n(u)| \leqslant c|b_n(u)|$ holds for all n and $u \in \mathcal{U}$. Without loss of generality, in the following we assume that $\mathbb{E}\{X_{ij}(u)\} \equiv 0$ and both estimators $\widehat{\Sigma}_{jk}(u,v)$ and $\widehat{\Theta}_{jk}(u,v)$ are defined as

$$\widehat{\Sigma}_{jk}(u,v) = \frac{1}{n}\sum_{i=1}^{n} X_{ij}(u)X_{ik}(v) \text{ and } \widehat{\Theta}_{jk}(u,v) = \frac{1}{n}\sum_{i=1}^{n} X_{ij}(u)^2X_{ik}(v)^2 - \widehat{\Sigma}_{jk}(u,v)^2,$$

respectively.

Lemma A1 *Suppose that Conditions 1–4 hold. Then for any $M > 0$, there exists some constant $\rho_1 > 0$ such that*

$$P\left\{\max_{j,k}\left\|\frac{\widehat{\Theta}_{jk} - \Theta_{jk}}{\Theta_{jk}}\right\|_\infty \geqslant \rho_1\frac{\log^2 p}{n^{1/2}}\right\} = O(p^{-M}).$$

Proof. Denote $\widetilde{\Theta}_{jk}(u,v) = \mathbb{E}\{X_{ij}(u)^2X_{ik}(v)^2\}$. We decompose $\widehat{\Theta}_{jk}(u,v) - \Theta_{jk}(u,v)$ as

$$\begin{aligned}&\widehat{\Theta}_{jk}(u,v) - \Theta_{jk}(u,v)\\ =\ &\Sigma_{jk}(u,v)^2 - \widehat{\Sigma}_{jk}(u,v)^2 + \frac{1}{n}\sum_{i=1}^{n}\Big\{X_{ij}(u)^2X_{ik}(v)^2 - \widetilde{\Theta}_{jk}(u,v)\Big\}.\end{aligned}$$

By Condition 3, $\Theta_{jk}(u,v) \geqslant \tau\sigma_j(u)\sigma_k(v)$ for each $j,k = 1,\ldots,p$. Hence,

$$\begin{aligned}
&\left|\frac{\widehat{\Theta}_{jk}(u,v) - \Theta_{jk}(u,v)}{\Theta_{jk}(u,v)}\right| \\
\leqslant &\left|\frac{\Sigma_{jk}(u,v)^2 - \widehat{\Sigma}_{jk}(u,v)^2}{\tau\sigma_j(u)\sigma_k(v)}\right| + \left|\frac{1}{n}\sum_{i=1}^{n}\frac{X_{ij}(u)^2X_{ik}(v)^2 - \widetilde{\Theta}_{jk}(u,v)}{\tau\sigma_j(u)\sigma_k(v)}\right| \\
= &\, H_{jk}^{(1)}(u,v) + H_{jk}^{(2)}(u,v).
\end{aligned}$$

First, consider the concentration bound for $\|H_{jk}^{(1)}\|_\infty$. Denote $\widetilde{Y}_{ijk}(u,v) = Y_{ij}(u)Y_{ik}(v) - \Sigma_{jk}(u,v)/\{\sigma_j(u)^{1/2}\sigma_k(v)^{1/2}\}$ and let $d_{jk}((u,v),(u',v')) = d_j(u,u') + d_k(v,v')$. Applying Theorem 8.4 in Kosorok (2008) under Conditions 1 and 2, we obtain that, there exists some constant $C_1 > 0$ such that $\left\|\sup_{u\in\mathcal{U}}|Y_{1j}(u)|\right\|_{\psi_2} \leqslant C_1$ for all $j = 1,\ldots,p$. By the property of ψ_1-norm, we have that

$$\begin{aligned}
&\|Y_{ij}(u)Y_{ik}(v) - Y_{ij}(u')Y_{ik}(v')\|_{\psi_1} \\
\leqslant &\|Y_{ij}(u)\{Y_{ik}(v) - Y_{ik}(v')\}\|_{\psi_1} + \|\{Y_{ij}(u) - Y_{ij}(u')\}Y_{ik}(v')\|_{\psi_1} \\
\leqslant &\|Y_{ij}(u)\|_{\psi_2}\|Y_{ik}(v) - Y_{ik}(v')\|_{\psi_2} + \|Y_{ik}(v')\|_{\psi_2}\|Y_{ij}(u) - Y_{ij}(u')\|_{\psi_2} \\
\lesssim &\{d_j(u,u') + d_k(v,v')\} = d_{jk}((u,v),(u',v')),
\end{aligned}$$

which implies that

$$\left\|\widetilde{Y}_{ijk}(u,v) - \widetilde{Y}_{ijk}(u',v')\right\|_{\psi_1} \lesssim d_{jk}((u,v),(u',v')). \tag{S.1}$$

Note that

$$\bar{Z}_{jk}(u,v) = \frac{\widehat{\Sigma}_{jk}(u,v) - \Sigma_{jk}(u,v)}{\sigma_j(u)^{1/2}\sigma_k(v)^{1/2}} = \frac{1}{n}\sum_{i=1}^{n}\left\{Y_{ij}(u)Y_{ik}(v) - \frac{\Sigma_{jk}(u,v)}{\sigma_j(u)^{1/2}\sigma_k(v)^{1/2}}\right\},$$

and for a random variable X and any integer $m \geqslant 1$, $\mathbb{E}\|X\|^m \leqslant m!\|X\|_{\psi_1}^m$. By Bernstein's inequality and Lemma 8.3 of Kosorok (2008), we have that for $u,v,u',v' \in \mathcal{U}$,

$$\left\|n^{1/2}\left\{\bar{Z}_{jk}(u,v) - \bar{Z}_{jk}(u',v')\right\}\right\|_{\psi_1} \lesssim d_{jk}((u,v),(u',v')).$$

For the semimetric d_{jk}, $D(\epsilon, d_{jk}) \leqslant D(\epsilon/2, d_j)D(\epsilon/2, d_k) \lesssim \epsilon^{-2r}$. Applying Theorem 8.4 in Kosorok (2008) with Conditions 1 and 2 again, we obtain that, there exists some constant

$C_2 > 0$ such that

$$\max_{1\leqslant j,k\leqslant p} \left\| \sup_{u,v\in\mathcal{U}} |n^{1/2}\bar{Z}_{jk}(u,v)| \right\|_{\psi_1} \leqslant C_2.$$

This immediately implies that there exist some universal constant $C_3 > 0$ such that for any $x > 0$,

$$P\left\{ \max_{j,k} \sup_{u,v\in\mathcal{U}} \left| \frac{\widehat{\Sigma}_{jk}(u,v) - \Sigma_{jk}(u,v)}{\sigma_j(u)^{1/2}\sigma_k(v)^{1/2}} \right| > x \right\} \lesssim p^2 \exp\{-C_3 n^{1/2} x\}.$$

As a result, for any $M > 0$, there exists some constant $\tilde{\rho}_1 > 0$ such that

$$P\left\{ \max_{j,k} \sup_{u,v\in\mathcal{U}} \left| \frac{\widehat{\Sigma}_{jk}(u,v) - \Sigma_{jk}(u,v)}{\sigma_j(u)^{1/2}\sigma_k(v)^{1/2}} \right| > \tilde{\rho}_1 \frac{\log p}{n^{1/2}} \right\} \lesssim p^{-M}. \tag{S.2}$$

Observe that

$$\left| \frac{\widehat{\Sigma}_{jk}(u,v)^2 - \Sigma_{jk}(u,v)^2}{\sigma_j(u)\sigma_k(v)} \right| \leqslant \left| \frac{\widehat{\Sigma}_{jk}(u,v) - \Sigma_{jk}(u,v)}{\sigma_j(u)^{1/2}\sigma_k(v)^{1/2}} \right|^2 + 2\left| \frac{\widehat{\Sigma}_{jk}(u,v) - \Sigma_{jk}(u,v)}{\sigma_j(u)^{1/2}\sigma_k(v)^{1/2}} \right|,$$

since $|\Sigma_{jk}(u,v)| \leqslant \sigma_j(u)^{1/2}\sigma_k(v)^{1/2}$. By the inequality (S.2), we have that

$$P\left\{ \max_{j,k} \|H_{jk}^{(1)}\|_\infty > 2\tilde{\rho}_1 \frac{\log p}{n^{1/2}} + \tilde{\rho}_1^2 \frac{\log^2 p}{n} \right\} \lesssim p^{-M}. \tag{S.3}$$

We next control the bound for $\|H_{jk}^{(2)}\|_\infty$ through the truncation technique. Note that

$$\frac{X_{ij}(u)^2 X_{ik}(v)^2 - \widetilde{\Theta}_{jk}(u,v)}{\sigma_j(u)\sigma_k(v)} = Y_{ij}(u)^2 Y_{ik}(v)^2 - \frac{\widetilde{\Theta}_{jk}(u,v)}{\sigma_j(u)\sigma_k(v)}.$$

Define that $Y_{ij}^*(u) = Y_{ij}(u) I\left\{ \|Y_{ij}\|_\infty \leqslant C_4 \log^{1/2}(p\vee n) \right\}$ and

$$Z_{ijk}^*(u,v) = Y_{ij}^*(u)^2 Y_{ik}^*(v)^2 - \mathbb{E}\{Y_{ij}^*(u)^2 Y_{ik}^*(v)^2\}.$$

By the property of ψ_1-norm and $|Y_{ij}^*(u)^2 - Y_{ij}^*(u')^2| \leqslant 2C_4 \log^{1/2}(p\vee n)|Y_{ij}^*(u) - Y_{ij}^*(u')|$, we have that

$$\begin{aligned}
&\left\|Y_{ij}^*(u)^2 Y_{ik}^*(v)^2 - Y_{ij}^*(u')^2 Y_{ik}^*(v')^2\right\|_{\psi_1} \\
&\leqslant \left\|Y_{ij}^*(u)^2\{Y_{ik}^*(v)^2 - Y_{ik}^*(v')^2\}\right\|_{\psi_1} + \left\|\{Y_{ij}^*(u)^2 - Y_{ij}^*(u')^2\}Y_{ik}^*(v')^2\right\|_{\psi_1} \\
&\lesssim \log(p\vee n)\Big\{ \|Y_{ij}^*(u)\|_{\psi_2} \left\|Y_{ik}^*(v) - Y_{ik}^*(v')\right\|_{\psi_2} + \|Y_{ik}^*(v')\|_{\psi_2} \left\|Y_{ij}^*(u) - Y_{ij}^*(u')\right\|_{\psi_2} \Big\} \\
&\lesssim \log(p\vee n)\{d_j(u,u') + d_k(v,v')\} \lesssim \log(p\vee n) d_{jk}((u,v),(u',v')),
\end{aligned}$$

which implies that, similar to (S.1),

$$\left\|Z^*_{ijk}(u,v) - Z^*_{ijk}(u',v')\right\|_{\psi_1} \lesssim \log(p\vee n) d_{jk}((u,v),(u',v')).$$

Let $\bar{Z}^*_{jk}(u,v) = n^{-1}\sum_{i=1}^n Z^*_{ijk}(u,v)$. We apply the similar technique of $\bar{Z}_{jk}$ above to the term $\bar{Z}^*_{jk}$ and obtain that there exists some universal constant $C_5 > 0$ such that for any $x > 0$,

$$P\left\{\max_{j,k}\sup_{u,v\in\mathcal{U}}\left|\frac{\bar{Z}^*_{jk}(u,v)}{\log(p\vee n)}\right| > x\right\} \lesssim p^2\exp(-C_5 n^{1/2}x).$$

As a result, for any $M > 0$, there exists some constant $\tilde{\rho}_2 > 0$ such that

$$P\left\{\max_{j,k}\sup_{u,v\in\mathcal{U}}\left|\bar{Z}^*_{jk}(u,v)\right| > \tilde{\rho}_2\frac{\log^2(p\vee n)}{n^{1/2}}\right\} \lesssim p^{-M}.$$

Now we consider the bound of the term $\|Y_{ij}\|_\infty$. By Conditions 1-2 and Theorem 8.4 of Kosorok (2008), we immediately have that there exists some constant $C_6 > 0$

$$\max_{1\leqslant i\leqslant n, 1\leqslant j\leqslant p}\left\|\sup_{u\in\mathcal{U}}|Y_{ij}(u)|\right\|_{\psi_2} \leqslant C_6,$$

which also implies that there exists some constant $C_7 > 0$ such that for any $x > 0$,

$$P\left\{\max_{1\leqslant i\leqslant n, 1\leqslant j\leqslant p}\|Y_{ij}(u)\|_\infty > x\right\} \lesssim np\exp(-C_7 x^2).$$

Hence we obtain that for any $M > 0$, there exists some constant $C_4 > 0$ such that

$$P\left\{\max_{1\leqslant i\leqslant n, 1\leqslant j\leqslant p}\|Y_{ij}\|_\infty > C_4\log^{1/2}(p\vee n)\right\} \lesssim (p\vee n)^{-M}. \tag{S.4}$$

On the event

$$\Omega_{n0} = \left\{\max_{1\leqslant i\leqslant n, 1\leqslant j\leqslant p}\|Y_{ij}\|_\infty \leqslant C_4\log^{1/2}(p\vee n)\right\},$$

we find that

$$\begin{aligned} Y_{ij}(u)^2Y_{ik}(v)^2 - \frac{\widetilde{\Theta}_{jk}(u,v)}{\sigma_j(u)\sigma_k(v)} = {} & Y^*_{ij}(u)^2Y^*_{ik}(v)^2 - \mathbb{E}\Big\{Y^*_{ij}(u)^2Y^*_{ik}(v)^2\Big\} \\ & + \mathbb{E}\Big\{Y^*_{ij}(u)^2Y^*_{ik}(v)^2 - Y_{ij}(u)^2Y_{ik}(v)^2\Big\}. \end{aligned}$$

Note that $Y^*_{ij}(u)^2 - Y_{ij}(u)^2 = Y_{ij}(u)^2 I\{\|Y_{ij}\|_\infty > C_4\log^{1/2}(p\vee n)\}$. By the inequality (S.4), we can obtain that

$$\left|\mathbb{E}\Big\{Y^*_{ij}(u)^2Y^*_{ik}(v)^2 - Y_{ij}(u)^2Y_{ik}(v)^2\Big\}\right| \lesssim (p\vee n)^{-M}.$$

Therefore, for any $M > 0$, there exist some constant $\tilde{\rho}_3 > 0$ such that

$$P\left\{\max_{1\leqslant j\leqslant p}\|H_{jk}^{(2)}\|_\infty > \tilde{\rho}_3\frac{\log^2(p\vee n)}{n^{1/2}}\right\} \lesssim p^{-M}. \tag{S.5}$$

Combining (S.3) and (S.5), we obtain that for any $M > 0$, there exists some constant $\rho_1 > 0$ such that

$$P\left\{\max_{j,k}\left\|\frac{\widehat{\Theta}_{jk}-\Theta_{jk}}{\Theta_{jk}}\right\|_\infty \geqslant \rho_1\frac{\log^2(p\vee n)}{n^{1/2}}\right\} \lesssim p^{-M}.$$

The proof is complete. ▫

Lemma A2 *Suppose that Conditions 1–4 hold. Then for any $M > 0$, there exist some constant $\rho_2 > 0$ such that*

$$\max_{j,k}\left\|\frac{\Theta_{jk}^{1/2}-\widehat{\Theta}_{jk}^{1/2}}{\widehat{\Theta}_{jk}^{1/2}}\right\|_\infty \leqslant \rho_2\frac{\log^2 p}{n^{1/2}} \tag{S.6}$$

with probability greater than $1 - O(p^{-M})$.

Proof. Let the event $\Omega_n(s) = \{\|(\widehat{\Theta}_{jk}-\Theta_{jk})/\Theta_{jk}\|_\infty \leqslant s\log^2 p/n^{1/2} \leqslant 1/2\}$. For any $M > 0$, it follows from Lemma A1 that there exists some constant $\rho_1 > 0$ such that $P\{\Omega_n(\rho_1)\} \geqslant 1 - O(p^{-M})$. Since

$$\left\|\frac{\Theta_{jk}}{\widehat{\Theta}_{jk}}\right\|_\infty = \left\|\frac{\Theta_{jk}-\widehat{\Theta}_{jk}}{\widehat{\Theta}_{jk}}+1\right\|_\infty \leqslant \left\|\frac{\Theta_{jk}-\widehat{\Theta}_{jk}}{\Theta_{jk}}\right\|_\infty\left\|\frac{\Theta_{jk}}{\widehat{\Theta}_{jk}}\right\|_\infty + 1,$$

hence, on the event $\Omega_n(\rho_1)$, we have that $\|\Theta_{jk}/\widehat{\Theta}_{jk}\|_\infty \leqslant 2$. As a result, on the event $\Omega_n(\rho_1)$, it follows that

$$\left\|\frac{\Theta_{jk}^{1/2}-\widehat{\Theta}_{jk}^{1/2}}{\widehat{\Theta}_{jk}^{1/2}}\right\|_\infty = \left\|\frac{\Theta_{jk}-\widehat{\Theta}_{jk}}{\widehat{\Theta}_{jk}+\widehat{\Theta}_{jk}^{1/2}\Theta_{jk}^{1/2}}\right\|_\infty \leqslant \left\|\frac{\Theta_{jk}-\widehat{\Theta}_{jk}}{\Theta_{jk}}\right\|_\infty\left\|\frac{\Theta_{jk}}{\widehat{\Theta}_{jk}}\right\|_\infty \leqslant 2\rho_1\frac{\log^2 p}{n^{1/2}}.$$

Take $\rho_2 = 2\rho_1$ and the proof is complete. ▫

Lemma A3 *Suppose that Conditions 1–4 holds. Then for any $M > 0$, there exist some positive constant $\rho_3 > 0$ such that*

$$\max_{j,k}\left\|\frac{\widehat{\Sigma}_{jk}-\Sigma_{jk}}{\widehat{\Theta}_{jk}^{1/2}}\right\|_{\mathcal{S}} \leqslant \rho_3\left(\frac{\log p}{n}\right)^{1/2}$$

with probability greater than $1 - O(p^{-M})$.

Proof. Let $\widetilde{Y}_{ijk}(u,v) = Y_{ij}(u)Y_{ik}(v) - \Sigma_{jk}(u,v)/\{\sigma_j(u)^{1/2}\sigma_k(v)^{1/2}\}$ and

$$\bar{Z}_{jk}(u,v) = \frac{\widehat{\Sigma}_{jk}(u,v) - \Sigma_{jk}(u,v)}{\sigma_j(u)^{1/2}\sigma_k(v)^{1/2}} = \frac{1}{n}\sum_{i=1}^{n}\widetilde{Y}_{ijk}(u,v).$$

We first derive the concentration bound of $\|\bar{Z}_{jk}\|_{\mathcal{S}}$. It follows from the proof of Lemma A1 that there exists some constant $C_8 > 0$ such that

$$\max_{j,k}\left\|\sup_{u,v\in\mathcal{U}}\left|\widetilde{Y}_{1jk}(u,v)\right|\right\|_{\psi_1} \leqslant C_8.$$

which further implies that $\max_{j,k}\left\|\left\|\widetilde{Y}_{1jk}\right\|_{\mathcal{S}}\right\|_{\psi_1} \leqslant C_8$. As a result, it follows from Theorem 2.5 of Bosq (2000) that there exists some universal constant $C_9 > 0$ such that for any $x > 0$

$$P\left(\left\|\bar{Z}_{jk}\right\|_{\mathcal{S}} \geqslant x\right) \leqslant 2\exp\{-C_9 n\min(x^2, x)\}.$$

For any $M > 0$, there exists some constant $\tilde{\rho} > 0$ that

$$\left\|\bar{Z}_{jk}\right\|_{\mathcal{S}} \leqslant \tilde{\rho}\left(\frac{\log p}{n}\right)^{1/2} \tag{S.7}$$

with probability greater than $1 - O(p^{-M})$.

Now we derive the bound of $\left\|\left(\widehat{\Sigma}_{jk} - \Sigma_{jk}\right)/\widehat{\Theta}_{jk}^{1/2}\right\|_{\mathcal{S}}$. Note that Condition 3 implies that $\Theta_{jk}(u,v) \geqslant \tau\sigma_j(u)\sigma_k(v)$. We obtain that

$$\left\|\frac{\widehat{\Sigma}_{jk} - \Sigma_{jk}}{\widehat{\Theta}_{jk}^{1/2}}\right\|_{\mathcal{S}} \leqslant \left\|\frac{\widehat{\Sigma}_{jk} - \Sigma_{jk}}{\Theta_{jk}^{1/2}}\right\|_{\mathcal{S}}\left\|\frac{\Theta_{jk}^{1/2}}{\widehat{\Theta}_{jk}^{1/2}}\right\|_{\infty} \leqslant \left\|\tau^{-1/2}\bar{Z}_{jk}\right\|_{\mathcal{S}}\left(\left\|\frac{\Theta_{jk}^{1/2} - \widehat{\Theta}_{jk}^{1/2}}{\widehat{\Theta}_{jk}^{1/2}}\right\|_{\infty} + 1\right).$$

Hence, together with (S.7) and Lemma A2, the lemma follows. The proof is complete. ▫

Proof of Theorem 1. For easy representation, define

$$\widehat{\Phi}_{jk}(u,v) = \frac{\widehat{\Sigma}_{jk}(u,v)}{\widehat{\Theta}_{jk}(u,v)^{1/2}}, \quad \widetilde{\Phi}_{jk}(u,v) = \frac{\Sigma_{jk}(u,v)}{\widehat{\Theta}_{jk}(u,v)^{1/2}} \quad \text{and} \quad \Phi_{jk}(u,v) = \frac{\Sigma_{jk}(u,v)}{\Theta_{jk}(u,v)^{1/2}}.$$

Let

$$\Omega_{n1} = \left\{\max_{j,k}\|\widehat{\Phi}_{jk} - \widetilde{\Phi}_{jk}\|_{\mathcal{S}} \leqslant \lambda\right\}, \Omega_{n2} = \left\{\max_{j,k}\left\|\frac{\widehat{\Theta}_{jk} - \Theta_{jk}}{\Theta_{jk}}\right\|_{\infty} \leqslant \frac{1}{2}\right\}.$$

It is immediate to see that under the event Ω_{n2}, $2^{-1}\|\Theta_{jk}\|_{\infty} \leqslant \|\widehat{\Theta}_{jk}\|_{\infty} \leqslant 2\|\Theta_{jk}\|_{\infty}$ for all j and k. By Conditions 1–3, we have $\Theta_{jk}(u,v) \leqslant C'\sigma_j(u)\sigma_k(v)$ and $\Theta_{jk}(u,v) \geqslant \tau\sigma_j(u)\sigma_k(v)$

Then under the event $\Omega_{n1} \cap \Omega_{n2}$ and Conditions (i)-(iii) on $S_\lambda(Z)$, we obtain that

$$
\begin{aligned}
&\sum_{k=1}^{p} \|\widehat{\Sigma}^{\mathrm{A}}_{jk} - \Sigma_{jk}\|_{\mathcal{S}} \\
&= \sum_{k=1}^{p} \|\widehat{\Sigma}^{\mathrm{A}}_{jk} - \Sigma_{jk}\|_{\mathcal{S}} I\{\|\widehat{\Phi}_{jk}\|_{\mathcal{S}} \geqslant \lambda\} + \sum_{k=1}^{p} \|\Sigma_{jk}\|_{\mathcal{S}} I\{\|\widehat{\Phi}_{jk}\|_{\mathcal{S}} < \lambda\} \\
&\leqslant \sum_{k=1}^{p} \Big\{ \|s_\lambda(\widehat{\Phi}_{jk}) - \widehat{\Phi}_{jk}\|_{\mathcal{S}} + \|\widehat{\Phi}_{jk} - \widetilde{\Phi}_{jk}\|_{\mathcal{S}} \Big\} \big\|\widehat{\Theta}^{1/2}_{jk}\big\|_\infty I\{\|\widehat{\Phi}_{jk}\|_{\mathcal{S}} \geqslant \lambda, \|\widetilde{\Phi}_{jk}\|_{\mathcal{S}} \geqslant \lambda\} \\
&\quad + \sum_{k=1}^{p} \Big\| \big[s_\lambda(\widehat{\Phi}_{jk}) - \widetilde{\Phi}_{jk}\big] \widehat{\Theta}^{1/2}_{jk} \Big\|_{\mathcal{S}} I\{\|\widehat{\Phi}_{jk}\|_{\mathcal{S}} \geqslant \lambda, \|\widetilde{\Phi}_{jk}\|_{\mathcal{S}} < \lambda\} + \sum_{k=1}^{p} \|\Sigma_{jk}\|_{\mathcal{S}} I\{\|\widetilde{\Phi}_{jk}\|_{\mathcal{S}} < 2\lambda\} \\
&\leqslant \sum_{k=1}^{p} 2\lambda \big\|\widehat{\Theta}^{1/2}_{jk}\big\|_\infty I\{\|\widetilde{\Phi}_{jk}\|_{\mathcal{S}} \geqslant \lambda\} + \sum_{k=1}^{p} (1+c) \|\widetilde{\Phi}_{jk}\|_{\mathcal{S}} \|\widehat{\Theta}^{1/2}_{jk}\|_\infty I\{\|\widetilde{\Phi}_{jk}\|_{\mathcal{S}} < \lambda\} \\
&\quad + \sum_{k=1}^{p} \|\widetilde{\Phi}_{jk}\|_{\mathcal{S}} \big\|\widehat{\Theta}^{1/2}_{jk}\big\|_\infty I\{\|\widetilde{\Phi}_{jk}\|_{\mathcal{S}} < 2\lambda\} \\
&\lesssim \lambda^{1-q} \sum_{k=1}^{p} \big\|\widehat{\Theta}_{jk}\big\|^{1/2}_\infty \|\widetilde{\Phi}_{jk}\|^q_{\mathcal{S}} \lesssim \lambda^{1-q} \sum_{k=1}^{p} \big\|\sigma_j\big\|^{(1-q)/2}_\infty \big\|\sigma_k\big\|^{(1-q)/2}_\infty \|\Sigma_{jk}\|^q_{\mathcal{S}} \lesssim s_0(p) \left(\frac{\log p}{n}\right)^{\frac{1-q}{2}}.
\end{aligned}
$$

Since there exists some constant $\delta > 0$ such that $P\{\Omega^C_{n1}\} + P\{\Omega^C_{n2}\} \lesssim p^{-M}$, the theorem follows. ▫

Proof of Theorem 2. We consider two sets: $S_{n1} = \{(j,k) : \|\widehat{\Sigma}^{\mathrm{A}}_{jk}\|_{\mathcal{S}} \neq 0 \text{ and } \|\Sigma_{jk}\|_{\mathcal{S}} = 0\}$ and $S_{n2} = \{(j,k) : \|\widehat{\Sigma}^{\mathrm{A}}_{jk}\|_{\mathcal{S}} = 0 \text{ and } \|\Sigma_{jk}\|_{\mathcal{S}} \neq 0\}$. It suffices to prove that

$$P\big(|S_{n1}| > 0\big) + P\big(|S_{n2}| > 0\big) \to 0,$$

as $n, p \to \infty$. By Conditions (i)-(iii) on $S_\lambda(Z)$,

$$S_{n1} = \left\{ (j,k) : \left\| \frac{\widehat{\Sigma}_{jk}}{\widehat{\Theta}^{1/2}_{jk}} \right\|_{\mathcal{S}} > \lambda \text{ and } \|\Sigma_{jk}\|_{\mathcal{S}} = 0 \right\} \subset \left\{ (j,k) : \left\| \frac{\widehat{\Sigma}_{jk} - \Sigma_{jk}}{\widehat{\Theta}^{1/2}_{jk}} \right\|_{\mathcal{S}} > \lambda \right\}$$

Therefore, with the choice $\lambda = \delta(\log p/n)^{1/2}$, we obtain

$$P(|S_{n1}| > 0) \leqslant P\left\{ \max_{j,k} \left\| \frac{\widehat{\Sigma}_{jk} - \Sigma_{jk}}{\widehat{\Theta}^{1/2}_{jk}} \right\|_{\mathcal{S}} > \lambda \right\} \lesssim p^{-M}. \tag{S.8}$$

for some prespecified $M > 0$. Similarly, we have

$$S_{n2} = \left\{ (j,k) : \left\| \frac{\widehat{\Sigma}_{jk}}{\widehat{\Theta}^{1/2}_{jk}} \right\|_{\mathcal{S}} \leqslant \lambda \text{ and } \|\Sigma_{jk}\|_{\mathcal{S}} \neq 0 \right\}.$$

Note that $\|\Sigma_{jk}\|_{\mathcal{S}} \neq 0$ implies that

$$(2\delta+\gamma)\left(\frac{\log p}{n}\right)^{1/2} < \left\|\frac{\Sigma_{jk}}{\Theta_{jk}^{1/2}}\right\|_{\mathcal{S}} \leqslant \left[\left\|\frac{\Sigma_{jk}-\widehat{\Sigma}_{jk}}{\widehat{\Theta}_{jk}^{1/2}}\right\|_{\mathcal{S}} + \left\|\frac{\widehat{\Sigma}_{jk}}{\widehat{\Theta}_{jk}^{1/2}}\right\|_{\mathcal{S}}\right]\left\|\frac{\widehat{\Theta}_{jk}^{1/2}}{\Theta_{jk}^{1/2}}\right\|_{\infty}. \tag{S.9}$$

Let $\Omega_{n3} = \left\{\|(\widehat{\Theta}_{jk}^{1/2}-\Theta_{jk}^{1/2})/\widehat{\Theta}_{jk}^{1/2}\|_{\infty} \leqslant \epsilon\right\}$ for some small constant $0 < \epsilon < \gamma/(4\delta+2\gamma)$. Conditioned on the event of Ω_{n3}, the inequality

$$\left\|\frac{\widehat{\Theta}_{jk}^{1/2}}{\Theta_{jk}^{1/2}}\right\|_{\infty} \leqslant \left\|\frac{\widehat{\Theta}_{jk}^{1/2}-\Theta_{jk}^{1/2}}{\widehat{\Theta}_{jk}^{1/2}}\right\|_{\infty}\left\|\frac{\widehat{\Theta}_{jk}^{1/2}}{\Theta_{jk}^{1/2}}\right\|_{\infty} + 1$$

implies that $\|\widehat{\Theta}_{jk}^{1/2}/\Theta_{jk}^{1/2}\|_{\infty} \leqslant 1/(1-\epsilon)$. This together with (S.9) shows that

$$S_{n2} \cap \Omega_{n3} \subset \left\{(j,k) : \left\|\frac{\widehat{\Sigma}_{jk}-\Sigma_{jk}}{\widehat{\Theta}_{jk}^{1/2}}\right\|_{\mathcal{S}} > \delta\left(\frac{\log p}{n}\right)^{1/2}\right\}.$$

As a result,

$$P(|S_{n2}|>0) \leqslant P(\Omega_{n3}^{C}) + P\left\{\max_{j,k}\left\|\frac{\widehat{\Sigma}_{jk}-\Sigma_{jk}}{\widehat{\Theta}_{jk}^{1/2}}\right\|_{\mathcal{S}} > \delta\left(\frac{\log p}{n}\right)^{1/2}\right\} \lesssim p^{-M}. \tag{S.10}$$

Combining (S.8) and (S.10), we complete our proof. ▫

B Examples of functional thresholding operators

In Section B.1, we verify that our proposed soft, SCAD and adaptive lasso functional thresholding rules satisfy conditions (i)–(iii) in Section 2. We then present the derivations of these three functional thresholding rules in Section B.2.

B.1 Verification of conditions (i)–(iii)

It is directly implied from the thresholding rules that the soft, SCAD and adaptive lasso functional methods satisfy condition (ii). Since the soft functional thresholding has the largest amount of functional shrinkage in the Hilbert–Schmidt norm compared with SCAD and adaptive lasso methods, it suffices to show that the soft functional thresholding satisfies condition (iii). For $\|Z\|_{\mathcal{S}} \leqslant \lambda$, the thresholding effect leads to $\|0-Z\|_{\mathcal{S}} \leqslant \lambda$. When $\|Z\|_{\mathcal{S}} > \lambda$, we obtain that $\|Z\lambda/\|Z\|_{\mathcal{S}}\|_{\mathcal{S}} = \lambda$.

We next show that the above three thresholding methods satisfy condition (i). By the triangle inequality, $\|Z - Y\|_{\mathcal{S}} \leqslant \lambda$ in condition (i) implies that $\big|\|Z\|_{\mathcal{S}} - \lambda\big| \leqslant \|Y\|_{\mathcal{S}}$.

- Soft functional thresholding: If $\|Z\|_{\mathcal{S}} \leqslant \lambda$, $0 \leqslant c\|Y\|_{\mathcal{S}}$ directly holds for all $Y \in \mathbb{S}$ and $c > 0$. When $\|Z\|_{\mathcal{S}} > \lambda$, we have $\|s_\lambda^{\mathrm{S}}(Z)\|_{\mathcal{S}} = \|Z\|_{\mathcal{S}} - \lambda \leqslant \|Y\|_{\mathcal{S}}$ with the choice of $c = 1$.

- SCAD functional thresholding: When $\|Z\|_{\mathcal{S}} \leqslant 2\lambda$, $s_\lambda^{\mathrm{SC}}(Z)$ is the same as the soft functional thresholding rule. For $\|Z\|_{\mathcal{S}} > 2\lambda$, we have $\|s_\lambda^{\mathrm{SC}}(Z)\|_{\mathcal{S}} \leqslant \|Z\|_{\mathcal{S}} \leqslant \|Y\|_{\mathcal{S}} + \lambda \leqslant \|Y\|_{\mathcal{S}} + \|Z\|_{\mathcal{S}}/2$ and hence $\|s_\lambda^{\mathrm{SC}}(Z)\|_{\mathcal{S}} \leqslant \|Z\|_{\mathcal{S}} \leqslant 2\|Y\|_{\mathcal{S}}$. Combining the above results, we take $c = 2$.

- Adaptive lasso functional thresholding: Let $\lceil \eta \rceil$ denote the smallest integer greater than or equal to η. For $\|Z\|_{\mathcal{S}} \leqslant \lambda$, this condition holds for all $Y \in \mathbb{S}$ and $c > 0$. For $\|Z\|_{\mathcal{S}} > \lambda$, we have that $\|s_\lambda^{\mathrm{AL}}(Z)\|_{\mathcal{S}} = \|Z(1 - \lambda^{\eta+1}/\|Z\|_{\mathcal{S}}^{\eta+1})\|_{\mathcal{S}} = (\|Z\|_{\mathcal{S}}^{\eta+1} - \lambda^{\eta+1})/\|Z\|_{\mathcal{S}}^{\eta} \leqslant (\|Z\|_{\mathcal{S}}^{\lceil\eta\rceil+1} - \lambda^{\lceil\eta\rceil+1})/\|Z\|_{\mathcal{S}}^{\lceil\eta\rceil} = (\|Z\|_{\mathcal{S}} - \lambda)(\|Z\|_{\mathcal{S}}^{\lceil\eta\rceil} + \|Z\|_{\mathcal{S}}^{\lceil\eta\rceil-1}\lambda + \cdots + \lambda^{\lceil\eta\rceil})/\|Z\|_{\mathcal{S}}^{\lceil\eta\rceil} \leqslant (\lceil\eta\rceil + 1)\|Y\|_{\mathcal{S}}$. Hence, for any $\eta \geqslant 0$, we can find $c = \lceil\eta\rceil + 1$. In the special case of $\eta = 0$, $s_\lambda^{\mathrm{AL}}(Z)$ degenerates to the soft functional thresholding rule with $c = 1$, which is consistent with our finding for the soft functional thresholding.

B.2 Derivations of the functional thresholding rules from various penalty functions

Soft functional thresholding can be obtained via

$$s_\lambda^{\mathrm{S}}(Z) = \arg\min_{\theta \in \mathbb{S}} \left\{ \frac{1}{2}\|\theta - Z\|_{\mathcal{S}}^2 + \lambda\|\theta\|_{\mathcal{S}} \right\}. \tag{S.11}$$

First, we show that if $\|Z\|_{\mathcal{S}} \leqslant \lambda$, then $\|s_\lambda^{\mathrm{S}}(Z)\|_{\mathcal{S}} = 0$ and hence $s_\lambda^{\mathrm{S}}(Z) = 0$. This results from the fact that, for any θ,

$$\begin{aligned}\frac{1}{2}\|\theta - Z\|_{\mathcal{S}}^2 + \lambda\|\theta\|_{\mathcal{S}} &\geqslant \frac{1}{2}\big(\|\theta\|_{\mathcal{S}} - \|Z\|_{\mathcal{S}}\big)^2 + \lambda\|\theta\|_{\mathcal{S}} \\ &= \frac{1}{2}\|\theta\|_{\mathcal{S}}^2 + (\lambda - \|Z\|_{\mathcal{S}})\|\theta\|_{\mathcal{S}} + \frac{1}{2}\|Z\|_{\mathcal{S}}^2 \geqslant \frac{1}{2}\|Z\|_{\mathcal{S}}^2.\end{aligned}$$

Second, we show that if $\|Z\|_{\mathcal{S}} > \lambda$, then $\|s_\lambda^{\mathrm{s}}(Z)\|_{\mathcal{S}} \neq 0$. In fact, we can find $\theta_c = cZ$ with $c = 1 - \lambda/\|Z\|_{\mathcal{S}} > 0$ such that

$$\frac{1}{2}\|\theta_c - Z\|_{\mathcal{S}}^2 + \lambda\|\theta_c\|_{\mathcal{S}} = \frac{1}{2}(1-c)^2\|Z\|_{\mathcal{S}}^2 + \lambda c\|Z\|_{\mathcal{S}} < \frac{1}{2}\|Z\|_{\mathcal{S}}^2.$$

As a result, we are able to take the first derivative of (S.11) with respect to θ and set $p'_\lambda(\theta) = \theta - Z + \lambda\theta/\|\theta\|_{\mathcal{S}} = 0$. Thus, $\widehat{\theta} = Z\|\widehat{\theta}\|_{\mathcal{S}}/\big(\|\widehat{\theta}\|_{\mathcal{S}} + \lambda\big)$, which implies that $\|\widehat{\theta}\|_{\mathcal{S}} = \|Z\|_{\mathcal{S}} - \lambda$. Combining the above results, we have that $\widehat{\theta} = Z(1 - \lambda/\|Z\|_{\mathcal{S}})_+$.

The SCAD and adaptive lasso functional thresholding rules can be derived in a similar fashion. Hence, we only present their penalty functions here. The functional version of SCAD penalty takes the form of

$$p_\lambda(\theta) = \lambda\|\theta\|_{\mathcal{S}} I(\|\theta\|_{\mathcal{S}} \leqslant \lambda) + \frac{2a\lambda\|\theta\|_{\mathcal{S}} - \|\theta\|_{\mathcal{S}}^2 - \lambda^2}{2(a-1)} I(\lambda < \|\theta\|_{\mathcal{S}} \leqslant a\lambda) + \frac{\lambda^2(a+1)}{2} I(\|\theta\|_{\mathcal{S}} > a\lambda),$$

for $a > 2$. For the functional version of adaptive lasso penalty, we use $p_\lambda(\theta) = \lambda^{\eta+1}\|Z\|_{\mathcal{S}}^{-\eta}\|\theta\|_{\mathcal{S}}$, for $\eta \geqslant 0$. A similar adaptive lasso penalty function operating on $|\cdot|$ for the univariate scalar case can be found in Rothman et al. (2009).

C Further discussion

C.1 Supremum-norm-based class of functional thresholding operators

In general, conditions (i)–(iii) are satisfied by a number of solutions to (1), where the presence of $\|\cdot\|_{\mathcal{S}}$ in both the loss and various penalty functions leads to the solutions as functions of $\|Z\|_{\mathcal{S}}$. Such connection demonstrates the rationale of imposing Hilbert–Schmidt-norm based conditions (i)–(iii). For examples of functional data with some local spikes, one may suggest another class of functional thresholding operators $\tilde{s}_\lambda(Z)$ satisfying three supremum-norm based conditions analogous to conditions (i)–(iii), where, for any $Q \in \mathbb{S}$, we denote its supremum norm by $\|Q\|_\infty = \sup_{u,v\in\mathcal{U}}|Q(u,v)|$. In this case, $\tilde{s}_\lambda(Z)$

can not be directly derived as the solution to (1) with $p_\lambda(\theta) = \tilde{p}_\lambda(\|\theta\|_\infty)$. However, by substituting $\|\cdot\|_{\mathcal{S}}$ in $s_\lambda^{\mathrm{S}}(Z), s_\lambda^{\mathrm{SC}}(Z)$ and $s_\lambda^{\mathrm{AL}}(Z)$ with $\|\cdot\|_\infty$, the corresponding supremum-norm based functional thresholding rules can be presented and checked to satisfy three conditions for $\tilde{s}_\lambda(Z)$ in a similar fashion. To study theoretical properties analogous to Theorems 1 and 2 in Section 3, the main challenge is to establish concentration bounds on some standardized processes in the supremum norm, where our tools and results in Section A of the Supplementary Material can be applied accordingly. In this regard, the $\|\cdot\|_{\mathcal{S}}$ that we adopt in $s_\lambda(Z)$ is not necessarily the unique choice, but serves as the building block for the sparse covariance function estimation problem.

C.2 Additional applications

The fourth interesting application considers estimating *functional graphical models* targeting at identifying the conditional dependence structure among components in $\mathbf{X}_i(\cdot)$. Qiao et al. (2019) proposed to estimate a block sparse inverse covariance matrix by treating dimensions of $X_{ij}(\cdot)$'s as approaching infinity. However, to deal with truly infinite-dimensional objects, it is desirable to avoid the estimation of the unbounded inverse of $\mathbf{\Sigma}$. For Gaussian graphical models, an innovative transformation (Fan and Lv, 2016) converts the problem of estimating sparse inverse covariance matrix to that of sparse covariance matrix estimation. It is interesting to generalize this transformation strategy to the functional domain and hence our sparse covariance function estimation approach can be applied.

The fifth potential application considers the functional classification problem when estimating the covariance function plays a key role, e.g., functional linear discriminant analysis to classify univariate functional data (Park et al., 2021). One natural way to deal with classification for multivariate functional data $\{\mathbf{X}_i(\cdot)\}_{i=1}^n$ is to concatenate multiple functions directly and then generalize the univariate functional classification methods to the multivariate setting by making use of the estimation of $\mathbf{\Sigma}$. When p is large, it is thus of interest to incorporate the proposed sparse covariance function estimation framework into

the development of the classification approach for a large bundle of curves.

D Partially observed functional data

Section D.1 gives the expression of the local linear surface smoother for the cross-covariance estimation. Section D.2 presents the details of pre-smoothing for densely sampled functional data. Section D.3 provides all technical proofs for the partially observed functional scenario. Section D.4 presents the heuristic verification of I_{jk} in (12) and Condition 8.

D.1 Local linear surface smoother

We use (7) to derive the expression of its minimizer. Recall $T_{ab,ijk}$ and $S_{ab,jk}$ in (8) and (10), respectively, for $a, b = 0, 1, 2$, $i = 1, \ldots, n$ and $j, k = 1, \ldots, p$. To minimize the objective in (7), some calculations lead to the resulting estimator

$$\begin{aligned}\widehat{\Sigma}_{jk} =& \sum_{i=1}^{n} \frac{(S_{20}S_{02} - S_{11}^2)T_{00,ijk} - (S_{10}S_{02} - S_{01}S_{11})T_{10,ijk} + (S_{10}S_{11} - S_{01}S_{20})T_{01,ijk}}{(S_{20}S_{02} - S_{11}^2)S_{00} - (S_{10}S_{02} - S_{01}S_{11})S_{10} + (S_{10}S_{11} - S_{01}S_{20})S_{01}} \\ :=& \sum_{i=1}^{n} \left(W_{1,jk}T_{00,ijk} + W_{2,jk}T_{10,ijk} + W_{3,jk}T_{01,ijk}\right),\end{aligned} \tag{S.12}$$

where we drop subscripts j, k in $S_{ab,jk}$'s to simplify the notation. Note that, under Model (19), $S_{ab,jk}$'s no longer depend on j, k, and hence subscripts j, k in $S_{ab,jk}$'s can be dropped.

D.2 Pre-smoothing

When each random function $X_{ij}(\cdot)$ is densely observed with errors satisfying Model (6), the commonly adopted pre-smoothing approach applies local linear smoother to estimate each $X_{ij}(\cdot)$ before subsequent analysis. The reconstructed individual function is obtained by $\widehat{X}_{ij}(u) = \hat{a}_0$, where

$$(\hat{a}_0, \hat{a}_1) = \underset{a_0, a_1}{\operatorname{argmin}} \sum_{l=1}^{L_{ij}} \left\{Z_{ijl} - a_0 - a_1(U_{ijl} - u)\right\}^2 K_{h_X}(U_{il} - u).$$

Let $T_{a,ij}(u) = \sum_{l=1}^{L_{ij}} K_{h_X}(U_{ijl} - u)(U_{ijl} - u)^a Z_{ijl}$ and $S_{a,ij}(u) = \sum_{l=1}^{L_{ij}} K_{h_X}(U_{ijl} - u)(U_{ijl} - u)^b$ for $a = 0, 1, 2$. Solving the minimization problem above yields that

$$\widehat{X}_{ij}(u) = \frac{S_{2,ij}(u)T_{0,ij}(u) - S_{1,ij}(u)T_{1,ij}(u)}{S_{2,ij}(u)S_{0,ij}(u) - \{S_{1,ij}(u)\}^2}.$$

Under the simplified model in (19), we drop the subscript j in L_{ij} and $S_{a,ij}$ in the expression of $\widehat{X}_{ij}(u)$ above. For an equally-spaced grid of R points $u_1 < \cdots < u_R \in \mathcal{U}$, the binned approximation of $\widehat{X}_{ij}(u)$ is

$$\breve{X}_{ij}(u) = \frac{\breve{S}_{2,i}(u)\breve{T}_{0,ij}(u) - \breve{S}_{1,i}(u)\breve{T}_{1,ij}(u)}{\breve{S}_{2,i}(u)\breve{S}_{0,i}(u) - \{\breve{S}_{1,i}(u)\}^2},$$

where $\breve{T}_{a,ij}(u) = \sum_{r=1}^{R} K_{h_X}(u_r - u)(u_r - u)^a \mathcal{D}_{r,ij}$ and $\breve{S}_{a,i}(u) = \sum_{r=1}^{R} K_{h_X}(u_r - u)(u_r - u)^a \varpi_{r,i}$. See also Table 7 for the computational complexity analysis of the pre-smoothing based on local linear smoother and its binned implementation, denoted as LLS-P and BinLLS-P respectively, under Models (6) and (19).

Table 7: The computational complexity analysis of LLS- and BinLLS-based pre-smoothings under Models (6) and (19) when evaluating the reconstructed functions at a grid of R points.

Method	Model	Number of kernel evaluations	Number of operations (additions and multiplications)
LLS-P	(6)	$O(R\sum_{i=1}^{n}\sum_{j=1}^{p} L_{ij})$	$O(R\sum_{i=1}^{n}\sum_{j=1}^{p} L_{ij})$
LLS-P	(19)	$O(R\sum_{i=1}^{n} L_i)$	$O(pR\sum_{i=1}^{n} L_i)$
BinLLS-P	(19)	$O(R)$	$O(npR^2 + p\sum_{i=1}^{n} L_i)$

D.3 Technical proofs

Proof of Theorem 3. Define

$$\widetilde{\Lambda}_{jk}(u,v) = \frac{\widetilde{\Sigma}_{jk}(u,v)}{\widetilde{\Psi}_{jk}(u,v)^{1/2}}, \quad \breve{\Lambda}_{jk}(u,v) = \frac{\Sigma_{jk}(u,v)}{\widetilde{\Psi}_{jk}(u,v)^{1/2}} \quad \text{and} \quad \Lambda_{jk}(u,v) = \frac{\Sigma_{jk}(u,v)}{\Psi_{jk}(u,v)^{1/2}}.$$

Let

$$\widetilde{\Omega}_{n1} = \left\{ \max_{j,k} \|\widetilde{\Lambda}_{jk} - \breve{\Lambda}_{jk}\|_{\mathcal{S}} \leqslant \lambda \right\}, \quad \widetilde{\Omega}_{n2} = \left\{ \max_{j,k} \left\| \frac{\widetilde{\Psi}_{jk} - \Psi_{jk}}{\Psi_{jk}} \right\|_{\infty} \leqslant \frac{1}{2} \right\}.$$

First, we can obtain from Condition 8 that $P(\widetilde{\Omega}_{n2}^C) = o(1)$. Note that

$$\left\|\frac{\widetilde{\Sigma}_{jk}-\Sigma_{jk}}{\widetilde{\Psi}_{jk}^{1/2}}\right\|_{\mathcal{S}} \leqslant \left\|\frac{\widetilde{\Sigma}_{jk}-\Sigma_{jk}}{\Psi_{jk}^{1/2}}\right\|_{\mathcal{S}} \left\|\frac{\Psi_{jk}^{1/2}}{\widetilde{\Psi}_{jk}^{1/2}}\right\|_{\infty} \lesssim \left\|\widetilde{\Sigma}_{jk}-\Sigma_{jk}\right\|_{\mathcal{S}} \left(\left\|\frac{\Psi_{jk}^{1/2}-\widetilde{\Psi}_{jk}^{1/2}}{\widetilde{\Psi}_{jk}^{1/2}}\right\|_{\infty}+1\right).$$

It follows from Condition 7 that there exists some constant $\tilde{\delta} > 0$ such that $P\{(\widetilde{\Omega}_{n1})^C\} = o(1)$. We also can see that under the event $\widetilde{\Omega}_{n2}$, $2^{-1}\|\Psi_{jk}\|_\infty \leqslant \|\widetilde{\Psi}_{jk}\|_\infty \leqslant 2\|\Psi_{jk}\|_\infty$ for all j and k. Then on the event $\widetilde{\Omega}_{n1} \cap \widetilde{\Omega}_{n2}$ and Conditions (i)-(iii) on $S_\lambda(Z)$, we obtain that

$$\begin{aligned}
&\sum_{k=1}^{p} \|\widetilde{\Sigma}_{jk}^{\mathrm{A}} - \Sigma_{jk}\|_{\mathcal{S}} \\
&= \sum_{k=1}^{p} \|\widetilde{\Sigma}_{jk}^{\mathrm{A}} - \Sigma_{jk}\|_{\mathcal{S}} I\{\|\widetilde{\Lambda}_{jk}\|_{\mathcal{S}} \geqslant \lambda\} + \sum_{k=1}^{p} \|\Sigma_{jk}\|_{\mathcal{S}} I\{\|\widetilde{\Lambda}_{jk}\|_{\mathcal{S}} < \lambda\} \\
&\leqslant \sum_{k=1}^{p} \left\{\|s_\lambda(\widetilde{\Lambda}_{jk}) - \widetilde{\Lambda}_{jk}\|_{\mathcal{S}} + \|\widetilde{\Lambda}_{jk} - \check{\Lambda}_{jk}\|_{\mathcal{S}}\right\} \|\widetilde{\Psi}_{jk}^{1/2}\|_\infty I\{\|\widetilde{\Lambda}_{jk}\|_{\mathcal{S}} \geqslant \lambda, \|\check{\Lambda}_{jk}\|_{\mathcal{S}} \geqslant \lambda\} \\
&\quad + \sum_{k=1}^{p} \left\|\left[s_\lambda(\widetilde{\Lambda}_{jk}) - \check{\Lambda}_{jk}\right]\widetilde{\Psi}_{jk}^{1/2}\right\|_{\mathcal{S}} I\{\|\widetilde{\Lambda}_{jk}\|_{\mathcal{S}} \geqslant \lambda, \|\check{\Lambda}_{jk}\|_{\mathcal{S}} < \lambda\} + \sum_{k=1}^{p} \|\Sigma_{jk}\|_{\mathcal{S}} I\{\|\check{\Lambda}_{jk}\|_{\mathcal{S}} < 2\lambda\} \\
&\leqslant \sum_{k=1}^{p} 2\lambda \|\widetilde{\Psi}_{jk}^{1/2}\|_\infty I\{\|\check{\Lambda}_{jk}\|_{\mathcal{S}} \geqslant \lambda\} + \sum_{k=1}^{p} (1+c)\|\check{\Lambda}_{jk}\|_{\mathcal{S}} \|\widetilde{\Psi}_{jk}^{1/2}\|_\infty I\{\|\check{\Lambda}_{jk}\|_{\mathcal{S}} < \lambda\} \\
&\quad + \sum_{k=1}^{p} \|\check{\Lambda}_{jk}\|_{\mathcal{S}} \|\widetilde{\Psi}_{jk}^{1/2}\|_\infty I\{\|\check{\Lambda}_{jk}\|_{\mathcal{S}} < 2\lambda\} \\
&\lesssim \lambda^{1-q} \sum_{k=1}^{p} \|\widetilde{\Psi}_{jk}\|_\infty^{1/2} \|\check{\Lambda}_{jk}\|_{\mathcal{S}}^q \lesssim \lambda^{1-q} \sum_{k=1}^{p} \|\Psi_{jk}\|_\infty^{(1-q)/2} \|\Sigma_{jk}\|_{\mathcal{S}}^q \lesssim \tilde{s}_0(p) \left(\frac{\log p}{n^{2\gamma_1}}\right)^{\frac{1-q}{2}}.
\end{aligned}$$

Theorem 3 follows. ▫

Proof of Theorem 4. Consider two sets: $\widetilde{S}_{n1} = \{(j,k) : \|\widetilde{\Sigma}_{jk}^{\mathrm{A}}\|_{\mathcal{S}} \neq 0 \text{ and } \|\Sigma_{jk}\|_{\mathcal{S}} = 0\}$ and $\widetilde{S}_{n2} = \{(j,k) : \|\widetilde{\Sigma}_{jk}^{\mathrm{A}}\|_{\mathcal{S}} = 0 \text{ and } \|\Sigma_{jk}\|_{\mathcal{S}} \neq 0\}$. It suffices to prove that

$$P\big(|\widetilde{S}_{n1}| > 0\big) + P\big(|\widetilde{S}_{n2}| > 0\big) \to 0,$$

as $n, p \to \infty$. By Conditions (i)-(iii) on $S_\lambda(Z)$,

$$\widetilde{S}_{n1} = \left\{(j,k) : \left\|\frac{\widetilde{\Sigma}_{jk}}{\widetilde{\Psi}_{jk}^{1/2}}\right\|_{\mathcal{S}} > \lambda \text{ and } \|\Sigma_{jk}\|_{\mathcal{S}} = 0\right\} \subset \left\{(j,k) : \left\|\frac{\widetilde{\Sigma}_{jk}-\Sigma_{jk}}{\widetilde{\Psi}_{jk}^{1/2}}\right\|_{\mathcal{S}} > \lambda\right\}$$

Therefore, with the choice $\lambda = \tilde{\delta}(\log p/n^{2\gamma_1})^{1/2}$, we obtain

$$P(|\widetilde{S}_{n1}| > 0) \leqslant P\left\{\max_{j,k} \left\|\frac{\widetilde{\Sigma}_{jk}-\Sigma_{jk}}{\widetilde{\Psi}_{jk}^{1/2}}\right\|_{\mathcal{S}} > \lambda\right\} = o(1), \tag{S.13}$$

as stated in the proof of Theorem 3. Similarly, we have

$$\widetilde{S}_{n2} = \left\{ (j,k) : \left\| \frac{\widetilde{\Sigma}_{jk}}{\widetilde{\Psi}_{jk}^{1/2}} \right\|_{\mathcal{S}} \leqslant \lambda \text{ and } \|\Sigma_{jk}\|_{\mathcal{S}} \neq 0 \right\}.$$

Note that $\|\Sigma_{jk}\|_{\mathcal{S}} \neq 0$ implies that

$$(2\tilde{\delta} + \tilde{\gamma}) \left(\frac{\log p}{n^{2\gamma_1}} \right)^{1/2} < \left\| \frac{\Sigma_{jk}}{\Psi_{jk}^{1/2}} \right\|_{\mathcal{S}} \leqslant \left[\left\| \frac{\Sigma_{jk} - \widetilde{\Sigma}_{jk}}{\widetilde{\Psi}_{jk}^{1/2}} \right\|_{\mathcal{S}} + \left\| \frac{\widetilde{\Sigma}_{jk}}{\widetilde{\Psi}_{jk}^{1/2}} \right\|_{\mathcal{S}} \right] \left\| \frac{\widetilde{\Psi}_{jk}^{1/2}}{\Psi_{jk}^{1/2}} \right\|_{\infty}. \tag{S.14}$$

Let $\widetilde{\Omega}_{n3} = \left\{ \|(\widetilde{\Psi}_{jk}^{1/2} - \Psi_{jk}^{1/2})/\widetilde{\Psi}_{jk}^{1/2}\|_{\infty} \leqslant \tilde{\epsilon} \right\}$ for some small constant $0 < \tilde{\epsilon} < \tilde{\gamma}/(4\tilde{\delta} + 2\tilde{\gamma})$. By Condition 8, $P\{(\widetilde{\Omega}_{n3})^C\} = o(1)$. Conditioning on the event of $\widetilde{\Omega}_{n3}$, we can see that $\|\widetilde{\Psi}_{jk}^{1/2}/\Psi_{jk}^{1/2}\|_{\infty} \leqslant 1/(1-\tilde{\epsilon})$. This together with (S.14) shows that

$$\widetilde{S}_{n2} \cap \widetilde{\Omega}_{n3} \subset \left\{ (j,k) : \left\| \frac{\widetilde{\Sigma}_{jk} - \Sigma_{jk}}{\widetilde{\Psi}_{jk}^{1/2}} \right\|_{\mathcal{S}} > \tilde{\delta} \left(\frac{\log p}{n^{2\gamma_1}} \right)^{1/2} \right\}.$$

As a result,

$$P(|\widetilde{S}_{n2}| > 0) \leqslant P\{(\widetilde{\Omega}_{n3})^C\} + P\left\{ \max_{j,k} \left\| \frac{\widetilde{\Sigma}_{jk} - \Sigma_{jk}}{\widetilde{\Psi}_{jk}^{1/2}} \right\|_{\mathcal{S}} > \tilde{\delta} \left(\frac{\log p}{n^{2\gamma_1}} \right)^{1/2} \right\} = o(1). \tag{S.15}$$

Combining (S.13) and (S.15), we complete our proof. ▫

D.4 Heuristic verification of I_{jk} in (12) and Condition 8

In this section we provide the heuristic verification of I_{jk} in (12) and Condition 8 as their detailed proofs are not only long and challenging but also largely deviate from the current focus of the paper.

Recall that

$$\widetilde{\Psi}_{jk} = I_{jk} \sum_{i=1}^{n} \left(W_{1,jk} V_{00,ijk} + W_{2,jk} V_{10,ijk} + W_{3,jk} V_{01,ijk} \right)^2,$$

where, for $a, b = 0, 1, 2$,

$$V_{ab,ijk}(u,v) = \sum_{i=1}^{L_{ij}} \sum_{m=1}^{L_{ik}} g_{ab}\big(h_C, (u,v), (U_{ijl}, U_{ikm})\big) \big\{ Z_{ijl} Z_{ikm} - \widetilde{\Sigma}_{jk}(u,v) \big\},$$
$$g_{ab}\big\{h, (u,v), (U_{ijl}, U_{ikm})\big\} = K_h(U_{ijl} - u) K_h(U_{ikm} - v) (U_{ijl} - u)^a (U_{ikm} - v)^b.$$

The expression of $\widetilde{\Psi}_{jk}$ in (11) can be decomposed as

$$\begin{aligned}\widetilde{\Psi}_{jk} &= I_{jk}W_{1,jk}^2\sum_{i=1}^{n}V_{00,ijk}^2 + I_{jk}W_{2,jk}^2\sum_{i=1}^{n}V_{10,ijk}^2 + I_{jk}W_{3,jk}^2\sum_{i=1}^{n}V_{01,ijk}^2 \\ &\quad +2I_{jk}W_{1,jk}W_{2,jk}\sum_{i=1}^{n}V_{00,ijk}V_{10,ijk} + 2I_{jk}W_{1,jk}W_{3,jk}\sum_{i=1}^{n}V_{00,ijk}V_{01,ijk} \\ &\quad +2I_{jk}W_{2,jk}W_{3,jk}\sum_{i=1}^{n}V_{10,ijk}V_{01,ijk} \\ &= \widetilde{\Psi}_{jk}^{(1)} + \widetilde{\Psi}_{jk}^{(2)} + \cdots + \widetilde{\Psi}_{jk}^{(5)} + \widetilde{\Psi}_{jk}^{(6)}. \end{aligned} \tag{S.16}$$

We first focus on the term $\widetilde{\Psi}_{jk}^{(1)}$. For $a, b = 0, 1, 2$, define

$$\begin{aligned}V_{ab,ijk}^{(1)}(u,v) &= \sum_{l=1}^{L_{ij}}\sum_{m=1}^{L_{ik}} g_{ab}\big(h_C,(u,v),(U_{ijl},U_{ikm})\big)\big\{Z_{ijl}Z_{ikm} - \Sigma_{jk}(u,v)\big\}, \\ V_{ab,ijk}^{(2)}(u,v) &= \sum_{l=1}^{L_{ij}}\sum_{m=1}^{L_{ik}} g_{ab}\big(h_C,(u,v),(U_{ijl},U_{ikm})\big)\big\{\Sigma_{jk}(u,v) - \widetilde{\Sigma}_{jk}(u,v)\big\}.\end{aligned}$$

The term $\widetilde{\Psi}_{jk}^{(1)}$ can be re-expressed as

$$\begin{aligned}\widetilde{\Psi}_{jk}^{(1)} &= I_{jk}W_{1,jk}^2\sum_{i=1}^{n}\big\{V_{00,ijk}^{(1)}\big\}^2 + I_{jk}W_{1,jk}^2\sum_{i=1}^{n}\big\{V_{00,ijk}^{(2)}\big\}^2 + 2I_{jk}W_{1,jk}^2\sum_{i=1}^{n}V_{00,ijk}^{(1)}V_{00,ijk}^{(2)} \\ &= D_{jk,1} + D_{jk,2} + D_{jk,3}. \end{aligned} \tag{S.17}$$

(a) *Verification of* I_{jk}. To show the rationale of imposing the rate I_{jk} in (12), we need to verify that

$$I_{jk}\sum_{i=1}^{n}\big(W_{1,jk}V_{00,ijk} + W_{2,jk}V_{10,ijk} + W_{3,jk}V_{01,ijk}\big)^2 \asymp 1 + o_P(1). \tag{S.18}$$

for each $u, v \in \mathcal{U}$. Denote by $\tilde{n}_{jk} = \sum_{i=1}^{n} L_{ij}L_{ik}$. Recall that

$$S_{ab,jk}(u,v) = \sum_{i=1}^{n}\sum_{l=1}^{L_{ij}}\sum_{m=1}^{L_{ik}} g_{ab}\big\{h_C,(u,v),(U_{ijl},U_{ikm})\big\}.$$

It can be shown that $S_{ab,jk}(u,v) \asymp \tilde{n}_{jk}h_C^{a+b}\{1+o_P(1)\}$ for $a, b = 0, 1, 2$, which together with (S.12) implies that

$$W_{1,jk}(u,v) \asymp \tilde{n}_{jk}^{-1}\big\{1+o_P(1)\big\}, \quad W_{2,jk}(u,v) \asymp W_{3,jk}(u,v) \asymp \tilde{n}_{jk}^{-1}h_C^{-1}\big\{1+o_P(1)\big\}. \tag{S.19}$$

Similarly, we can also show that

$$
\begin{aligned}
&\sum_{i=1}^{n}\Big\{\sum_{l=1}^{L_{ij}}\sum_{m=1}^{L_{ik}} g_{00}\big(h_C,(u,v),(U_{ijl},U_{ikm})\big)\Big\}^2 \\
=\;& \sum_{i=1}^{n}\sum_{l=1}^{L_{ij}}\sum_{m=1}^{L_{ik}} K_{h_C}^2(U_{ijl}-u)K_{h_C}^2(U_{ikm}-v) \\
&+\sum_{i=1}^{n}\sum_{l=1}^{L_{ij}}\sum_{m'\neq m}^{L_{ik}} K_{h_C}^2(U_{ijl}-u)K_{h_C}(U_{ikm}-v)K_{h_C}(U_{ikm'}-v) \\
&+\sum_{i=1}^{n}\sum_{l\neq l'}\sum_{m=1}^{L_{ik}} K_{h_C}(U_{ijl}-u)K_{h_C}(U_{ijl'}-u)K_{h_C}^2(U_{ikm}-v) \\
&+\sum_{i=1}^{n}\sum_{l\neq l'}^{L_{ij}}\sum_{m\neq m'}^{L_{ik}} K_{h_C}(U_{ijl}-u)K_{h_C}(U_{ikm}-v)K_{h_C}(U_{ijl'}-u)K_{h_C}(U_{ikm'}-v) \\
\asymp\;& \Big\{\sum_{i=1}^{n}\big(L_{ij}L_{ik}h_C^{-2}+L_{ij}^2L_{ik}h_C^{-1}+L_{ij}L_{ik}^2h_C^{-1}+L_{ij}^2L_{ik}^2\big)\Big\}\big\{1+o_P(1)\big\}. \qquad \text{(S.20)}
\end{aligned}
$$

By (S.19) and (S.20), we obtain that

$$
W_{1,jk}^2\sum_{i=1}^{n}\Big\{\sum_{l=1}^{L_{ij}}\sum_{m=1}^{L_{ik}} g_{00}\big(h_C,(u,v),(U_{ijl},U_{ikm})\big)\Big\}^2 \asymp I_{jk}^{-1}\big\{1+o_P(1)\big\}, \qquad \text{(S.21)}
$$

which together with $\Sigma_{jk}(u,v)-\widetilde{\Sigma}_{jk}(u,v)=o_P(1)$ implies that $D_{jk,2}=o_P(1)$ and $D_{jk,3}=o_P(1)$. Note that $\mathbb{E}\big\{Z_{ijl}Z_{ikm}-\Sigma_{jk}(u,v)\big\}^2$ is bounded. Together with (S.19) and (S.20), we can also show that $D_{jk,1}(u,v)\asymp 1+o_P(1)$. Combining the above results yields that $\widetilde{\Psi}_{jk}^{(1)}\asymp 1+o_P(1)$. In a similar fashion, we can also show that $\widetilde{\Psi}_{jk}^{(i)}\asymp 1+o_P(1)$ for $i=2,\ldots,6$ in (S.16) and hence (S.18) follows.

(b) *Verification of Condition 8.* To verify the uniform convergence rate in Condition 8, we need to refine our analysis above to construct the exponential type of tail bounds on $\widetilde{\Psi}_{jk}(u,v)-\Psi_{jk}(u,v)$ at each $(u,v)\in\mathcal{U}^2$ rather than the consistency results in (a).

Consider the first term $D_{jk,1}(u,v)=(\tilde{n}_{jk}W_{1,jk})^2\times I_{jk}\tilde{n}_{jk}^{-2}\sum_{i=1}^{n}\{V_{00,ijk}^{(1)}\}^2$ in (S.17). Note that by (S.19) $\tilde{n}_{jk}|W_{1,jk}|$ is bounded with an overwhelming probability. Suppose that $X_{ij}(\cdot)$'s are sub-Gaussian processes and ε_{ijl}'s are independent sub-Gaussian errors. Since $\{V_{00,ijk}^{(1)}(u,v),i=1,\ldots,n\}$ forms an independent sequence, we can obtain the tail bound on $D_{jk,1}(u,v)-\mathbb{E}\{D_{jk,1}(u,v)\}$ by calculating all q-th moments of $\zeta_{ijk}=\big\{V_{00,ijk}^{(1)}(u,v)\big\}^2-$

$\mathbb{E}\big[\{V^{(1)}_{00,ijk}(u,v)\}^2\big]$ for $q = 2, 3, 4, \ldots$ under regularity conditions. Since ζ_{ijk}'s are either sub-Gaussian or sub-exponential, we can follow the similar techniques to prove Lemma 5 of Qiao et al. (2020) by adopting a truncation technique and then applying Bernstein inequality (Boucheron et al., 2014) to establish a rough exponential type of concentration inequality (i.e., the equipped tail bound is in the same form of the exponential tail bound in (17)) for $D_{jk,1}(u,v)$ at each $(u,v) \in \mathcal{U}^2$. Similarly, we can also derive the exponential type of concentration inequality for the third term $D_{jk,3}(u,v)$ in (S.17).

Consider the second term $D_{jk,2}(u,v)$ in (S.17), which can be re-expressed as

$$D_{jk,2}(u,v) = 2I_{jk}W^2_{1,jk}\sum_{i=1}^{n}\Big\{\sum_{l=1}^{L_{ij}}\sum_{m=1}^{L_{ik}} g_{ab}\big(h_C,(u,v),(U_{ijl},U_{ikm})\big)\Big\}^2\Big\{\Sigma_{jk}(u,v)-\widetilde{\Sigma}_{jk}(u,v)\Big\}^2.$$

Note that it follows from (S.21) that $I_{jk}W^2_{1,jk}\sum_{i=1}^{n}\big\{\sum_{l=1}^{L_{ij}}\sum_{m=1}^{L_{ik}} g_{ab}(h_C,(u,v),(U_{ijl},U_{ikm}))\big\}^2$ is bounded with an overwhelming probability. Then the exponential type of concentration bound on $D_{jk,2}(u,v)$ at each $(u,v)\in\mathcal{U}^2$ can be obtained through the exponential type tail bound on $\Sigma_{jk}(u,v)-\widetilde{\Sigma}_{jk}(u,v)$, which has been established in Qiao et al. (2020), see details in proofs of its Lemmas 4 and 5 under the sparse and dense designs, respectively.

To derive the uniform (i.e., over $\mathcal{U}^2$) concentration inequality for $\widetilde{\Psi}^{(1)}_{jk}(u,v)$ in (S.17), we can apply the partition technique that reduces the problem from supremum over $\mathcal{U}^2$ to the maximum over a grid of pairs and then follow the similar developments to prove the uniform concentration inequalities in Lemmas 4 and 5 of Qiao et al. (2020). In a similar fashion to the above procedure, we can develop the corresponding exponential type of uniform concentration inequality for $\widetilde{\Psi}^{(i)}_{jk}(u,v)$ for $i = 2,\ldots,6$. As a result, the exponential type of uniform concentration inequality for $\widetilde{\Psi}_{jk}(u,v)$ can be obtained.

The uniform convergence rate in Condition 8 is implied by the exponential type of uniform concentration inequalities for $\widetilde{\Psi}_{jk}(u,v)$ for each j,k, which partially depend on the uniform concentration bounds on $\widetilde{\Sigma}_{jk}(u,v)$'s. In a similar spirit to the L_2 concentration bounds on $\widetilde{\Sigma}_{jk}(u,v)$'s implied by Condition 7, we consider the uniform convergence rate of

$\widetilde{\Sigma}_{jk}(u,v)$,

$$\max_{1\leqslant j,k\leqslant p}\sup_{u,v\in\mathcal{U}}\left|\widetilde{\Sigma}_{jk}(u,v)-\Sigma_{jk}(u,v)\right| = O_P\left(\sqrt{\frac{\log p}{n^{2\gamma_1}}}+h^2\right), \tag{S.22}$$

which is satisfied if there exists some positive constants c_i for $i=6,\ldots,9$ and $\gamma_1\in(0,1/2]$ such that for each $j,k=1,\ldots,p$ and $t\in(0,1]$,

$$P\Big\{\sup_{u,v\in\mathcal{U}}|\widetilde{\Sigma}_{jk}(u,v)-\Sigma_{jk}(u,v)|\geqslant t+c_8h^2\Big\}\leqslant c_7n^{c_9}\exp(-c_6n^{2\gamma_1}t^2). \tag{S.23}$$

Larger values of γ_1 correspond to a more frequent measurement schedule and hence faster rate in (S.22). For sparsely sampled functional data, it follows from Lemma 4 of Qiao et al. (2020) and the same proof technique for $j\neq k$ that (S.23) holds by choosing $\gamma_1=1/2-a$ and $c_9=1+2a$ with $h\asymp n^{-a}$ for some positive constant $a<1/2$. For densely sampled functional data, it follows from Lemma 5 of Qiao et al. (2020) and more efforts for $j\neq k$ that (S.23) holds with the choice of $\gamma_1=\min(1/2,1/3+b/6-\epsilon'/2-2a/3)$ and $c_9=\max(1,2/3-\epsilon'-b/3+4a/3)$ for some small constant $\epsilon'>0$ when $h\asymp n^{-a}$ and $L\asymp n^b$ for some positive constants a,b.

Following the proof procedure described above, we can establish exponential type of uniform concentration inequality for $\widetilde{\Psi}_{jk}(u,v)$ for each j,k in the same form as (S.23) but with different positive constants and in particular $\gamma_2\in(0,1/2]$, which will result in the uniform convergence rate in Condition 8. It is worth mentioning that such heuristic analysis can only help us establish uniform concentration inequalities for $\widetilde{\Psi}_{jk}(u,v)$'s leading to the sub-optimal rate. Investigating the corresponding optimal rate through the precise specification of the largest values of γ_2 under different measurement schedules or more generally through n,h and possibly L for the dense case is quite challenging and remains an open topic to be pursued in the future.

E Additional empirical results

E.1 Applications of sparse covariance function estimation

In Section 2.3, we summarize the usefulness of our functional thresholding proposals in more general statistical frameworks involving the sparse estimation of $\mathbf{\Sigma}$ to handle high-dimensional functional data. In Sections E.1.1 and E.1.2, we conduct simulations to demonstrate the significantly improved finite-sample performance of functional-thresholding-based estimators using two applications of the sparse covariance function estimation including multivariate FPCA (Happ and Greven, 2018) and multivariate functional linear regression (Chiou et al., 2016), respectively.

E.1.1 Multivariate FPCA

Before presenting the methodology, we first solidify some notation. Denote the p-fold Cartesian product defined on $\mathcal{U}$ by $\mathbb{H} = L_2(\mathcal{U}) \times \cdots \times L_2(\mathcal{U})$. For any $\mathbf{f}, \mathbf{g} \in \mathbb{H}$, we denote the inner product by $\langle \mathbf{f}, \mathbf{g} \rangle = \int_{\mathcal{U}} \mathbf{f}(u)^{\mathrm{T}} \mathbf{g}(u) \mathrm{d}u$ and the induced norm $\|\cdot\| = \langle \cdot, \cdot \rangle^{1/2}$. Following the proposal of Chiou et al. (2014) and Happ and Greven (2018), we consider a normalized version of the Karhunen-Loéve expansion for multivariate functional data in (3), which accounts for differences in degrees of variability among the components of the multivariate random functions, i.e.

$$\mathbf{X}_i(\cdot) = \sum_{l=1}^{\infty} \xi_{il} \mathbf{W} \boldsymbol{\phi}_l(\cdot),$$

where $\mathbb{E}\{\mathbf{X}_i(\cdot)\} = 0$, the weight matrix $\mathbf{W} = \mathrm{diag}(w_1, \cdots, w_p)$ with each $w_j = \|\Sigma_{jj}\|_{\mathcal{N}}^{1/2}$, $\{\boldsymbol{\phi}_l(\cdot)\}_{l \geqslant 1}$ are a sequence of orthonormal functions with $\langle \boldsymbol{\phi}_l, \boldsymbol{\phi}_{l'} \rangle = I(l = l')$, and ξ_{il}'s are mean-zero principal component scores obtained via $\xi_{il} = \langle \mathbf{W}^{-1} \mathbf{X}_i, \boldsymbol{\phi}_l \rangle$ for $l \geqslant 1$ with $\mathrm{cov}(\xi_{il}, \xi_{il'}) = \pi_l I(l = l')$. Here $\{(\pi_l, \boldsymbol{\phi}_l(\cdot))\}_{l=1}^{\infty}$ are eigenvalue/eigenfunction pairs satisfying $\int_{\mathcal{U}} \mathbf{W}^{-1} \mathbf{\Sigma}(u, v) \mathbf{W}^{-1} \boldsymbol{\phi}_l(v) \mathrm{d}v = \pi_l \boldsymbol{\phi}_l(u)$ and eigenvalues are sorted in descending order $\pi_1 \geqslant \pi_2 \geqslant \cdots > 0$. Let $\widehat{\mathbf{\Sigma}}$ be some legitimate estimator of $\mathbf{\Sigma}$, for example, we can take $\widehat{\mathbf{\Sigma}}$ as adaptive functional thresholding estimator $\widehat{\mathbf{\Sigma}}_{\mathrm{A}}$ (denoted as AdaFT) or univer-

sal functional thresholding estimator $\widehat{\mathbf{\Sigma}}_{\mathrm{U}}$ (denoted as UniFT) or sample covariance function estimator $\widehat{\mathbf{\Sigma}}_{\mathrm{S}}$ (denoted as Sam). Define $\widehat{\mathbf{W}} = \mathrm{diag}(\hat{w}_1, \cdots, \hat{w}_p)$ with $\hat{w}_j = \|\widehat{\Sigma}_{jj}\|_{\mathcal{H}}^{1/2}$. Performing eigen-decomposition on $\widehat{\mathbf{\Pi}}(u,v) = \widehat{\mathbf{W}}^{-1}\widehat{\mathbf{\Sigma}}(u,v)\widehat{\mathbf{W}}^{-1}$, we obtain the estimated eigenvalue/eigenvector pairs $\{(\hat{\pi}_l, \widehat{\boldsymbol{\phi}}_l(\cdot))\}_{l=1}^m$ with $\hat{\pi}_1 \geqslant \cdots \geqslant \hat{\pi}_m$. It is worth mentioning that the proposed functional thresholding estimators may not be positive-definite. To guarantee the positive-definiteness of $\widehat{\mathbf{\Pi}}$, we follow the technique adopted in Chen and Leng (2016) to replace each $\hat{\pi}_l$ with its adjusted version $\hat{\pi}_l - \hat{\pi}_m$ if $\hat{\pi}_m < 0$ before subsequent analysis.

We next present the data generating process. The multivariate functional data $\{\mathbf{X}_i(\cdot)\}_{i=1}^n$ for $p = 50$ and $n = 100, 200$ are generated by the same procedure as in Section 5. Specifically, we generate functional variables by $X_{ij}(u) = \mathbf{s}(u)^{\mathrm{T}}\boldsymbol{\theta}_{ij}$ for $i = 1, \ldots, n, j = 1, \ldots, p$ and $u \in \mathcal{U} = [0,1]$, where $\mathbf{s}(u)$ is a 10-dimensional Fourier basis function and each $\boldsymbol{\theta}_i = (\boldsymbol{\theta}_{i1}^{\mathrm{T}}, \ldots, \boldsymbol{\theta}_{ip}^{\mathrm{T}})^{\mathrm{T}} \in \mathbb{R}^{10p}$ is sampled from a mean zero multivariate Gaussian distribution with block covariance matrix $\mathbf{\Omega} \in \mathbb{R}^{10p \times 10p}$. The (j,k)-th block of $\mathbf{\Omega}$ is $\mathbf{\Omega}_{jk} = \omega_{jk}\mathbf{D} \in \mathbb{R}^{10\times 10}$ with $\mathbf{D} = (D_{ll'})_{10\times 10} = \mathrm{diag}(2, 1, 3^{-2} \ldots, 10^{-2})$ for $j, k = 1, \ldots, p$, where ω_{jk}'s are generated according to Model 2 as specified in Section 5.

We examine the performance of multivariate FPCA based on AdaFT, UniFT and Sam in terms of their relative estimation errors, i.e., $|\hat{\pi}_l - \pi_l|/\pi_l$ (with $m = 300$) for eigenvalues and $\|\widehat{\boldsymbol{\phi}}_l - \mathrm{sign}(\langle \boldsymbol{\phi}_l, \widehat{\boldsymbol{\phi}}_l \rangle) \cdot \boldsymbol{\phi}_l\|$ for eigenfunctions. To be specific, $\widehat{\mathbf{\Sigma}}_{\mathrm{A}}$ and $\widehat{\mathbf{\Sigma}}_{\mathrm{U}}$ are computed using the adaptive lasso (with $\eta = 3$) functional thresholding rule with the associated $\hat{\lambda}$'s selected by fivefold cross-validation. The numerical results are summarized over 100 Monte Carlo runs. Figure 3 displays boxplots of the relative estimation errors for the top 7 eigenvalues and the corresponding eigenfunctions. A few trends are apparent. First, two functional-thresholding methods give substantially improved accuracies for estimated eigenpairs compared to the baseline Sam, which fails to detect the functional sparsity pattern of $\mathbf{\Sigma}$, and hence results in elevated estimation errors. Second, even with the implementation of the weight matrix $\widehat{\mathbf{W}}$, AdaFT provides better overall performance than UniFT especially for sufficiently large n. This again demonstrates the superiority of AdaFT

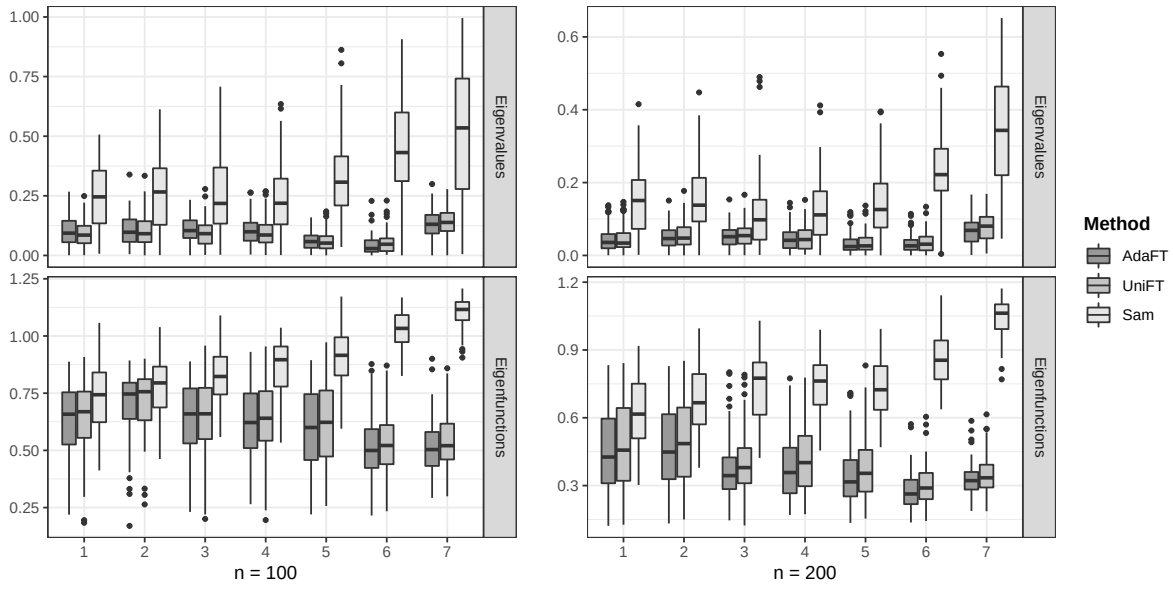


Figure 3: Relative estimation errors of the top 7 eigenvalues (top row) and the corresponding eigenfunctions (bottom row) for $n = 100$ and 200 over 100 simulation runs.

over UniFT in the sense that $\widehat{\mathbf{\Sigma}}_{\mathrm{A}}$ can capture the pointwise variability of $\widehat{\Sigma}_{jk}(u, v)$ more precisely.

E.1.2 Multivariate functional linear regression

We first present the methodology for multivariate functional linear regression in (4). Under orthonormal basis functions $\{\boldsymbol{\phi}_l(\cdot)\}_{l\geqslant 1}$, we expand the functional coefficient vector $\boldsymbol{\beta}(\cdot) = \sum_{l=1}^{\infty} \kappa_l \mathbf{W}^{-1}\boldsymbol{\phi}_l(\cdot)$, where $\kappa_l = \langle \mathbf{W}\boldsymbol{\beta}, \boldsymbol{\phi}_l\rangle$, and rewrite (4) as

$$Y_i = \int_{\mathcal{U}} \mathbf{X}_i(u)^{\mathrm{T}}\boldsymbol{\beta}(u)\mathrm{d}u + \epsilon_i = \sum_{l=1}^{q} \xi_{il}\kappa_l + \sum_{l=q+1}^{\infty} \xi_{il}\kappa_l + \epsilon_i \equiv \boldsymbol{\xi}_i^{\mathrm{T}}\boldsymbol{\kappa} + R_i + \epsilon_i,$$

where q is the truncated dimension, $\boldsymbol{\xi}_i = (\xi_{i1}, \ldots, \xi_{iq})^{\mathrm{T}}$ and $\boldsymbol{\kappa} = (\kappa_1, \ldots, \kappa_q)^{\mathrm{T}}$. Let $\mathbf{Y} = (Y_1, \ldots, Y_n)^{\mathrm{T}} \in \mathbb{R}^n$ and $\mathbf{\Xi} \in \mathbb{R}^{n\times q}$ with its row vectors given by $\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_n$. We further represent (4) as

$$\mathbf{Y} = \mathbf{\Xi}\boldsymbol{\kappa} + \mathbf{R} + \boldsymbol{\epsilon}, \tag{S.24}$$

where $\mathbf{R} = (R_1, \ldots, R_n)^{\mathrm{T}} \in \mathbb{R}^n$ and $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)^{\mathrm{T}} \in \mathbb{R}^n$ correspond to the truncation and the random errors, respectively. We implement standard three-step procedure to estimate

$\boldsymbol{\beta}(\cdot)$.

Step 1. Perform multivariate FPCA on $\{\mathbf{X}_i(\cdot)\}_{i=1}^n$, thus obtaining estimated eigenfunctions $\{\widehat{\boldsymbol{\phi}}_l(\cdot)\}$ and estimated principal component scores $\widehat{\xi}_{il} = \langle \widehat{\mathbf{W}}^{-1}\mathbf{X}_i, \widehat{\boldsymbol{\phi}}_l \rangle$. Select q such that the cumulative percentage of the largest q estimated eigenvalues exceeds 90%.

Step 2. Replace $\boldsymbol{\Xi}$ in (S.24) by $\widehat{\boldsymbol{\Xi}} = (\widehat{\xi}_{il})_{n\times q}$ and obtain the least-squares estimator of $\boldsymbol{\kappa}$ as $\widehat{\boldsymbol{\kappa}} = (\widehat{\boldsymbol{\Xi}}^{\mathrm{T}}\widehat{\boldsymbol{\Xi}})^{-1}\widehat{\boldsymbol{\Xi}}^{\mathrm{T}}\mathbf{Y} \equiv (\hat{\kappa}_1, \ldots, \hat{\kappa}_q)^{\mathrm{T}}$.

Step 3. Recover the functional coefficient vector by $\widehat{\boldsymbol{\beta}}(\cdot) = \sum_{l=1}^q \hat{\kappa}_l \widehat{\mathbf{W}}^{-1}\widehat{\boldsymbol{\phi}}_l(\cdot)$.

We next generate simulated data for model (4) using the same multivariate functional predictors $\{\mathbf{X}_i(\cdot)\}_{i=1}^n$ as in Section E.1.1. For each j, the associated functional coefficient is generated by $\beta_j(u) = \mathbf{s}(u)^{\mathrm{T}}\mathbf{b}_j$, where components in $\mathbf{b}_j = (b_{j1}, \ldots, b_{j10})^{\mathrm{T}} \in \mathbb{R}^{10}$ are sampled from the uniform distribution with support $[-r_l, -0.5r_l] \cup [0.5r_l, r_l]$ and $r_l = D_{ll}$ for $l = 1, \ldots, 10$. The scalar responses $\{Y_i\}_{i=1}^n$ are generated according to (4), in which the random errors ϵ_t's are sampled independently from $\mathcal{N}(0, 1)$.

We assess the performance of AdaFT-, UniFT- and Sam-based methods in terms of both estimation and prediction accuracies. We implement Step 1 following the same procedure as in Section E.1.1 to estimate $\boldsymbol{\Sigma}(\cdot, \cdot)$ and then Steps 2–3 to estimate $\boldsymbol{\beta}(\cdot)$. The estimation and prediction accuracies are measured by the relative estimation error $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|/\|\boldsymbol{\beta}\|$ and root mean squared prediction error (rMSPE) of an independent test set $\{(\widetilde{\mathbf{X}}_i(\cdot), \widetilde{Y}_i)\}_{i=1}^{200}$ generated by the same model, i.e. $\{\sum_{i=1}^{200}(\langle \widetilde{\mathbf{X}}_i, \widehat{\boldsymbol{\beta}} \rangle - \widetilde{Y}_i)^2/200\}^{1/2}$, respectively. The simulation is repeated over 100 runs. Table 8 reports the relative estimation errors of $\boldsymbol{\beta}(\cdot)$ and rMSPEs for all three methods. For comparison, we also consider the oracle case, where $\boldsymbol{\beta}(\cdot)$ is estimated by assuming that the true eigenpairs $\{\pi_l, \boldsymbol{\phi}_l(\cdot)\}$ are known in advance. The observable trends from Table 8 are consistent to those from Figure 3. These results demonstrate that AdaFT not only provides more accurate estimation of the covariance function itself but also largely improves the accuracies of other covariance-function-based estimation and prediction that arise from a range of high-dimensional functional data analysis prob-

Table 8: The mean and standard error (in parentheses) of $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|/\|\boldsymbol{\beta}\|$ and rMSPEs over 100 simulation runs.

	AdaFT	UniFT	Sam	Oracle
		$\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|/\|\boldsymbol{\beta}\|$		
$n = 100$	0.607(0.007)	0.612(0.007)	0.724(0.004)	0.551(0.007)
$n = 200$	0.506(0.003)	0.510(0.003)	0.617(0.004)	0.483(0.004)
		rMSPE		
$n = 100$	11.211(0.186)	11.300(0.198)	16.223(0.195)	9.418(0.192)
$n = 200$	7.278(0.059)	7.335(0.059)	10.755(0.112)	5.818(0.069)

lems, for example, multivariate FPCA and multivariate functional linear regression with large p.

E.2 Simulation studies

E.2.1 Fully observed functional data

Figure 4 displays the simulated trajectories of $X_{ij}(\cdot)$ for $i = 1$ at a selection of j's for Models 1 and 2 with $p = 50$. Figures 5 and 6 plot the heat maps of the frequency of the zeros identified for the Hilbert–Schimidt norm of each entry of the estimated covariance function, when $p = 50$, out of 100 simulation runs. The true nonzero patterns of Models 1 and 2, and the corresponding Hilbert–Schmidt-norm based pseudo correlation matrix, defined as $\left\{\|\Sigma_{jk}\|_{\mathcal{S}}/(\|\Sigma_{jj}\|_{\mathcal{S}}\|\Sigma_{kk}\|_{\mathcal{S}})^{1/2}\right\}_{p\times p}$, are presented in Figures 5(a), 6(a) and 7, respectively. Tables 9 and 10 present numerical results in terms of estimation accuracy and support recovery consistency of all competing approaches under Model 1. Figure 8 plots some selected entries of $\boldsymbol{\Sigma}$ under Model 1 with $p = 50$ together with their corresponding $\widehat{\boldsymbol{\Sigma}}_{\mathrm{A}}$ and $\widehat{\boldsymbol{\Sigma}}_{\mathrm{U}}$ (the corresponding $\hat{\lambda}$'s are selected by fivefold cross-validation using hard functional thresholding rule), and $\widehat{\boldsymbol{\Sigma}}_{\mathrm{S}}$ of one simulation run. It is worth mentioning that such design of $\boldsymbol{\Sigma}$ is able to mimic the positive and negative banding patterns in the HCP

data analysis. See Figure 16 for details. Figure 9 displays the average receiver operating characteristic (ROC) curves (plots of true positive rates versus false positive rates over a sequence of λ values) for both the adaptive functional thresholding and universal functional thresholding methods. These results again demonstrate the uniform superiority of the adaptive functional thresholding method.

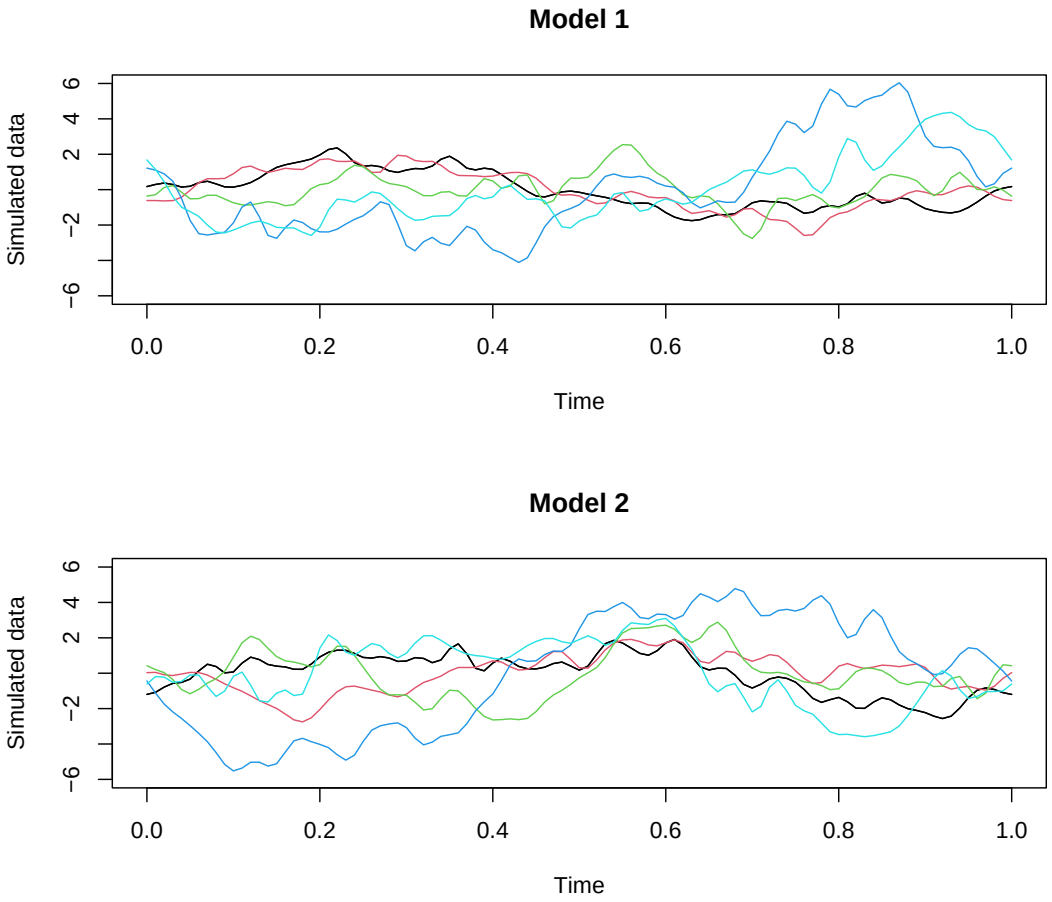
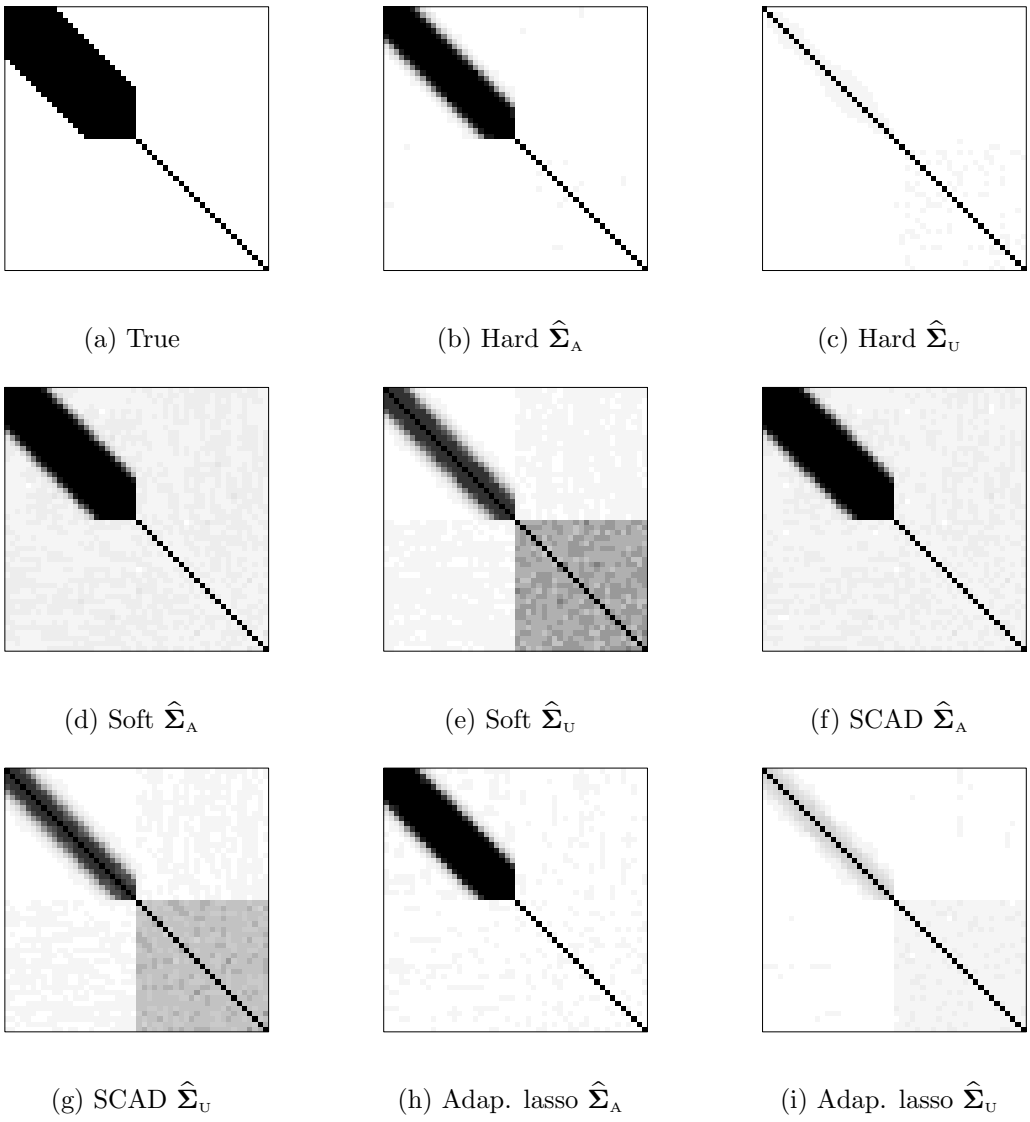


Figure 4: Simulated dataset: $X_{1j}(\cdot)$ at $j = 10, 20, 30, 40, 50$ for $p = 50$.

E.2.2 Partially observed functional data

Tables 11 and 12 provide the estimation and support recovery performance of BinLLS-based adaptive and universal functional thresholding estimators for the setting of $p = 50$ under Model 1. Tables 13 and 14 summarize the performance under both Models 1 and 2 for

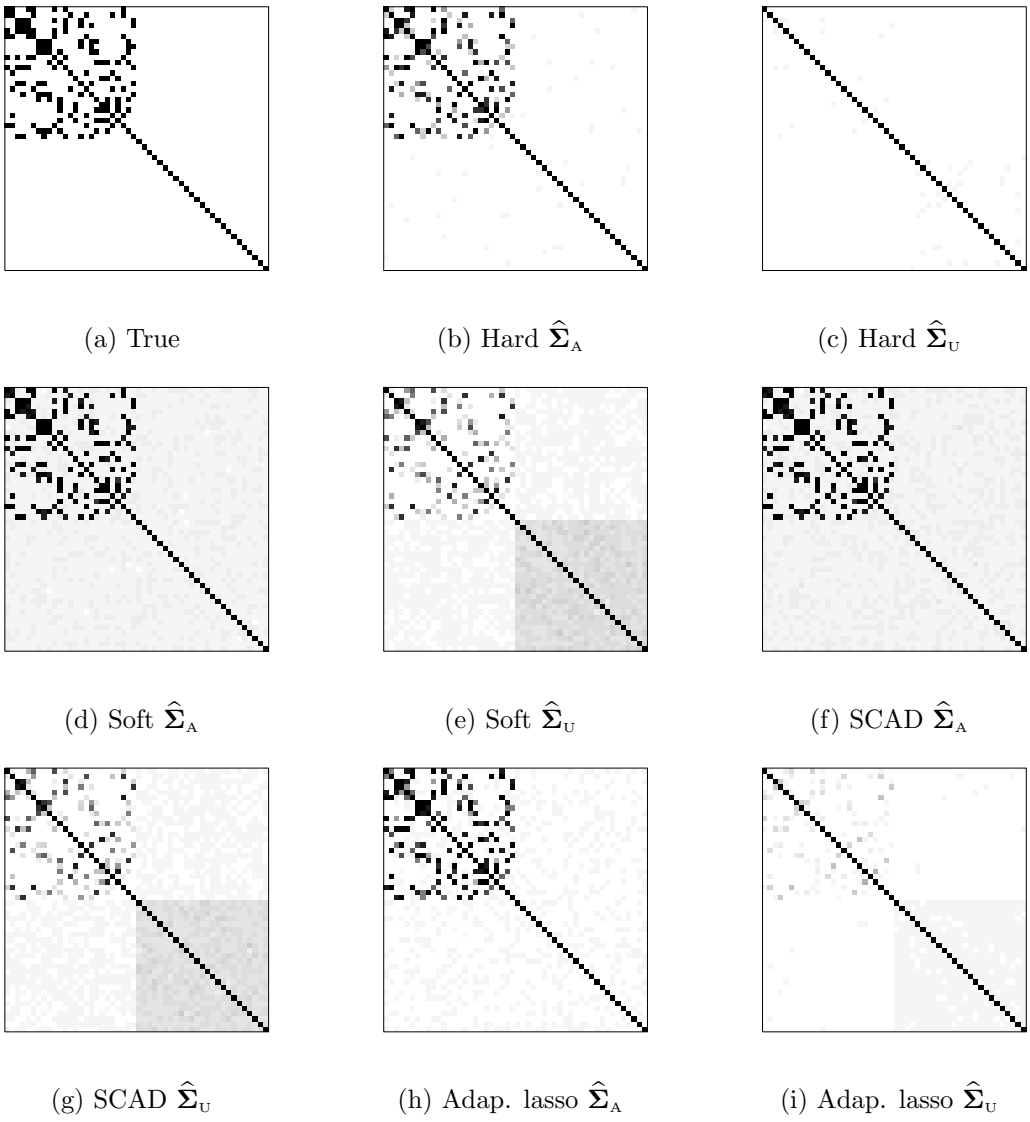


(a) True (b) Hard $\widehat{\mathbf{\Sigma}}_{\text{A}}$ (c) Hard $\widehat{\mathbf{\Sigma}}_{\text{U}}$

(d) Soft $\widehat{\mathbf{\Sigma}}_{\text{A}}$ (e) Soft $\widehat{\mathbf{\Sigma}}_{\text{U}}$ (f) SCAD $\widehat{\mathbf{\Sigma}}_{\text{A}}$

(g) SCAD $\widehat{\mathbf{\Sigma}}_{\text{U}}$ (h) Adap. lasso $\widehat{\mathbf{\Sigma}}_{\text{A}}$ (i) Adap. lasso $\widehat{\mathbf{\Sigma}}_{\text{U}}$

Figure 5: Heat maps of the frequency of the zeros identified for the Hilbert–Schimidt norm of each entry of the estimated covariance function (when $p = 50$) for Model 1 out of 100 simulation runs. White and black correspond to 100/100 and 0/100 zeros identified, respectively.



(a) True (b) Hard $\widehat{\mathbf{\Sigma}}_{\text{A}}$ (c) Hard $\widehat{\mathbf{\Sigma}}_{\text{U}}$

(d) Soft $\widehat{\mathbf{\Sigma}}_{\text{A}}$ (e) Soft $\widehat{\mathbf{\Sigma}}_{\text{U}}$ (f) SCAD $\widehat{\mathbf{\Sigma}}_{\text{A}}$

(g) SCAD $\widehat{\mathbf{\Sigma}}_{\text{U}}$ (h) Adap. lasso $\widehat{\mathbf{\Sigma}}_{\text{A}}$ (i) Adap. lasso $\widehat{\mathbf{\Sigma}}_{\text{U}}$

Figure 6: Heat maps of the frequency of the zeros identified for the Hilbert–Schimidt norm of each entry of the estimated covariance function (when $p = 50$) for Model 2 out of 100 simulation runs. White and black correspond to 100/100 and 0/100 zeros identified, respectively.

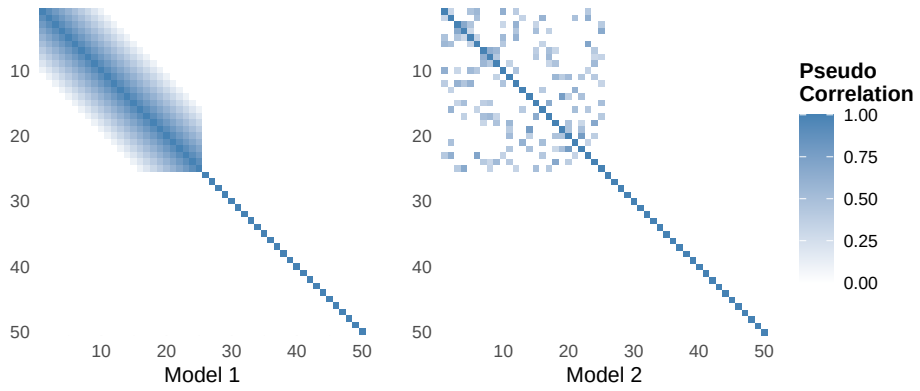


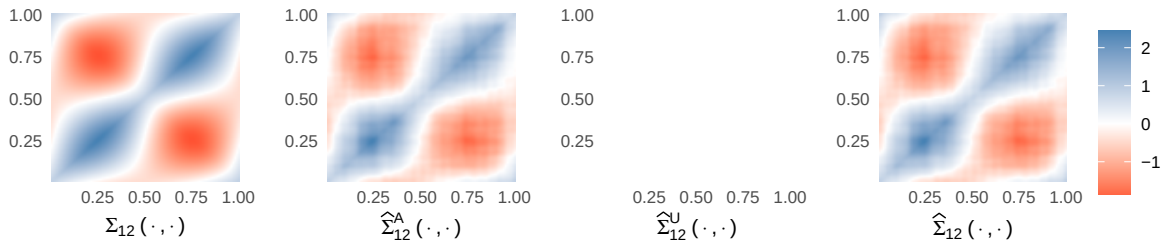
Figure 7: Heat maps of the pseudo correlation matrix (with $p = 50$) for Models 1 and 2. The color from white to blue corresponds to the value of $\|\Sigma_{jk}\|_{\mathcal{S}}/(\|\Sigma_{jj}\|_{\mathcal{S}}\|\Sigma_{kk}\|_{\mathcal{S}})^{1/2}$ from small to large.

Table 9: The average (standard error) functional matrix losses over 100 simulation runs.

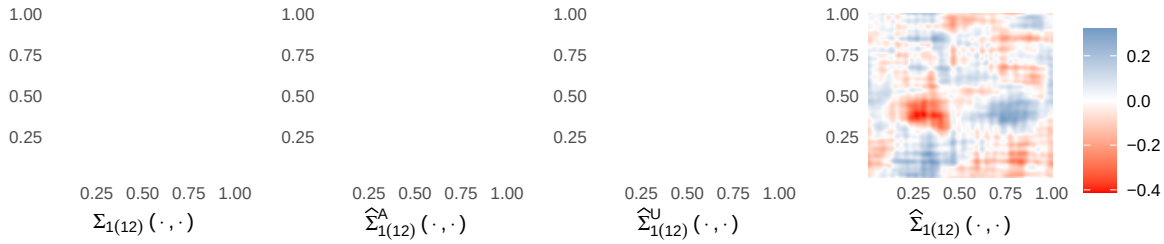
		$p = 50$		$p = 100$		$p = 150$	
Model	Method	$\widehat{\mathbf{\Sigma}}_{\text{A}}$	$\widehat{\mathbf{\Sigma}}_{\text{U}}$	$\widehat{\mathbf{\Sigma}}_{\text{A}}$	$\widehat{\mathbf{\Sigma}}_{\text{U}}$	$\widehat{\mathbf{\Sigma}}_{\text{A}}$	$\widehat{\mathbf{\Sigma}}_{\text{U}}$
		Functional Frobenius norm					
1	Hard	5.40(0.04)	11.90(0.02)	7.91(0.03)	17.27(0.01)	9.94(0.04)	21.36(0.01)
	Soft	6.28(0.05)	10.40(0.08)	9.41(0.05)	16.53(0.07)	11.85(0.06)	21.16(0.04)
	SCAD	5.68(0.05)	10.56(0.08)	8.53(0.05)	16.59(0.07)	10.80(0.06)	21.19(0.04)
	Adap. lasso	5.28(0.04)	11.42(0.07)	7.76(0.04)	17.26(0.01)	9.72(0.04)	21.36(0.01)
	Sample	19.82(0.04)		39.54(0.05)		59.28(0.06)	
		Functional matrix ℓ_1 norm					
	Hard	3.96(0.06)	9.23(0.01)	4.49(0.05)	9.31(0.01)	4.78(0.05)	9.34(0.01)
	Soft	5.04(0.07)	8.14(0.08)	5.88(0.05)	9.15(0.02)	6.21(0.04)	9.31(0.01)
	SCAD	4.40(0.08)	8.32(0.07)	5.35(0.06)	9.18(0.02)	5.75(0.05)	9.31(0.01)
	Adap.lasso	3.85(0.06)	8.91(0.07)	4.52(0.05)	9.30(0.01)	4.83(0.06)	9.34(0.01)
	Sample	26.60(0.13)		52.65(0.18)		78.69(0.22)	

Table 10: The average TPRs/ FPRs over 100 simulation runs.

Model	Method	$p = 50$		$p = 100$		$p = 150$	
		$\widehat{\mathbf{\Sigma}}_{\mathrm{A}}$	$\widehat{\mathbf{\Sigma}}_{\mathrm{U}}$	$\widehat{\mathbf{\Sigma}}_{\mathrm{A}}$	$\widehat{\mathbf{\Sigma}}_{\mathrm{U}}$	$\widehat{\mathbf{\Sigma}}_{\mathrm{A}}$	$\widehat{\mathbf{\Sigma}}_{\mathrm{U}}$
1	Hard	0.71/0.00	0.00/0.00	0.66/0.00	0.00/0.00	0.64/0.00	0.00/0.00
	Soft	0.89/0.08	0.47/0.17	0.85/0.04	0.22/0.05	0.84/0.03	0.06/0.01
	SCAD	0.89/0.07	0.42/0.13	0.85/0.04	0.20/0.04	0.84/0.03	0.05/0.01
	Adap. lasso	0.78/0.00	0.11/0.02	0.74/0.00	0.00/0.00	0.73/0.00	0.00/0.00



(a) $(1, 2)$th entry



(b) $(1, 12)$th entry

Figure 8: Selected entries of $\mathbf{\Sigma}$, $\widehat{\mathbf{\Sigma}}_{\mathrm{A}}$, $\widehat{\mathbf{\Sigma}}_{\mathrm{U}}$ and $\widehat{\mathbf{\Sigma}}_{\mathrm{S}}$ of one simulation run (when $p = 50$) for Model 1. The corresponding $\hat{\lambda}$'s of $\widehat{\mathbf{\Sigma}}_{\mathrm{A}}$ and $\widehat{\mathbf{\Sigma}}_{\mathrm{U}}$ are selected by fivefold cross-validation using hard functional thresholding rule.

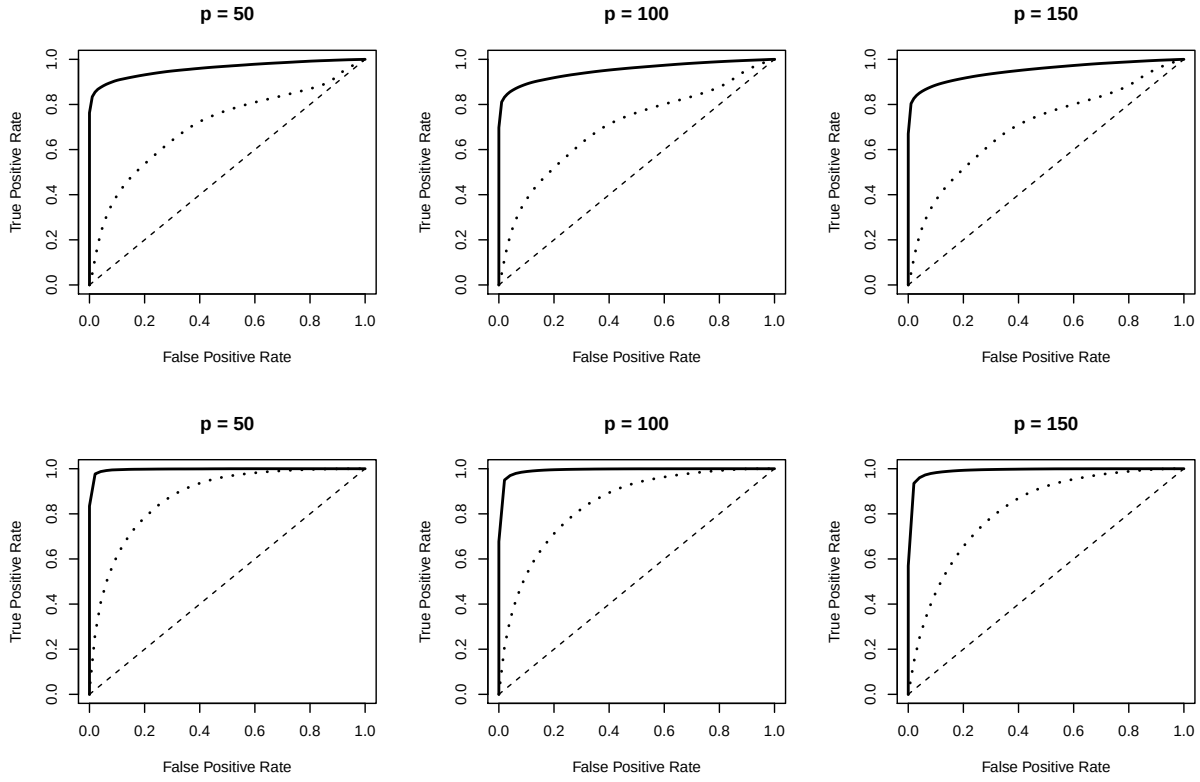


Figure 9: Model 1 (top row) and Model 2 (bottom row) for $p = 50, 100, 150$: Comparison of the average ROC curves for adaptive functional thresholding (solid line) and universal functional thresholding (dotted line) over 100 simulation runs.

$p = 100$. The same patterns as those from Tables 5–6 can be observed from Tables 11–14.

Table 11: The average (standard error) functional matrix losses for partially observed functional scenarios and $p = 50$ over 100 simulation runs.

		$L_i = 11$		$L_i = 21$		$L_i = 51$		$L_i = 101$	
Model	Method	$\check{\Sigma}_{\text{A}}$	$\check{\Sigma}_{\text{U}}$	$\check{\Sigma}_{\text{A}}$	$\check{\Sigma}_{\text{U}}$	$\check{\Sigma}_{\text{A}}$	$\check{\Sigma}_{\text{U}}$	$\check{\Sigma}_{\text{A}}$	$\check{\Sigma}_{\text{U}}$
		Functional Frobenius norm							
1	Hard	7.78(0.03)	12.65(0.01)	6.61(0.04)	12.26(0.01)	5.83(0.04)	12.04(0.02)	5.57(0.04)	11.89(0.04)
	Soft	8.69(0.04)	12.63(0.01)	7.64(0.05)	11.75(0.06)	6.94(0.05)	10.51(0.07)	6.71(0.05)	10.05(0.07)
	SCAD	8.36(0.05)	12.63(0.01)	7.13(0.05)	11.80(0.06)	6.28(0.05)	10.67(0.07)	5.99(0.05)	10.27(0.07)
	Adap. lasso	7.69(0.04)	12.64(0.01)	6.57(0.04)	12.21(0.02)	5.83(0.04)	11.54(0.08)	5.57(0.04)	11.05(0.10)
		Functional matrix ℓ_1 norm							
	Hard	5.35(0.05)	9.36(0.01)	4.68(0.06)	9.30(0.01)	4.09(0.06)	9.24(0.02)	3.87(0.06)	9.13(0.05)
	Soft	6.38(0.06)	9.35(0.01)	5.86(0.07)	8.94(0.05)	5.43(0.07)	8.13(0.08)	5.29(0.07)	7.84(0.08)
	SCAD	6.12(0.07)	9.35(0.01)	5.40(0.08)	8.99(0.05)	4.78(0.08)	8.32(0.07)	4.56(0.08)	8.09(0.07)
	Adap.lasso	5.31(0.07)	9.36(0.01)	4.71(0.07)	9.28(0.02)	4.15(0.07)	8.89(0.07)	3.98(0.07)	8.59(0.09)

Table 12: The average TPRs/ FPRs for partially observed functional scenarios and $p = 50$ over 100 simulation runs.

		$L_i = 11$		$L_i = 21$		$L_i = 51$		$L_i = 101$	
Model	Method	$\check{\Sigma}_{\text{A}}$	$\check{\Sigma}_{\text{U}}$	$\check{\Sigma}_{\text{A}}$	$\check{\Sigma}_{\text{U}}$	$\check{\Sigma}_{\text{A}}$	$\check{\Sigma}_{\text{U}}$	$\check{\Sigma}_{\text{A}}$	$\check{\Sigma}_{\text{U}}$
1	Hard	0.63/0.00	0.00/0.00	0.66/0.00	0.00/0.00	0.69/0.00	0.01/0.00	0.71/0.00	0.03/0.00
	Soft	0.85/0.05	0.01/0.00	0.87/0.07	0.22/0.09	0.89/0.08	0.5/0.17	0.89/0.08	0.57/0.18
	SCAD	0.86/0.06	0.01/0.00	0.87/0.07	0.2/0.07	0.88/0.07	0.45/0.14	0.89/0.07	0.51/0.14
	Adap. lasso	0.72/0.00	0.00/0.00	0.75/0.00	0.01/0.00	0.77/0.00	0.12/0.02	0.78/0.00	0.20/0.03

E.3 ADHD dataset

In this section, we illustrate our adaptive functional thresholding estimation using the ADHD-200 Sample, collected by New York University Medical Center. This dataset consists of resting-state fMRI scans with Blood Oxygenation Level-Dependent (BOLD) signals recorded every 2 seconds in the whole brain with $L = 172$ locations in total, for $n_{\text{ADHD}} = 90$

Table 13: The average (standard error) functional matrix losses for partially observed functional scenarios and $p = 100$ over 100 simulation runs.

		$L_i = 11$		$L_i = 21$		$L_i = 51$		$L_i = 101$	
Model	Method	$\breve{\mathbf{\Sigma}}_{\text{A}}$	$\breve{\mathbf{\Sigma}}_{\text{U}}$	$\breve{\mathbf{\Sigma}}_{\text{A}}$	$\breve{\mathbf{\Sigma}}_{\text{U}}$	$\breve{\mathbf{\Sigma}}_{\text{A}}$	$\breve{\mathbf{\Sigma}}_{\text{U}}$	$\breve{\mathbf{\Sigma}}_{\text{A}}$	$\breve{\mathbf{\Sigma}}_{\text{U}}$
		Functional Frobenius norm							
1	Hard	11.40(0.03)	18.34(0.01)	9.63(0.03)	17.80(0.01)	8.55(0.04)	17.51(0.01)	8.17(0.04)	17.42(0.01)
	Soft	12.79(0.05)	18.33(0.01)	11.28(0.05)	17.71(0.02)	10.33(0.05)	16.68(0.07)	10.01(0.05)	16.06(0.07)
	SCAD	12.41(0.05)	18.33(0.01)	10.58(0.05)	17.72(0.02)	9.42(0.05)	16.77(0.06)	9.01(0.05)	16.23(0.07)
	Adap. lasso	11.22(0.04)	18.33(0.01)	9.59(0.04)	17.79(0.01)	8.54(0.04)	17.49(0.01)	8.19(0.04)	17.34(0.03)
		Functional matrix ℓ_1 norm							
	Hard	5.97(0.05)	9.41(0.01)	5.15(0.05)	9.35(0.01)	4.70(0.05)	9.33(0.01)	4.53(0.05)	9.32(0.01)
	Soft	7.06(0.04)	9.41(0.01)	6.55(0.05)	9.34(0.01)	6.23(0.05)	9.19(0.02)	6.12(0.05)	9.02(0.03)
	SCAD	6.93(0.05)	9.41(0.01)	6.20(0.05)	9.34(0.01)	5.74(0.05)	9.23(0.02)	5.56(0.05)	9.11(0.03)
	Adap.lasso	6.00(0.05)	9.41(0.01)	5.32(0.06)	9.35(0.01)	4.89(0.06)	9.32(0.01)	4.74(0.06)	9.32(0.01)
		Functional Frobenius norm							
2	Hard	13.21(0.04)	17.03(0.01)	11.33(0.04)	16.40(0.01)	10.06(0.04)	16.06(0.01)	9.60(0.04)	15.96(0.01)
	Soft	13.54(0.04)	17.01(0.01)	12.06(0.04)	16.26(0.02)	11.10(0.04)	15.32(0.05)	10.75(0.04)	14.86(0.05)
	SCAD	13.50(0.04)	17.01(0.01)	11.90(0.04)	16.26(0.02)	10.78(0.04)	15.35(0.05)	10.36(0.04)	14.93(0.05)
	Adap. lasso	12.61(0.04)	17.01(0.01)	10.94(0.04)	16.39(0.01)	9.80(0.04)	15.99(0.02)	9.37(0.04)	15.81(0.03)
		Functional matrix ℓ_1 norm							
	Hard	6.14(0.04)	7.27(0.01)	5.49(0.04)	7.19(0.01)	5.01(0.05)	7.16(0.01)	4.83(0.05)	7.15(0.01)
	Soft	6.22(0.02)	7.26(0.01)	5.90(0.03)	7.16(0.01)	5.65(0.03)	7.03(0.02)	5.55(0.03)	6.97(0.02)
	SCAD	6.21(0.02)	7.26(0.01)	5.87(0.03)	7.16(0.01)	5.58(0.03)	7.04(0.02)	5.45(0.03)	6.99(0.02)
	Adap. lasso	5.88(0.04)	7.26(0.01)	5.42(0.04)	7.19(0.01)	5.04(0.04)	7.15(0.01)	4.87(0.04)	7.14(0.01)

Table 14: The average TPRs/ FPRs for partially observed functional scenarios and $p = 100$ over 100 simulation runs.

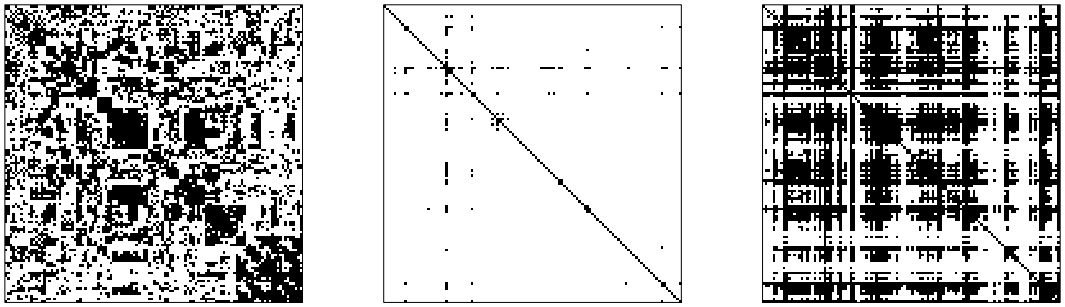
		$L_i = 11$		$L_i = 21$		$L_i = 51$		$L_i = 101$	
Model	Method	$\breve{\mathbf{\Sigma}}_{\text{A}}$	$\breve{\mathbf{\Sigma}}_{\text{U}}$	$\breve{\mathbf{\Sigma}}_{\text{A}}$	$\breve{\mathbf{\Sigma}}_{\text{U}}$	$\breve{\mathbf{\Sigma}}_{\text{A}}$	$\breve{\mathbf{\Sigma}}_{\text{U}}$	$\breve{\mathbf{\Sigma}}_{\text{A}}$	$\breve{\mathbf{\Sigma}}_{\text{U}}$
1	Hard	0.57/0.00	0.00/0.00	0.62/0.00	0.00/0.00	0.65/0.00	0.00/0.00	0.66/0.00	0.00/0.00
	Soft	0.80/0.03	0.00/0.00	0.83/0.04	0.03/0.01	0.85/0.04	0.24/0.05	0.85/0.04	0.36/0.07
	SCAD	0.81/0.03	0.00/0.00	0.84/0.04	0.03/0.01	0.85/0.04	0.22/0.04	0.85/0.04	0.32/0.06
	Adap. lasso	0.67/0.00	0.00/0.00	0.71/0.00	0.00/0.00	0.73/0.00	0.00/0.00	0.74/0.00	0.01/0.00
2	Hard	0.48/0.00	0.00/0.00	0.57/0.00	0.00/0.00	0.65/0.00	0.00/0.00	0.68/0.00	0.00/0.00
	Soft	0.90/0.03	0.00/0.00	0.94/0.04	0.07/0.01	0.96/0.04	0.29/0.04	0.97/0.04	0.40/0.05
	SCAD	0.90/0.03	0.00/0.00	0.95/0.04	0.06/0.01	0.96/0.05	0.28/0.03	0.97/0.05	0.37/0.04
	Adap. lasso	0.70/0.00	0.00/0.00	0.78/0.00	0.00/0.00	0.83/0.00	0.02/0.00	0.85/0.00	0.03/0.00

patients diagnosed with attention-deficit/hyperactivity disorder (ADHD) and $n_{\text{TDC}} = 87$ typically-developing controls (TDC). The preprocessing of the raw fMRI data is performed by Neuro Bureau using the Athena pipeline (Bellec et al., 2017). See Figure 11 in Section E.4 for plots of pre-smoothed BOLD signals at a selection of ROIs. Following Li and Solea (2018) based on the same dataset, we treat the signals at different ROIs as multivariate functional data. Our goal is to construct resting state functional connectivity networks among $p = 116$ ROIs (Tzourio-Mazoyer et al., 2002), with the first 90 ROIs from the cerebrum and the last 26 ROIs from the cerebellum, for ADHD and TDC groups, respectively. To this end, we implement adaptive and universal functional thresholding methods to discover the networks for two groups.

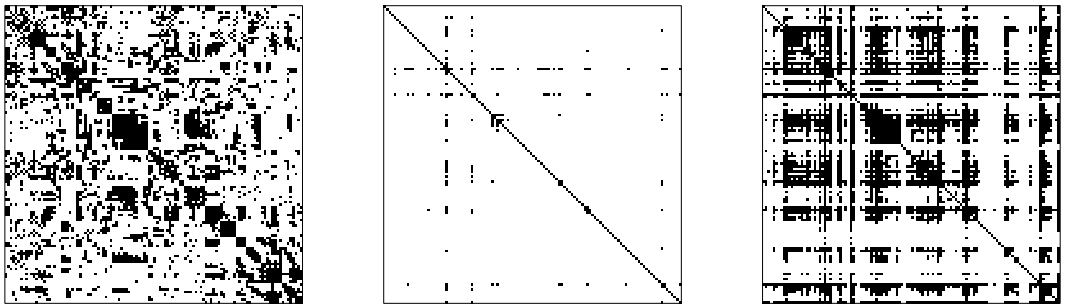
Figure 10 plots the sparsity patterns in estimated covariance functions corresponding to identified functional connectivity networks. We observe several interesting patterns. First, with $\hat{\lambda}$ selected by the cross-validation, $\widehat{\mathbf{\Sigma}}_{\text{A}}$ in Fig. 10(a)–(b) reveal clear blockwise connectivity structures with two blocks coinciding with the regions of the cerebrum and the cerebellum, while $\widehat{\mathbf{\Sigma}}_{\text{U}}$ in Fig. 10(c)–(d) result in very sparse networks. Second, under the same sparsity levels as those of $\widehat{\mathbf{\Sigma}}_{\text{A}}$ in Fig. 10(a)–(b), $\widehat{\mathbf{\Sigma}}_{\text{U}}$ in Fig. 10(e)–(f) only retain edges related to large marginal-covariance functions but fail to identify some essential within-network connections, e.g., those of the cerebellar region (Dobromyslin et al., 2012) on the bottom right corner. Third, the ADHD group has increased connections relative to the TDC group, which is in line with the finding in Konrad and Eickhof (2010) that ADHD patients tend to exhibit abnormal spontaneous functional connectivity patterns.

E.4 Additional real data results

Figures 11 and 12 display the pre-smoothed BOLD signal trajectories at a selection of ROIs of subjects from the ADHD and HCP datesets, respectively. Moreover, we zoom in on a randomly selected subinterval $(0.5, 0.6)$ of time $[0, 1]$ in Figure 12 and plot the pre-smoothed BOLD signals during this subinterval in Figure 13. It is evident that the



(a) ADHD: $\widehat{\mathbf{\Sigma}}_{\text{A}}$ (57.50% zeros) (c) ADHD: $\widehat{\mathbf{\Sigma}}_{\text{U}}$ (98.94% zeros) (e) ADHD: $\widehat{\mathbf{\Sigma}}_{\text{U}}$ (57.50% zeros)



(b) TDC: $\widehat{\mathbf{\Sigma}}_{\text{A}}$ (71.24% zeros) (d) TDC: $\widehat{\mathbf{\Sigma}}_{\text{U}}$ (98.85% zeros) (f) TDC: $\widehat{\mathbf{\Sigma}}_{\text{U}}$ (71.24% zeros)

Figure 10: The sparsity structures in $\widehat{\mathbf{\Sigma}}_{\text{A}}$ and $\widehat{\mathbf{\Sigma}}_{\text{U}}$ for ADHD and TDC groups: (a)–(d) with the corresponding $\hat{\lambda}$ selected by fivefold cross-validation using soft functional thresholding rule; (e)–(f) with the same sparsity levels as those in (a)–(b). Black corresponds to non-zero entries of $\widehat{\mathbf{\Sigma}}_{\text{A}}$ and $\widehat{\mathbf{\Sigma}}_{\text{U}}$ (identified edges connecting a subset of ROIs).

smoothed signal trajectories are directly available after the standard preprocessing step following the existing neuroscience literature. Figures 14 and 15 plot the connectivity strengths at fluid intelligence $gF \leqslant 8$ and $gF \geqslant 23$ in Fig. 1(a)–(b) and Fig. 1(c)–(d), respectively. We observe that as gF increases, the connectivity strengths in the medial frontal and frontoparietal modules tend to increase while those in the default mode module decrease, which is consistent with our finding in Section 6. Finally, we present in Figure 16 some randomly selected entries of $\widehat{\mathbf{\Sigma}}_{\text{A}}$ in Fig. 1 (c)–(d) during the same subinterval $(0.5, 0.6)$. The dynamic structures of the estimated covariance surfaces are observable in the sense that the values of $\widehat{\Sigma}^{\text{A}}_{jk}$ change from positive to negative (or from negative to positive) as a function of (u, v) along some directions. In particular, there seems to exist an interesting common positive and negative banding pattern as $|u - v|$ varies in the presented entries.

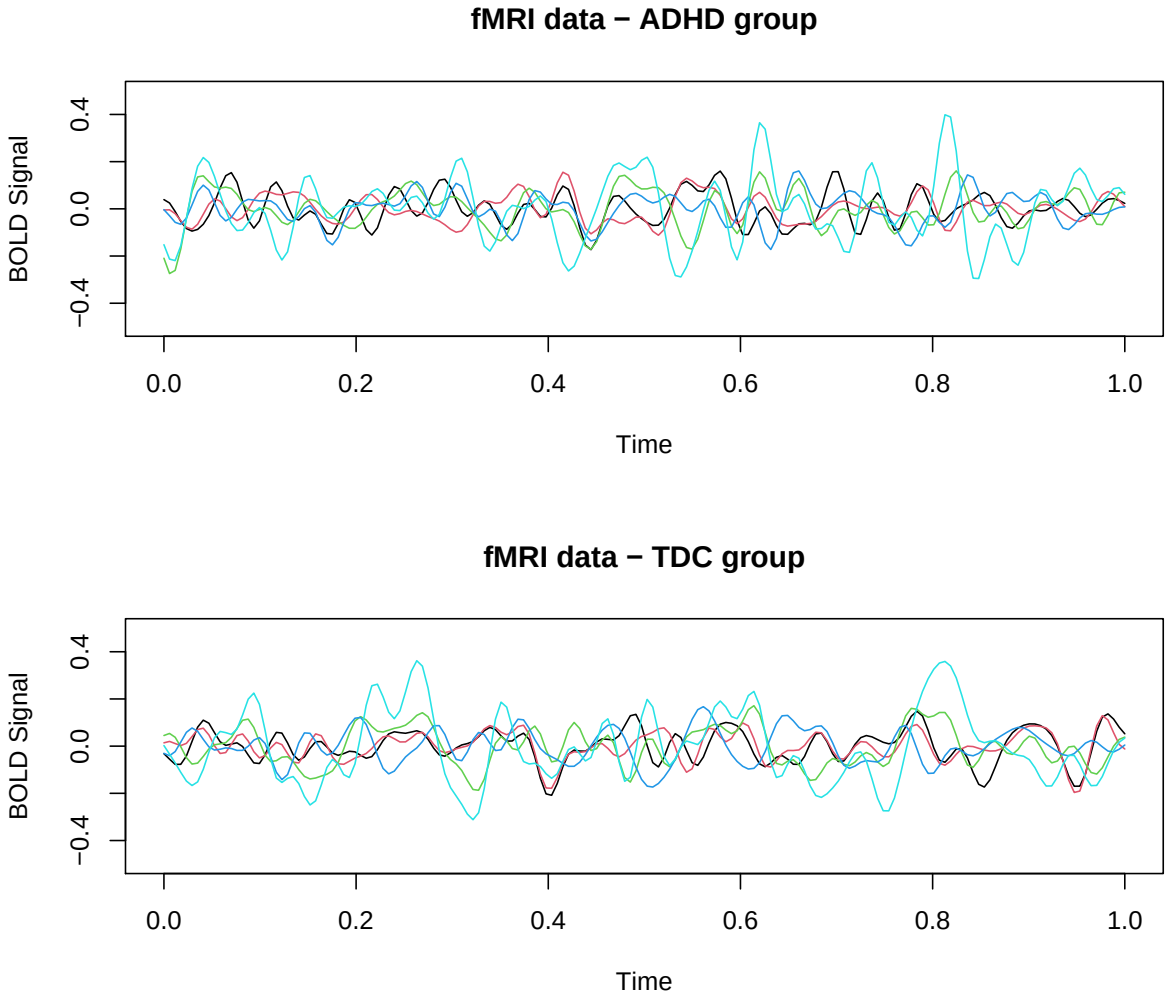


Figure 11: ADHD dataset: the smoothed BOLD signals at the first 5 ROIs of two subjects in ADHD and TDC groups respectively. The 5.73-minute interval with 172 scanning points is rescaled to $[0, 1]$.

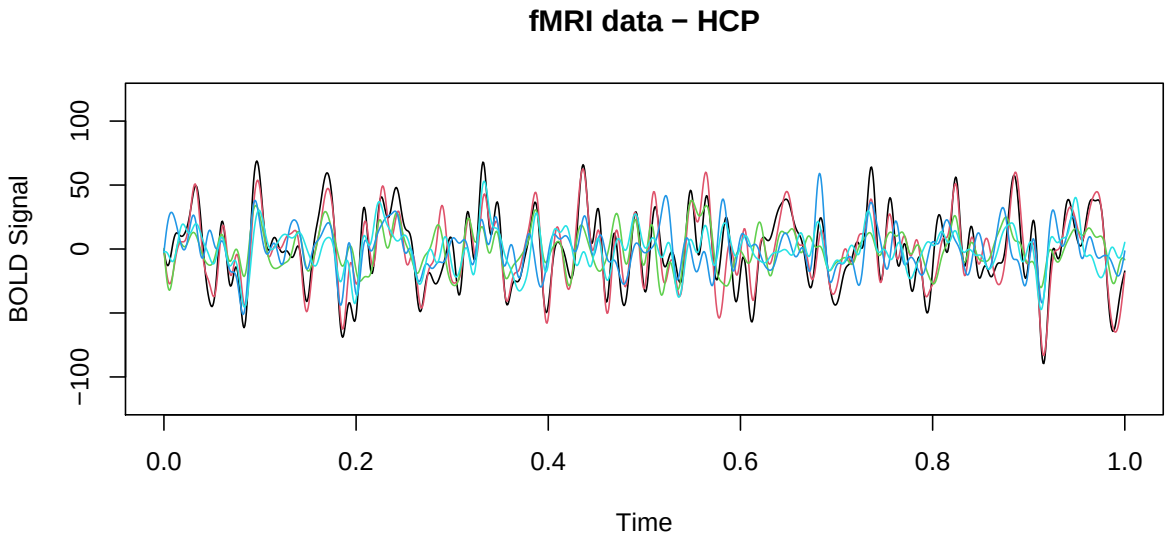


Figure 12: HCP dataset: the smoothed BOLD signals at the first 5 ROIs of one subject. The 14.40-minute interval with 1200 scanning points (14.40 mins) is rescaled to $[0, 1]$.

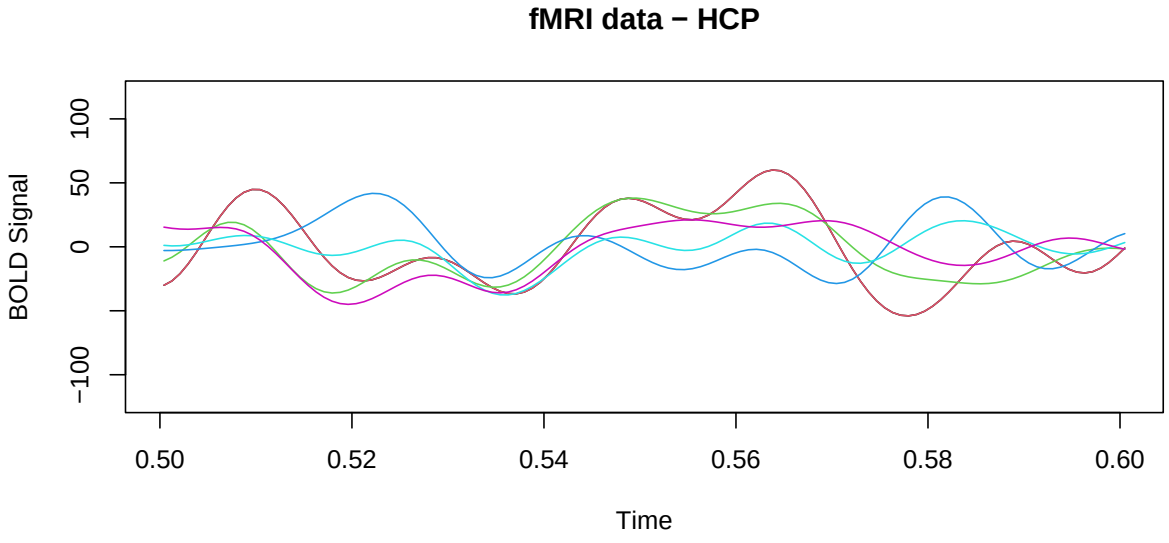
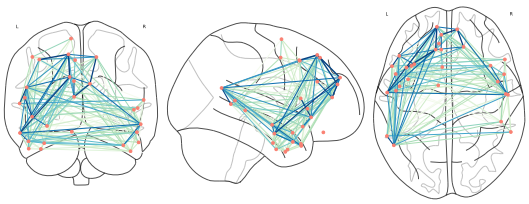
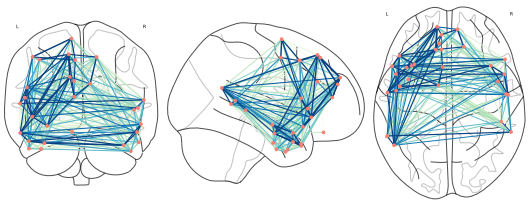


Figure 13: HCP dataset: the smoothed BOLD signals in Fig. 12 during (0.5,0.6).

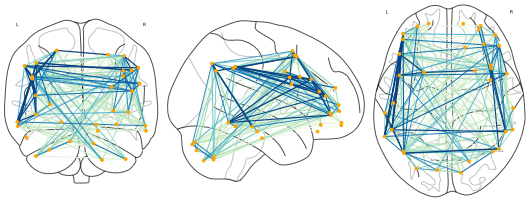


(a) $gF \leqslant 8$: the medial frontal module in Fig. 1(a)

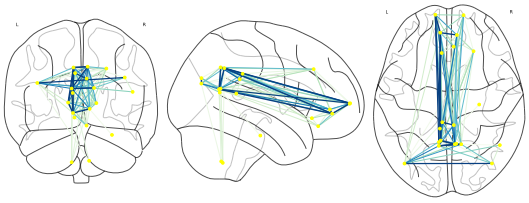


(d) $gF \geqslant 23$: the medial frontal module in Fig. 1(b)

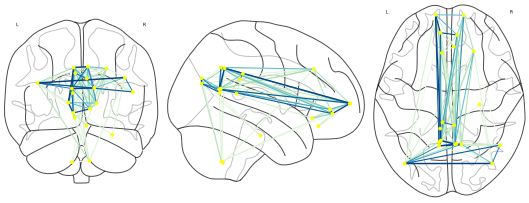
(b) $gF \leqslant 8$: the frontoparietal module in Fig. 1(a)



(e) $gF \geqslant 23$: the frontoparietal module in Fig. 1(b)

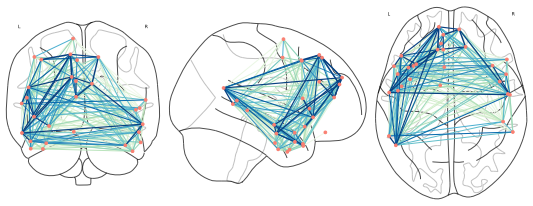


(c) $gF \leqslant 8$: the default mode module in Fig. 1(a)

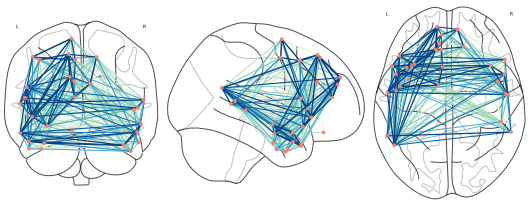


(f) $gF \geqslant 23$: the default mode module in Fig. 1(b)

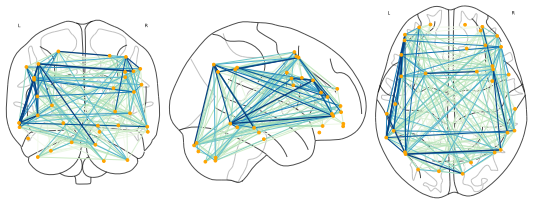
Figure 14: The connectivity strengths in Fig. 1(a)–(b) at fluid intelligence $gF \leqslant 8$ and $gF \geqslant 23$. Salmon, orange and yellow nodes represent the ROIs in the medial frontal, frontoparietal and default mode modules, respectively. The edge color from cyan to blue corresponds to the value of $\|\widehat{\Sigma}^{\mathrm{A}}_{jk}\|_{\mathcal{S}}/(\|\widehat{\Sigma}^{\mathrm{A}}_{jj}\|_{\mathcal{S}}\|\widehat{\Sigma}^{\mathrm{A}}_{kk}\|_{\mathcal{S}})^{1/2}$ from small to large.



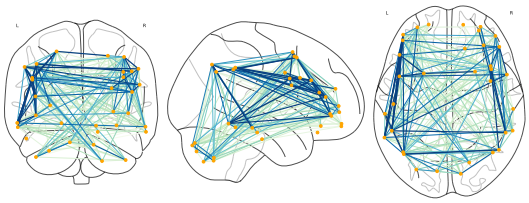
(a) $gF \leqslant 8$: the medial frontal module in Fig. 1(c)



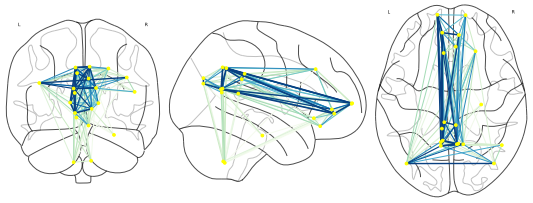
(d) $gF \geqslant 23$: the medial frontal module in Fig. 1(d)



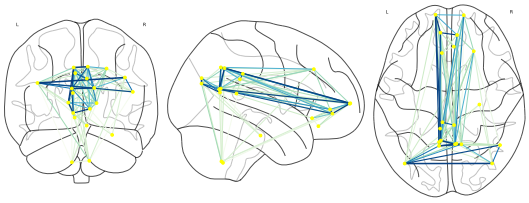
(b) $gF \leqslant 8$: the frontoparietal module in Fig. 1(c)



(e) $gF \geqslant 23$: the frontoparietal module in Fig. 1(d)



(c) $gF \leqslant 8$: the default mode module in Fig. 1(c)



(f) $gF \geqslant 23$: the default mode module in Fig. 1(d)

Figure 15: The connectivity strengths in Fig. 1(c)–(d) at fluid intelligence $gF \leqslant 8$ and $gF \geqslant 23$. Salmon, orange and yellow nodes represent the ROIs in the medial frontal, frontoparietal and default mode modules, respectively. The edge color from cyan to blue corresponds to the value of $\|\widehat{\Sigma}^{\mathrm{A}}_{jk}\|_{\mathcal{S}}/(\|\widehat{\Sigma}^{\mathrm{A}}_{jj}\|_{\mathcal{S}}\|\widehat{\Sigma}^{\mathrm{A}}_{kk}\|_{\mathcal{S}})^{1/2}$ from small to large.

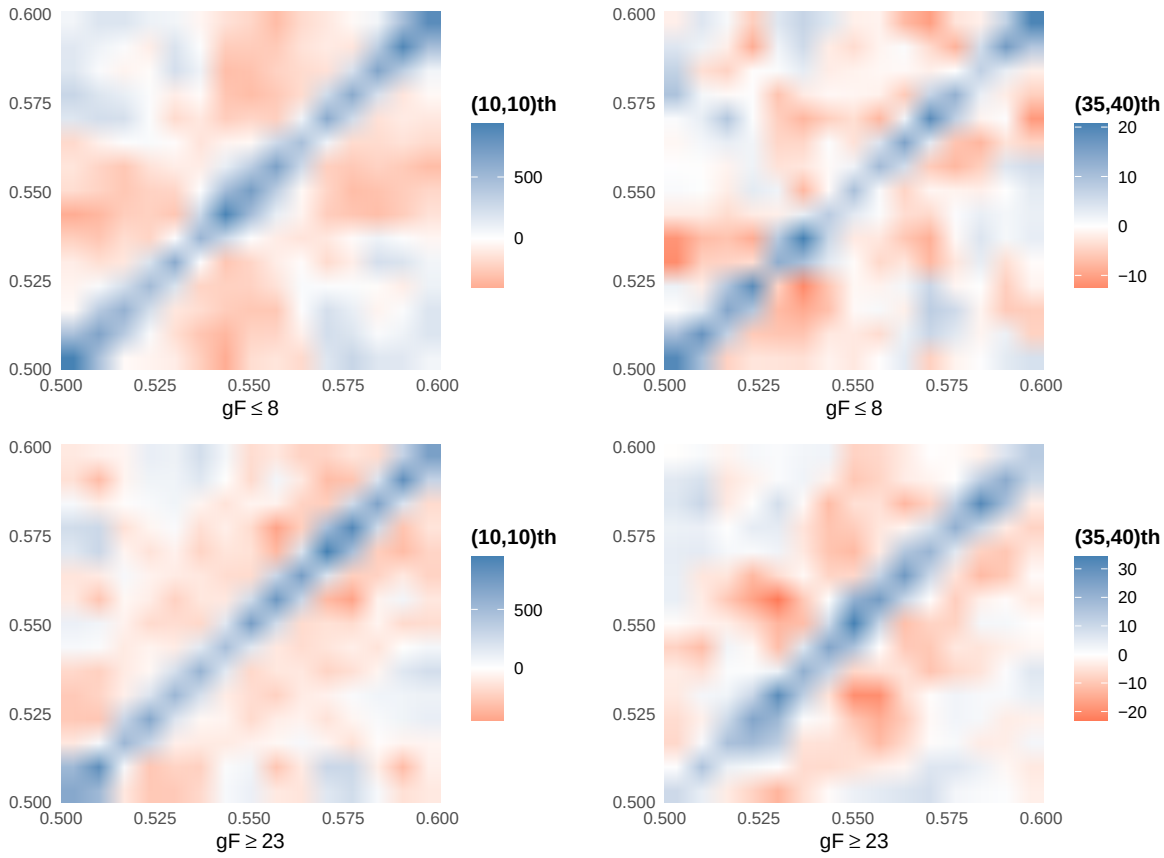


Figure 16: Selected entries of $\widehat{\mathbf{\Sigma}}_{\mathrm{A}}$ in Fig. 1 (c)–(d) during (0.5,0.6).

References

Bellec, P., Chu, C., Chouinard-Decorte, F., Benhajali, Y., Margulies, D. S. and Craddock, R. C. (2017). The neuro bureau ADHD-200 preprocessed repository, *Neuroimage*, **144**: 275–286.

Bosq, D. (2000). *Linear Process in Function Spaces*. New York: Springer.

Boucheron, S., Lugosi, G. and Massart, P. (2014). *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press.

Chen, Z. and Leng, C. (2016). Dynamic covariance models, *Journal of the American Statistical Association*, **111**: 1196-1207.

Chiou, J.-M., Chen, Y.-T. and Yang, Y.-F. (2014). Multivariate functional principal component analysis: a normalization approach, *Statistica Sinica*, **24**: 1571-1596.

Chiou, J.-M., Yang, Y.-F. and Chen, Y.-T. (2016). Multivariate functional linear regression and prediction, *Journal of Multivariate Analysis*, **146**: 301-312.

Dobromyslin, V. I., Salat, D. H., Fortier, C. B., Leritz, E. C., Beckmann, C. F., Milberg, W. P. and McGlinchey, R. E. (2012). Distinct functional networks within the cerebellum and their relation to cortical systems assessed with independent component analysis, *Neuroimage*, **60**: 2073–2085.

Happ, C. and Greven, S. (2018). Multivariate functional principal component analysis for data observed on different (dimensional) domains, *Journal of the American Statistical Association*, **113**: 649–659.

Fan, Y. and Lv, J. (2016). Innovated scalable efficient estimation in ultra-large Gaussian graphical models, *The Annals of Statistics*, **44**: 2098–2126.

Konrad, K. and Eickhoff, S. B. (2010). Is the ADHD brain wired differently? A review on structural and functional connectivity in attention deficit hyperactivity disorder, *Human Brain Mapping*, **31**: 904–916.

Kosorok, M. R. (2008). *Introduction to Empirical Processes and Semiparametric Inference.* Springer Series in Statistics. Springer, New York.

Li, B. and Solea, E. (2018). A nonparametric graphical model for functional data with application to brain networks based on fMRI, *Journal of the American Statistical Association*, **113**: 1637–1655.

Park, J., Ahn, J. and Jeon, Y. (2021). Sparse functional linear discriminant analysis, *Biometrika*, **109**: 209–226.

Qiao, X., Guo, S. and James, G. (2019). Functional graphical models, *Journal of the American Statistical Association*, **114**: 211–222.

Qiao, X., Qian, C., James, G. M. and Guo, S. (2020). Doubly functional graphical models in high dimensions. *Biometrika*, **107**: 415–431.

Rothman, A. J., Levina, E. and Zhu, J. (2009). Generalized thresholding of large covariance matrices. *Journal of the American Statistical Association*, **104**: 177–186.

Tzourio-Mazoyer, N., Landeau, B., Papathanassiou, D., Crivello, F., Etard, O., Delcroix, N., Mazoyer, B. and Joliot, M. (2002). Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain, *Neuroimage*, **15**: 273–289.