Title

Partial separability and functional graphical models for multivariate Gaussian processes

Authors

Zapata, J
Oh, SY
Petersen, A

Peer reviewed

Partial Separability and Functional Graphical Models for Multivariate Gaussian Processes

Javier Zapata[*1], Sang-Yun Oh[2], and Alexander Petersen[3]

[1,2]Department of Statistics and Applied Probability
University of California Santa Barbara
[2]Lawrence Berkeley Lab, Berkeley, CA
[3]Department of Statistics, Brigham Young University, Provo, Utah, U.S.A.

January 6, 2022

Abstract

The covariance structure of multivariate functional data can be highly complex, especially if the multivariate dimension is large, making extensions of statistical methods for standard multivariate data to the functional data setting challenging. For example, Gaussian graphical models have recently been extended to the setting of multivariate functional data by applying multivariate methods to the coefficients of truncated basis expansions. However, a key difficulty compared to multivariate data is that the covariance operator is compact, and thus not invertible. The methodology in this paper addresses the general problem of covariance modeling for multivariate functional data, and functional Gaussian graphical models in particular. As a first step, a new notion of separability for the covariance operator of multivariate functional data is proposed, termed partial separability, leading to a novel Karhunen-Loève-type expansion for such data. Next, the partial separability structure is shown to be particularly useful in order to provide a well-defined functional Gaussian graphical model that can be identified with a sequence of finite-dimensional graphical models, each of identical fixed dimension. This motivates a simple and efficient estimation procedure through application of the joint graphical lasso. Empirical performance of the method for graphical model estimation is assessed through simulation and analysis of functional brain connectivity during a motor task.

Keywords: Functional Data, Functional Brain Connectivity; Inverse Covariance, Separability

---
[*]Corresponding Author

1 Introduction

The analysis of functional data continues to be an important field for statistical development given the abundance of data collected over time via sensors or other tracking equipment. Frequently, such time-dependent signals are vector-valued, resulting in multivariate functional data. Prominent examples include longitudinal behavioral tracking (Chiou & Müller 2014), blood protein levels (Dubin & Müller 2005), traffic measurements (Chiou et al. 2014, 2016), and neuroimaging data (Petersen & Müller 2016, Happ & Greven 2018), for which dimensionality reduction and regression have been the primary methods investigated. As for standard multivariate data, the nature of dependencies between component functions of multivariate functional data constitute an important question requiring careful consideration.

Dependencies between functional magnetic resonance imaging (fMRI) signals for a large number of regions across the brain during a motor task experiment are the motivating example for this paper. Since fMRI signals are collected simultaneously, it is natural to model these as a multivariate process $\{X(t) \in \mathbb{R}^p : t \in \mathcal{T}\}$, where $\mathcal{T} \subset \mathbb{R}$ is a time interval over which the scans are taken (Qiao et al. 2019). The dual multivariate and functional aspects of the data make the covariance structure of X complex, particularly if the multivariate dimension p is large. This leads to difficulties in extending highly useful multivariate analysis techniques, such as graphical models, to multivariate functional data without further structural assumptions. For example, in the analogous setting of spatio-temporal data, it is common to impose further structure to the covariance, usually assuming that the spatial and temporal effects can be separated in some way. However, similar notions for multivariate functional data have not yet been considered.

As for ordinary multivariate data, the conditional independence properties of X are perhaps of greater interest than the marginal covariance, leading to the consideration of inverse covariance operators and graphical models for functional data. If X is Gaussian, each component function X_j corresponds to a node in the functional Gaussian graphical model, which is a single network of p nodes. This is inherently different from the estimation of time-dependent graphical models (Zhou et al. 2010, Kolar & Xing 2011, Qiu et al. 2016, Qiao et al. 2020), in which the graph is dynamic and has nodes corresponding to scalar random variables. In this paper, the graph is considered to be static while each node represents an infinite-dimensional functional object. This is an important distinction, as covariance operators for functional data are compact and thus not invertible in the usual sense, so that presence or absence of edges cannot in general be identified immediately with zeros in any precision operator. In the past few years, there has been some investigation into this problem. Zhu et al. (2016) developed a Bayesian framework for graphical models on product function spaces, including the extension of Markov laws and appropriate prior distributions. Qiao et al. (2019) implemented a truncation approach, whereby each function is represented by the coefficients of a truncated basis expansion using functional principal components analysis, and a finite-dimensional graphical model is estimated by a modified graphical lasso criterion. Li & Solea (2018) developed a

non-Gaussian variant, where conditional independence was replaced by a notion of so-called additive conditional independence.

The methodology proposed in this paper is within the setting of multivariate Gaussian processes as in Qiao et al. (2019), and exploits a notion of separability for multivariate functional data to develop efficient estimation of suitable inverse covariance objects. There are at least three novel contributions of this methodology to the fields of functional data analysis and Gaussian graphical models. First, a structure termed partial separability is defined for the covariance operator of multivariate functional data, yielding a novel Karhunen-Loève type representation. The second contribution is to show that, when the process is indeed partially separable, the functional graphical model is well-defined and can be identified with a sequence of finite-dimensional graphical models. In particular, the assumption of partial separability overcomes the problem of noninvertibility of the covariance operator when X is infinite-dimensional, in contrast with Zhu et al. (2016), Qiao et al. (2019) which assumed that the functional data were concentrated on finite-dimensional subspaces. Third, an intuitive estimation procedure is developed based on simultaneous estimation of multiple graphical models. Furthermore, theoretical properties are derived under the regime of fully observed functional data. Empirical performance of the proposed method is then compared to that of Qiao et al. (2019) through simulations involving dense and noisily observed functional data, including a setting where partial separability is violated. Finally, the method is applied to the study of brain connectivity (also known as functional connectivity in the neuroscience literature) using data from the Human Connectome Project corresponding to a motor task experiment. Through these practical examples, our proposed method is shown to provide improved efficiency in estimation and computation. An R package `fgm` implementing the proposed methods is freely available via the CRAN repository.

2 Preliminaries

2.1 Gaussian Graphical Models

Consider a p-variate random variable $\theta = (\theta_1, \ldots, \theta_p)^{\mathrm{T}}$, $p > 2$. For any distinct indices $j, k = 1, \ldots, p$, let $\theta_{-(j,k)} \in \mathcal{R}^{p-2}$ denote the subvector of θ obtained by removing its jth and kth entries. A graphical model (Lauritzen 1996) for θ is an undirected graph $G = (V, E)$, where $V = \{1, \ldots, p\}$ is the node set and $E \subset V \times V \setminus \{(j, j) : j \in V\}$ is called the edge set. The edges in E encode the presence or absence of conditional independencies amongst the distinct components of θ by excluding (j, k) from E if and only if $\theta_j \perp\!\!\!\perp \theta_k \mid \theta_{-(j,k)}$. In the case that $\theta \sim \mathcal{N}_p(0, \Sigma)$, the corresponding Gaussian graphical model is intimately connected to the positive definite covariance matrix Σ through its inverse $\Omega = \Sigma^{-1}$, known as the precision matrix of θ. Specifically, the edge set E can be readily obtained from Ω by the relation $(j, k) \in E$ if and only if $\Omega_{jk} \neq 0$ (Lauritzen 1996). This identification of edges in E with the non-zero off-diagonal entries of Ω is due to the simple fact that the latter are proportional to the conditional covariance between components. Thus, the zero/non-zero structure of Ω serves as an adjacency matrix of the graph G, making disposable a vast

number of statistical tools for sparse inverse covariance estimation in order to recover a sparse graph structure from data.

2.2 Functional Gaussian Graphical Models

We first introduce some notation. Let $L^2[0,1]$ denote the space of square-integrable measurable functions on $[0,1]$ endowed with the standard inner product $\langle g_1, g_2 \rangle = \int_0^1 g_1(t)g_2(t)\,\mathrm{d}t$ and associated norm $\|\cdot\|$. $(L^2[0,1])^p$ is its p-fold Cartesian product or direct sum, endowed with inner product $\langle f_1, f_2 \rangle_p = \sum_{j=1}^p \langle f_{1j}, f_{2j} \rangle$ and its associated norm $\|\cdot\|_p$. For a generic compact covariance operator $\mathcal{A}$ defined on an arbitrary Hilbert space, let $\lambda_j^{\mathcal{A}}$ denote its j-th largest eigenvalue. Suppose $f \in (L^2[0,1])^p$, $g \in L^2[0,1]$, $a \in \mathcal{R}^p$, Δ is a $p \times p$ matrix, and $\mathcal{B} : L^2[0,1] \to L^2[0,1]$ is a linear operator. Then $ag \in (L^2[0,1])^p$ takes values $\{g(x)\}a \in \mathcal{R}^p$, $\Delta f \in (L^2[0,1])^p$ takes values $\Delta\{f(x)\} \in \mathcal{R}^p$, $\mathcal{B}(f) = (\mathcal{B}(f_1), \ldots, \mathcal{B}(f_p)) \in (L^2[0,1])^p$, and $(\Delta \mathcal{B})(f) = \mathcal{B}(\Delta f)$. The tensor products $g \otimes g$ and $f \otimes_p f$ signify the operators $(g \otimes g)(\cdot) = \langle g, \cdot \rangle g$ and $(f \otimes_p f)(\cdot) = \langle f, \cdot \rangle_p f$ on $L^2[0,1]$ and $(L^2[0,1])^p$, respectively.

In this paper, multivariate functional data constitute a random sample from a multivariate process $\{X(t) \in \mathbb{R}^p : t \in [0,1]\}$, which, for the moment, is assumed to be zero-mean such that $X \in (L^2[0,1])^p$ almost surely and $E\left(\|X\|_p^2\right) < \infty$. If X is also assumed to be Gaussian, then its distribution is uniquely characterized by its covariance operator $\mathcal{G}$, the infinite-dimensional counterpart of the covariance matrix for standard multivariate data. In fact, one can think of it as a matrix of operators $\mathcal{G} = \{\mathcal{G}_{jk} : j, k \in V\}$, where each entry $\mathcal{G}_{jk}$ is a linear, trace class integral operator on $L^2[0,1]$ (Hsing & Eubank 2015) with kernel $G_{jk}(s,t) = \mathrm{cov}\{X_j(s), X_k(t)\}$. That is, for any $g \in L^2[0,1]$, $\mathcal{G}_{jk}(g)(\cdot) = \int_0^1 G_{jk}(\cdot, t)g(t)\mathrm{d}t$. Then $\mathcal{G}$ is an integral operator on $(L^2[0,1])^p$ with $\{\mathcal{G}(f)\}_j = \sum_{k=1}^p \mathcal{G}_{jk}(f_k)$ ($f \in (L^2[0,1])^p$, $j \in V$).

A functional Gaussian graphical model for X is a graph $G = (V, E)$ that encodes the conditional independency structure amongst its components. As in the finite-dimensional case, the edge set can be recovered from the conditional covariance functions

$$C_{jk}(s,t) = \mathrm{cov}\{X_j(s), X_k(t) \mid X_{-(j,k)}\} \quad (j, k \in V, j \neq k), \tag{1}$$

through the relation $(j,k) \in E$ if and only if $C_{jk}(s,t) = 0$ for all $s, t \in [0,1]$. However, unlike the finite-dimensional case, the covariance operator $\mathcal{G}$ is compact and thus not invertible, with the consequence that the connection between conditional independence and an inverse covariance operator is lost, as the latter does not exist. This is an established issue for infinite-dimensional functional data, for instance in linear regression models with functional predictors; see Wang et al. (2016) and references therein. Thus, a common approach is to regularize the problem by first performing dimensionality reduction, most commonly through a truncated basis expansion of the functional data. Specifically, one chooses an orthonormal functional basis $\{\phi_{jl}\}_{l=1}^\infty$ of $L^2[0,1]$ for

each j, and expresses each component of X as

$$X_j(t) = \sum_{l=1}^{\infty} \xi_{jl}\phi_{jl}(t), \quad \xi_{jl} = \int_0^1 X_j(t)\phi_{jl}(t)\mathrm{d}t. \tag{2}$$

These expansions are then truncated at a finite number of basis functions to perform estimation, and the basis size is allowed to diverge with the sample size to obtain asymptotic properties.

Previous work related to functional Gaussian graphical models include Zhu et al. (2016) and Qiao et al. (2019). The former rigorously considered the notion of conditional independence for functional data, and proposed a family of priors for the covariance operator $\mathcal{G}$. The latter truncated (2) at L terms using the functional principal component basis (Hsing & Eubank 2015), and set $\xi_j = (\xi_{j1}, \ldots, \xi_{jL})^{\mathrm{T}}$ ($j \in V$). Qiao et al. (2019) define a $pL \times pL$ covariance matrix Γ blockwise for the concatenated vector $(\xi_1^{\mathrm{T}}, \ldots, \xi_p^{\mathrm{T}})^{\mathrm{T}}$, as $\Gamma = (\Gamma_{jk})_{j,k=1}^p, (\Gamma_{jk})_{lm} = \mathrm{cov}(\xi_{jl}, \xi_{km}), \ (l, m = 1, \ldots, L)$. Then, a functional graphical lasso algorithm was developed to estimate Γ^{-1} with sparse off-diagonal blocks in order to estimate the edge set.

The method of Qiao et al. (2019) constitutes an intuitive approach to functional graphical model estimation, but encounters some difficulties that we seek to address in this paper. From a theoretical point of view, even when p is fixed, consistent estimation of the graphical model requires that one permit L to diverge, so that the number of covariance parameters needing to be estimated is $(pL)^2$. Additionally, the identification of zero off-diagonal blocks in Γ^{-1} was only shown to be linked to the true functional graphical model under the strict assumption that each X_j take values in a finite-dimensional space almost surely. In many practical applications, the dimension p can be high, the number of basis functions L may need to be large in order to retain a suitable representation of the observed data, or both of these may occur simultaneously. It is thus desirable to introduce structure on $\mathcal{G}$ in order to provide a parsimonious basis expansion for multivariate functional data that is amenable to graphical model estimation.

3 Partial Separability

3.1 A Parsimonious Basis for Multivariate Functional Data

Functional principal component analysis is a commonly used tool for functional data, with one of its most useful features being the parsimonious reduction of each univariate component X_j to a countable sequence of uncorrelated random variables through the Karhunen-Loève expansion (Hsing & Eubank 2015), taking the form of (2) when the basis is chosen as the eigenbasis of $\mathcal{G}_{jj}$. Chiou et al. (2014) extended this to multivariate functional data via the spectral decomposition $\mathcal{G} = \sum_{m=1}^{\infty} \lambda_m^{\mathcal{G}} \rho_m \otimes_p \rho_m$, leading to the multivariate Karhunen-Loève expansion $X(t) = \sum_{l=1}^{\infty} \langle X, \rho_l \rangle_p \rho_l(t)$ where $\{\rho_l\}_{l=1}^{\infty}$ is an orthonormal basis of $(L^2[0, 1])^p$. While this decomposition is indeed parsimonious, the multivariate aspect of the data is lost since the random coefficients $\langle X, \rho_l \rangle_p$ are scalar. As a consequence, one cannot readily apply tools from finite-

dimensional Gaussian graphical model estimation to these coefficients. We begin by proposing a novel structural assumption on the eigenfunctions of $\mathcal{G}$, termed *partial separability*, and will then demonstrate its advantages for defining and estimating functional Gaussian graphical models.

Definition 1. A covariance operator $\mathcal{G}$ on $(L^2[0,1])^p$ is *partially separable* if there exist orthonormal bases $\{e_{lj}\}_{j=1}^p$ ($l \in \mathbb{N}$) of $\mathcal{R}^p$ and $\{\varphi_l\}_{l=1}^\infty$ of $L^2[0,1]$ such that the eigenfunctions of $\mathcal{G}$ take the form $e_{lj}\varphi_l$ ($l \in \mathbb{N}$, $j \in V$).

We first draw a connection to separability of covariance operators as they appear in spatiotemporal analyses, after which the implications of partial separability will be further explored. Dependent functional data arise naturally in the context of a spatiotemporal random field that is sampled at p discrete spatial locations. In many instances, it is assumed that the covariance of X is *separable*, meaning that there exists a $p \times p$ covariance matrix Δ and covariance operator $\mathcal{B}$ on $L^2[0,1]$ such that $\mathcal{G} = \Delta \mathcal{B}$ (Gneiting et al. 2006, Genton 2007, Aston et al. 2017). Letting $\{e_j\}_{j=1}^p$ and $\{\varphi_l\}_{l=1}^\infty$ be the orthonormal eigenbases of Δ and $\mathcal{B}$ respectively, it is clear that $e_j\varphi_l$ are the eigenfunctions of $\mathcal{G}$. Hence, a separable covariance operator $\mathcal{G}$ satisfies the conditions of Definition 1. It should also be noted that the property of $\mathcal{G}$ having eigenfunctions of the form $e_j\varphi_l$ has also been referred to as *weak separability* (Lynch & Chen 2018), and is a consequence and not a characterization of separability. The connections between these three separability notions are summarized in the following result, whose proof is simple, and thus omitted.

Proposition 1. *Suppose $\mathcal{G}$ is partially separable according to Definition 1. Then $\mathcal{G}$ is also weakly separable if and only if the bases $\{e_{lj}\}_{j=1}^p$ do not depend on l. If $\mathcal{G}$ is weakly separable, then it is also separable if and only if the eigenvalues take the form $\langle \mathcal{G}(e_j\varphi_l), e_j\varphi_l \rangle_p = c_j d_l$ for positive sequences $\{c_j\}_{j=1}^p$, $\{d_l\}_{l=1}^\infty$.*

The next result gives several characterizations of partial separablity. The proof of this and all remaining theoretical results can be found in the Appendix.

Theorem 1. *Let $\{\varphi_l\}_{l=1}^\infty$ by an orthonormal basis of $L^2[0,1]$. The following are equivalent:*

1. *$\mathcal{G}$ is partially separable with $L^2[0,1]$ basis $\{\varphi_l\}_{l=1}^\infty$.*

2. *There exists a sequence of $p \times p$ covariance matrices $\{\Sigma_l\}_{l=1}^\infty$ such that*

$$\mathcal{G} = \sum_{l=1}^\infty \Sigma_l \varphi_l \otimes \varphi_l.$$

3. *The covariance operator of each X_j can be written as $\mathcal{G}_{jj} = \sum_{l=1}^\infty \sigma_{ljj}\varphi_l \otimes \varphi_l$, with $\sigma_{ljj} > 0$ and $\sum_{l=1}^\infty \sigma_{ljj} < \infty$, and $\operatorname{cov}(\langle X_j, \varphi_l \rangle, \langle X_k, \varphi_{l'} \rangle) = 0$ ($j, k \in V$, $l \neq l'$).*
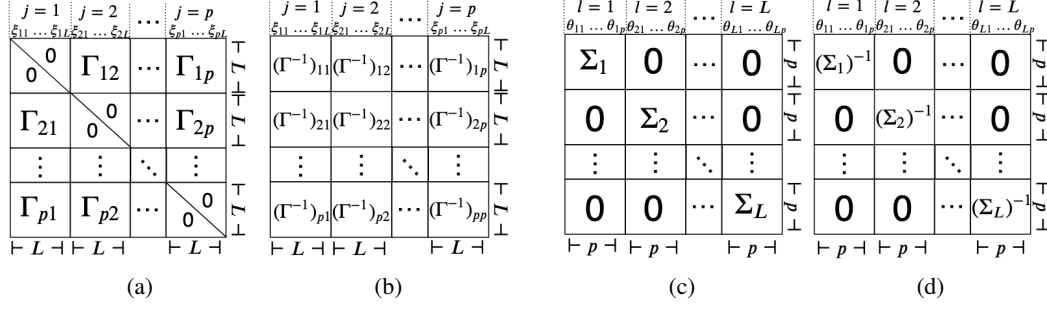
6

Figure 1: Covariance structures of $\mathbb{R}^{Lp}$-valued random coefficients from different L-truncated Karhunen-Loève type expansions. (a) and (b): covariance and precision matrices, respectively, of functional principal component coefficients $(\xi_1^T, \ldots, \xi_p^T)^T$ in (2) as in Qiao et al. (2019). (c) and (d): block diagonal covariance and precision matrices, respectively, of coefficients $(\theta_1{}^{\mathrm{T}}, \ldots, \theta_L{}^{\mathrm{T}})^{\mathrm{T}}$ under partial separability in (3).

4. The expansion

$$X = \sum_{l=1}^{\infty} \theta_l \varphi_l, \quad \theta_l = (\langle X_1, \varphi_l \rangle, \ldots, \langle X_p, \varphi_l \rangle)^{\mathrm{T}}, \tag{3}$$

holds almost surely in $(L^2[0,1])^p$, where the θ_l are mutually uncorrelated random vectors.

As will be seen in Section 3.2, the matrices Σ_l in point 2 of Theorem 1 contain all of the necessary information to form the functional graphical model when X is Gaussian and $\mathcal{G}$ is partially separable. For clarity, when $\mathcal{G}$ is partially separable, the expansion in point 2 is assumed to be ordered according to decreasing values of $\mathrm{tr}\,(\Sigma_l)$. Property 3 reveals that the $\mathcal{G}_{jj}$ share common eigenfunctions and are thus simultaneously diagonalizable, with projections of any features onto different eigenfunctions being uncorrelated. This is related to the concept known as *coregionalization* (Banerjee & Johnson 2006) but is more general in that it allows for different eigenvalues for each $\mathcal{G}_{jj}$. Consequently, one obtains the vector Karhunen-Loève type expansion in (3). If one truncates (3) at L components, the covariance matrix of the concatenated vector $(\theta_1{}^{\mathrm{T}}, \ldots, \theta_L{}^{\mathrm{T}})^{\mathrm{T}}$ is block diagonal, with the $p \times p$ matrices $\Sigma_l = \mathrm{var}(\theta_l)$ constituting the diagonal blocks. Figure 1 visualizes this covariance structure in comparison with that of Qiao et al. (2019), along with comparisons of the inverse covariance structure. The latter comparison is the more striking and relevant one, since the model of Qiao et al. (2019) possesses a potentially full inverse covariance structure, whereas that under partial separability remains block diagonal. As a consequence, the model of Qiao et al. (2019) has $\mathcal{O}(L^2 p^2)$ free parameters, while the corresponding model under partial separability has only $\mathcal{O}(L p^2)$ free parameters.

Lastly, we establish optimality and uniqueness properties for the basis $\{\varphi_l\}_{l=1}^{\infty}$ of a partially

separable $\mathcal{G}$. A key object is the trace class covariance operator

$$\mathcal{H} = \frac{1}{p}\sum_{j=1}^{p}\mathcal{G}_{jj}. \tag{4}$$

Let $\lambda_l = \lambda_l^{\mathcal{H}}$ ($l \in \mathbb{N}$) denote the eigenvalues of $\mathcal{H}$, in nonincreasing order.

Theorem 2. *Suppose the eigenvalues of $\mathcal{H}$ in (4) have multiplicity one.*

1. *For any $L \in \mathbb{N}$, and for any orthonormal set $\{\tilde{\varphi}_l\}_{l=1}^{L}$ in $L^2[0,1]$, $\sum_{l=1}^{L}\sum_{j=1}^{p}\mathrm{var}(\langle X_j, \tilde{\varphi}_l\rangle) \le \sum_{l=1}^{L}\lambda_l$, with equality if and only if $\{\tilde{\varphi}\}_{l=1}^{L}$ span the first L eigenspaces of $\mathcal{H}$.*

2. *If $\mathcal{G}$ is partially separable with $L^2[0,1]$ basis $\{\varphi_l\}_{l=1}^{\infty}$, then*

$$\mathcal{H} = \sum_{l=1}^{\infty}\lambda_l\varphi_l \otimes \varphi_l, \quad \lambda_l = \frac{1}{p}\mathrm{tr}(\Sigma_l). \tag{5}$$

Part 1 states that, independent of partial separability, the eigenbasis of $\mathcal{H}$ is optimal in terms of retaining the greatest amount of total variability in vectors of the form $(\langle X_1, \tilde{\varphi}_l\rangle, \ldots, \langle X_p, \tilde{\varphi}_l\rangle)^{\mathrm{T}}$, subject to orthogonality constraints. Part 2 indicates that, if $\mathcal{G}$ is partially separable, the unique basis of $L^2[0,1]$ that makes Definition 1 hold corresponds to this optimal basis.

3.2 Consequences for Functional Gaussian Graphical Models

Assume that $\mathcal{G}$ is partially separable according to Definition 1, so that the partially separable Karhunen-Loève expansion in (3) holds. If we further assume that X is Gaussian, then $\theta_l \sim \mathcal{N}(0, \Sigma_l)$, $l \in \mathbb{N}$, are independent, where Σ_l is positive definite for each l. These facts follow from Theorem 1. Recall that, in order to define a coherent functional Gaussian graphical model, one needs that the conditional covariance functions C_{jk} in (1) between component functions X_j and X_k be well-defined. The expansion in (3) facilitates a simple connection between the C_{jk} and the inverse covariance matrices $\Omega_l = \Sigma_l^{-1}$, as follows. Let $\Sigma_l = (\sigma_{ljk})_{j,k=1}^{p}$. For any fixed $j, k \in V$, define the partial covariance between θ_{lj} and θ_{lk} as

$$\tilde{\sigma}_{ljk} = \sigma_{ljk} - \mathrm{cov}\{\theta_{lj}, \theta_{l,-(j,k)}\}\mathrm{var}\{\theta_{l,-(j,k)}\}^{-1}\mathrm{cov}\{\theta_{l,-(j,k)}, \theta_{lk}\}. \tag{6}$$

It is well-known that these partial covariances are directly related to the precision matrix $\Omega_l = (\omega_{ljk})_{j,k=1}^{p}$ by $\tilde{\sigma}_{ljk} = -\omega_{ljk}/(\omega_{ljj}\omega_{lkk} - \omega_{ljk}^2)$, so that $\tilde{\sigma}_{ljk} = 0$ if and only if $\omega_{ljk} = 0$. The next result establishes that the conditional covariance functions C_{jk} can be expanded in the partial separability basis $\{\varphi_l\}_{l=1}^{\infty}$ with coefficients $\tilde{\sigma}_{ljk}$.

Theorem 3. *If $\mathcal{G}$ is partially separable, then the cross-covariance kernel between X_j and X_k, conditional on the multivariate subprocess $\{X_{-(j,k)}(u) : u \in [0,1]\}$, is*

$$C_{jk}(s,t) = \sum_{l=1}^{\infty} \tilde{\sigma}_{ljk} \varphi_l(s) \varphi_l(t) \quad (j,k \in V, \ j \neq k, \ s,t \in [0,1]). \tag{7}$$

Now, the conditional independence graph for the multivariate Gaussian process can be defined by $(j,k) \notin E$ if and only if $C_{jk}(s,t) \equiv 0$. Due to the above result, the edge set E is connected to the sequence of edge sets $\{E_l\}_{l=1}^{\infty}$, for which $(j,k) \notin E_l$ if and only if $\tilde{\sigma}_{ljk} = \omega_{ljk} = 0$, corresponding to the sequence of Gaussian graphical models (V, E_l) for each θ_l.

Corollary 1. *Under the setting of Theorem 3, the functional graph edge set E is related to the sequence of edge sets E_l by $E = \bigcup_{l=1}^{\infty} E_l$.*

We have thus established that, under partial separability, the problem of functional graphical model estimation can be simplified to estimation of a sequence of decoupled graphical models. When partial separability fails, the edge sets remain meaningful. Recall from Theorem 2 that the eigenbasis of $\mathcal{H}$ is optimal in a sense independent of partial separability, so that the vectors $\theta_l = (\langle X_1, \varphi_l \rangle, \dots, \langle X_p, \varphi_l \rangle)^{\mathrm{T}}$ are still the coefficients of X in an optimal expansion. Although one loses a direct connection between the E_l and the edge set of the functional graph, each E_l remains the edge set of the Gaussian graphical model for the coefficient vector θ_l in this optimal expansion. Moreover, the equivalence $E = \bigcup_{l=1}^{\infty} E_l$ may hold independent of partial separability. For instance, Proposition 2 in Section A1 of the Appendix gives sufficient conditions, based on a Markov-type property and a edge coherence assumption, under which the equivalence holds.

4 Graph Estimation and Theory

4.1 Joint Graphical Lasso Estimator

Consider a p-variate process X, with means $\mu_j(t) = E\{X_j(t)\}$ and covariance operator $\mathcal{G}$. Let $\{\varphi_l\}_{l=1}^{\infty}$ be an orthonormal eigenbasis of $\mathcal{H}$ in (4), and set $\theta_{lj} = \langle X_j, \varphi_l \rangle$, $\Sigma_l = \mathrm{var}(\theta_l)$. The targets are the edge sets E_l, where $(j,k) \in E_l$ if and only if $(\Sigma_l^{-1})_{jk} \neq 0$, as motivated by the developments of Section 3.2. Specifically, when X is Gaussian and $\mathcal{G}$ is partially separable, the conditional independence graph of X has edge set $E = \bigcup_{l=1}^{\infty} E_l$. When partial separability fails, these targets still provide useful information about the conditional independencies of X when projected onto the eigenbasis of $\mathcal{H}$, which is optimal in the sense of Theorem 2. Furthermore, when X is not Gaussian, rather than representing conditional independence, the E_l represent the sparsity structure of the partial correlations of θ_l, which may still be of interest. By Theorem 1, $\mathrm{tr}(\Sigma_l) = \lambda_l \downarrow 0$ as $l \to \infty$. As a practical consideration, this makes estimators of Σ_l progressively more unstable to work with as l increases. To avoid this, we work with $\Xi_l = R_l^{-1}$, where R_l is the correlation matrix corresponding

to Σ_l. Ξ_l and Ω_l share the same edge information as entries in these two matrices are either zero or nonzero simultaneously.

We first define the estimation procedure with targets Ξ_l, from a random sample $X_1, \ldots, X_n$, each distributed as X. X is not required to be Gaussian, nor $\mathcal{G}$ to be partially separable, in developing the theoretical properties of the estimators, which also allow the dimension p to diverge with n. In order to make these methods applicable to any functional data set, it is assumed that preliminary mean and covariance estimates $\hat{\mu}_j$ and $\hat{\mathcal{G}}_{jk}$, $j, k = 1, \ldots, p$, have been computed for each component. As an example, if the X_i are fully observed, cross-sectional estimates

$$\hat{\mu}_j = \frac{1}{n} \sum_{i=1}^{n} X_{ij}, \quad \hat{\mathcal{G}}_{jk} = \frac{1}{n} \sum_{i=1}^{n} (X_{ij} - \hat{\mu}_j) \otimes (X_{ik} - \hat{\mu}_k), \tag{8}$$

can be used. For practical observational designs, smoothing can be applied to the pooled data to estimate these quantities (Yao et al. 2005, Yang et al. 2011). Given such preliminary estimates, the estimate of $\mathcal{H}$ is $\hat{\mathcal{H}} = p^{-1} \sum_{j=1}^{p} \hat{\mathcal{G}}_{jj}$, leading to empirical eigenfunctions $\hat{\varphi}_l$, which are uniquely defined only up to a sign and for $1 \le l \le np = \mathrm{rank}(\mathcal{H})$. These quantities produce estimates of $\sigma_{ljk} = \langle \mathcal{G}_{jk}(\varphi_l), \varphi_l \rangle$ by plugin as

$$s_{ljk} = (S_l)_{jk} = \langle \hat{\mathcal{G}}_{jk}(\hat{\varphi}_l), \hat{\varphi}_l \rangle. \tag{9}$$

A group graphical lasso approach (Danaher et al. 2014) will be used to estimate the Ξ_l. Let $(\hat{R}_l)_{jk} = \hat{r}_{ljk} = s_{ljk}/[s_{ljj}s_{lkk}]^{1/2}$ be the estimated correlations. The estimates target the first $L \le np$ inverse correlation matrices Ξ_l by

$$(\hat{\Xi}_1, \ldots, \hat{\Xi}_L) = \arg \min_{\Upsilon_l \succ 0, \Upsilon_l = \Upsilon_l^{\mathrm{T}}} \sum_{l=1}^{L} \left\{ \mathrm{tr}(\hat{R}_l \Upsilon_l) - \log(|\Upsilon_l|) \right\} + P(\Upsilon_1, \ldots, \Upsilon_L), \tag{10}$$

In the Gaussian case, these are penalized likelihood estimators with penalty

$$P(\Upsilon_1, \ldots, \Upsilon_L) = \gamma \left\{ \alpha \sum_{l=1}^{L} \sum_{j \ne k} |v_{ljk}| + (1 - \alpha) \sum_{j \ne k} \left(\sum_{l=1}^{L} v_{ljk}^2 \right)^{1/2} \right\}, \quad (\Upsilon_l)_{jk} = v_{ljk}. \tag{11}$$

The parameter $\gamma > 0$ controls the overall penalty level, while $\alpha \in [0, 1]$ distributes the penalty between the two penalty terms. Then the estimated edge set is $(j, k) \in \hat{E}_l$ if and only if $\hat{\Xi}_{ljk} \ne 0$. The joint graphical lasso was chosen to borrow structural information across multiple bases instead of multiple classes as was done in Danaher et al. (2014). If $\alpha = 1$, the first penalty will encourage sparsity in each $\hat{\Xi}_l$ and the corresponding edge set $\hat{E}_l$, but the overall estimate $\hat{E} = \bigcup_{l=1}^{L} \hat{E}_l$ may not be sparse. While consistent graph recovery is still possible with $\alpha = 1$ as demonstrated below

in Theorem 5, the influence of the second penalty term when $\alpha < 1$ ensures that the overall graph estimate is sparse, enhancing interpretation.

In practice, tuning parameters γ and α can be chosen with cross-validation to minimize (10) for out-of-sample data. Specifically, the procedure would select γ and α that minimize the average of (10) evaluated over each fold, where $\Upsilon_1, \ldots, \Upsilon_L$ are computed with the training set and $\hat{R}_l$ are from the validation set. Another practically useful, and less computationally intensive, approach is to choose these parameters to yield a desired sparsity level of the estimated graph (Qiao et al. 2019). This latter approach is implemented in the data example of Section 6.

4.2 Asymptotic Properties

The goal of the current section is to provide lower bounds on the sample size n so that, with high probability, $\hat{E}_l = E_l$ ($l = 1, \ldots, L$). The approach follows that of Ravikumar et al. (2011), adapting the results to the case of functional graphical model estimation in which multiple graphs are estimated simultaneously. For simplicity, and to facilitate comparisons with the asymptotic properties of Qiao et al. (2019), the results are derived under the setting of fully observed functional data, so that $\hat{\mu}$ and $\hat{\mathcal{G}}_{jk}$ are as in (8). As a preliminary result, we first derive a concentration inequality for the estimated covariances s_{ljk} in (9), requiring the following mild assumptions.

Assumption 1. The eigenvalues λ_l of $\mathcal{H}$ have multiplicity one, and are thus strictly decreasing.

Assumption 2. There exists $\varsigma^2 > 0$ such that $E\left(e^{s\theta_{lj}}\right) \leq e^{s^2\varsigma^2\sigma_{ljj}/2}$ for all $l \in \mathbb{N}$, $j \in V$, and $s \in \mathcal{R}$; that is, the standardized scores $\theta_{lj}/\sigma_{ljj}^{1/2}$ are sub-Gaussian random variables with parameter ς^2. Furthermore, there is M independent of p such that $\sup_{j \in V} \sum_{l=1}^{\infty} \sigma_{ljj} < M < \infty$.

Assumption 2 can be relaxed to accommodate eigenvalues with multiplicity greater than 1, at the cost of an increased notational burden. The eigenvalue spacings play a key role through the quantities $\tau_1 = 2\sqrt{2}(\lambda_1 - \lambda_2)^{-1}$ and $\tau_l = 2\sqrt{2}\max\left\{(\lambda_{l-1} - \lambda_l)^{-1}, (\lambda_l - \lambda_{l+1})^{-1}\right\}$, for $l \geq 2$. Assumption 2 clearly holds in the Gaussian case, and can be relaxed to accommodate different parameters ς_l^2 for each l, though for simplicity these are assumed uniform.

Theorem 4. *Suppose that Assumptions 1 and 2 hold. Then there exist constants $C_1, C_2, C_3 > 0$ such that, for any $0 < \delta \leq C_3$ and for all $l \leq np$ and $j, k \in V$,*

$$\mathrm{pr}\left(|s_{ljk} - \sigma_{ljk}| \geq \delta\right) \leq C_2 \exp\left(-C_1\tau_l^{-2}n\delta^2\right). \tag{12}$$

Concentrations inequalities such as (12) are generally required in penalized estimation problems where the dimension diverges to infinity. For the current problem, even if the dimension p of the process remains fixed, the dimension still diverges since one requires L to diverge with n. Furthermore, in contrast to standard multivariate scenarios, the bound in Theorem 4 contains the additional factor τ_l^{-2}. Since $\lambda_l \downarrow 0$, τ_l diverges to infinity with l, so that (12) reflects the increased difficulty of estimating covariances corresponding to eigenfunctions with smaller eigenvalue gaps.

11

Remark 1. A similar result to Theorem 4 was obtained by Qiao et al. (2019) under a specific eigenvalue decay rate and truncation parameter scheme. Imposing similar assumptions on the eigenvalues of $\mathcal{H}$, we have $\tau_l = O(l^{1+\beta_1})$ for some $\beta_1 > 1$, so that for any $0 < \beta_2 < 1/(4\beta_1)$ and $L = n^{\beta_2}$, (12) implies

$$\max_{l=1,\ldots,L} \max_{j,k \in V} \mathrm{pr}\left(|s_{ljk} - \sigma_{ljk}| \geq \delta\right) \leq C_2 \exp\{-C_1 n^{1-2\beta_2(1+\beta_1)}\delta^2\},$$

matching the rate of Qiao et al. (2019). In addition to establishing the concentration inequality for a general eigenvalue decay rate, our proof is greatly simplified by using the inequality

$$|s_{ljk} - \sigma_{ljk}| \leq 2\tau_l \|\mathcal{G}_{jk}\|_{\mathrm{HS}} \|\hat{\mathcal{H}} - \mathcal{H}\|_{\mathrm{HS}} + \|\hat{\mathcal{G}}_{jk} - \mathcal{G}_{jk}\|_{\mathrm{HS}}, \tag{13}$$

where $\|\cdot\|_{\mathrm{HS}}$ is the Hilbert-Schmidt operator norm.

Remark 2. The bound in (13) utilizes a basic eigenfunction inequality found, for example, in Lemma 4.3 of Bosq (2000); see also Bhatia et al. (1983). However, using expansions instead of geometric inequalities, Jirak (2016) and other authors cited therein established stonger results for differences between true and estimated eigenfunctions in the form of limiting distributions and moment bounds. Thus, it is likely that the bound in (12) is suboptimal, although improvements along the lines of Jirak (2016) would require further challenging work in order to establish the required exponential tail bounds.

As the objective (10) utilizes the correlations $\hat{r}_{ljk}$, the following corollary is needed.

Corollary 2. *Under the assumptions of Theorem 4, there exists constants $D_1, D_2, D_3 > 0$ such that, for any $0 < \delta \leq D_3$ and for all $l \leq np$ and $j, k \in V$,*

$$\mathrm{pr}\left(|\hat{r}_{ljk} - r_{ljk}| \geq \delta\right) \leq D_2 \exp\left(-D_1 n m_l^2 \delta^2\right), \quad m_l = \tau_l^{-1}\pi_l, \ \pi_l = \min_{j \in V} \sigma_{ljj}. \tag{14}$$

To establish consistency of $\hat{E}_l$, some additional notation will be introduced. Let $\Psi_l = R_l \tilde{\otimes} R_l$, where $\tilde{\otimes}$ is the Kronecker product, and $\overline{E}_l = E_l \cup \{(1,1),\ldots,(p,p)\}$. For $B \subset V \times V$, let $\Psi_{l,BB}$ denote the submatrix of Ψ_l indexed by sets of pairs $(j,k) \in B$, where $\Psi_{l,(j,k),(j',k')} = R_{ljj'}R_{lkk'}$. For a $p \times p$ matrix Δ, let $\|\|\Delta\|\|_\infty = \max_{j=1,\ldots,p} \sum_{k=1}^p |\Delta_{jk}|$. The following assumption corresponds to the irrepresentability or neighborhood stability condition often seen in sparse matrix and regression estimation (Ravikumar et al. 2011, Meinshausen et al. 2006).

Assumption 3. For $l = 1,\ldots,L$, there exists $\eta_l \in (0,1]$ such that

$$\left\|\left\|\Psi_{l,\overline{E}_l^c\overline{E}_l}\left(\Psi_{l,\overline{E}_l\overline{E}_l}\right)^{-1}\right\|\right\|_\infty \leq 1 - \eta_l.$$

For fixed l, Assumption 3 was employed by Ravikumar et al. (2011) as sufficient for model selection consistency. As Theorem 5 below implies simultaneous consistency of the first L edge sets, we require the assumption for each l.

Set $\kappa_{R_l} = \|\|R_l\|\|_\infty$, $\kappa_{\Psi_l} = \|\|(\Psi_{l,\overline{E}_l\overline{E}_l})^{-1}\|\|_\infty$, let $y_l = \max_{j\in V}|\{k \in V : \Xi_{ljk} \neq 0\}|$ be the maximum degree of the graph (V, E_l), and $\xi_{\min,l} = \min\{|\Xi_{ljk}| : \Xi_{ljk} \neq 0\}$. Finally, when Assumptions 1–3 hold, for any $\alpha > 1 - \min_{l=1,\ldots,L}\eta_l$, define $\eta'_l = \alpha + \eta_l - 1 > 0$ and $\epsilon_L = \min_{1\leq l\leq L}\eta'_l m_l$. Then, with D_3 as in Corollary 2, set

$$\mathfrak{a}_L = D_3 \min_{l=1,\ldots,L} m_l,$$

$$\mathfrak{b}_L = \min_{l=1,\ldots,L}\left\{6y_l m_l \max(\kappa_{\Psi_l}^2\kappa_{R_l}^3, \kappa_{\Psi_l}\kappa_{R_l})(m_l^{-1} + 8\epsilon_L^{-1})^2\right\}^{-1},$$

$$\mathfrak{c}_L = \min_{l=1,\ldots,L} \xi_{\min,l}\left\{4\kappa_{\Psi_l}(m_l^{-1} + 8\epsilon_L^{-1})\right\}^{-1}.$$

Tracking $\mathfrak{a}_L$ and $\mathfrak{b}_L$, including the maximal degrees y_l, allows one to obtain uniform consistency of the estimates $\hat{\Xi}_l$ in (10), and to conclude that $\hat{E}_l \subset E_l$ with high probability; see Lemma 2 in Section A2.1 of the Appendix. The quantity $\mathfrak{c}_L$ involves the weakest nonzero signal $\xi_{\min,l}$ of each graph, with weaker signals requiring larger n for recovery. The quantities $\mathfrak{b}_L$ and $\mathfrak{c}_L$ decrease with α, so that smaller values of α also require larger n.

Theorem 5. *Suppose Assumptions 1–3 hold, $(1 - \min_{l=1,\ldots,L}\eta_l) < \alpha \leq 1$, $L \leq np$, and that, for some $\varrho > 2$, $\gamma = 8\epsilon_L^{-1}\left\{(D_1n)^{-1}\log\left(D_2L^{\varrho-1}p^\varrho\right)\right\}^{1/2}$, where D_1 and D_2 are constants from Corollary 2. If the sample size n satisfies*

$$n\min(\mathfrak{a}_L, \mathfrak{b}_L, \mathfrak{c}_L)^2 > D_1^{-1}\left\{\log(D_2) + (\varrho - 1)\log(n) + (2\varrho - 1)\log(p)\right\}, \qquad (15)$$

then, with probability at least $1 - (Lp)^{2-\varrho}$, $\hat{E}_l = E_l$ for all $l = 1, \ldots, L$.

Remark 3. If L grows sufficiently slowly with respect to n, (15) becomes $n\min(\mathfrak{a}_L, \mathfrak{b}_L, \mathfrak{c}_L)^2 \gtrsim \varrho\log(p)$ for large n, $\gtrsim$ denoting inequality up to a multiplicative constant. Hence, the conclusion of Theorem 5 holds for large n when $\log(p) = o(n)$ by suitably limiting the growth of L.

Remark 4. To understand how the graph properties affect the lower bound, assume κ_{Ψ_l}, κ_{R_l}, and η_l do not depend on l, n, or p, and $\min_{1\leq l\leq n} m_l \gtrsim n^{-d}$, $0 < d < 1/4$. Then (15) becomes

$$n \gtrsim \left[\left\{\left(\max_{1\leq l\leq L}\xi_{\min,l}^{-2}\right) + \left(\max_{1\leq l\leq L}y_l^2\right)\right\}\varrho\log(p)\right]^{1-4d}$$

asymptotically. In particular, if L remains bounded so that $d = 0$, the above bound is consistent with that of Ravikumar et al. (2011), where the maxima over l reflect the need to satisfy the bound for the edge set E_l that is most difficult to estimate.

If the conditional independence graph of X is $E = \bigcup_{l=1}^{\infty} E_l$, as is the case when X is Gaussian and $\mathcal{G}$ partially separable, Theorem 5 can lead to edge selection consistency in the functional graphical model. For a given n, there exists a finite L_p^* such that $E = \bigcup_{l=1}^{L_p^*} E_l$, where L_p^* can only diverge with n if p does so. The following corollary is then immediate.

Corollary 3. *Under the assumptions of Theorem 5, suppose* $\log(p) = o(n)$, $L \to \infty$ *such that* $\min(\mathfrak{a}_L, \mathfrak{b}_L, \mathfrak{c}_L)[n/\log\{\max(n,p)\}]^{1/2} \to \infty$ *and, for large n, $L \geq L_p^*$. Then, for large n, $E = \bigcup_{l=1}^{L} \hat{E}_l$ with probability at least $1 - (Lp)^{2-\varrho}$.*

5 Numerical Experiments

5.1 Simulation Settings

The simulations in this section compare the proposed method for partially separable functional Gaussian graphical models, with that of Qiao et al. (2019). Throughout this section we denote these methods as psFGGM and FGGM, respectively. Throughout this section we denote these methods as psFGGM and FGGM, respectively. Other potentially competing non-functional based approaches are not included since they are clearly outperformed by the latter (Qiao et al. 2019). An initial conditional independence graph $G = (V, E)$ is generated from a power law distribution with parameter $\pi = \mathrm{pr}\{(j, k) \in E\}$. Then, for a fixed M, a sequence of edge sets $E_1, \ldots, E_M$ is generated so that $E = \bigcup_{l=1}^{M} E_l$. A set of common edges to all edge sets is computed for a given proportion of common edges $\tau \in [0, 1]$. Next, $p \times p$ precision matrices Ω_l are generated for each E_l based on the algorithm of Peng et al. (2009). A fully detailed description of this step is included in the Section A4 of the Appendix.

Random vectors $\theta_i \in \mathbb{R}^{Mp}$ are then generated from a mean zero multivariate normal distribution with covariance matrix Σ, yielding discrete and noisy functional data

$$Y_{ijk} = X_{ij}(t_k) + \varepsilon_{ijk}, \quad \varepsilon_{ijk} \sim N(0, \sigma_\varepsilon^2) \quad (i = 1, \ldots, n;\ j = 1, \ldots, p;\ k = 1, \ldots, M).$$

Here, $\sigma_\epsilon^2 = 0.05 \sum_{l=1}^{M} \mathrm{tr}(\Sigma_l)/p$ and $X_{ij}(t_k) = \sum_{l=1}^{M} \theta_{ilj}\varphi_l(t_k)$ according to the partially separable Karhunen-Loève expansion in (3). Fourier basis functions $\varphi_1, \ldots, \varphi_M$ evaluated on an equally spaced time grid of $t_1, \ldots, t_T$, with $t_1 = 0$ and $t_T = 1$, were used to generate the data. In all settings, 100 simulations were conducted. To resemble real data example from Section 6 below, we set $T = 30$, $M = 20$ and $\pi = 5\%$ for a sparse graph.

We consider two models for Σ, corresponding to partially separable and non-partially separable X, respectively. In the first, the covariance Σ_{ps} is formed as a block diagonal matrix with $p \times p$ diagonal blocks $\Sigma_l = a_l \Omega_l^{-1}$. The decaying factors $a_l = 3l^{-1.8}$ guarantee that $\mathrm{tr}(\Sigma_l)$ decreases monotonically in l. In the second, Σ_{ps} is modified to violate partial separability. Specifically, a block-banded precision matrix Ω is computed with $p \times p$ blocks $\Omega_{l,l} = \Omega_l$ and $\Omega_{l+1,l} = \Omega_{l,l+1} = 0.5(\Omega_l^* + \Omega_{l+1}^*)$ with $\Omega_l^* = \Omega_l - \mathrm{diag}(\Omega_l)$. Then, the non-partially separable covariance is computed as $\Sigma_{\mathrm{non\text{-}ps}} = \mathrm{diag}(\Sigma_{\mathrm{ps}})^{1/2}\Omega^{-1}\mathrm{diag}(\Sigma_{\mathrm{ps}})^{1/2}$.

14

5.2 Comparison of Results

Comparisons between the proposed method and that of Qiao et al. (2019), implemented using code provided by the authors, are presented here. Additional comparisons obtained by thresholding correlations are provided in Section A4.4 of the Appendix. Although this alternative method does not estimate a sparse inverse covariance structure, its graph recovery is competitive with that of the proposed method in some settings. As performance metrics, the true and false positive rates of correctly identifying edges in graph G are computed over a range of γ values and a coarse grid of five evenly spaced points $\alpha \in [0, 1]$. The value of α maximizing the area under the receiver operating characteristic curve is considered for the comparison. In all cases, we set $\pi = 0.05$ and $\tau = 0$. The two methods are compared using L principal components explaining at least 90% of the variance. For all simulations and both methods, this threshold results in the choice of $L = 5$ or $L = 6$ components. For higher variance explained thresholds, however, we see a sharp contrast. While the proposed method consistently converges to a solution, that of Qiao et al. (2019) does not, due to increasing numerical instability. The reason for the instability is the need to estimate $L = 5$ or 6 times more parameters compared to the proposed method. The proposed method can thus accommodate larger L, thereby incorporating more information from the data. In the figures and tables, additional results are available for the proposed method when L is increased to explain at least 95% of the variance.

Figure 2a shows average true/false positive rate curves for the high-dimensional case $n = p/2$. The smoothed curves are computed using the `supsmu` R package that implements SuperSmoother (Friedman 1984), a variable bandwidth smoother that uses cross-validation to find the best bandwidth. Table 1 shows the mean and standard deviation of area under the curve estimates for various settings. When partial separability holds, $\Sigma = \Sigma_{\text{ps}}$, the proposed method exhibits uniformly higher true positive rates across the full range of false positive rates. Even when partial separability is violated, $\Sigma = \Sigma_{\text{non-ps}}$, the two methods perform comparably. More importantly, and in all cases, the proposed method is able to leverage 95% level of variance explained, owing to the numerical stability mentioned above. Figure 2b and Table 1 summarize results for the large sample case $n = 1.5p$ with similar conclusions. Comparisons under additional simulation settings can be found in Section A4 of the Appendix.

6   Application to Functional Brain Connectivity

In this section, the proposed method is used to reconstruct the brain connectivity structure using functional magnetic resonance imaging (fMRI) data from the Human Connectome Project. We analyze the ICA-FIX preprocessed data variant that controls for spatial distortions and alignments across both subjects and modalities (Glasser et al. 2013). In particular, we use the 1200 Subjects 3T MR imaging data available at `https://db.humanconnectome.org` that consists of fMRI scans of individuals performing basic body movements. During each scan, a three-second visual cue
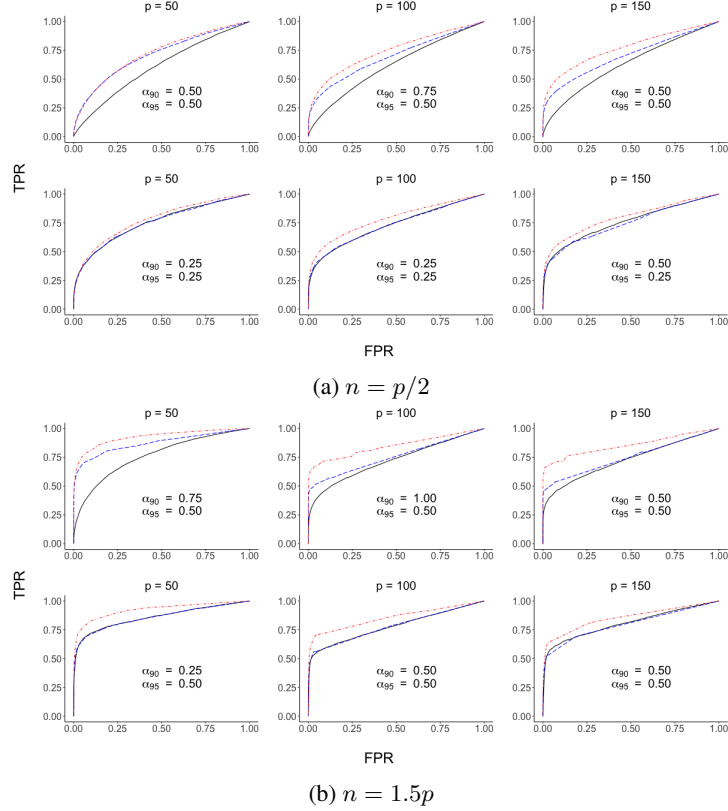
Figure 2: Mean receiver operating characteristic curves for the proposed method (psFGGM) and that of Qiao et al. (2019) (FGGM). In subfigures (a) and (b), Σ_{ps} (top) and $\Sigma_{\text{non-ps}}$ (bottom) were used for $p = 50, 100, 150$. Curves are coded as psFGGM (- - - -) and FGGM (——) at 90% of variance and psFGGM (— · —) at 95% of variance explained. For psFGGM, the values of α used to compute the curve values are printed in each panel.

signals the subject to move a specific body part, which is then recorded for 12 seconds at a temporal resolution of 0.72 seconds. For this work, we considered only the data from left- and right-hand finger movements.

The left- and right-hand tasks data for $n = 1054$ subjects with complete meta-data were preprocessed by averaging the blood oxygen level dependent signals over $p = 360$ regions of interest (ROIs) (Glasser et al. 2016). After removing cool down and ramp up observations, $T = 16$ time points of pure movement tasks remained. Motivated by part 3 of Theorem 1, diagnostics were performed to assess the plausibility of the partial separability assumption, with no indications to the contrary; see Section A5 of the Appendix for more details. Penalty parameters $\gamma = 0.91$ and $\alpha = 0.95$ were used to estimate very sparse graphs in both tasks.

16

Table 1: Mean area under the curve (and standard error) values for Figures $2a$ and $2b$

n			$\Sigma = \Sigma_{\text{ps}}$			$\Sigma = \Sigma_{\text{non-ps}}$		
			$p = 50$	$p = 100$	$p = 150$	$p = 50$	$p = 100$	$p = 150$
$p/2$	AUC	FGGM$_{90\%}$	0.60(0.03)	0.62(0.02)	0.63(0.01)	0.75(0.03)	0.72(0.02)	0.75(0.02)
		psFGGM$_{90\%}$	0.71(0.04)	0.69(0.02)	0.70(0.01)	0.75(0.03)	0.73(0.02)	0.74(0.03)
		psFGGM$_{95\%}$	0.72(0.04)	0.74(0.02)	0.77(0.02)	0.77(0.03)	0.78(0.02)	0.79(0.02)
	AUC15†	FGGM$_{90\%}$	0.15(0.04)	0.18(0.02)	0.20(0.01)	0.39(0.04)	0.40(0.02)	0.45(0.03)
		psFGGM$_{90\%}$	0.30(0.05)	0.35(0.02)	0.37(0.02)	0.39(0.04)	0.42(0.03)	0.44(0.04)
		psFGGM$_{95\%}$	0.29(0.05)	0.40(0.03)	0.46(0.03)	0.41(0.05)	0.48(0.03)	0.51(0.03)
$1.5p$	AUC	FGGM$_{90\%}$	0.76(0.02)	0.72(0.02)	0.73(0.01)	0.86(0.02)	0.78(0.02)	0.80(0.03)
		psFGGM$_{90\%}$	0.87(0.03)	0.75(0.02)	0.75(0.01)	0.85(0.02)	0.78(0.02)	0.79(0.03)
		psFGGM$_{95\%}$	0.92(0.02)	0.84(0.02)	0.85(0.02)	0.92(0.03)	0.85(0.02)	0.85(0.02)
	AUC15†	FGGM$_{90\%}$	0.37(0.04)	0.41(0.02)	0.44(0.02)	0.66(0.03)	0.55(0.03)	0.57(0.04)
		psFGGM$_{90\%}$	0.69(0.04)	0.52(0.02)	0.52(0.02)	0.65(0.04)	0.56(0.04)	0.55(0.05)
		psFGGM$_{95\%}$	0.75(0.04)	0.68(0.03)	0.69(0.03)	0.76(0.06)	0.68(0.03)	0.64(0.03)

†AUC15 is AUC computed for FPR in the interval $[0, 0.15]$, normalized to have maximum area 1.

(a) Left-hand task (b) Right-hand task

(c) Activated ROIs unique to left-hand task (d) Activated ROIs unique to right-hand task

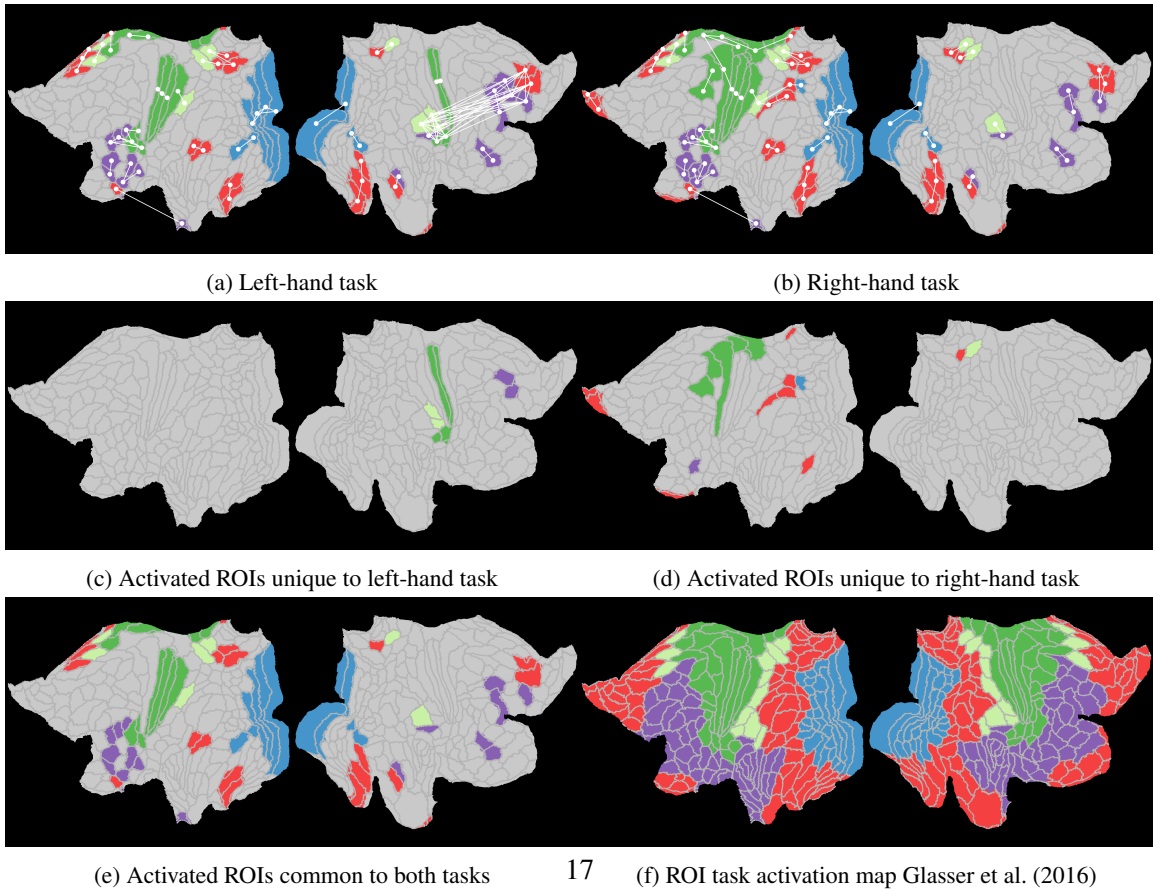(e) Activated ROIs common to both tasks (f) ROI task activation map Glasser et al. (2016)

Figure 3: psFGGM estimated functionally connected cortical ROIs for the left- and right-hand motor tasks. Each sub-figure shows a flat brain map of the left and right hemispheres (in that order). ROIs having a positive degree of connectivity in each estimated graph are colored based on their functionality Glasser et al. (2016): visual (**blue**), motor (**green**), mixed motor (**light green**), mixed other (**red**) and other (**purple**).

17

Figure 3 shows comparison of activation patterns from left and right-hand task datasets. Figures $3a$ and $3b$ show the recovered ROI graph on a flat brain map, and only those ROIs with positive degree of connectivity are colored. Figures $3c$ and $3d$ show connected ROIs that are unique to each task, whereas Figure $3e$ show only those that are common to both tasks. In this map, one can see that almost all of the visual cortex ROIs in the occipital lobe are shared by both maps. This is expected as both tasks require individuals to watch visual cues. Furthermore, the primary sensory cortex, corresponding to touch and motor sensory inputs, and intraparietal sulcus, corresponding to perceptual motor coordination, are activated during both left and right-hand tasks. On the other hand, the main difference between these motor tasks lies at the motor cortex near the central sulcus. In Figure $3c$ and $3d$ the functional maps for the left- and right-hand tasks present particular motor-related cortical areas in the right and left hemisphere, respectively. These results are in line with the motor task activation maps obtained by Barch et al. (2013).

7 Discussion

Partial separability for multivariate functional data is a novel structural assumption with further potential applications beyond graphical models. For example, it is well-known that the functional linear model is simplified by univariate functional principal component analysis, which parses out the problem into a sequence of simple linear regressions; see Wang et al. (2016) and references therein. The partially separable Karhunen-Loève expansion in (3) demonstrates a similar potential, namely to break down a problem involving multivariate functional data into a sequence of standard multivariate problems. This potential was shown in this paper by decomposing a functional graphical model into a sequence of standard multivariate graphical models.

Motivated by the brain connectivity example, we have presented partial separability for multivariate processes with components X_j defined on the same domain. However, this restriction is not necessary in order to define partial separability. If X_j are elements of $L^2(\mathcal{T}_j)$, $j = 1, \ldots, p$, a more general definition of partial separability would be the existence of orthonormal bases $\{\varphi_{jl}\}_{l=1}^{\infty}$ of $L^2(\mathcal{T}_j)$, $j = 1, \ldots, p$, such that the vectors $\theta_l = (\theta_{l1}, \ldots, \theta_{lp})^{\mathrm{T}}$, where $\theta_{lj} = \int_{\mathcal{T}_j} X_j(t)\varphi_{jl}(t)\mathrm{d}t$ are mutually uncorrelated across l. Such a generalization is highly desirable, as many multivariate functional data sets consist of functions on different domains. In fact, the above notion is even applicable when the domains $\mathcal{T}_j$ are of different dimension (Happ & Greven 2018) or even a complex manifold, such as the surface of the brain.

The proposed method for functional graphical model estimation is equally applicable to dense or sparse functional data, observed with or without noise. However, rates of convergence will inevitably suffer as observations become more sparse or are contaminated with higher levels of noise. The results in Theorem 4 of this paper or Theorem 1 of Qiao et al. (2019) have been derived under the setting of fully observed functional data, so that future work will include similar derivations under more general observation schemes.

Acknowledgement

Appendix

Sections A1 and A2 of the Appendix contain proofs of theoretical results from the main paper, while Section A3 contains an alternative edge selection consistency result. Section A4 contains details for the algorithm used in Section 5 and includes additional figures for very sparse graphs. Finally, Section A5 assesses the asssumption of partial separability for the neuroimaging data example.

A1   Proofs for Section 3

Proof of Theorem 1.

(1 $\Leftrightarrow$ 2) Suppose 1 holds and let $\lambda_{jl}^{\mathcal{G}} = \langle \mathcal{G}(e_{lj}\varphi_l), e_{lj}\varphi_l \rangle_p$ be the eigenvalues of $\mathcal{G}$, and set $\Sigma_l = \sum_{j=1}^{p} \lambda_{jl}^{\mathcal{G}} e_{lj} e_{lj}^{\mathrm{T}}$. Since $(e_{lj}\varphi_l) \otimes_p (e_{lj}\varphi_l) = (e_{lj}e_{lj}^{\mathrm{T}})\varphi_l \otimes \varphi_l$,

$$\mathcal{G} = \sum_{l=1}^{\infty} \sum_{j=1}^{p} \lambda_{jl}^{\mathcal{G}} (e_{lj}\varphi_l) \otimes_p (e_{lj}\varphi_l) = \sum_{l=1}^{\infty} \left(\sum_{j=1}^{p} \lambda_{jl}^{\mathcal{G}} e_{lj} e_{lj}^{\mathrm{T}} \right) \varphi_l \otimes \varphi_l = \sum_{l=1}^{\infty} \Sigma_l \varphi_l \otimes \varphi_l,$$

and 2 holds. If 2 holds, let $\{e_{lj}\}_{j=1}^{p}$ be an orthonormal basis for Σ_l. Then

$$\mathcal{G}(e_{lj}\varphi_l) = \sum_{l'=1}^{\infty} \varphi_{l'} \otimes \varphi_{l'} \{\Sigma_{l'}(e_{lj}\varphi_l)\} = \sum_{l'=1}^{\infty} (\Sigma_{l'} e_{lj}) \varphi_{l'} \otimes \varphi_{l'}(\varphi_l)$$

$$= (\Sigma_l e_{lj})\varphi_l = \lambda_j^{\Sigma_l} e_{lj}\varphi_l,$$

so $e_{lj}\varphi_l$ are the eigenfunctons of $\mathcal{G}$.

(2 $\Leftrightarrow$ 3) If 2 holds, set $\sigma_{ljj} = (\Sigma_l)_{jj}$, so the expression for $\mathcal{G}_{jj}$ clearly holds. Next, for $l \neq l'$,

$$\mathrm{cov}\left(\langle X_j, \varphi_l \rangle, \langle X_k, \varphi_{l'} \rangle\right) = \langle \varphi_l, \mathcal{G}_{jk}(\varphi_{l'}) \rangle = \langle \varphi_l, \sigma_{l'jk}\varphi_{l'} \rangle = 0,$$

so 3 holds. If 3 holds, then define $\sigma_{ljk} = \mathrm{cov}(\langle X_j, \varphi_l \rangle, \langle X_k, \varphi_l \rangle)$ for $j \neq k$, and set $(\Sigma_l)_{jk} = \sigma_{ljk}$ $(j, k \in V)$. Then 2 clearly holds.

(1 $\Leftrightarrow$ 4) Suppose 1 holds. By Theorem 7.2.7 of Hsing & Eubank (2015), and since $\sum_{j=1}^{p} e_{lj} e_{lj}^{\mathrm{T}}$ is the identity matrix,

$$X = \sum_{l=1}^{\infty} \sum_{j=1}^{p} \langle X, e_{lj}\varphi_l \rangle_p e_{lj}\varphi_l = \sum_{l=1}^{\infty} \sum_{j=1}^{p} \left(\sum_{k=1}^{p} e_{ljk} \langle X_k, \varphi_l \rangle \right) e_{lj}\varphi_l = \sum_{l=1}^{\infty} \left(\sum_{j=1}^{p} e_{lj} e_{lj}^{\mathrm{T}} \right) \theta_l \varphi_l,$$

so that 4 holds. If 4 holds, let Σ_l be the covariance matrix of θ_l, and $\{e_{lj}\}_{j=1}^{p}$ an orthonormal

20

eigenbasis for Σ_l. Then

$$
\mathcal{G} = \sum_{l=1}^{\infty} \sum_{l'=1}^{\infty} \mathrm{cov}(\theta_l, \theta_{l'}) \varphi_l \otimes \varphi_{l'} = \sum_{l=1}^{\infty} \mathrm{var}(\theta_l) \varphi_l \otimes \varphi_l = \sum_{l=1}^{\infty} \left(\sum_{j=1}^{p} \lambda_j^{\Sigma_l} e_{lj} e_{lj}^{\mathrm{T}} \right) \varphi_l \otimes \varphi_l
$$

$$
= \sum_{l=1}^{\infty} \sum_{j=1}^{p} \lambda_j^{\Sigma_l} (e_{lj} \varphi_l) \otimes_p (e_{lj} \varphi_l),
$$

so that $e_{lj}\varphi_l$ are the eigenfunctions of $\mathcal{G}$.

Proof of Theorem 2.

To prove 1, for any orthonormal basis $\{\tilde{\varphi}_l\}_{l=1}^{\infty}$ of $L^2[0,1]$,

$$
\sum_{l=1}^{L} \sum_{j=1}^{p} \mathrm{var}\left(\langle X_j, \tilde{\varphi}_l \rangle \right) = \sum_{l=1}^{L} \sum_{j=1}^{p} \langle \mathcal{G}_{jj}(\tilde{\varphi}_l), \tilde{\varphi}_l \rangle = p \sum_{l=1}^{L} \langle \mathcal{H}(\tilde{\varphi}_l), \tilde{\varphi}_l \rangle.
$$

Because the eigenvalues of $\mathcal{H}$ have multiplicity one, its eigenspaces are one-dimensional, and equality is obtained if and only if the $\tilde{\varphi}_l$ span the first L eigenspaces of $\mathcal{H}$, as claimed.

For the second claim, using part 2 of Theorem 1, we have $\mathcal{H} = p^{-1} \sum_{l=1}^{\infty} \mathrm{tr}(\Sigma_l) \varphi_l \otimes \varphi_l$, and $\{\varphi_l\}_{l=1}^{\infty}$ is an orthonormal eigenbasis of $\mathcal{H}$. Since it was assumed that the eigenvalues of $\mathcal{H}$ are unique, and the Σ_l are assumed to be ordered so that $\mathrm{tr}(\Sigma_l)$ is nonincreasing, this yields the spectral decomposition of $\mathcal{H}$.

Proof of Theorem 3.

We have

$$
\mathrm{cov}\left\{ X_j(s), X_k(t) \mid X_{-(j,k)} \right\}
$$

$$
= \mathrm{cov}\left\{ \sum_{l=1}^{\infty} \theta_{lj} \varphi_l(s), \sum_{l'=1}^{\infty} \theta_{l'j} \varphi_{l'}(s) \mid X_{-(j,k)} \right\}
$$

$$
= \sum_{l,l'=1}^{\infty} \mathrm{cov}\left\{ \theta_{lj}, \theta_{l'k} \mid X_{-(j,k)} \right\} \varphi_l(s) \varphi_{l'}(t)
$$

$$
= \sum_{l=1}^{\infty} \mathrm{cov}\left\{ \theta_{lj}, \theta_{lk} \mid \theta_{l,-(j,k)} \right\} \varphi_l(s) \varphi_l(t)
$$

$$
= \sum_{l=1}^{\infty} \tilde{\sigma}_{ljk} \varphi_l(s) \varphi_l(t)
$$

21

Convergence of the sum in the last line follows since $\sum_{l=1}^{\infty} \tilde{\sigma}_{ljk}^2 \le \sum_{l=1}^{\infty} \sigma_{ljj}\sigma_{lkk} < \infty$.

Proof of Corollary 1.

The result follows immediately, since $(j,k) \notin E$ if and only if $\tilde{\sigma}_{ljk} = 0$ for all $l \in \mathbb{N}$, which holds if and only if $(j,k) \notin E_l$ for all $l \in \mathbb{N}$.

Proposition 2. *Let $\theta_{lj} = \langle X_j, \varphi_l \rangle$, and E_l be the edge set of the Gaussian graphical model for θ_l. Suppose that the following properties hold for each $j, k \in V$ and $l \in \mathbb{N}$.*

- $E(\theta_{lj}|X_{-(j,k)}) = E(\theta_{lj}|\theta_{l,-(j,k)})$ *and* $E(\theta_{lj}\theta_{lk}|X_{-(j,k)}) = E(\theta_{lj}\theta_{lk}|\theta_{l,-(j,k)})$

- $(j,k) \notin E_l$ *and* $(j,k) \notin E_l'$ *implies* $\mathrm{cov}(\theta_{lj}, \theta_{l'k}|X_{-(j,k)}) = 0$.

Then $E = \bigcup_{l=1}^{\infty} E_l$.

Proof of Proposition 2.

In general, we may write $X_j = \sum_{l=1}^{\infty} \theta_{lj}\varphi_l$, though the coefficients $\theta_{lj} = \langle X_j, \varphi_l \rangle$ need not be uncorrelated across l when $\mathcal{G}$ is not partially separable. Then, under the first assumption of the proposition,

$$
\begin{aligned}
\mathrm{cov}\left\{ X_j(s), X_k(t) \mid X_{-(j,k)} \right\} &= \sum_{l=1}^{\infty} \mathrm{cov}\{\theta_{lj}, \theta_{lk} \mid X_{-(j,k)}\}\varphi_l(s)\varphi_l(t) \\
&\quad + \sum_{l \ne l'} \mathrm{cov}\{\theta_{lj}, \theta_{l'k} \mid X_{-(j,k)}\}\varphi_l(s)\varphi_{l'}(t) \\
&= \sum_{l=1}^{\infty} \mathrm{cov}\{\theta_{lj}, \theta_{lk} \mid \theta_{l,-(j,k)}\}\varphi_l(s)\varphi_l(t) \\
&\quad + \sum_{l \ne l'} \mathrm{cov}\{\theta_{lj}, \theta_{l'k} \mid X_{-(j,k)}\}\varphi_l(s)\varphi_{l'}(t).
\end{aligned}
$$

Now, since $\{\varphi_l \otimes \varphi_{l'}\}_{l,l'=1}^{\infty}$ is an orthonormal basis of $L^2([0,1]^2)$, $(j,k) \notin E$ if and only if all of the coefficients in the above expansion are zero. Hence, if $(j,k) \notin E$, we have $\mathrm{cov}\{\theta_{lj}, \theta_{lk} \mid \theta_{l,-(j,k)}\} = 0$ for all $l \in \mathbb{N}$, hence $(j,k) \notin \bigcup_{l=1}^{\infty} E_l$. On the other hand, if $(j,k) \notin E_l$ for all l, the second assumption of the proposition implies that $\mathrm{cov}\{\theta_{lj}, \theta_{l'k} \mid X_{-(j,k)}\} = 0$ for all $l \ne l'$, whence all of the coefficients in the above display are zero, and $(j,k) \notin E$.

A2   Proof of results from Section 4

Recall that $\|\cdot\|$ is the ordinary norm on $L^2[0,1]$. For a linear operator $\mathcal{A}$ on $L^2[0,1]$ and an orthonormal basis $\{\phi_l\}_{l=1}^\infty$ for this space, the Hilbert-Schmidt norm of $\mathcal{A}$ is $\|\mathcal{A}\|_{\mathrm{HS}} = \left(\sum_{l=1}^\infty \|\mathcal{A}(\phi_l)\|^2\right)^{1/2}$, where this definition is independent of the chosen orthonormal basis. In particular, for any $f \in L^2[0,1]$, $\|f \otimes f\|_{\mathrm{HS}} = \|f\|^2$.

Lemma 1. *Suppose Assumption 2 holds, and let $\hat{\mu}_j$ and $\hat{\mathcal{G}}_{jk}$ ($j, k \in V$) be the mean and covariance estimates in* (8) *for a sample of fully observed functional data $X_i \sim X$. Then there exist constants $\tilde{C}_1, \tilde{C}_2, \tilde{C}_3 > 0$ such that, for any $0 < \delta \le \tilde{C}_3$ and for all $j, k \in V$,*

$$\mathrm{pr}\left(\|\hat{\mathcal{G}}_{jk} - \mathcal{G}_{jk}\|_{\mathrm{HS}} \ge \delta\right) \le \tilde{C}_2 \exp\left(-\tilde{C}_1 n\delta^2\right).$$

Proof of Lemma 1.

Without loss of generality, assume $\mu_j(t) = E\{X_{1j}(t)\} \equiv 0$ and set $Y_{ijk} = X_{ij} \otimes X_{ik}$, $\overline{Y}_{jk} = n^{-1}\sum_{i=1}^n Y_{ijk}$. Then the triangle inequality implies

$$\mathrm{pr}\left(\|\hat{\mathcal{G}}_{jk} - \mathcal{G}_{jk}\|_{\mathrm{HS}} \ge \delta\right) \le \mathrm{pr}\left(\|\overline{Y}_{jk} - \mathcal{G}_{jk}\|_{\mathrm{HS}} \ge \frac{\delta}{2}\right) + 2\max_{j \in V}\mathrm{pr}\left(\|\hat{\mu}_j\|^2 \ge \frac{\delta}{2}\right). \quad \text{(A1)}$$

We begin with the first term on the right-hand side of (A1), and will apply Theorem 2.5 of Bosq (2000). Specifically, we need to find $L_1, L_2 > 0$ such that

$$E\left(\|Y_{ijk} - \mathcal{G}_{jk}\|_{\mathrm{HS}}^b\right) \le \frac{b!}{2}L_1 L_2^{b-2}, \quad (b = 2, 3, \ldots),$$

which will then imply that

$$\mathrm{pr}\left(\|\overline{Y}_{jk} - \mathcal{G}_{jk}\|_{\mathrm{HS}} \ge \frac{\delta}{2}\right) \le 2\exp\left(-\frac{n\delta^2}{8L_1 + 4L_2\delta}\right). \quad \text{(A2)}$$

Let $M_j = \sum_{l=1}^\infty \sigma_{ljj} < M$ and write $X_{ij} = \sum_{l=1}^\infty \sigma_{ljj}^{1/2}\xi_{ilj}\varphi_l$, where $\xi_{ilj} = \theta_{ilj}/\sigma_{ljj}^{1/2}$ are standardized random variables with mean zero and variance one, independent across i. Then, for any $b = 2, 3, \ldots$, by Jensen's inequality,

$$\|Y_{ijk} - \mathcal{G}_{jk}\|_{\mathrm{HS}}^b = \left\{\sum_{l,l'=1}^\infty \sigma_{ljj}\sigma_{l'kk}\left(\xi_{ilj}\xi_{il'k} - \delta_{ll'}r_{ljk}\right)^2\right\}^{b/2}$$

$$= (M_j M_k)^{b/2}\left\{\sum_{l,l'=1}^\infty \frac{\sigma_{ljj}\sigma_{l'kk}}{M_j M_k}\left(\xi_{ilj}\xi_{il'k} - \delta_{ll'}r_{ljk}\right)^2\right\}^{b/2}$$

$$\le (M_j M_k)^{b/2-1}\sum_{l,l'=1}^\infty \sigma_{ljj}\sigma_{l'kk}\left|\xi_{ilj}\xi_{il'k} - \delta_{ll'}r_{ljk}\right|^b,$$

23

where $\delta_{ll'}$ is the Kronecker delta. By Assumption 2, one has $E(|\xi_{ilj}|^{2b}) \leq 2(2\varsigma^2)^b b!$ where, without loss of generality, we may assume $\varsigma^2 \geq 1$. The fact that $|r_{ljk}| < 1$ combined with the C_r inequality implies that

$$\sup_{l,l'} E\left(|\xi_{ilj}\xi_{il'k} - \delta_{ll'}r_{ljk}|\right) \leq 2^{b-1}\sup_l\left\{E(|\xi_{ilj}|^{2b}) + 1\right\} \leq 2^{b+1}\{(2\varsigma^2)^b b!\}.$$

Thus,

$$E\left(\|Y_{ijk} - \mathcal{G}_{jk}\|_{\mathrm{HS}}^b\right) \leq \frac{b!}{2}(4M\varsigma^2)^{b-2}(8M\varsigma^2)^2,$$

and we can take $L_2 = 4M\varsigma^2$ and $L_1 = 2L_2^2$ in (A2).

By similar reasoning, we can find constants $\tilde{L}_1, \tilde{L}_2 > 0$ such that

$$E(\|X_{1j}\|^b) \leq \frac{b!}{2}\tilde{L}_1\tilde{L}_2^{b-2}, \quad (b = 2, 3, \ldots),$$

whence

$$\mathrm{pr}\left(\|\hat{\mu}_j\| \geq \frac{\delta}{2}\right) \leq 2\exp\left(-\frac{n\delta^2}{8\tilde{L}_1 + 4\tilde{L}_2\delta}\right) \tag{A3}$$

Now, setting $\tilde{C}_3 \leq 2$ and $0 < \delta < \tilde{C}_3$, we find that $\mathrm{pr}(\|\hat{\mu}_j\|^2 \geq \delta/2)$ is also bounded by the right hand side of (A3), since $\delta/2 < 1$. Hence, choosing $\tilde{C}_1^{-1} = \max\{8L_1 + 4L_2\tilde{C}_3, 8\tilde{L}_1 + 4\tilde{L}_2\tilde{C}_3\}$ and $\tilde{C}_2 = 6$, (A1)–(A3) together imply the result.

Proof of Theorem 4.

Recall that $\sigma_{ljk} = \langle\mathcal{G}_{jk}(\varphi_l), \varphi_l\rangle$ and $s_{ljk} = \langle\hat{\mathcal{G}}_{jk}(\hat{\varphi}_l), \hat{\varphi}_l\rangle$. Thus, by Lemma 4.3 of Bosq (2000) and Assumption 2,

$$|s_{ljk} - \sigma_{ljk}| \leq |\langle\mathcal{G}_{jk}(\varphi_l), \varphi_l - \hat{\varphi}_l\rangle| + |\langle\mathcal{G}_{jk}(\varphi_l - \hat{\varphi}_l), \hat{\varphi}_l\rangle| + |\langle[\mathcal{G}_{jk} - \hat{\mathcal{G}}_{jk}](\hat{\varphi}_l), \hat{\varphi}_l\rangle|$$

$$\leq \|\mathcal{G}_{jk}(\varphi_l)\|\|\varphi_l - \hat{\varphi}_l\| + \|\mathcal{G}_{jk}(\varphi_l - \hat{\varphi}_l)\|\|\hat{\varphi}_l\| + \|[\mathcal{G}_{jk} - \hat{\mathcal{G}}_{jk}](\hat{\varphi}_l)\|\|\hat{\varphi}_l\|$$

$$\leq 2M\tau_l\|\hat{\mathcal{H}} - \mathcal{H}\|_{\mathrm{HS}} + \|\hat{\mathcal{G}}_{jk} - \mathcal{G}_{jk}\|_{\mathrm{HS}}. \tag{A4}$$

Now, by applying similar reasoning as in the proof of Lemma 1, there exist $C_1^*, C_2^*, C_3^* > 0$ such that, for all $0 < \delta \leq C_3^*$,

$$\mathrm{pr}\left(\|\hat{\mathcal{H}} - \mathcal{H}\|_{\mathrm{HS}} \geq \delta\right) \leq C_2^*\exp\left\{-C_1^*n\delta^2\right\}.$$

Next, let $\tau_{\min} = \min_{l\in\mathbb{N}}\tau_l > 0$, and $\tilde{C}_j$ as in Lemma 1. Set $C_3 = \min\{4M\tau_{\min}C_3^*, 2\tilde{C}_3\}$ and observe that $0 < \delta < C_3$ implies that $\delta(4M\tau_l)^{-1} < C_3^*$ $(l \in \mathbb{N})$ and $\delta/2 < \tilde{C}_3$. Hence, by applying

(A4), when $0 < \delta < C_3$, for any $l \in \mathbb{N}$ and any $j, k \in V$,

$$\text{pr}\left(|s_{ljk} - \sigma_{ljk}| > \delta\right) \leq \text{pr}\left(\|\hat{\mathcal{H}} - \mathcal{H}\|_{\text{HS}} > \frac{\delta}{4M\tau_l}\right) + \text{pr}\left(\|\hat{\mathcal{G}}_{jk} - \mathcal{G}_{jk}\|_{\text{HS}} > \frac{\delta}{2}\right)$$

$$\leq C_2^* \exp\{-C_1^* n\delta^2 (4M\tau_l)^{-2}\} + \tilde{C}_2 \exp\{-\tilde{C}_1 n(\delta/2)^2\}.$$

Hence, setting $C_2 = C_2^* + \tilde{C}_2$ and $C_1 = \min\{C_1^*(4M)^{-2}, \tilde{C}_1 \min(\tau_{\min}^2, 1)/4\}$, the result holds.

Proof of Corollary 2.

Define $\hat{c}_{lj} = \sqrt{s_{ljj}/\sigma_{ljj}}$ and, for $\epsilon \in (0, 1)$, the events

$$A_{lj}(\epsilon) = \{|1 - \hat{c}_{lj}| \leq \epsilon\} \quad (l \in \mathbb{N}, j \in V).$$

Note that

$$|\hat{r}_{ljk} - r_{ljk}| \leq \frac{|s_{ljk} - \sigma_{ljk}|}{\hat{c}_{lj}\hat{c}_{lk}\pi_l} + |1 - (\hat{c}_{lj}\hat{c}_{lk})^{-1}|.$$

Suppose $0 < \delta \leq \epsilon$. Then

$$\text{pr}\left(|\hat{r}_{ljk} - r_{ljk}| \geq 2\delta\right) \leq 2\max_{j \in V}\text{pr}\{A_{lj}(\epsilon)^c\} + \text{pr}\left[A_{lj}(\epsilon) \cap A_{lk}(\epsilon) \cap \{|\hat{r}_{ljk} - r_{ljk}| \geq 2\delta\}\right]$$

$$\leq 2\max_{j \in V}\text{pr}\{A_{lj}(\epsilon)^c\} + \text{pr}\left\{|s_{ljk} - \sigma_{ljk}| \geq \delta(1-\epsilon)^2\pi_l\right\}$$

$$+ \text{pr}\left\{|1 - \hat{c}_{lj}\hat{c}_{lk}| \geq \delta(1-\epsilon)^2\right\}. \tag{A5}$$

We next obtain bounds for the first and last terms of the last line above. Let C_j be as in Theorem 4, and $\bar{\pi} = \max_{l \in \mathbb{N}} \pi_l$. If $D_3 = \min(\epsilon, C_3\bar{\pi}^{-1})$ and $0 < \delta \leq D_3$, then

$$\text{pr}\{A_{lj}(\epsilon)^c\} \leq \text{pr}(|1 - \hat{c}_{lj}^2| > \epsilon) = \text{pr}(|s_{ljj} - \sigma_{ljj}| > \epsilon\sigma_{ljj}) \leq \text{pr}(|s_{ljj} - \sigma_{ljj}| > \delta\pi_l)$$

$$\leq C_2 \exp\{-C_1 n\tau_l^{-2}\pi_l^2\delta^2\}.$$

Next, for any $a, b, c > 0$ such that $|1 - ab| \geq 3c$, we must have either $|1 - a| \geq c$ or $|1 - b| \geq c$. By Theorem 4,

$$\text{pr}\left\{|1 - \hat{c}_{lj}| > \frac{\delta(1-\epsilon)^2}{3}\right\} \leq \text{pr}\left\{|s_{ljj} - \sigma_{ljk}| > \frac{\delta(1-\epsilon)^2\pi_l}{3}\right\}$$

$$\leq C_2 \exp\left\{-C_1 n\pi_l^2\delta^2\tau_l^{-2}(1-\epsilon)^4/9\right\}.$$

Putting these facts together, (A5) becomes

$$\text{pr}\left(|\hat{r}_{ljk} - r_{ljk}| > 2\delta\right) \leq 2C_2 \exp\{-C_1 n\tau_l^{-2}\pi_l^2\delta^2\} + C_2 \exp\{-C_1(1-\epsilon)^4 n\tau_l^{-2}\pi_l^2\delta^2\}$$

$$+ 2C_2 \exp\left\{-C_1 n\pi_l^2\delta^2\tau_l^{-2}(1-\epsilon)^4/9\right\}.$$

Taking D_3 as already stated, $D_2 = 5C_2$ and $D_1 = C_1(1-\epsilon)^4/9$, the result holds.

A2.1 Lemma 2 and Proof of Theorem 5

We introduce some additional notation. First, let $\|\cdot\|_E$ denote the usual Euclidean norm on $\mathcal{R}^m$ of any dimension, where the dimension will be clear from context. Recall that, for a $p \times p$ matrix Δ, $\|\|\Delta\|\|_\infty = \max_{j=1,\ldots,p} \sum_{k=1}^p |\Delta_{jk}|$, $\|\|\Delta\|\|_1 = \|\|\Delta^{\mathrm{T}}\|\|_\infty$, and define the vectorized norm $\|\Delta\|_\infty = \max_{j,k=1,\ldots,p} |\Delta|_{jk}$. Additionally, for $S \subset V \times V$, Δ_S is the vector formed by the elements Δ_{jk}, $(j,k) \in S$. Finally, with D_j and m_l as in Corollary 2, define functions

$$\bar{n}(\delta; c) = \frac{\log(D_2 c)}{D_1 \delta^2}, \quad \bar{\delta}(n; c) = \left\{ \frac{\log(D_2 c)}{D_1 n} \right\}^{1/2} \quad (c, \delta > 0, n \in \mathbb{N}). \tag{A6}$$

Before proceeding to the results, we describe our primal-dual witness approach as a modification of that of Ravikumar et al. (2011) to account for the presence of the group Lasso penalty in (11). Of importance are the sub-differentials of each of the penalty terms in (11), omitting the tuning parameter factor, evaluated at a generic set of inputs $(\Upsilon_1, \ldots, \Upsilon_L)$. Let $v_{ljk} = (\Upsilon_l)_{jk}$. The sub-differential contains a restricted set of stacked matrices $Z = (Z_l)_{l=1}^L$, $Z_l \in \mathcal{R}^{p \times p}$. For the Lasso penalty, these satisfy

$$(Z_l)_{jk} = \begin{cases} 0 & \text{if } j = k \\ \mathrm{sgn}(v_{ljk}) & \text{if } j \neq k, \, v_{ljk} \neq 0 \\ \in [-1, 1] & \text{if } j \neq k, \, v_{ljk} = 0. \end{cases} \tag{A7}$$

In the case of the group penalty, define $v_{\cdot jk} = (v_{1jk}, \ldots, v_{Ljk})^{\mathrm{T}}$ and $z_{\cdot jk} = \{(Z_1)_{jk}, \ldots, (Z_L)_{jk}\}^{\mathrm{T}}$. Then, for the group penalty, Z must satisfy

$$z_{\cdot jk} = \begin{cases} 0 & \text{if } j = k \\ \frac{v_{\cdot jk}}{\|v_{\cdot jk}\|_E} & \text{if } j \neq k, \, \|v_{\cdot jk}\|_E \neq 0 \\ \in \{y \in \mathcal{R}^L : \|y\|_E \leq 1\} & \text{if } j \neq k, \, \|v_{\cdot jk}\|_E = 0. \end{cases} \tag{A8}$$

We construct the so-called primal-dual witness solutions $\{(\tilde{\Xi}_l, \tilde{Z}_l) : l = 1, \ldots, L\}$ as follows.

(a) With $\overline{E}_l = E_l \cup (1,1) \cup \cdots \cup (p,p)$, define

$$(\tilde{\Xi}_1, \ldots, \tilde{\Xi}_L) = \arg \min_{\Upsilon_l \succ 0, \Upsilon_l = \Upsilon_l^{\mathrm{T}}, \Upsilon_{l, \overline{E}_l^c} = 0} \sum_{l=1}^L \left\{ \mathrm{tr}(\hat{R}_l \Upsilon_L) - \log(|\Upsilon_l|) \right\}$$

$$+ \gamma \left\{ \alpha \sum_{l=1}^L \sum_{j \neq k} |v_{ljk}| + (1 - \alpha) \sum_{j \neq k} \left(\sum_{l=1}^L v_{ljk}^2 \right)^{1/2} \right\} \tag{A9}$$

(b) Select elements $\tilde{Z}_1$ and $\tilde{Z}_2$ of the Lasso and group penalty sub-differentials evaluated at $(\tilde{\Xi}_1, \ldots, \tilde{\Xi}_L)$, respectively, that satisfy the optimality condition

$$\left[\hat{R}_l - \tilde{\Xi}_l^{-1} + \gamma \left\{ \alpha \tilde{Z}_{1l} + (1 - \alpha) \tilde{Z}_{2l} \right\} \right]_{\overline{E}_l} = 0 \quad (l = 1, \ldots, L). \tag{A10}$$

(c) Update

$$\left(\tilde{Z}_{1,l}\right)_{jk} = \frac{1}{\gamma\alpha}\left\{\left(\tilde{\Xi}_l^{-1}\right)_{jk} - \hat{r}_{ljk}\right\}, \quad \left(\tilde{Z}_{2,l}\right)_{jk} = 0, \qquad \{(j,k) \in \overline{E}_l^c, l = 1, \ldots, L\}.$$

(A11)

(d) Verify strict dual feasibility condition

$$\left|\left(\tilde{Z}_{1,l}\right)_{jk}\right| < 1, \quad \{(j,k) \in \overline{E}_l^c, l = 1, \ldots, L\}.$$

(A12)

Lemma 2. *Suppose Assumptions 1–3 hold and that $\gamma = 8\epsilon_L^{-1}\overline{\delta}(n; L^{\varrho-1}p^\varrho)$ for some $\varrho > 2$. If the sample size satisfies the lower bound*

$$n > \overline{n}\left(\min\{\mathfrak{a}_L, \mathfrak{b}_L\}; L^{\varrho-1}p^\varrho\right),$$

(A13)

then, with probability at least $1 - (Lp)^{2-\varrho}$, the bounds

$$\|\hat{\Xi}_l - \Xi_l\|_\infty \le 2\kappa_{\Psi_l}\left(m_l^{-1} + 8\epsilon_L^{-1}\right)\overline{\delta}(n; L^{\varrho-1}p^\varrho)$$

(A14)

hold simultaneously for $l = 1, \ldots, L$.

Proof of Lemma 2.

Define D_j and m_l as in Corollary 2. Observe that $n > \overline{n}(\delta; c)$ implies $\overline{\delta}(n; c) < \delta$. Since (A13) implies that $n > \overline{n}(D_3 m_l; L^{\varrho-1}p^\varrho)$ for each l, $1 \le l \le L$, we have $m_l^{-1}\overline{\delta}(n; L^{\varrho-1}p^\varrho) \le D_3$, for $1 \le l \le L$. Define

$$\mathcal{A}_l = \left\{\|R_l - \hat{R}_l\|_\infty \le m_l^{-1}\overline{\delta}(n; L^{\varrho-1}p^\varrho)\right\}.$$

(A15)

Applying Corollary 2 with $\delta = m_l^{-1}\overline{\delta}(n; L^{\varrho-1}p^\varrho)$ together with the union bound, we obtain $\mathrm{pr}(\mathcal{A}_l^c) \le L^{1-\varrho}p^{2-\varrho}$, $1 \le l \le L$, so that $\mathrm{pr}\left(\bigcap_{l=1}^L \mathcal{A}_l\right) \ge 1 - (Lp)^{2-\varrho}$. Let $\{(\tilde{\Xi}_l, \tilde{Z}_l), l = 1, \ldots, L\}$ be the primal-dual witness solutions constructed in steps (a)–(d) preceding the lemma statement. The result in (A14) will follow once we have established that, on $\bigcap_{l=1}^L \mathcal{A}_l$, we have $\tilde{\Xi}_l = \hat{\Xi}_l$ ($l = 1, \ldots, L$), and that (A14) holds with $\hat{\Xi}_l$ replaced by $\tilde{\Xi}_l$.

When $\mathcal{A}_l$ holds, we apply the condition $\gamma = 8\epsilon_L^{-1}\overline{\delta}(n; L^{\varrho-1}p^\varrho)$ to conclude that

$$\|\hat{R}_l - R_l\|_\infty \le m_l^{-1}\overline{\delta}(n; L^{\varrho-1}p^\varrho) = \left(\frac{\epsilon_L}{m_l\eta_l'}\right)\left\{8\epsilon_L^{-1}\overline{\delta}(n; L^{\varrho-1}p^\varrho)\right\}\frac{\eta_l'}{8} \le \frac{\gamma\eta_l'}{8},$$

(A16)

27

as $\epsilon_L = \min_{l=1,\dots,L} m_l \eta_l'$. Additionally, since $n > \overline{n}(\mathfrak{b}_L; L^{\varrho-1}p^\varrho)$, it must be that

$$2\kappa_{\Psi_l}\left(\|\hat{R}_l - R_l\|_\infty + \gamma\right) \le 2\kappa_{\Psi_l}(m_l^{-1} + 8\epsilon_L^{-1})\overline{\delta}(n; L^{\varrho-1}p^\varrho)$$

$$\le \frac{2\kappa_{\Psi_l}(m_l^{-1} + 8\epsilon_L^{-1})}{6y_l m_l \max(\kappa_{\Psi_l}^2\kappa_{R_l}^3, \kappa_{\Psi_l}\kappa_{R_l})(m_l^{-1} + 8\epsilon_L^{-1})^2} \quad \text{(A17)}$$

$$\le \min\left(\frac{1}{3y_l\kappa_{\Psi_l}\kappa_{R_l}^3}, \frac{1}{3y_l\kappa_{R_l}}\right)$$

whenever $\mathcal{A}_l$ holds, where the last line follows since $m_l(m_l^{-1} + 8\epsilon_L^{-1}) > 1$. Hence, the assumptions of Lemma 6 in Ravikumar et al. (2011) are satisfied whenever $\mathcal{A}_l$ holds, so that

$$\|\tilde{\Xi}_l - \Xi_l\|_\infty \le 2\kappa_{\Psi_l}(\|\hat{R}_l - R_l\|_\infty + \gamma) \le 2\kappa_{\Psi_l}(m_l^{-1} + 8\epsilon_L^{-1})\overline{\delta}(n; L^{\varrho-1}p^\varrho). \quad \text{(A18)}$$

Define $\mathcal{W}_l = \tilde{\Xi}_l^{-1} - R_l + R_l(\tilde{\Xi}_l - \Xi_l)R_l$. Having established (A17) and (A18), we apply Lemma 5 of Ravikumar et al. (2011) to conclude that

$$\|\mathcal{W}_l\|_\infty \le \frac{3}{2}y_l\|\tilde{\Xi}_l - \Xi_l\|_\infty^2\kappa_{R_l}^3$$

$$\le 6\kappa_{R_l}^3\kappa_{\Psi_l}^2 y_l(m_l^{-1} + 8\epsilon_L^{-1})^2\left\{\overline{\delta}(n; L^{\varrho-1}p^\varrho)\right\}^2$$

$$= \left\{6\kappa_{R_l}^3\kappa_{\Psi_l}^2 y_l(m_l^{-1} + 8\epsilon_L^{-1})^2\overline{\delta}(n; L^{\varrho-1}p^\varrho)\right\}\left(\frac{\epsilon_L}{\eta_l'}\right)\frac{\gamma\eta_l'}{8} \quad \text{(A19)}$$

$$\le \left\{6\kappa_{R_l}^3\kappa_{\Psi_l}^2 y_l m_l(m_l^{-1} + 8\epsilon_L^{-1})^2\overline{\delta}(n; L^{\varrho-1}p^\varrho)\right\}\left(\frac{\epsilon_L}{m_l\eta_l'}\right)\frac{\gamma\eta_l'}{8}$$

$$\le \frac{\gamma\eta_l'}{8}.$$

The last line follows by (A13) and because $\epsilon_L \le m_l\eta_l'$.

Together, (A16) and (A19) imply that, when $\bigcap_{l=1}^L \mathcal{A}_l$ holds,

$$\max\{\|\hat{R}_l - R_l\|_\infty, \|\mathcal{W}_l\|_\infty\} \le \frac{\gamma\eta_l'}{8} \quad (l = 1, \dots, L).$$

Following similar derivations to those of Lemma 4 of Ravikumar et al. (2011), for any $l = 1, \dots, L$ and $(j, k) \in \overline{E}_l^c$,

$$\left|\left(\tilde{Z}_{1,l}\right)\right|_{jk} \le \frac{\eta_l'}{4\alpha} + \frac{1}{\gamma\alpha}\left\|\Psi_{l,\overline{E}_l^c\overline{E}_l}\left(\Psi_{l,\overline{E}_l\overline{E}_l}^{-1}\right)\right\|_1\frac{2\gamma\eta_l'}{8}$$

$$+ \frac{1}{\alpha}\left\|\Psi_{l,\overline{E}_l^c\overline{E}_l}\left(\Psi_{l,\overline{E}_l\overline{E}_l}^{-1}\right)\right\|_1\left\|\alpha\left(\tilde{Z}_{1,l}\right)_{\overline{E}_l} + (1 - \alpha)\left(\tilde{Z}_{2,l}\right)_{\overline{E}_l}\right\|_\infty.$$

28

Using Assumption 3, the definitions of the sub-differentials in (A7) and (A8), and the fact that $\eta_l' = \alpha - (1 - \eta_l) > 0$, the bound then becomes

$$\left|\left(\tilde{Z}_{1,l}\right)\right|_{jk} \leq \frac{\eta_l'(2 - \eta_l)}{4\alpha} + \frac{1 - \eta_l}{\alpha} < 1,$$

and strict dual feasibility holds. Therefore, $\tilde{\Xi}_l = \hat{\Xi}_l$ for each l when $\bigcap_{l=1}^{L} \mathcal{A}_l$ holds. Together with (A18), this completes the proof.

Proof of Theorem 5.

By construction of the primal witness in (A9), it is clear that $(j, k) \notin E_l$ implies $\tilde{\Xi}_{ljk} = 0$. Under the given constraint on the sample size, we have $\tilde{\Xi}_l = \hat{\Xi}_l$ with probaility at least $1 - (Lp)^{2-\varrho}$ by Lemma 2, so that $\hat{E}_l \subset E_l$ $(1 \leq l \leq L)$ with at least the same probability.

Furthermore, (15) implies $n > \overline{n}(\mathfrak{c}_l; L^{\varrho-1}p^\varrho)$, so that $\overline{\delta}(n; L^{\varrho-1}p^\varrho) < \xi_{\min,l} \left\{4\kappa_{\Psi_l}(m_l^{-1} + 8\epsilon_L^{-1})\right\}^{-1}$ $(1 \leq l \leq L)$. Hence, for any $(j, k) \in E_l$,

$$\begin{aligned}
|\hat{\Xi}_{ljk}| &\geq |\Xi_{ljk}| - |\hat{\Xi}_{ljk} - \Xi_{ljk}| \\
&\geq \xi_{\min,l} - 2\kappa_{\Psi_l}\left(\tau_l^{-2}\pi_l^2 + 8\epsilon_L^{-1}\right)\overline{\delta}(n; L^{\varrho-1}p^\varrho) \\
&\geq \xi_{\min,l}/2 > 0.
\end{aligned}$$

It follows that, with probability at least $1 - (Lp)^{2-\varrho}$, $E_l \subset \hat{E}_l$, $(1 \leq l \leq L)$, and the proof is complete.

A3 Additional Edge Consistency Result

In this section, we prove a second result on edge selection consistency that is not restrictive on the value of the tuning parameter α. As a trade-off for removing this restriction, we obtain the slightly weaker result that $\bigcup_{l=1}^{L} \hat{E}_l = \bigcup_{l=1}^{L} E_l$ with high probability, rather than accurate recovery of each individual edge set simultaneously. Unlike the proof of Theorem 5, the result deals more explicitly with the group Lasso penalty, and requires an adapted version of the irrepresentability condition. However, the constraints on the sample size and divergence of the parameter L are slightly weakened as a result.

Recall the definition of Ψ_l from Section 4.2. Define the block matrix $\tilde{\Psi} = \{\tilde{\Psi}_{e,e'}\}_{e,e' \in V \times V}$, where $\Psi_{(j,k),(j'k')}$ is an $L \times L$ diagonal matrix with diagonal equal to $\{\Psi_{1,(j,k),(j',k')}, \ldots, \Psi_{L,(j,k),(j',k')}\}$. Thus, $\tilde{\Psi}$ groups the elements of each of the Ψ_l within the same edge pairs rather than the same basis. Letting $S = \bigcup_{l=1}^{L} \overline{E}_l$, we can define the submatrix $\tilde{\Psi}_{S^c S}$, with row and column blocks indexed by S^c and S, respectively. Similarly, define Ψ_{SS}.

We next define an alternative operator norm on $\tilde{\Psi}_{S^c S}\tilde{\Psi}_{SS}^{-1}$ tailored to the group Lasso sub-differential defined in (A8). Let A be an $(|S^c|^2 L) \times (|S|^2 L)$ matrix consisting of $L \times L$ blocks

$A_{(j,k),(j',k')}$ that are themselves diagonal. Where as the norm in Assumption 3 corresponds to the ℓ_∞/ℓ_∞ matrix operator norm, due to the more restricted set of matrices in the group Lasso sub-differential, we define the blockwise ℓ_∞/ℓ_2 norm

$$\|A\|_{\infty,2} = \max_{e \in S^c} \max_{\|z_{e'}\|_E \leq 1} \left\|\sum_{e' \in S} A_{ee'} z_{e'}\right\|_E = \max_{e \in S^c} \left\{\sum_{l=1}^{L} \left(\sum_{e' \in S} |A_{ee'll}|\right)^2\right\}^{1/2}. \qquad (A20)$$

We require the following group irrepresentability condition.

Assumption 4. For some $\eta \in (0,1]$, $\left\|\tilde{\Psi}_{S^c S} \tilde{\Psi}_{SS}^{-1}\right\|_{\infty,2} \leq 1 - \eta$.

Next, define $\tilde{\kappa}_{\Psi_l} = \left\|(\Psi_{l,SS})^{-1}\right\|_\infty$, $y = \max_{j \in V} |\{k \in V : \sum_{l=1}^{L} \Xi_{ljk}^2 \neq 0\}|$, and $\tilde{\xi}_{\min} = \min_{(j,k) \in \bigcup_{l=1}^{L} E_l} \{\max_{l=1,\dots,L} |\Xi_{ljk}|\}$. With D_j as in Corollary 2 and $\tilde{\epsilon}_L = \eta \min_{l=1,\dots,L} m_l$, set

$$\tilde{\mathfrak{a}}_L = D_3 \min_{l=1,\dots,L} m_l,$$

$$\tilde{\mathfrak{b}}_L = (6y)^{-1} \min_{l=1,\dots,L} \left\{m_l(m_l^{-1} + 8\tilde{\epsilon}_L^{-1})^2 \max\left(\tilde{\kappa}_{\Psi_l}^2 \kappa_{R_l}^2, \tilde{\kappa}_{\Psi_l} \kappa_{R_l}\right)\right\}^{-1} \qquad (A21)$$

$$\tilde{\mathfrak{c}}_L = \tilde{\xi}_{\min}\{4\tilde{\kappa}_{\Psi_l}(m_l^{-1} + 8\tilde{\epsilon}_L^{-1})\}^{-1}.$$

Theorem 6. *Suppose Assumptions 1, 2, and 4 hold, $L \leq np$, $\gamma = 8\tilde{\epsilon}_L^{-1}\left\{(D_1 n)^{-1}\log(D_2 L^{\varrho-1}p^\varrho)\right\}^{1/2}$ for some $\varrho > 2$ and that the sample size satisfies the lower bound*

$$n \min(\tilde{\mathfrak{a}}_L, \tilde{\mathfrak{b}}_L, \tilde{\mathfrak{c}}_L)^2 > D_1^{-1}\left\{\log(D_2) + (\varrho - 1)\log(n) + (2\varrho - 1)\log(p)\right\}. \qquad (A22)$$

Then, for any $\alpha \in (0,1)$ and with probability at least $1 - (Lp)^{2-\varrho}$, $\bigcup_{l=1}^{L} \hat{E}_l = \bigcup_{l=1}^{L} E_l$.

Before giving the proof, a few remarks are in order. For large n, the bound in (A22) once again becomes $n \min(\tilde{\mathfrak{a}}_L, \tilde{\mathfrak{b}}_L, \tilde{\mathfrak{c}}_L)^2 \gtrsim \varrho \log(p)$, so that one can achieve selection consistency of the union of the first L edge sets so long as $\log(p) = o(n)$ and L grows slowly enough with n. Second, if we regard $\tilde{\kappa}_{\Psi_l}$, κ_{R_l}, and η to be fixed as n, p, and L diverge, if $\min_{l=1,\dots,L} m_l \gtrsim n^{-d}$ for $0 < d < 1/4$, (A22) becomes

$$n \gtrsim \left[\left\{\tilde{\xi}_{\min}^{-2} + y^2\right\} \varrho \log(p)\right]^{1-4d}.$$

Compared to the bound given in Remark 4 under analagous settings, it is weaker in the first term since $\tilde{\xi}_{\min} > \min_{l=1,\dots,L} \xi_{\min,l}$. However, since $y \geq y_l$ for any l, the second term here can be more restrictive. Practically speaking, this is not as much of a concern as it will only be apparent when the individual edge sets E_l all have a much smaller maximal degree than their union. Finally, one can deduce edge selection consistency if L is capable of growing faster than $\tilde{L}_p^* = \min\left\{L : \bigcup_{l=1}^{L} E_l = \bigcup_{l=1}^{\infty} E_l\right\}$ while still satisfying $L \leq np$ and $\min(\tilde{\mathfrak{a}}_L, \tilde{\mathfrak{b}}_L, \tilde{\mathfrak{c}}_L) \{n/\log(n)\}^{1/2} \to \infty$.

Corollary 4. *Under the assumptions of Theorem 6, suppose $L \to \infty$ such that $\min(\tilde{\mathfrak{a}}_L, \tilde{\mathfrak{b}}_L, \tilde{\mathfrak{c}}_L)\{n/\log(n)\}^{1/2} \to \infty$ and, for large n, $L \geq L_p^*$. Then, for large n, $E = \bigcup_{l=1}^L \hat{E}_l$ with probability at least $1 - (Lp)^{2-\varrho}$.*

Proof of Theorem 6.

As the proof follows the same logical flow as that of Theorem 5, we will sketch the proof while outlining major differences. First of all, steps 1–3 from Section A2.1 that detail the construction of the primal/dual witness pairs $(\tilde{\Xi}_l, \tilde{Z}_l)$ are modified as follows. In the first step, one computes the penalized estimator as in (A9) except that one only restricts $\Upsilon_{l,S^c} = 0$. In step 2, the elements of the sub-differential are chosen to satisfy the optimality condition as in (A10), but over the entire set S rather than $\overline{E}_l$. In step 3, one updates the sub-differential elements for all $(j,k) \in S^c$ as

$$\left(\tilde{Z}_{1,l}\right)_{jk} = \left(\tilde{Z}_{2,l}\right)_{jk} = \left(\tilde{Z}_l\right)_{jk} := \frac{1}{\gamma}\left\{\left(\tilde{\Xi}_l^{-1}\right)_{jk} - \hat{r}_{ljk}\right\} \quad (l = 1, \ldots, L).$$

With these amendments, one proceeds by showing that, when

$$b_l := 2\tilde{\kappa}_{\Psi_l}\left(\|\hat{R}_l - R_l\|_\infty + \gamma\right) \leq \min\left(\frac{1}{3\kappa_{R_l}y}, \frac{1}{3\kappa_{R_l}^3 \tilde{\kappa}_{\Psi_l}y}\right),$$

one has $\|\tilde{\Xi}_l - \Xi_l\|_\infty \leq b_l$. This result can be proven using the same logic used in the proofs of Lemmas 5 and 6 of Ravikumar et al. (2011) using the sub-differential properties in (A7) and (A8). Next, one uses the fact that $n > \overline{n}(\tilde{\mathfrak{a}}_L; L^{\varrho-1}p^\varrho)$ to show that, with probability at least $1 - (Lp)^{2-\varrho}$, the event $\bigcap_{l=1}^L \mathcal{A}_l$ holds, where $\mathcal{A}_l$ is defined in (A15). The rest of the proof is then conditional on this event.

Using the facts that $\gamma = 8\tilde{\epsilon}_L^{-1}\overline{\delta}(n; L^{\varrho-1}p^\varrho)$ and $n > \overline{n}(\tilde{\mathfrak{b}}_L; L^{\varrho-1}p^\varrho)$ one can then establish that

$$\max_{l=1,\ldots,L}\max\left(\|\hat{R}_l - R_l\|_\infty, \|\mathcal{W}_l\|_\infty\right) < \frac{\gamma\eta}{8},$$

so that

$$\|\tilde{\Xi}_l - \Xi_l\|_\infty \leq 2\tilde{\kappa}_{\Psi_l}(m_l^{-1} + 8\tilde{\epsilon}_L^{-1})\overline{\delta}(n; L^{\varrho-1}p^\varrho) \quad (l = 1, \ldots, L).$$

Then, using arguments similar to Lemma 1 of Ravikumar et al. (2011), the bound

$$\|\{(\tilde{Z}_1)_{jk}, \ldots, (\tilde{Z}_L)_{jk}\}^{\mathsf{T}}\|_E \leq \frac{2}{\gamma}\left(\left\|\tilde{\Psi}_{S^cS}\tilde{\Psi}_{SS}^{-1}\right\|_{2,\infty} + 1\right)\left(\frac{\gamma\eta}{8}\right) + \left\|\tilde{\Psi}_{S^cS}\tilde{\Psi}_{SS}^{-1}\right\|_{2,\infty}$$

$$\leq \frac{\eta(2-\eta)}{4} + 1 - \eta < 1$$

by Assumption 4. Hence, for each $l = 1, \ldots, L$, we have $\hat{\Xi}_l = \tilde{\Xi}_l$, so that $(j,k) \notin \bigcup_{l=1}^L \hat{E}_l$ for any $(j,k) \in S^c$ and $\bigcup_{l=1}^L \hat{E}_l \subset \bigcup_{l=1}^L E_l$. Finally, using the bound $n > \overline{n}(\tilde{\mathfrak{c}}_L; L^{\varrho-1}p^\varrho)$, for

31

$(j, k) \in \bigcup_{l=1}^{L} E_l$, one has

$$
\begin{aligned}
\max_{l=1,\ldots,L} |\hat{\tilde{\Xi}}_{ljk}| &\geq \tilde{\xi}_{\min} - \max_{l=1,\ldots,L} \|\hat{\tilde{\Xi}}_l - \Xi_l\|_\infty \\
&\geq \tilde{\xi}_{\min} - \max_{l=1,\ldots,L} 2\tilde{\kappa}_{\Psi_l}(m_l^{-1} + 8\tilde{\epsilon}_L^{-1})\bar{\delta}(n; L^{\varrho-1}p^\varrho) \\
&\geq \tilde{\xi}_{\min} - \frac{\tilde{\xi}_{\min}}{2} > 0.
\end{aligned}
$$

This implies $\bigcup_{l=1}^{L} E_l \subset \bigcup_{l=1}^{L} \hat{E}_l$, and the proof is complete.

A4 Simulation Details and Additional Figures for Section 5

A4.1 Simulation Details for Section 5

This section describes the generation of edge sets $E_1, \ldots, E_M$ and precision matrices $\Omega_1, \ldots, \Omega_M$ for the simulation settings in section 5. An initial conditional independence graph $G = (V, E)$ is generated from a power law distribution with parameter $\pi = \text{pr}\{(j, k) \in E\}$. Then, for a fixed M, a sequence of edge sets $E_1, \ldots, E_M$ is generated so that $E = \bigcup_{l=1}^{M} E_l$. This process has two main steps. First, a set of common edges to all edge sets is computed and denoted as E_c for a given proportion of common edges $\tau \in [0, 1]$. Next, the set of edges $E \setminus E_c$ is partitioned into $\tilde{E}_1, \ldots, \tilde{E}_M$ where $|\tilde{E}_l| \geq |\tilde{E}_{l'}|$ for $l < l'$ and set $E_l = E_c \cup \tilde{E}_l$. The details for this process are described in Algorithm 1.

Algorithm 1 Pseudocode to create the edge sets $E_1, \ldots, E_M$

1:   Inputs: graph $G = (V, E)$ with nodes $V = \{1, \ldots, p\}$ and edge set E
               number of basis M
               proportion of common edges τ
2:   Set $E_c \leftarrow$ random subset of size $\tau|E|$ from E
3:   Set $E_l \leftarrow E_c$ for $l = 1, \ldots, M$
4:   Set $l \leftarrow 1; B \leftarrow 1$
5:   For $e \in E \setminus E_c$
        $E_l \leftarrow E_l \bigcup e$
        $l \leftarrow l + 1$
        If $(l > B)$
            $l \leftarrow 1$
            $B \leftarrow (B + 1) \mod M$
6:   Return $E_1, \ldots, E_M$

Next, $p \times p$ precision matrices Ω_l are generated for each E_l based on the algorithm of Peng et al.

(2009). Let $\tilde{\Omega}_l$ be a $p \times p$ matrix with entries

$$\left(\tilde{\Omega}_l\right)_{jk} = \begin{cases} 1 & j = k \\ 0 & (j, k) \notin E_l \text{ or } j < k \\ \sim \mathcal{U}(\mathcal{D}) & (j, k) \in E_l \end{cases} \qquad (j, k = 1, \ldots, p)$$

where $\mathcal{D} = [-2/3, -1/3] \cup [1/3, 2/3]$. Then, rowwise, we sum the absolute value of the off-diagonal entries and divide each row by 1.5 times this sum componentwise. Finally, $\tilde{\Omega}_l$ is averaged with its transpose and has its diagonal entries set to one. The output is a precision matrix Ω_l which is guaranteed to be symmetric and diagonally dominant.

A4.2 Results for Very Sparse Graphs for Section 5

A comparison of the two methods under the very sparse case is included. For this case one has $\pi = 0.025$ with a proportion of common edges $\tau = 0$. Finally, we check the robustness of our conclusions under other settings including $\tau \in \{0, 0.1, 0.2\}$ and $\pi \in \{2.5\%, 5\%, 10\%\}$, as well as p greater, equal or smaller than n.

Table A1: Mean area under the curve (and standard error) values for Figures $A1a$ and $A1b$

			$\Sigma = \Sigma_{\text{ps}}$			$\Sigma = \Sigma_{\text{non-ps}}$		
			$p = 50$	$p = 100$	$p = 150$	$p = 50$	$p = 100$	$p = 150$
$n = p/2$	AUC	FGGM$_{90\%}$	0.61(0.06)	0.65(0.02)	0.67(0.02)	0.78(0.05)	0.77(0.02)	0.77(0.02)
		psFGGM$_{90\%}$	0.73(0.05)	0.75(0.03)	0.74(0.02)	0.81(0.05)	0.79(0.02)	0.77(0.02)
		psFGGM$_{95\%}$	0.70(0.05)	0.81(0.02)	0.82(0.02)	0.80(0.05)	0.85(0.03)	0.84(0.02)
	AUC15[†]	FGGM$_{90\%}$	0.15(0.06)	0.21(0.03)	0.26(0.02)	0.40(0.07)	0.47(0.03)	0.51(0.03)
		psFGGM$_{90\%}$	0.31(0.08)	0.44(0.03)	0.47(0.03)	0.46(0.08)	0.52(0.04)	0.52(0.03)
		psFGGM$_{95\%}$	0.27(0.07)	0.50(0.04)	0.57(0.03)	0.42(0.08)	0.59(0.05)	0.63(0.04)
$n = 1.5p$	AUC	FGGM$_{90\%}$	0.79(0.05)	0.79(0.02)	0.77(0.02)	0.94(0.03)	0.84(0.02)	0.79(0.02)
		psFGGM$_{90\%}$	0.93(0.03)	0.82(0.02)	0.78(0.02)	0.96(0.02)	0.85(0.02)	0.80(0.02)
		psFGGM$_{95\%}$	0.92(0.03)	0.92(0.02)	0.87(0.02)	0.96(0.02)	0.93(0.02)	0.87(0.02)
	AUC15[†]	FGGM$_{90\%}$	0.39(0.08)	0.52(0.03)	0.53(0.02)	0.82(0.05)	0.68(0.03)	0.60(0.03)
		psFGGM$_{90\%}$	0.78(0.06)	0.65(0.03)	0.58(0.02)	0.86(0.06)	0.69(0.04)	0.60(0.04)
		psFGGM$_{95\%}$	0.74(0.07)	0.83(0.03)	0.75(0.03)	0.86(0.05)	0.83(0.03)	0.74(0.03)

†AUC15 is AUC computed for FPR in the interval [0, 0.15], normalized to have maximum area 1.
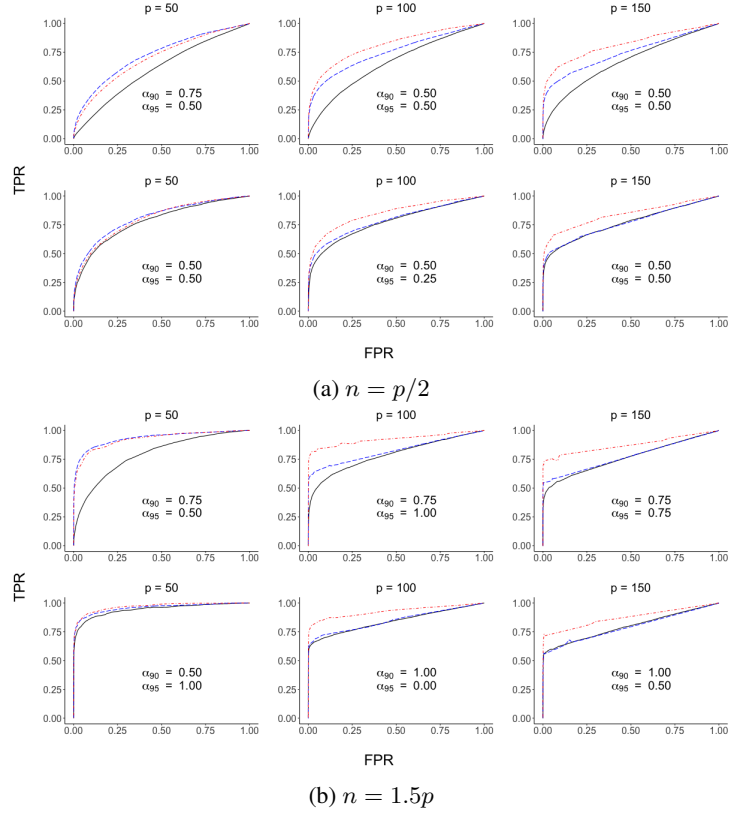
(a) $n = p/2$

(b) $n = 1.5p$

Figure A1: Mean receiver operating characteristic curves for the proposed method (psFGGM) with that of Qiao et al. (2019) (FGGM) under Σ_{ps} (top) and $\Sigma_{\text{non-ps}}$ (bottom) for $p = 50, 100, 150$. We see psFGGM (- - - -) and FGGM (——) at 90% of variance and psFGGM (— · —) at 95% of variance explained.

A4.3 Values of α for psFGGM During Simulations

Table A2: Values of α used for psFGGM in Figures $2a, 2b, A1a, A1b$

		Prop. of Variance Explained	$\Sigma = \Sigma_{\text{ps}}$			$\Sigma = \Sigma_{\text{non-ps}}$		
			$p=50$	$p=100$	$p=150$	$p=50$	$p=100$	$p=150$
Sparse	$n=0.5p$	90%	0.50	0.75	0.50	0.25	0.25	0.50
Case		95%	0.50	0.50	0.50	0.25	0.25	0.25
	$n=1.5p$	90%	0.75	1.00	0.50	0.25	0.50	0.50
		95%	0.50	0.50	0.50	0.50	0.50	0.50
Very Sparse	$n=0.5p$	90%	0.75	0.50	0.50	0.50	0.50	0.50
Case		95%	0.50	0.50	0.50	0.50	0.25	0.50
	$n=1.5p$	90%	0.75	0.75	0.75	0.50	1.00	1.00
		95%	0.50	1.00	0.75	1.00	0.00	0.50

A4.4 Comparison With An Independence Screening Procedure

In this section we compare our method (psFGGM) with another approach meant only for estimating a sparse graph identifying the conditionally independence pairs in a multivariate Gaussian process. This method, denoted as psSCREEN, is based on the sure independence screening procedure of Fan & Lv (2008). It assumes partial separability just like psFGGM, but the graph is estimated by thresholding the off-diagonal entries of the matrix $\left[\sum_{l=1}^{L} \hat{r}_{ljk}^2 \right]$ for $j, k = 1, \ldots, p$. Figures A2a and A2b follow the sparse case settings of section 5 with $\pi = 0.05$ and $\tau = 0$. And figures A3a and A3b follow the very sparse case settings of section A4.2 with $\pi = 0.025$ and $\tau = 0$.
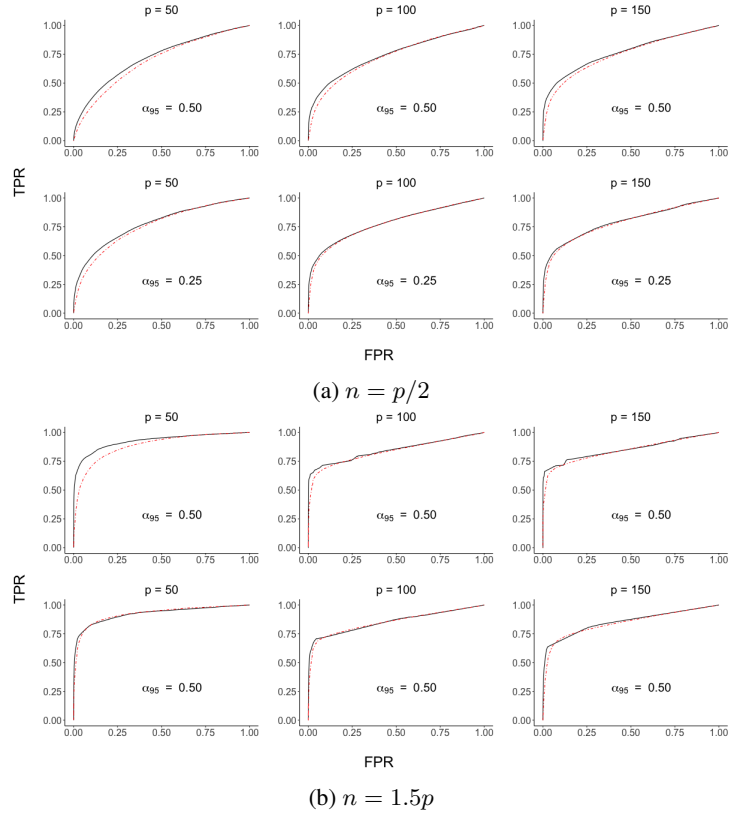
Figure A2: Mean receiver operating characteristic curves for the proposed method (psFGGM) and the independence screening procedure (psSCREEN) under Σ_{ps} (top) and $\Sigma_{\text{non-ps}}$ (bottom) for $p = 50, 100, 150$. We see psFGGM (——) and psSCREEN(— · —) both at 95% of variance for the sparse case.
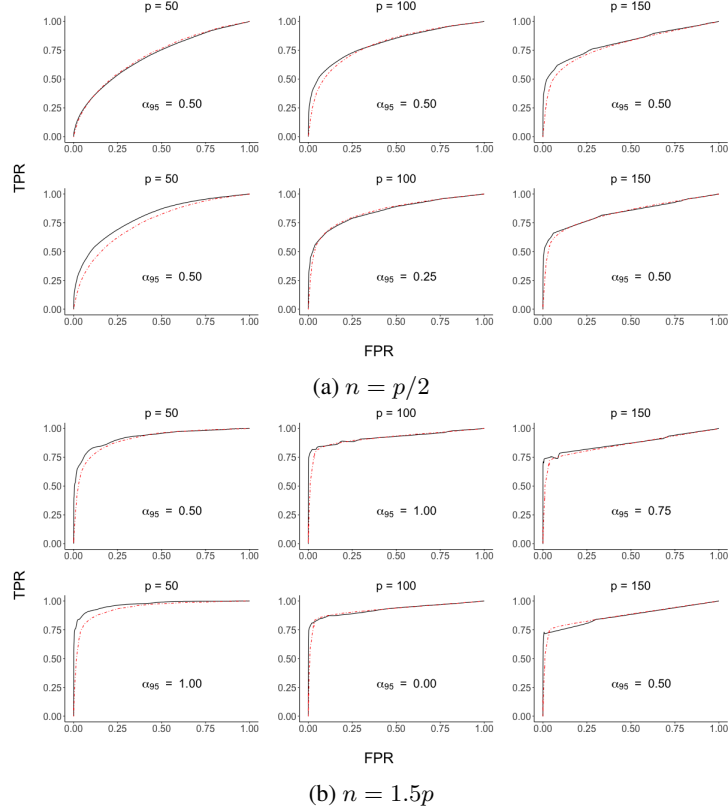
(a) $n = p/2$

(b) $n = 1.5p$

Figure A3: Mean receiver operating characteristic curves for the proposed method (psFGGM) and the independence screening procedure (psSCREEN) under Σ_{ps} (top) and $\Sigma_{\mathrm{non\text{-}ps}}$ (bottom) for $p = 50, 100, 150$. We see psFGGM (——) and psSCREEN(— · —) both at 95% of variance for the very sparse case.

A5   Right- and Left-Hand Task Partial Separability Estimates for Section 6

Motivated by Theorem 1.3 we compare the empirical performance of the partially separable and univariate Karhunen-Loève type expansions on the right-hand task fMRI curves under different number of eigenfunction estimates. Subjects are randomly assigned into training and validation sets of equal size. The training set is used to estimate the eigenfunctions of each expansion, whereas the validation set is used to compute out-of-sample variance explained percentages. Boxplots are computed on 100 simulations of this procedure. Figure A4a shows that the univariate Karhunen-Loève exhibits a better in-sample performance, a known optimality property of functional principal

37

Table A3: Mean area under the curve (and standard error) values for Figures $A2a$ and $A2b$

n			$\Sigma = \Sigma_{\text{ps}}$			$\Sigma = \Sigma_{\text{non-ps}}$		
			$p = 50$	$p = 100$	$p = 150$	$p = 50$	$p = 100$	$p = 150$
$p/2$	AUC	psFGGM$_{95\%}$	0.72(0.04)	0.74(0.02)	0.77(0.02)	0.77(0.03)	0.78(0.02)	0.79(0.02)
		psSCREEN$_{95\%}$	0.69(0.04)	0.73(0.02)	0.75(0.02)	0.75(0.03)	0.78(0.02)	0.79(0.02)
	AUC15†	psFGGM$_{95\%}$	0.29(0.05)	0.40(0.03)	0.46(0.03)	0.41(0.05)	0.48(0.03)	0.51(0.03)
		psSCREEN$_{95\%}$	0.23(0.05)	0.34(0.02)	0.40(0.02)	0.34 (0.05)	0.46(0.03)	0.49(0.03)
$1.5p$	AUC	psFGGM$_{95\%}$	0.92(0.02)	0.84(0.02)	0.85(0.02)	0.92(0.03)	0.85(0.02)	0.85(0.02)
		psSCREEN$_{95\%}$	0.89(0.02)	0.84(0.02)	0.85(0.02)	0.93(0.02)	0.86(0.02)	0.85(0.02)
	AUC15†	psFGGM$_{95\%}$	0.75(0.04)	0.68(0.03)	0.69(0.03)	0.76(0.06)	0.68(0.03)	0.64(0.03)
		psSCREEN$_{95\%}$	0.61(0.05)	0.65(0.03)	0.67(0.03)	0.76(0.05)	0.68(0.03)	0.63(0.03)

†AUC15 is AUC computed for FPR in the interval [0, 0.15], normalized to have maximum area 1.

Table A4: Mean area under the curve (and standard error) values for Figures $A3a$ and $A3b$

n			$\Sigma = \Sigma_{\text{ps}}$			$\Sigma = \Sigma_{\text{non-ps}}$		
			$p = 50$	$p = 100$	$p = 150$	$p = 50$	$p = 100$	$p = 150$
$p/2$	AUC	psFGGM$_{95\%}$	0.70(0.05)	0.81(0.02)	0.82(0.02)	0.80(0.05)	0.85(0.03)	0.84(0.02)
		psSCREEN$_{95\%}$	0.71(0.05)	0.80(0.02)	0.80(0.02)	0.75(0.04)	0.85(0.02)	0.84(0.02)
	AUC15†	psFGGM$_{95\%}$	0.27(0.07)	0.50(0.04)	0.57(0.03)	0.42(0.08)	0.59(0.05)	0.63(0.04)
		psSCREEN$_{95\%}$	0.26(0.07)	0.43(0.04)	0.49(0.03)	0.32(0.07)	0.58(0.04)	0.59(0.03)
$1.5p$	AUC	psFGGM$_{95\%}$	0.92(0.03)	0.92(0.02)	0.87(0.02)	0.96(0.02)	0.93(0.02)	0.87(0.02)
		psSCREEN$_{95\%}$	0.91(0.03)	0.92(0.02)	0.86(0.02)	0.94(0.03)	0.93(0.02)	0.88(0.02)
	AUC15†	psFGGM$_{95\%}$	0.74(0.07)	0.83(0.03)	0.75(0.03)	0.86(0.05)	0.83(0.03)	0.74(0.03)
		psSCREEN$_{95\%}$	0.66(0.07)	0.80(0.03)	0.72(0.03)	0.77(0.06)	0.83(0.03)	0.75(0.03)

†AUC15 is AUC computed for FPR in the interval [0, 0.15], normalized to have maximum area 1.

component analysis. On the other hand, Figures A4b and A4c show that the partially separable decomposition exhibits a better out-of-sample performance in both absolute terms and in relative comparison to its in-sample performance. Similar conclusion can be obtained for the left-hand task in Figure A5.

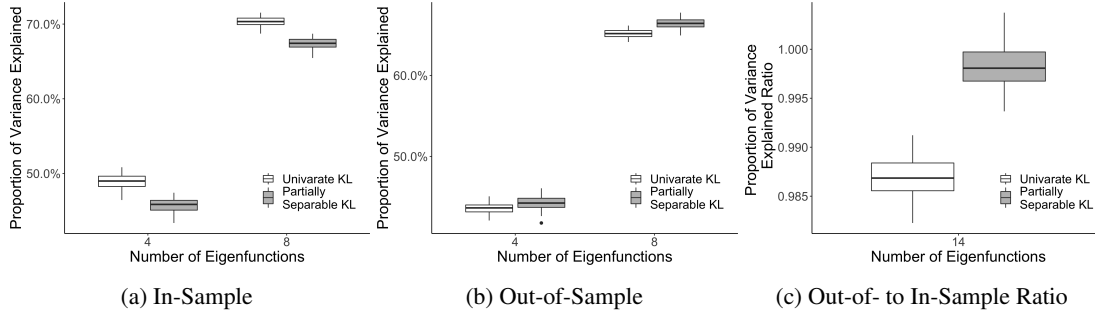| (a) In-Sample | (b) Out-of-Sample | (c) Out-of- to In-Sample Ratio |

Figure A4: Estimated variance explained for partially separable and univariate Karhunen-Loève type expansions for right-hand task fMRI curves: (a) In-Sample, (b) Out-of-Sample and (c) Out-of- to In-Sample Ratio.
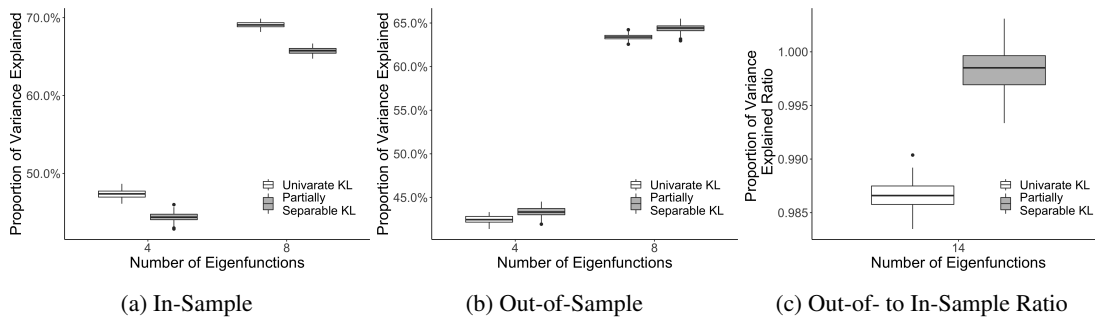

| (a) In-Sample | (b) Out-of-Sample | (c) Out-of- to In-Sample Ratio |

Figure A5: Estimated variance explained for partially separable and univariate Karhunen-Loève type expansions for left-hand task fMRI curves: (a) In-Sample, (b) Out-of-Sample and (c) Out-of- to In-Sample Ratio.

A second way to probe the partial separability assumption is to examine the correlations of the θ_{lj}. Part 3 of Theorem 1 suggest that these should be zero for different l. Figure A6 compares the correlation structures of the random coefficients in using both the functional principal component and partial separability bases, on a basis-first order (as illustrated in Figure 1(c)) using the entire right-hand task dataset. The partially separable expansion coefficients have a similar correlation structure to their univariate Karhunen-Loève type counterpart, and show that the strongest correlations exist between scores obtained from the same basis function. Similar conclusions can be drawn for the left-hand task in Figure A7. All in all, no contraindication of the partial separability structure was found in the dataset.
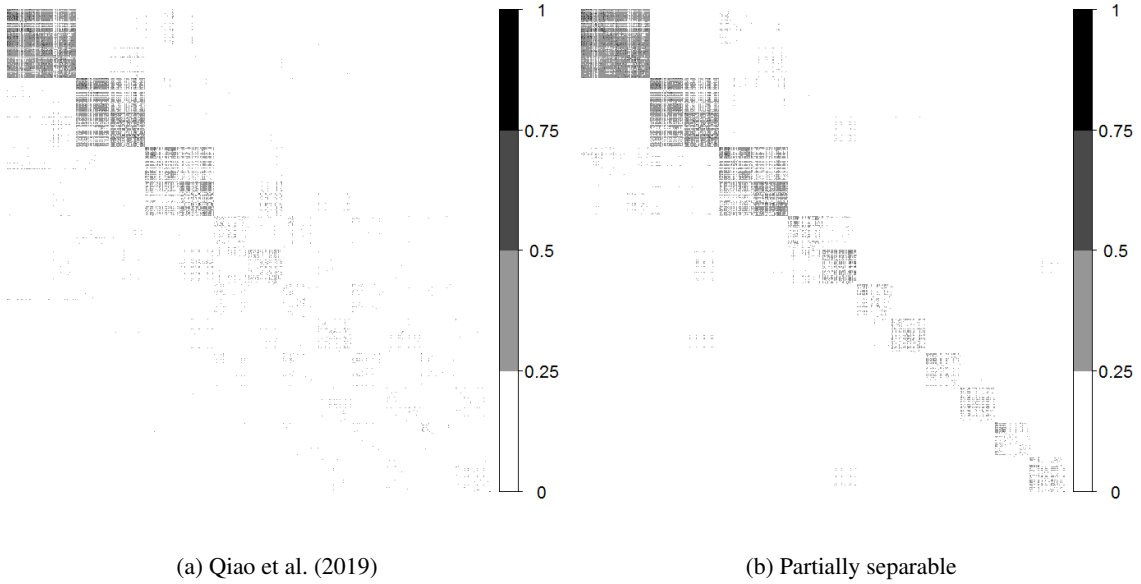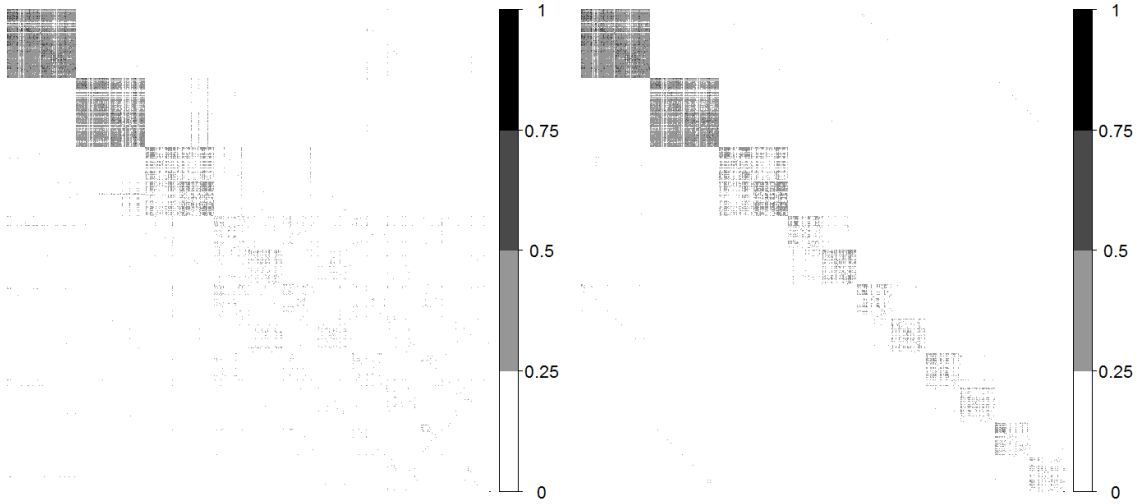
(a) Qiao et al. (2019)　　　　　　　　(b) Partially separable

Figure A6: Estimated correlation structures of $\mathbb{R}^{Lp}$-valued random coefficients from different L-truncated Karhunen-Loève type expansions for the right-hand task. The figure shows the upper left 7 x 7 basis blocks of the absolute correlation matrix in basis-first order for: (a) functional principal component coefficients $(\xi_1^T, \ldots, \xi_p^T)^T$ in (2) as in Qiao et al. (2019), and (b) random coefficients $(\theta_1^\mathrm{T}, \ldots, \theta_L^\mathrm{T})^\mathrm{T}$ under partial separability in (3).

(a) Qiao et al. (2019)

(b) Partially separable

Figure A7: Estimated correlation structures of $\mathbb{R}^{Lp}$-valued random coefficients from different L-truncated Karhunen-Loève type expansions for the left-hand task. The figure shows the upper left 7 x 7 basis blocks of the absolute correlation matrix in basis-first order for: (a) functional principal component coefficients $(\xi_1^T, \ldots, \xi_p^T)^T$ in (2) as in Qiao et al. (2019), and (b) random coefficients $(\theta_1^T, \ldots, \theta_L^T)^T$ under partial separability in (3).

41

References

Aston, J. A., Pigoli, D., Tavakoli, S. et al. (2017), 'Tests for separability in nonparametric covariance operators of random surfaces', *The Annals of Statistics* **45**(4), 1431–1461.

Banerjee, S. & Johnson, G. A. (2006), 'Coregionalized single-and multiresolution spatially varying growth curve modeling with application to weed growth', *Biometrics* **62**(3), 864–876.

Barch, D. M., Burgess, G. C., Harms, M. P., Petersen, S. E., Schlaggar, B. L., Corbetta, M., Glasser, M. F., Curtiss, S., Dixit, S., Feldt, C., Nolan, D., Bryant, E., Hartley, T., Footer, O., Bjork, J. M., Poldrack, R., Smith, S., Johansen-Berg, H., Snyder, A. Z. & Essen, D. C. V. (2013), 'Function in the human connectome: Task-fMRI and individual differences in behavior', *NeuroImage* **80**, 169 – 189.

Bhatia, R., Davis, C. & McIntosh, A. (1983), 'Perturbation of spectral subspaces and solution of linear operator equations', *Linear algebra and its applications* **52**, 45–67.

Bosq, D. (2000), *Linear Processes in Function Spaces: Theory and Applications*, Springer-Verlag, New York.

Chiou, J.-M., Chen, Y.-T. & Yang, Y.-F. (2014), 'Multivariate functional principal component analysis: A normalization approach', *Statistica Sinica* pp. 1571–1596.

Chiou, J.-M. & Müller, H.-G. (2014), 'Linear manifold modelling of multivariate functional data', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **76**, 605–626.

Chiou, J.-M., Yang, Y.-F. & Chen, Y.-T. (2016), 'Multivariate functional linear regression and prediction', *Journal of Multivariate Analysis* **146**, 301–312.

Danaher, P., Wang, P. & Witten, D. M. (2014), 'The joint graphical lasso for inverse covariance estimation across multiple classes', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **76**(2), 373–397.

Dubin, J. A. & Müller, H.-G. (2005), 'Dynamical correlation for multivariate longitudinal data', *Journal of the American Statistical Association* **100**(471), 872–881.

Fan, J. & Lv, J. (2008), 'Sure independence screening for ultrahigh dimensional feature space', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **70**(5), 849–911.

Friedman, J. H. (1984), A variable span scatterplot smoother, Technical Report 5, Stanford University.

Genton, M. G. (2007), 'Separable approximations of space-time covariance matrices', *Environmetrics: The official journal of the International Environmetrics Society* **18**(7), 681–695.

Glasser, M. F., Coalson, T. S., Robinson, E. C., Hacker, C. D., Harwell, J., Yacoub, E., Ugurbil, K., Andersson, J., Beckmann, C. F., Jenkinson, M., Smith, S. M. & Van Essen, D. C. (2016), 'A multi-modal parcellation of human cerebral cortex', *Nature* **536**, 171–178.

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., Essen, D. C. V. & Jenkinson, M. (2013), 'The minimal preprocessing pipelines for the human connectome project', *NeuroImage* **80**, 105–124.
URL: *http://www.sciencedirect.com/science/article/pii/S1053811913005053*

Gneiting, T., Genton, M. G. & Guttorp, P. (2006), 'Geostatistical space-time models, stationarity, separability, and full symmetry', *Monographs On Statistics and Applied Probability* **107**, 151.

Happ, C. & Greven, S. (2018), 'Multivariate functional principal component analysis for data observed on different (dimensional) domains', *Journal of the American Statistical Association* **113**(522), 649–659.

Hsing, T. & Eubank, R. (2015), *Theoretical Foundations of Functional Data Analysis, with an Introduction to Linear Operators*, John Wiley & Sons.

Jirak, M. (2016), 'Optimal eigen expansions and uniform bounds', *Probability Theory and Related Fields* **166**(3-4), 753–799.

Kolar, M. & Xing, E. (2011), On time varying undirected graphs, *in* 'Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics', pp. 407–415.

Lauritzen, S. L. (1996), *Graphical models*, Vol. 17, Clarendon Press.

Li, B. & Solea, E. (2018), 'A nonparametric graphical model for functional data with application to brain networks based on fMRI', *Journal of the American Statistical Association* pp. 1–19.

Lynch, B. & Chen, K. (2018), 'A test of weak separability for multi-way functional data, with application to brain connectivity studies', *Biometrika* **105**(4), 815–831.

Meinshausen, N., Bühlmann, P. et al. (2006), 'High-dimensional graphs and variable selection with the lasso', *Annals of statistics* **34**(3), 1436–1462.

Peng, J., Wang, P., Zhou, N. & Zhu, J. (2009), 'Partial correlation estimation by joint sparse regression models', *Journal of the American Statistical Association* **104**(486), 735–746.

Petersen, A. & Müller, H.-G. (2016), 'Fréchet integration and adaptive metric selection for interpretable covariances of multivariate functional data', *Biometrika* **103**, 103–120.

Qiao, X., Guo, S. & James, G. M. (2019), 'Functional graphical models', *Journal of the American Statistical Association* **114**(525), 211–222.

Qiao, X., Qian, C., James, G. M. & Guo, S. (2020), 'Doubly functional graphical models in high dimensions', *Biometrika* **107**(2), 415–431.

Qiu, H., Han, F., Liu, H. & Caffo, B. (2016), 'Joint estimation of multiple graphical models from high dimensional time series', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **78**(2), 487–504.

Ravikumar, P., Wainwright, M. J., Raskutti, G., Yu, B. et al. (2011), 'High-dimensional covariance estimation by minimizing $\ell 1$-penalized log-determinant divergence', *Electronic Journal of Statistics* **5**, 935–980.

Wang, J.-L., Chiou, J.-M. & Müller, H.-G. (2016), 'Functional data analysis', *Annual Review of Statistics and its Application* **3**, 257–295.

Yang, W., Müller, H.-G. & Stadtmüller, U. (2011), 'Functional singular component analysis', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **73**, 303–324.

Yao, F., Müller, H.-G. & Wang, J.-L. (2005), 'Functional data analysis for sparse longitudinal data', *Journal of the American Statistical Association* **100**(470), 577–590.

Zhou, S., Lafferty, J. & Wasserman, L. (2010), 'Time varying undirected graphs', *Machine Learning* **80**(2-3), 295–319.

Zhu, H., Strawn, N. & Dunson, D. B. (2016), 'Bayesian graphical models for multivariate functional data', *Journal of Machine Learning Research* **17**(204), 1–27.