University of California
Santa Barbara

Partial Separability and Graphical Models for High-Dimensional Functional Data

A dissertation submitted in partial satisfaction
of the requirements for the degree

Doctor of Philosophy
in
Statistics and Applied Probability

by

Javier Andres Zapata Ramirez

Committee in charge:

Professor Sang-Yun Oh, Co-Chair
Professor Alexander Petersen, Co-Chair
Professor Alexander Franks
Professor Wendy Meiring

June 2021

The Dissertation of Javier Andres Zapata Ramirez is approved.

Professor Alexander Franks

Professor Wendy Meiring

Professor Sang-Yun Oh, Committee Co-Chair

Professor Alexander Petersen, Committee Co-Chair

May 2021

Partial Separability and Graphical Models for High-Dimensional Functional Data

by

Javier Andres Zapata Ramirez

Dedicated to my parents Blanca Ramirez and Jose Zapata.

Acknowledgements

I want to thank my family, my friends and my Ph.D. advisors for their support during graduate school.

Curriculum Vitæ
Javier Andres Zapata Ramirez

Education

2020	Ph.D. in Statistics and Applied Probability, University of California, Santa Barbara.
2017	M.A. in Statistics, University of California, Santa Barbara.
2010	B.Eng.Sc. in Industrial Engineering, Universidad de Chile, Santiago, CHILE.

Publications

- J. Zapata, S. Oh & A. Petersen. Sparse Differential Functional Gaussian Graphical Models. *In Progress (2021)*
- J. Zapata, S. Oh & A. Petersen. Partial Separability and Functional Graphical Models for Multivariate Gaussian Processes. *Submitted (Under Review).*
- J. Zapata. **fgm** R-package for Partial Separability and Functional Graphical Models for Multivariate Gaussian Processes (2019)
- J. Zapata & A. Cifuentes. On the Stability of Synthetic CDO Credit Ratings. *International Finance, Vol. 19, No. 2, June, 2016.*

Abstract

Partial Separability and Graphical Models for High-Dimensional Functional Data

by

Javier Andres Zapata Ramirez

Functional data analysis (FDA) is the statistical methodology that analyzes datasets whose data points are functions measured over some domain, and is specially useful to model random processes over a continuum. This thesis develops a novel methodology to address the general problem of covariance modeling for multivariate functional data, and functional Gaussian graphical models in particular. The resulting methodology is applied to neuroimaging data from the Human Connectome Project (HCP).

First of all, a novel structural assumption for the covariance operator of multivariate functional data is introduced. The assumption, termed partial separability, leads to a novel Karhunen-Loève-type expansion for such data and is motivated by empirical results from the HCP data. The optimality and uniqueness of partial separability are discussed as well as an extension to multiclass datasets. The out-of-sample predictive performance of partial separability is assessed through the analysis of functional brain connectivity during a motor task.

Next, the partial separability structure is shown to be particularly useful to provide well-defined functional Gaussian graphical models. The first one is concerned with estimating conditional dependencies, while the second one estimates the difference between two functional graphical models. In each case, the models can be identified with a sequence of finite-dimensional graphical models, each of identical fixed dimension. Empirical performance of the methods for graphical model estimation is assessed through simulation and analysis of functional brain connectivity during motor tasks.

Contents

Chapter 1

Introduction

1.1 Graphical Models

Graphical models are a very powerful tool to describe the conditional dependence structure in a group of random variables through a network. The network consists of nodes representing the random variables, and edges representing conditional dependencies between a pair of variables. This thesis is concerned with undirected graphs for partial correlations where an edge represents a non-zero correlation between a pair of variables after controlling for all other remaining random variables. In the case where these random variables are jointly Gaussian then such edges can be identified by the non-zeros in the off-diagonal entries of the inverse covariance matrix, also known as the precision matrix.

The literature on Gaussian undirected graphical models has mostly focused on estimating sparse precision matrix for high-dimensional regimes where the number of observations is smaller than the number of random variables. In practice, the sparsity assumption is often used as it facilitates the interpretation of the graph. This literature can be divided into two main groups. The first one started with the neighborhood selection work of [1] where the goal is to solve a penalized regression problem where each variable is regressed on all the remaining variables. Thus non-zero entries in the precision matrix are identified column by column. And the second group corresponds to penalized likelihood methods starting with the seminal work of [2] known as the graphical lasso. In this branch of the literature the goal is to maximize a penalized Gaussian log-likelihood in terms of the precision matrix. The sparsity pattern in the resulting graph is achieved by means of a penalization on the entries of the precision matrix.

In particular, undirected graphical models have become widely used in many real life applications. They have been used in scientific domains such as computational biology [3], genetics [4], and neuroscience [5]. In particular, this work is motivated by applications of Gaussian graphical models to neuroimaging data where the goal is to estimate a functional connectivity map among the different regions of the brain.

1.2 Functional Data Analysis

Functional data analysis (FDA) is the statistical methodology that analyzes datasets whose data points are functions. In other words, each observation consists of a group of measurements observed on a discrete set of points of a continuous domain such as a time interval, a surface, etc. The main difference between FDA and traditional statistical methods is that allows function to be measured on irregular grids as opposed to an evenly spaced grid. Indeed, as the measurement points could be arbitrarily close the data-generating mechanism of interest is a random process over a continuum. For instance, consider a rainfall dataset with hourly observations over many years for multiple locations. For the purpose of FDA the dataset is as a collection of curves indexed by location and date.

With the development of new data technologies, functional datasets are becoming widely available in different fields. The list include chemometrics, medicine, biology, linguistics, ecology and finance [6] , e-commerce and marketing ([7] and [8]) to name a few.

One of the most important tools in FDA is functional principal component analysis (FPCA). It is a ubiquitous tool in functional linear regression and density estimation for functional data [9]. Its main goal is to facilitate the analysis of potentially infinite-dimensional functions by capturing the principal directions of variation of the data as well reducing its dimensionality. In essence, it summarizes the observed curves in terms of the basis provided by the principal components. In doing so, the structure of functional data can be analyzed without defining a probability measure on a functional space [10]. In particular, this thesis develops a parsimonious extension of FPCA to multivariate functional data. This extension is applied to neuroimaging data to formulate a novel and well-defined functional Gaussian graphical model.

1.3 Neuroimaging Data from the Human Connectome Project

This thesis is motivated by a large neuroimaging dataset from the Human Connectome Project (HCP). The HCP is a five-year project sponsored by sixteen components of the National Institutes of Health, split between two consortia of research institutions. And it is the first large-scale attempt to produce detailed data to understand the human connectional anatomy of the brain.

The neuroimaging data consists of functional magnetic resonance imaging (fMRI) data which measures blood-oxygen level dependent (BOLD) signals at multiple regions on the brain cortex. Variations in the blood oxygenation levels serve as a measurement of neural activity as they have a well-understood relationship with other biological processes in the cortex of the brain [11]. The BOLD signals are measured for volumetric pixels (or voxels) which represent a very small cube of brain tissue containing millions of brain cells.

In particular, this work focuses on a motor task dataset consisting of fMRI scans of individuals performing basic body movements. During each scan, a sequence of visual cues signals the subject to move one of five body parts: fingers of the left or right hand; toes of the left or right foot; or the tongue. After each three-second cue, a body movement lasts for 12 seconds with temporal resolution of 0.72 seconds. The data comes with observations for 1054 subjects with complete metadata and 91,282 voxels. In particular, this thesis analyzes the ICA-FIX pre-processed data variant as suggested by [12] that controls for spatial distortions and alignments across both subjects and modalities.

Using BOLD signals at the voxel level is not recommended in practice as they tend to be very noisy and extremely high dimensional [13]. For this reason the voxel-level BOLD signals are aggregated into averages using a parcellation (also known as atlas) of the different regions of the brain. Having removed cool down and ramp up observations, the dataset ends up with 16 time points of pure movement tasks.

Having this consideration in mind, this thesis makes use of the atlas in [14] because is a state-of-the-art parcellation with the highest number of regions among atlases using multiple MRI modalities [15]. It consists of 360 regions of interest (ROIs) delineated with respect to function, connectivity, cortical architecture, and topography, as well as, expert knowledge and meta-analysis results from the literature [13].

In Figure 1.1 we can see the different regions of the brain based on [14].

1.4 Summary of Chapters

The rest of the thesis is organized as follows. In Chapter 2 a novel structural assumption for the covariance operator of multivariate functional data is introduced. The assumption, termed partial separability, leads to a novel Karhunen-Loève-type expansion

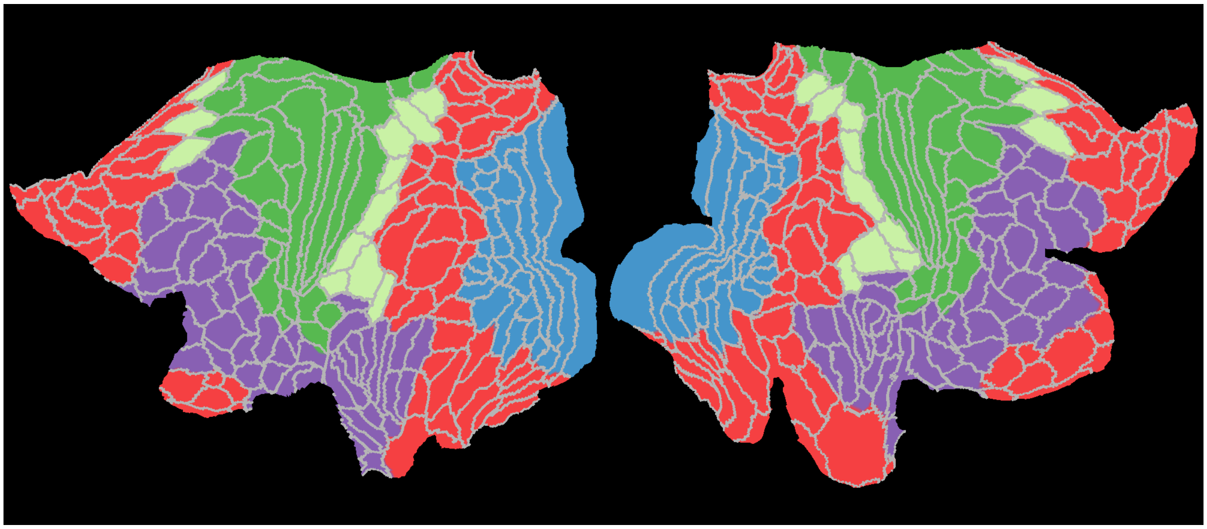


Figure 1.1: Cortical flat map of the brain for the left and right hemisphere using the brain parcellation of [14]. The regions of interest are colored based on their functionality [14]: visual (**blue**), motor (**green**), mixed motor (**light green**), mixed other (**red**) and other (**purple**).

for such data and is motivated by empirical results from the HCP data. The optimality and uniqueness of partial separability are discussed as well as its extension to multiple datasets. The out-of-sample predictive performance of partial separability is assessed through the analysis of functional brain connectivity during a motor task.

Next, in Chapter 3 the partial separability structure is shown to be particularly useful 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.

Finally, Chapter 4 introduces a differential functional graphical model based on the partial separability assumption. The differential functional Gaussian graphical model can be identified with a sequence of finite-dimensional differential graphical models, each of identical fixed dimension. This motivates a novel and efficient estimation procedure termed Joint Trace Loss. Empirical performance of the method for differential graphical model estimation is assessed through simulation and analysis of functional brain connectivity differences between two motor tasks.

Chapter 2

Partial Separability for Multivariate Functional Data

The extension of multivariate analysis techniques to multivariate functional data requires careful considerations. Examples include principal components analysis and graphical models, for which structural assumptions on the model can yield computational advantages or, in some cases, be necessary in order for the model to be well-defined.

The focus of this chapter is principal components analysis. One such extension is functional principal component analysis (FPCA) which is a commonly used tool for functional data. One of its salient features is the parsimonious reduction of a univariate stochastic process into a countable sequence of uncorrelated random variables through the Karhunen-Loève expansion [10]. This expansion holds under minimal assumptions and is especially useful in performing common functional data analysis tasks such as dimension reduction or regression [9]. However, FPCA does not have a unique multivariate extension and different approaches have been developed. For instance, [16] expands a multivariate random process into a sequence of scalar random variables. This is a very useful approach for clustering, but may not be useful for other multivariate analysis. For instance in graphical models the multivariate aspect of the data should be preserved by the expansion.

In this chapter, a novel structural assumption termed partial separability is proposed, yielding a new Karhunen-Loève type expansion for multivariate functional data. First of all, this assumption is motivated with an empirical analysis of the covariance structure of fMRI signals from a motor task experiment. Second, partial separability is defined and also compared to other separability principles available in the literature and is shown to rely on weaker assumptions. Third, the optimality and uniqueness properties of partial separability are discussed. Finally, an extension of partial separability to multiclass multivariate functional data is also provided, with an application to fMRI signals from a motor task.

2.1 Preliminaries

2.1.1 Notation

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 thesis, multivariate functional data constitute a random sample from a multivariate process

$$\{X(t) \in \mathbb{R}^p : t \in [0,1]\} \tag{2.1}$$

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 $\mathcal{G} = \{\mathcal{G}_{jk} : j,k \in \{1,\ldots,p\}\}$, where each entry $\mathcal{G}_{jk}$ is a linear, trace class integral operator on $L^2[0,1]$ [10] 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: $\{\mathcal{G}(f)\}_j = \sum_{k=1}^p \mathcal{G}_{jk}(f_k) \quad (f \in (L^2[0,1])^p, j \in V)$.

2.1.2 Dataset

To motivate and illustrate the proposed methods, a large data set consisting of functional magnetic resonance imaging (fMRI) data from the Human Connectome Project is used. A detailed description of the data can be found in the introductory chapter. For this section, data from left and right-hand finger movements task is considered.

The dataset consists of curves $\{X_{ij}(t) \in \mathbb{R}^p : t \in \tau\}$ where subjects are indexed by $i = 1,\ldots,n$, ROIs by $j = 1,\ldots,p$, and measurements are taken on a grid of equally spaced time points $\tau = \{t_1 = 0,\ldots,t_K = 1\}$. There are $n = 1054$ subjects, $p = 360$ ROIs and $K = 16$ timepoints. The curves have been centered for every ROI so that $n^{-1}\sum_{i=1}^n X_{ij}(t) = 0$ for $t \in \tau$ and $j = 1,\ldots,p$.

2.2 Empirical Motivation

For a multivariate functional process X as in (2.1), consider a univariate functional component X_j with $j \in 1, \dots, p$. The well-known Karhunen-Loève Theorem [10] provides the infinite dimensional expansion:

$$X_j(t) = \sum_{l=1}^{\infty} \xi_{jl}\phi_{jl}(t), \quad \xi_{jl} = \int_0^1 X_j(t)\phi_{jl}(t)\mathrm{d}t \tag{2.2}$$

where, the eigenfunctions $\{\phi_{jl}\}_{l=1}^{\infty}$ is an orthonormal basis of $L^2[0,1]$ and $\{\xi_{jl}\}_{l=1}^{\infty}$ are a sequence of uncorrelated zero-mean random variables also known as principal component scores or simply scores. In practice, this expansion is often truncated at a certain number of basis functions L for a given threshold of cumulative variance explained. This approach provides a regularization of the data by means of a dimensionality reduction for each component function X_j. Finally, set $\xi_j = (\xi_{j1}, \dots, \xi_{jL})^T$ $(j \in V)$ and define a $pL \times pL$ covariance matrix Γ blockwise for the concatenated vector $(\xi_1^T, \dots, \xi_p^T)^T$, as $\Gamma = (\Gamma_{jk})_{j,k=1}^p, (\Gamma_{jk})_{lm} = \mathrm{cov}(\xi_{jl}, \xi_{km}),\ (l, m = 1, \dots, L)$.

The structural assumption, that that will be proposed in Section 2.3, is motivated by an analysis of the empirical covariance structure of the scores random scores in the expansion (2.2). Using the fMRI data, the scores vector $\xi_{ij} = (\xi_{ij1}, \dots, \xi_{ijL})$ is computed for each component X_{ij}. Second, all the scores vectors are stacked into a single vector denoted $\xi_i = (\xi_{i1}^T, \dots, \xi_{ip}^T)^T \in \mathbb{R}^{Lp}$. Finally, the sample correlation matrix of ξ_i is computed with its element sorted in a basis-first ordering as: $\xi_i = (\xi_{i11}, \dots, \xi_{ip1}, \dots, \xi_{i1L}, \dots, \xi_{ipL})^T$. A detailed explanation on the computation of the univariate Karhunen-Loève expansion and the sample correlation for the random scores can be found in Section A.1 in the Appendix.

Figure 2.1 illustrates the sample correlation matrix of ξ. The first salient feature of the matrix is a block-diagonal structure with blocks of size p by p. Notice that these scores were computed independently for each univariate functional components and, in principle, they could exhibit a dense correlation structure throughout the matrix. However they only exhibit significant correlations in the diagonal blocks, especially for the first four principal components.

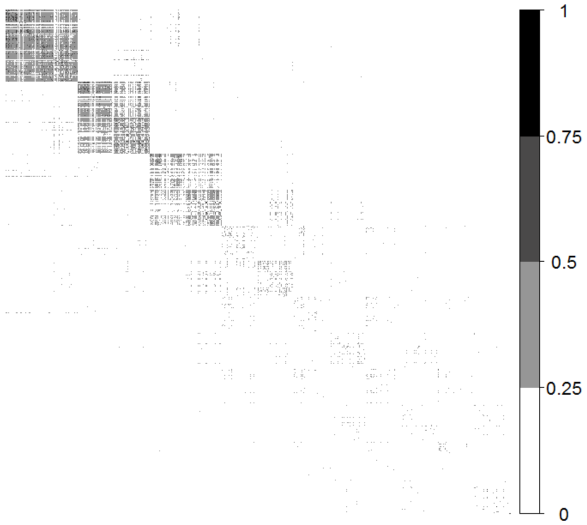
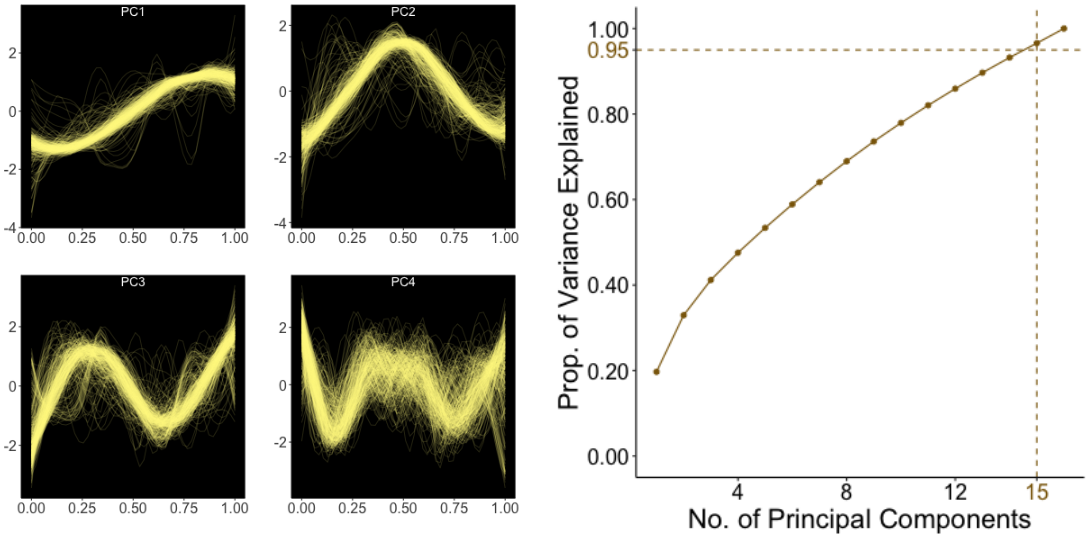


Figure 2.1: Estimated correlation structure of $\mathbb{R}^{Lp}$-valued random coefficients from an L-truncated Karhunen-Loève expansion for the right-hand task. The figure shows the upper left 7 x 7 basis blocks of the absolute correlation matrix in basis-first order for functional principal component coefficients $(\xi_1^T, \ldots, \xi_p^T)^T$ in (2.2) as in [17].

In addition, the off-diagonal blocks in Figure 2.1 exhibit a sparse pattern. These blocks are cross-correlation matrices between random scores corresponding to different principal components. For instance, block (1,2) corresponds to the sample cross-correlation matrix between score vectors $(\xi_{11}, \ldots, \xi_{p1})$ and $(\xi_{12}, \ldots, \xi_{p2})$. From the Karhunen-Loève expansion in (2.2) only the diagonal is expected to be zero but nothing prevents the remaining entries to be non-zeros.

Another motivation for an structural assumption comes from the eigenfunctions $\{\phi_{jl}\}_{l=1}^{\infty}$. Figure 2.2(a) shows the first four eigenfunctions for all the univariate components. At every principal component order l the eigenfunctions $\phi_{1l}, \ldots, \phi_{pl}$ exhibit a very similar shape. This especially interesting as they correspond to fMRI signals from different ROIs with different functionalities. On the other hand, the fMRI curves seem to be very complex and require a large number of principal components to capture a significant amount of variability. Indeed, 15 out of 16 components are needed to explain at least 95% of the variance as seen in Figure 2.2(b).



(a) (b)

Figure 2.2: Estimated functional principal components from an L-truncated Karhunen-Loève expansion for the right-hand task data set. (a): First four principal components functions $\phi_{1l}, \ldots, \phi_{pl}$ for $l = 1, \ldots, 4$ as in (2.2). (b): Proportion of variance explained by different number of principal components.

2.3 Partial Separability: A Parsimonious Basis for Multivariate Functional Data

Motivated by the empirical findings in the previous section a structural assumption for multivariate functional data is presented to obtain a novel Karhunen-Loève type decomposition. The goal in mind is to provide a parsimonious structure resulting in a functional principal component expansion that includes the main features observed in the data. That includes the block-diagonal correlation structure of the random scores, the similarity of the eigenfunctions across components, and finally, the proportion of variance explained by any number of principal components.

2.3.1 Definition and Characterization

First, a novel structural assumption on the eigenfunctions of $\mathcal{G}$ is presented, termed *partial separability.*

Definition 2.3.1. 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 (see e.g., [18, 19, 20]), 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 $\mathcal{G} = \Delta\mathcal{B}$. 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 2.3.1. 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* [21], 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 2.3.1. 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 separability. The proof of this and all remaining theoretical results can be found in the Appendix.

Theorem 2.3.1. *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 $\mathrm{cov}(\langle X_j, \varphi_l\rangle, \langle X_k, \varphi_{l'}\rangle) = 0$ $(j, k \in V,\ l \neq l')$.*

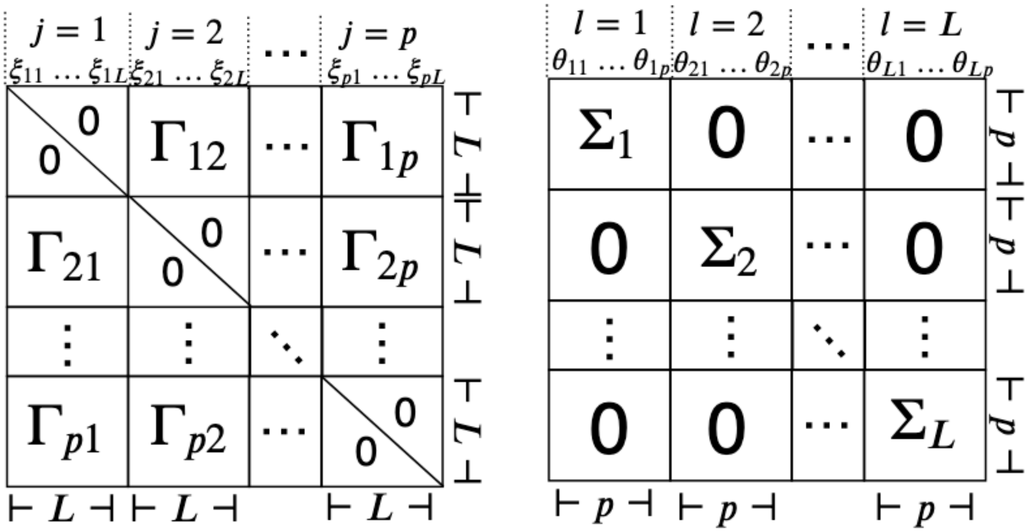
4. *The expansion*

$$X = \sum_{l=1}^{\infty} \theta_l \varphi_l, \quad \theta_l = (\langle X_1, \varphi_l\rangle, \ldots, \langle X_p, \varphi_l\rangle)^T, \tag{2.3}$$

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

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. Consequently, one obtains the vector Karhunen-Loève type expansion in (2.3). If one truncates (2.3) at L components, the covariance matrix of the concatenated vector $(\theta_1^T, \ldots, \theta_L^T)^T$ is block diagonal, with the $p \times p$ matrices $\Sigma_l = \mathrm{var}(\theta_l)$ constituting the diagonal blocks.

Figure 2.3 visualizes the partially separable covariance structure against the univariate Karhunen Loeve expansions in (2.2).


$j=1$
$j=2$
$\cdots$
$j=p$
$\xi_{11} \ldots \xi_{1L}$
$\xi_{21} \ldots \xi_{2L}$
$\xi_{p1} \ldots \xi_{pL}$
0
Γ_{12}
Γ_{1p}
Γ_{21}
Γ_{2p}
Γ_{p1}
Γ_{p2}
L
$l=1$
$l=2$
$l=L$
$\theta_{11} \ldots \theta_{1p}$
$\theta_{21} \ldots \theta_{2p}$
$\theta_{L1} \ldots \theta_{Lp}$
Σ_1
Σ_2
Σ_L
p

(a) (b)

Figure 2.3: Covariance structures of $\mathbb{R}^{Lp}$-valued random coefficients from different L-truncated Karhunen-Loève type expansions. (a): covariance of functional principal component coefficients $(\xi_1^T, \ldots, \xi_p^T)^T$ in (2.2). (b): block diagonal covariance of coefficients $(\theta_1^T, \ldots, \theta_L^T)^T$ under partial separability in (2.3).

2.3.2 Optimality and Uniqueness

Lastly, optimality and uniqueness properties are established for the basis $\{\varphi_l\}_{l=1}^{\infty}$ of a partially separable $\mathcal{G}$. A key object is the trace class covariance operator

$$\mathcal{H} = \frac{1}{p} \sum_{j=1}^{p} \mathcal{G}_{jj}. \tag{2.4}$$

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

Theorem 2.3.2. *Suppose the eigenvalues of $\mathcal{H}$ in* (2.4) *have multiplicity one.*

1. *For any $L \in \mathbb{N}$, and for any orthonormal set $\{\tilde{\varphi}_l\}_{l=1}^L$ in $L^2[0,1]$,*

$$\sum_{l=1}^{L}\sum_{j=1}^{p} \mathrm{var}(\langle X_j, \tilde{\varphi}_l \rangle) \leq \sum_{l=1}^{L} \lambda_l,$$

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*

$$\mathcal{H} = \sum_{l=1}^{\infty} \lambda_l \varphi_l \otimes \varphi_l, \quad \lambda_l = \frac{1}{p}\mathrm{tr}(\Sigma_l). \tag{2.5}$$

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 $(\langle X_1, \tilde{\varphi}_l \rangle, \ldots, \langle X_p, \tilde{\varphi}_l \rangle)^T$, 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 2.3.1 hold corresponds to this optimal basis. The proof is included in the Appendix.

2.3.3 Empirical Results

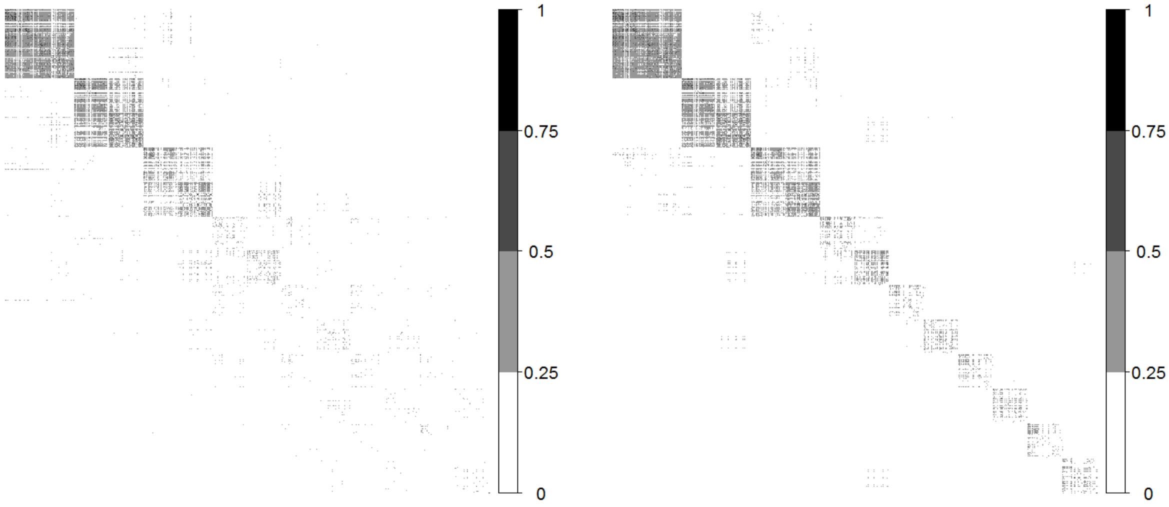
In this section the empirical performance of partial separability for multivariate functional data is analyzed. Throughout this section results for the right-hand task dataset are discussed, although similar conclusions can be obtained for the left-hand task dataset as seen in the Appendix.

First of all, the sample correlation matrices of the random scores vectors are compared for the right-hand task dataset. The random scores for the partially separable expansion are obtained from equation (2.3). The scores vector $\theta_{ij} = (\theta_{ij1}, \ldots, \theta_{ijL})$ is computed for each component X_{ij}. And then, all the scores vectors are stacked into a single vector denoted $\theta_i = (\theta_{i1}^T, \dots, \theta_{ip}^T)^T \in \mathbb{R}^{Lp}$. Finally, the sample correlation matrix of θ is computed with its element sorted in a basis-first ordering as: $\theta_i = (\theta_{i11}, \dots, \theta_{ip1}, \dots, \theta_{i1L}, \dots, \theta_{ipL})^T$.

Figures 2.4 (a) and (b) correspond to the sample correlation matrices of vectors ξ and θ. The two matrices exhibit a similar block diagonal sparse structure, with negligible entries in the off-diagonal blocks as suggested by Part 3 of Theorem 2.3.1. In particular for θ in Figure 2.4(b), the strongest correlations are concentrated within square sub-blocks of size $p/2$ by $p/2$ after the third diagonal block. Each one of these sub-blocks correspond to ROIs in the left and right hemispheres of the brain.

On the other hand, consider part 4 of Theorem 2.3.1. In principle, partial separability allows the ordering of the basis functions to be different for each component. By further assuming that $\{\varphi_l\}_{l=1}^{\infty}$ are ordered according to decreasing values of $\operatorname{tr}(\Sigma_l)$ for each component, then the expansion (2.3) has a further consequence. That is, all the univariate components of the partially separable process X should have the same eigenfunctions. In other words, the eigenfunctions ϕ_{jl} and φ_l as defined in equations (2.2) and (2.3) respectively should satisfy $\phi_{jl} = \varphi_l$ for $j = 1, \dots, p$.

Figure 2.5 compares these different eigenfunctions for every principal component. The partially separable eigenfunction φ_l has a similar behavior to the average across eigenfunction $\phi_{1l}, \dots, \phi_{pl}$ at every principal component.



(a) (b)

Figure 2.4: 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 (2.2), and (b) random coefficients $(\theta_1^T, \ldots, \theta_L^T)^T$ under partial separability in (2.3).

The similarity between the partially separable and univariate expansion eigenfunctions is explored using similarity measure in inner product spaces known as *cosine similarity.* Briefly, the cosine similarity between functions $f, g \in L^2[0,1]$ is defined as:

$$\frac{\langle f, g \rangle}{\|f\|_2 \|g\|_2}$$

taking values in the interval $[-1, 1]$. In particular, it is equal to 1 if $f = g$, -1 if $f = -g$ and 0 if f and g are orthogonal. Thus, is an appealing similarity measure to analyze part 4 of Theorem 2.3.1.

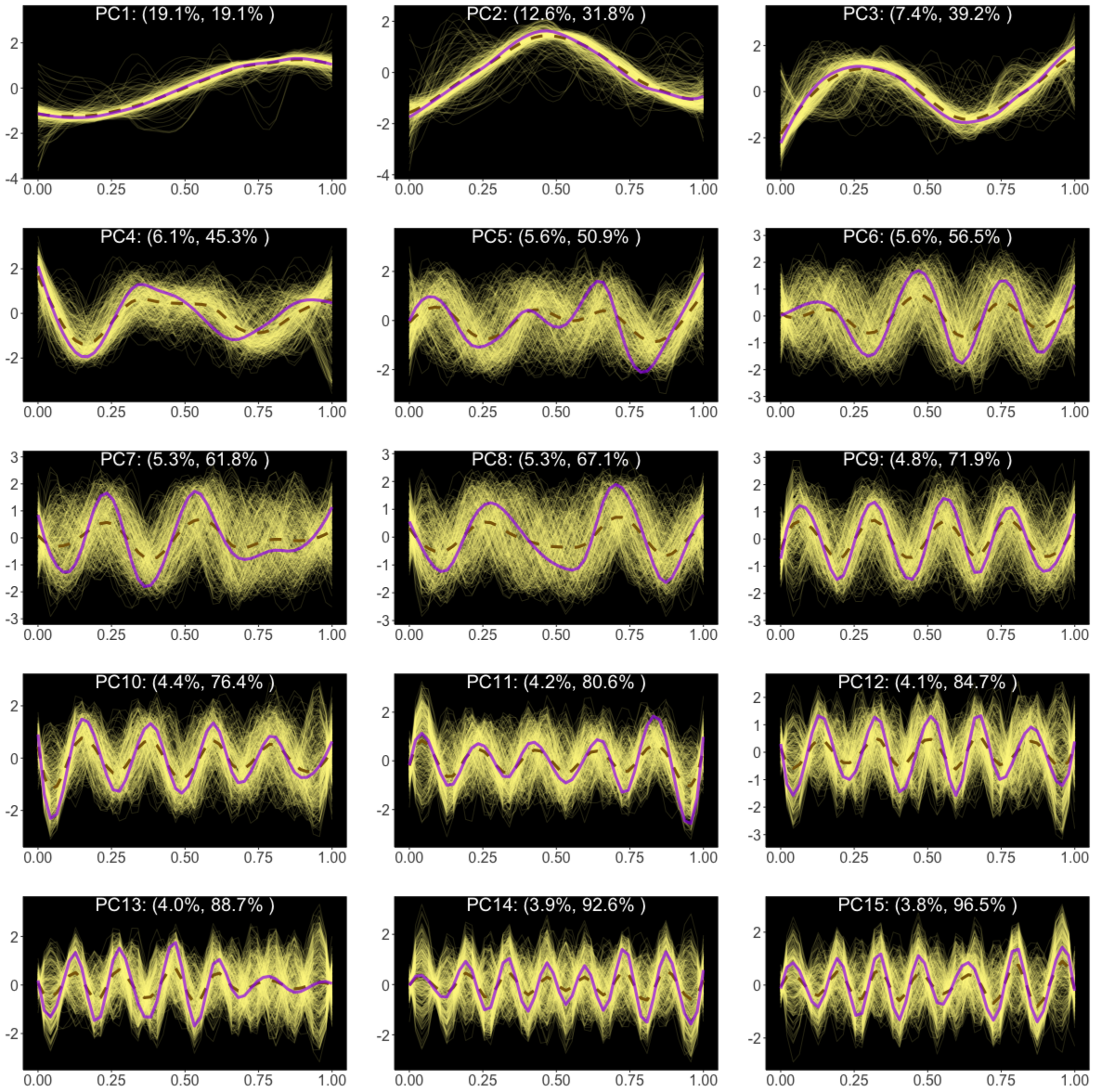
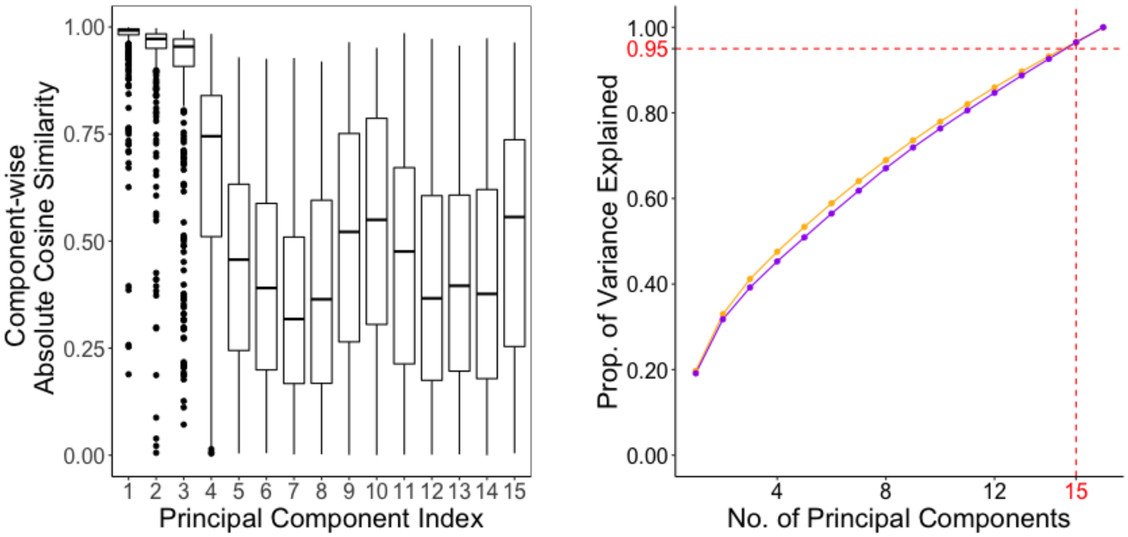


Figure 2.5: Estimated eigenfunctions from L-truncated Karhunen-Loève type expansions for the right-hand task. Curves are coded as: functional principal components eigenfunctions ϕ_{jl} (—) in (2.2), average univariate eigenfunction $p^{-1}\sum_{j=1}^{p}\phi_{jl}$ (- - -), and eigenfunctions φ_l (—) under partial separability in (2.3). Values on the top of each figure indicate the marginal and cumulative proportion of variance explained under partial separability.

Figure 2.6(a) shows the absolute cosine similarity between φ_l and each eigenfunction ϕ_{jl} for $j = 1, \ldots, p$. They look especially similar for the first four principal components with absolute cosine similarities very close to 1. And for higher order principal components, the median absolute cosine similarity does not exhibit a monotonic decreasing pattern towards zero in spite of the increasing instability of eigenfunctions estimators for higher order principal components.

Finally, Figure 2.6(b) compares the cumulative variance explained by both expansions. As expected, the univariate Karhunen-Loève exhibits a better in-sample performance, a known optimality property of functional principal component analysis. However, the partially separable expansion explains almost the same proportion of variance at every number of principal components. All in all, no strong contraindication of the partial separability structure was found in the dataset.



(a)

(b)

Figure 2.6: Comparison between L-truncated Karhunen-Loève (KL) type expansions. (a): Boxplots for component-wise absolute cosine similarity between eigenfunctions φ_l in (2.3) and ϕ_{jl} for $j = 1, \ldots, p$ in (2.2) for every principal component. (b): Proportion of variance explained for each expansion under different number of principal components. Curves are coded as: functional principal components eigenfunctions ϕ_{jl} (—) in (2.2), and eigenfunctions φ_l (—) under partial separability in (2.3).

2.3.4 Out-of-Sample Predictive Performance

This section compares the empirical performance of the partially separable and univariate Karhunen-Loève type expansions on the motor task fMRI dataset as motivated by Theorem 2.3.2. For this analysis 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. A detailed description of this procedure is given in the Appendix.

Figure 2.7 shows that the univariate Karhunen-Loève exhibits a better in-sample performance, a known optimality property of functional principal component analysis. On the other hand, Figures 2.7 and 2.8 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 conclusions can obtained for the left-hand task and are included in the Appendix.

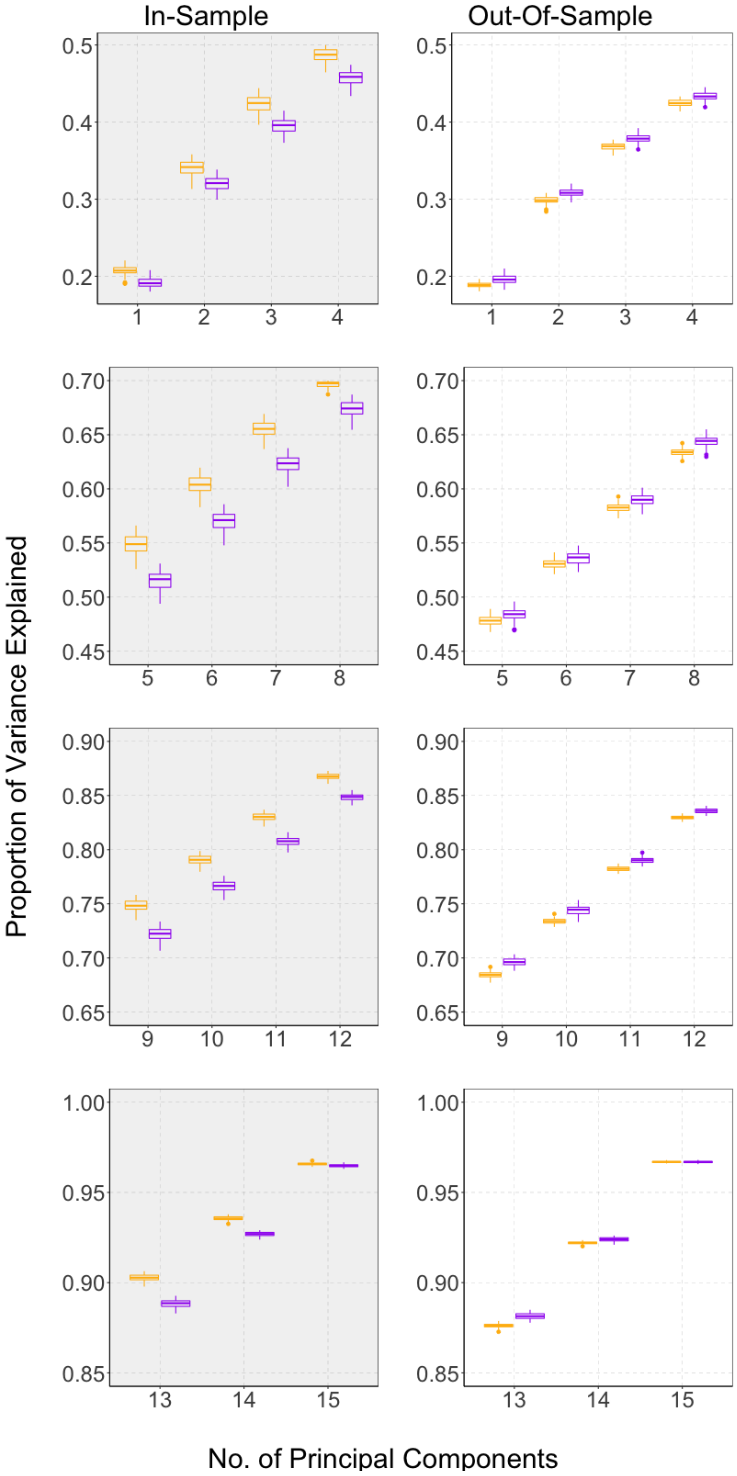


Figure 2.7: Estimated variance explained for partially separable and univariate Karhunen-Loève type expansions for right-hand task fMRI curves. Left: In-Sample. Right: Out-of-Sample. Boxplots are coded as: functional principal components in (2.2) as (—), and partially separable expansion in (2.3) as (—).

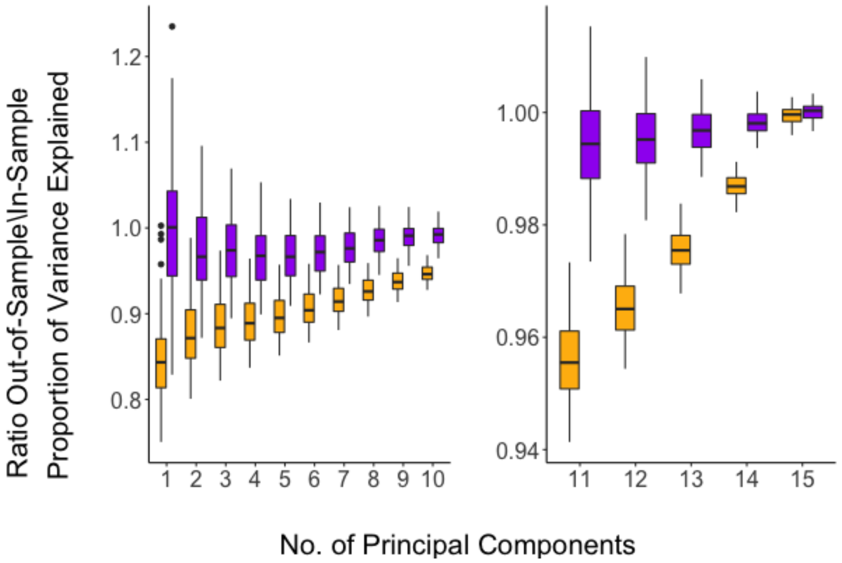


Figure 2.8: Estimated variance explained for partially separable and univariate Karhunen-Loève type expansions for right-hand task fMRI curves. The figure shows boxplots for the ratio out-of-sample over in-sample variance explained. Boxplots are coded as: functional principal components in in (2.2) as (—), and partially separable expansion in (2.3) as (—).

2.4 Joint Partial Separability

As for multivariate statistics, datasets incorporating multiple classes are also ubiquitous in functional data analysis. For instance, in neuroscience fMRI scans of healthy individuals are compared with those of patients with mental disorders [22, 23]. In particular, the empirical findings on the fMRI motor task data in Section 2.2 show that both the right- and left-hand tasks have a parsimonious covariance structure sharing similar characteristics. With this in mind, the goal in this section is to extend the partial separability principle to multiple classes to obtain a joint parsimonious structure for their covariance operators. This new formulation is applied to the two motor tasks datasets and their predictive performance is discussed.

2.4.1 Definition

Consider two multivariate functional datasets $\{X(t) \in \mathbb{R}^p, t \in \tau\}$ and $\{Y(t) \in \mathbb{R}^p, t \in \tau\}$ with covariance operators $\mathcal{G}_X$ and $\mathcal{G}_Y$, respectively. Both X and Y are assumed to be zero-mean and such that $X, Y \in (L^2[0,1])^p$ almost surely and finite moments $E(||X||_p^2)$ and $E(||Y||_p^2)$. This section begins by proposing a novel structural assumption on the eigenfunctions of $\mathcal{G}_X$ and $\mathcal{G}_Y$, termed *joint partial separability* as follows:

Definition 2.4.1. The covariance operators $\mathcal{G}_X$ and $\mathcal{G}_Y$ on $(L^2[0,1])^p$ are *joint partially separable* if there exist orthonormal bases $\{e_{lj}^X\}_{j=1,\ldots,p}$ and $\{e_{lj}^Y\}_{j=1,\ldots,p}$ of $\mathbb{R}^p$ for $l \in \mathbb{N}$ and $\{\psi_l\}_{l=1,\ldots,\infty}$ of $L^2[0,1]$ such that the eigenfunctions of $\mathcal{G}_X$ and $\mathcal{G}_Y$ take the form e_{lj}^X ψ_l and $e_{lj}^Y\psi_l(l \in \mathbb{N}, j = 1, \ldots, p)$.

In other words, definition 2.4.1 means that two multivariate functional datasets with a joint partially separable covariance structure share the same eigenfunctions. This is a stronger assumption than partial separability, but it has good empirical properties as it will be shown in Section 2.4.2. The following corollary follows directly definition 2.4.1 and provides a connection between the two partial separability notions:

Corollary 1. *The covariance operators $\mathcal{G}_X$ and $\mathcal{G}_Y$ on $(L^2[0,1])^p$ are joint partially separable if and only if they are individually partially separable with common $L^2[0,1]$ orthonormal basis.*

The main consequence from Corollary 1 is that X and Y have the same features of partially separable structures, including their characterizations and optimality and uniqueness properties. Thus, we can now write the functional principal component expansions of X and Y based on Theorem 2.3.1 part 4 as:

$$\begin{aligned} X &= \sum_{l=1}^{\infty} \vartheta_l^X \psi_l, \quad \vartheta_l^X = (\langle X_1, \psi_l \rangle, \dots, \langle X_p, \psi_l \rangle)^T \\ Y &= \sum_{l=1}^{\infty} \vartheta_l^Y \psi_l, \quad \vartheta_l^Y = (\langle Y_1, \psi_l \rangle, \dots, \langle Y_p, \psi_l \rangle)^T \end{aligned} \tag{2.6}$$

where $\{\psi_l\}_{l\geq 1}$ correspond to the eigenfunctions and $\vartheta_l^X, \vartheta_l^Y$ correspond to the random scores for X and Y respectively. The main advantage of the expansion in (2.6) is that the random scores for X and Y can be analyzed directly with multivariate statistics methods.

Lastly, optimality and uniqueness properties are established for the basis $\{\psi_l\}_{l=1}^{\infty}$ of a partially separable $\mathcal{G}^X$ and $\mathcal{G}^Y$. A key object is the trace class covariance operator

$$\mathcal{H} = \frac{1}{2p} \sum_{j=1}^{p} (\mathcal{G}_{jj}^X + \mathcal{G}_{jj}^Y) \tag{2.7}$$

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

Corollary 2. *Suppose the eigenvalues of $\mathcal{H}$ in (2.7) have multiplicity one.*

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

2. *If $\mathcal{G}^X$ and $\mathcal{G}^Y$ are joint partially separable with $L^2[0,1]$ basis $\{\psi_l\}_{l=1}^{\infty}$, then*

$$\mathcal{H} = \sum_{l=1}^{\infty} \lambda_l \psi_l \otimes \psi_l, \quad \lambda_l = \frac{1}{2p} \mathrm{tr}(\Sigma_l^X + \Sigma_l^Y) \tag{2.8}$$

Notice that Corollary 2 is analogous to Theorem 2.3.2 and the proof is omitted.

2.4.2 Empirical Results

Using the fMRI datasets for the right- and left-hand tasks, Figure 2.9 compares the eigenfunctions from the joint and single partial separability structures. Overall, the principal components from the joint expansion look very similar to their individually separable counterparts.

On the other hand, Figure 2.10 shows the resulting block correlation structures for the random score vectors $((\vartheta_1^X)^T, \ldots, (\vartheta_L^X)^T)^T$ and $((\vartheta_1^Y)^T, \ldots, (\vartheta_L^Y)^T)^T$ in (2.6). Again, a block diagonal structure is observed as expected. Finally, Figure 2.11 compares the proportion of variance explained between the functional principal components in (2.2) and ψ_l under joint partial separability in (2.6). For the left- and right-hand tasks, the curves look very similar.

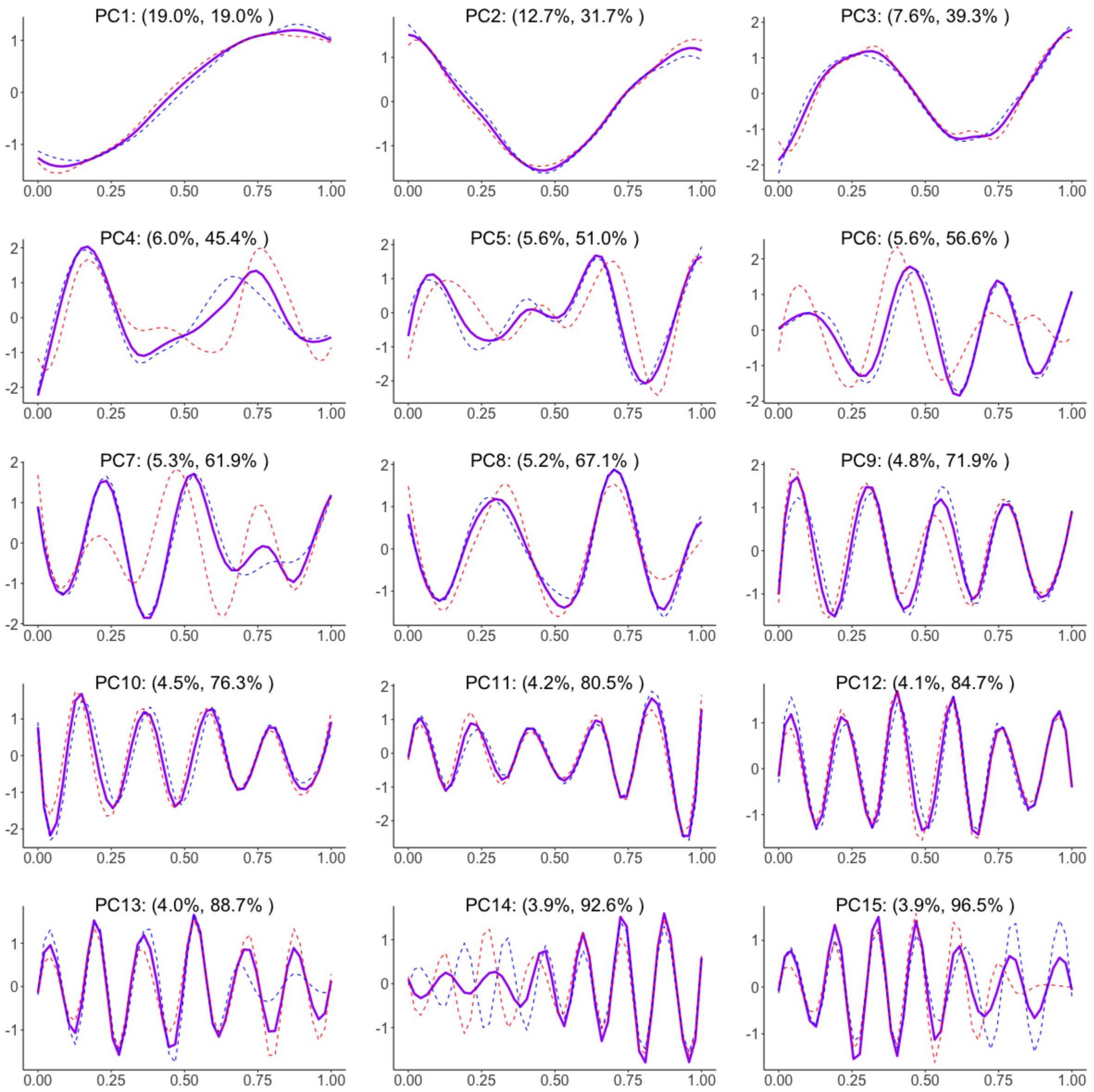
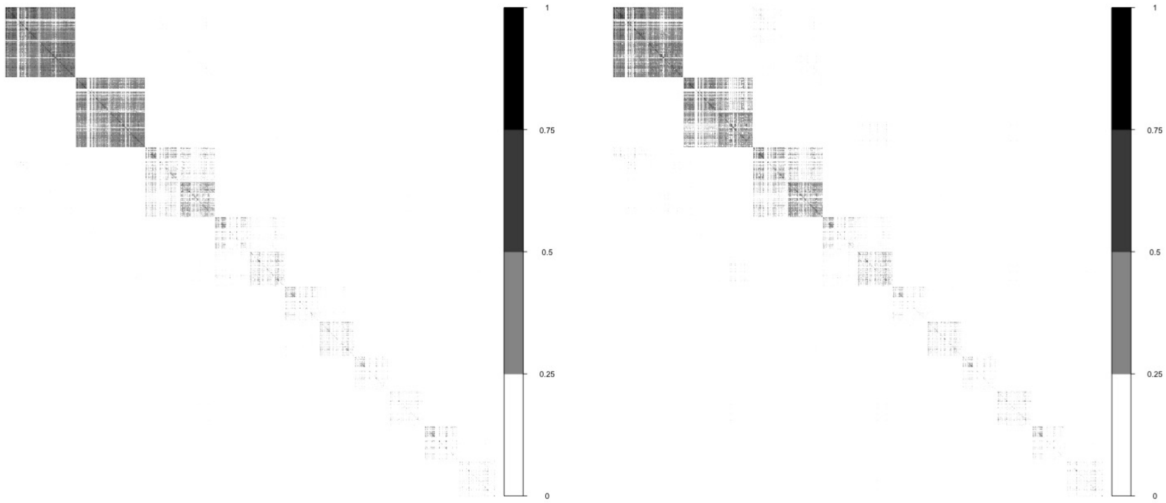


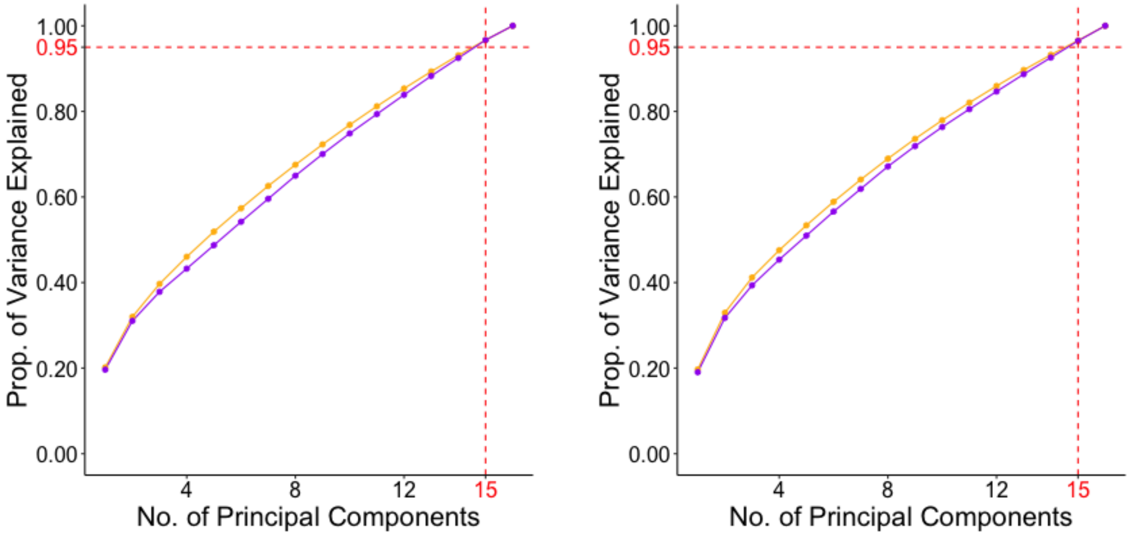
Figure 2.9: Estimated functional principal components from L-truncated partially separable Karhunen-Loève type expansions for the right- and left-hand task. Curves are coded as: partially separable functional principal components in (2.3) as right-hand (- - -) and left-hand (- - -), and joint partially separable expansion in (2.6) as (—). Values on the top of each figure indicate the marginal and cumulative proportion of variance explained under joint partial separability.



(a)

(b)

Figure 2.10: Estimated correlation structures of $\mathbb{R}^{Lp}$-valued random coefficients from the L-truncated Karhunen-Loève type expansion under joint partial separability in (2.6) for the fMRI motor task data. The figure shows the upper left 7 x 7 basis blocks of the absolute correlation matrix in basis-first order for: (a) left-hand task random scores vector $((\vartheta_1^X)^T, \ldots, (\vartheta_L^X)^T)^T$, and (b) right-hand task random scores vector $((\vartheta_1^X)^T, \ldots, (\vartheta_L^X)^T)^T$.



(a) (b)

Figure 2.11: Proportion of variance explained from different L-truncated Karhunen-Loève type expansions on the fMRI motor task dataset. (a): left-hand task. (b): right-hand task. Curves are coded as: functional principal components eigenfunctions ϕ_{jl} (—) in (2.2) and ψ_l (—) under joint partial separability eigenfunctions as in (2.6).

2.4.3 Out-of-Sample Predictive Performance

This section expands the analysis in Section 2.3.4 to include the joint partially separable Karhunen-Loève type expansions on the motor task fMRI dataset. The first column in Figure 2.12 shows that the joint partially separable expansion has the lowest in-sample performance when compared to the univariate and partially separable Karhunen-Loève expansions. On the other hand, the second column in Figure 2.12 and Figure 2.13 show that the joint partially separable expansion exhibits a better out-of-sample performance in both absolute terms and in relative comparison to its in-sample performance. Similar conclusions can be found for the left-hand task in Section A.3 in the Appendix.

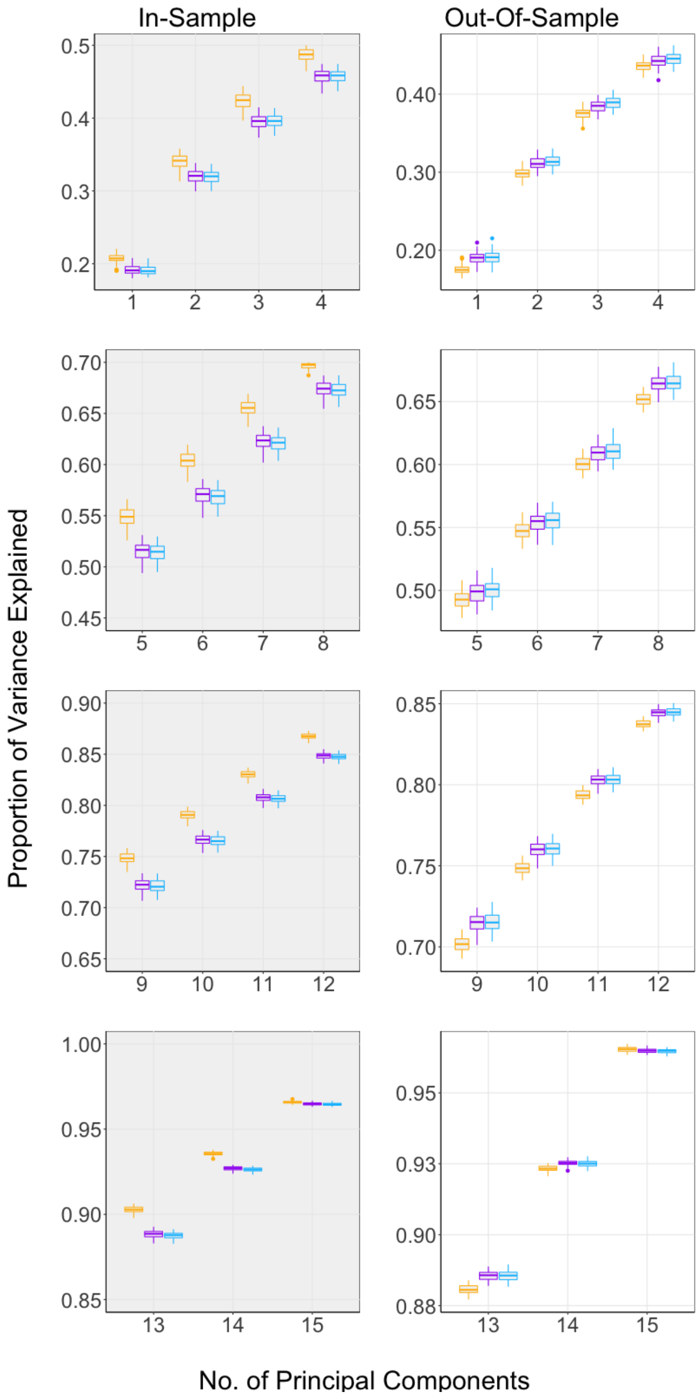


Figure 2.12: Estimated variance explained for different L-truncated Karhunen-Loève type expansions for right-hand task fMRI curves. Left: In-Sample. Right: Out--of-Sample. Boxplots are coded as: functional principal components (—) in (2.2), partially separable expansion (—) in (2.3), and joint partially separable expansion (—) in (2.6).

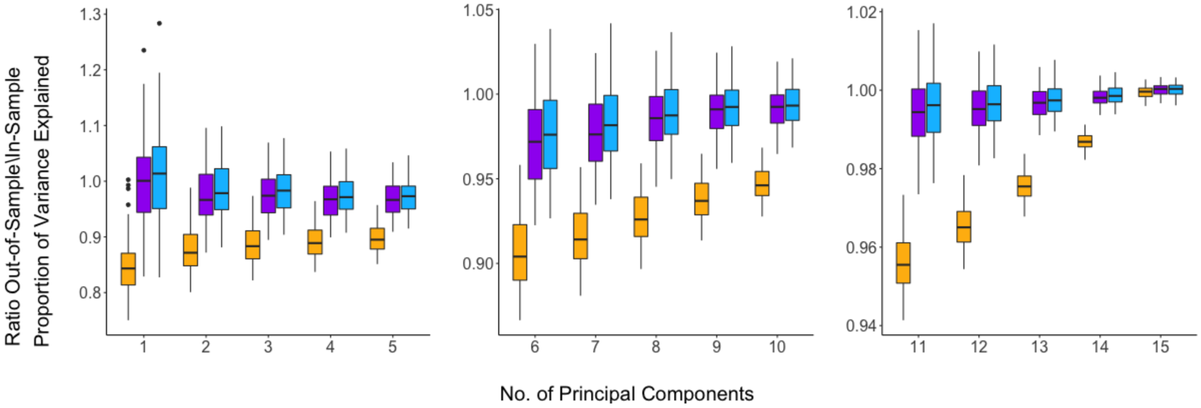


Figure 2.13: Estimated variance explained for different L-truncated Karhunen-Loève type expansions for right-hand task fMRI curves. The figure shows boxplots for the ratio out-of-sample over in-sample variance explained. Boxplots are coded as: functional principal components (🟧) in (2.2), partially separable expansion (🟪) in (2.3), and joint partially separable expansion (🟦) in (2.6).

2.5 Conclusions

The partially separable assumption provides a novel approach to analyze multivariate functional data. It enjoys several advantages. First of all, the partially separable covariance structure exhibits a better out-of-sample predictive performance than the true univariate Karhunen-Loève expansions for the motor-task fMRI data. This is a very powerful result and holds for the individual and joint partial separability assumptions. Second, partial separability is a weaker form of separability than other alternatives in the functional data literature. Third, it provides a parsimonious characterization for one of the most important neuroimaging open data repositories available. And finally, the decomposition into a vector of scores is potentially useful for statistical modeling, specifically the graphical models that will be introduced in the next chapter.

There exist several extensions for future research. First of all, partial separability is not restricted to multivariate Gaussian processes and can be extended to develop scalable methods to study other phenomena in the brain. For instance, multivariate count processes are ubiquitous in neural spike data where the goal is to understand the electrical activity of the neurons [24]. In this setting, a scalable structural assumption such as partial separability is needed to cope with functional datasets for over one million neurons [25]. Moreover, the applicability of partial separability for Poisson graphical models can be studied. And second, partial separability provides a useful framework for other functional data analysis methods. They include multilevel and mixed effects functional data where structural assumptions are needed to estimate functional fixed effects [26, 27], and multivariate functional linear regression where the predictors variables contain multiple random functions [28]. On the other hand, the joint partial separability principle provides a useful framework to develop classification and clustering algorithms for multivariate functional data. For instance, in neuroscience, differences in brain connectivity maps between healthy individuals and patients is a promising biomarker for mental disorders [22, 23].

Chapter 3

Functional Graphical Models for Partially Separable Multivariate Gaussian Processes

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 chapter. 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 [17]. The dual multivariate and functional aspects of the data make the covariance structure of X quite 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.

As for ordinary multivariate data, the conditional independence properties of X are perhaps of greater interest than 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 (e.g. [29, 30, 31, 32]), in which the graph is dynamic and has nodes corresponding to scalar random variables. In this chapter, 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. Under a Bayesian setting, [33] developed a framework for graphical models on product function spaces, including the extension of Markov laws and appropriate prior distributions. And in a frequentist formulation, [17] 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. On the other hand, [34] developed a non-Gaussian variant, where conditional independence was replaced by a notion of so-called additive conditional independence.

The methodology proposed in this chapter is within the setting of multivariate Gaussian processes as in [17], 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 [33, 17] 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 [17] 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.

3.1 Preliminaries

3.1.1 Gaussian Graphical Model

Consider a p-variate random variable $\theta = (\theta_1, \ldots, \theta_p)^T$, $p > 2$. 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 [35] 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 $\Omega_{jk} \neq 0$ [35]. 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.

One first approach for estimating Ω is maximum likelihood estimation. The multivariate normal distribution assumption on θ yields a Gaussian log likelihood (up to constants) given by:

$$\log(|\Omega|) - \operatorname{tr}(S\Omega)$$

with S the sample covariance matrix. For $p < n$, S is nonsingular and the log-likelihood can be maximized with $\hat{\Omega} = S^{-1}$. In order to add sparsity to $\hat{\Omega}$ and/or handle higher dimensional cases like $p \geq n$, $\hat{\Omega}$ needs to be regularized. The graphical lasso in [2] addresses these two limitations by adding a penalty term to the log likelihood function. Thus, the penalized likelihood problem becomes

$$\max_{\Omega} \log(|\Omega|) - \operatorname{tr}(S\Omega) - \lambda||\Omega||_1$$

where $\Omega \in \mathbb{R}^{p \times p}$ is symmetric positive definite, $\lambda \geq 0$ is a penalty parameter and $\|\Omega\|_1 = \sum_{i \neq j}^{p} |\Omega_{ij}|$. This penalty is similar to the standard lasso penalty and induces zeros in the off-diagonal entries of $\hat{\Omega}$.

3.1.2 Functional Gaussian Graphical Models

Using the notation for functional data adopted in Chapter 1, this section introduces graphical models for functional data. Consider 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 $E\left(\|X\|_p^2\right) < \infty$, and let $\mathcal{G}$ be the covariance operator of X. 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

$$C_{jk}(s,t) = \mathrm{cov}\{X_j(s), X_k(t) \mid X_{-(j,k)}\} \quad (j, k \in V, j \neq k), \tag{3.1}$$

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 [36] 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

$$X_j(t) = \sum_{l=1}^{\infty} \xi_{jl}\phi_{jl}(t), \quad \xi_{jl} = \int_0^1 X_j(t)\phi_{jl}(t)\mathrm{d}t. \tag{3.2}$$

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 [33] and [17]. In [33] the authors considered a rigorous notion of conditional independence for functional data, and proposed a family of priors for the covariance operator $\mathcal{G}$. On the other hand, in [17] the expansion in (2.2) is truncated at L terms using the functional principal component basis [10], and set $\xi_j = (\xi_{j1}, \ldots, \xi_{jL})^T$ $(j \in V)$. [17] then defined a $pL \times pL$ covariance matrix Γ for the concatenated vector $(\xi_1^T, \ldots, \xi_p^T)^T$, with $\Gamma = (\Gamma_{jk})_{j,k=1}^p, (\Gamma_{jk})_{lm} = \text{cov}(\xi_{jl}, \xi_{km}),$ $(l, m = 1, \ldots, L)$. 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 [17] constitutes an intuitive approach to functional graphical model estimation, but encounters some difficulties that are addressed in this chapter. 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.

3.2 Partial Separability and Functional Gaussian Graphical Models

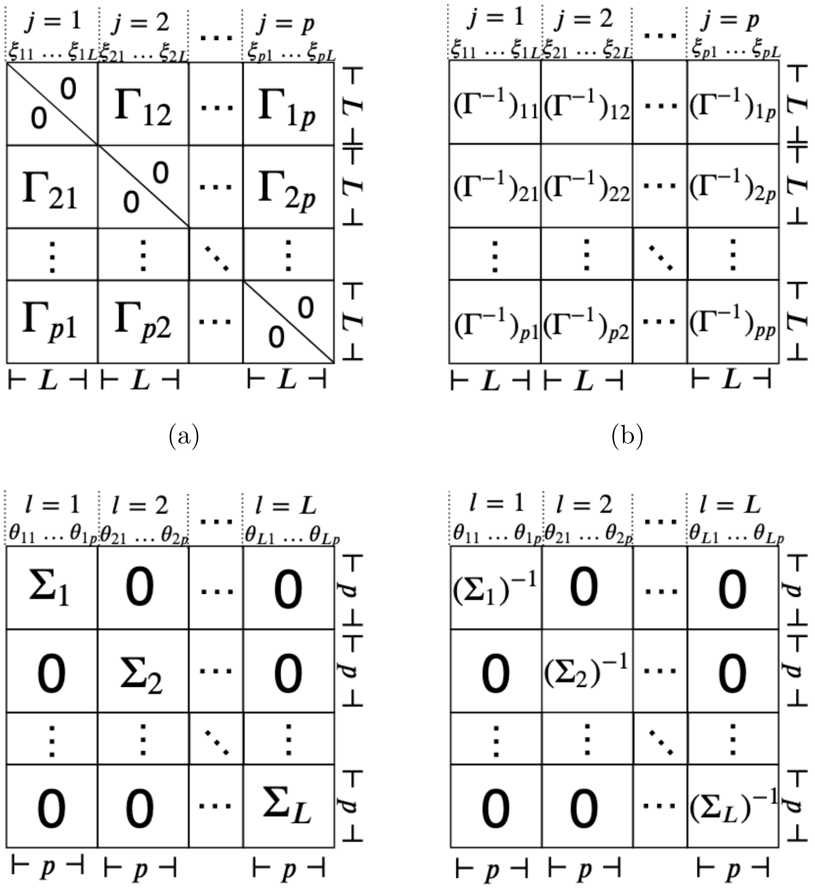
This section is based on the partial separability assumption for multivariate functional data. A detailed discussion and definitions of this assumption can be found in Chapter 1 of this work.

3.2.1 Consequences for Functional Gaussian Graphical Models

As a starting point for this section, consider the Karhunen-Loeve expansion for a partially separable Gaussian process X as introduced in Theorem 2.3.1 point 4 of Chapter 1:

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

where Σ_l is the covariance of θ_l. As will be seen in Section 3.2.1, the matrices Σ_l in point 2 of Theorem 2.3.1 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 $\mathrm{tr}\left(\Sigma_l\right)$. Consequently, one obtains the vector Karhunen-Loève type expansion in (2.3). If one truncates (2.3) at L components, the covariance matrix of the concatenated vector $(\theta_1^T, \ldots, \theta_L^T)^T$ is block diagonal, with the $p \times p$ matrices $\Sigma_l = \mathrm{var}(\theta_l)$ constituting the diagonal blocks. Figure 3.1 visualizes this covariance structure in comparison with that of [17], along with comparisons of the inverse covariance structure. The latter comparison is the more striking and relevant one, since the model of [17] possesses a potentially full inverse covariance structure, whereas that under partial separability remains block diagonal. As a consequence, the model of [17] has $\mathcal{O}(L^2p^2)$ free parameters, while the corresponding model under partial separability has only $\mathcal{O}(Lp^2)$ free parameters.


$j=1$ $j=2$ $\cdots$ $j=p$
$\xi_{11} \dots \xi_{1L}$ $\xi_{21} \dots \xi_{2L}$ $\xi_{p1} \dots \xi_{pL}$
0 0 Γ_{12} $\cdots$ Γ_{1p}
Γ_{21} 0 0 $\cdots$ Γ_{2p}
Γ_{p1} Γ_{p2} $\cdots$ 0 0
L L L
$(\Gamma^{-1})_{11}$ $(\Gamma^{-1})_{12}$ $\cdots$ $(\Gamma^{-1})_{1p}$
$(\Gamma^{-1})_{21}$ $(\Gamma^{-1})_{22}$ $\cdots$ $(\Gamma^{-1})_{2p}$
$(\Gamma^{-1})_{p1}$ $(\Gamma^{-1})_{p2}$ $\cdots$ $(\Gamma^{-1})_{pp}$
$l=1$ $l=2$ $\cdots$ $l=L$
$\theta_{11} \dots \theta_{1p}$ $\theta_{21} \dots \theta_{2p}$ $\theta_{L1} \dots \theta_{Lp}$
Σ_1 0 $\cdots$ 0
0 Σ_2 $\cdots$ 0
0 0 $\cdots$ Σ_L
p p p
$(\Sigma_1)^{-1}$ 0 $\cdots$ 0
0 $(\Sigma_2)^{-1}$ $\cdots$ 0
0 0 $\cdots$ $(\Sigma_L)^{-1}$

(a) (b)

(c) (d)

Figure 3.1: 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 $(\xi_1^T, \dots, \xi_p^T)^T$ in (2.2). (c) and (d): block diagonal covariance and precision matrices, respectively, of coefficients $(\theta_1^T, \dots, \theta_L^T)^T$ under partial separability in (2.3).

Assume that $\mathcal{G}$ is partially separable according to Definition 2.3.1, so that the partially separable Karhunen-Loève expansion in (2.3) 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 2.3.1. Recall that, in order to define a coherent functional Gaussian graphical model, one needs that the conditional covariance functions C_{jk} in (3.1) between component functions X_j and X_k be well-defined. The expansion in

(2.3) 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

$$\tilde{\sigma}_{ljk} = \sigma_{ljk} - \mathrm{cov}\{\theta_{lj}, \theta_{l,-(j,k)}\}\mathrm{var}\{\theta_{l,-(j,k)}\}^{-1}\mathrm{cov}\{\theta_{l,-(j,k)}, \theta_{lk}\}. \tag{3.3}$$

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 3.2.1. *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*

$$C_{jk}(s,t) = \sum_{l=1}^{\infty} \tilde{\sigma}_{ljk}\varphi_l(s)\varphi_l(t) \quad (j, k \in V,\ j \neq k,\ s, t \in [0,1]). \tag{3.4}$$

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 3. *Under the setting of Theorem 3.2.1, the functional graph edge set E is related to the sequence of edge sets E_l by $E = \bigcup_{l=1}^\infty E_l$.*

This result establishes 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 2.3.2 that the eigenbasis of $\mathcal{H}$ is optimal in a sense independent of partial separability, so that the vectors $\theta_l = (\langle X_1, \varphi_l\rangle, \ldots, \langle X_p, \varphi_l\rangle)^T$ 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 A.6 of the Appendix gives sufficient conditions, based on a Markov-type property and a edge coherence assumption, under which the equivalence holds.

3.2.2 Violations of Partial Separability: Consequences for the True Edge Set

An important question is how different violations of partial separability affect the true edge sets. In general, even for the Gaussian case, there is no straightforward formula, for connecting the partially separable graphical model with the true one if partial separability does not hold.

For simplicity, consider a generic example. Let X have three components with each one lying on a common two-dimensional space with probability one, and let $\mathcal{H} = 3^{-1}(\mathcal{G}_{11} + \mathcal{G}_{22} + \mathcal{G}_{33})$ have eigenfunctions (φ_1, φ_2), where $\mathcal{G}_{jj}$ is the covariance operator of X_j. Then one has $X_j = \theta_{1j}\varphi_1 + \theta_{2j}\varphi_2$ for $j = 1, 2, 3$. If X is Gaussian, the conditional dependence is completely determined by the block covariance matrix $\Sigma = \{\Sigma_{ll'}\}_{l,l'=1}^{2}$, $\Sigma_{ll'} = \{\sigma_{ll'jk}\}_{j,k=1}^{3}$, where $\Sigma_{ll'} = 0$ for $l \neq l'$ if and only if X is partially separable. Thus, let the partially separable edge set be $E^* = E_1 \cup E_2$, where $(j,k) \in E_l$ if and only if $(\Sigma_{ll}^{-1})_{jk} \neq 0$. On the other hand, the true edge set E has $(j,k) \in E$ if and only if the (j,k)-th element in at least one of the blocks in Σ^{-1} is nonzero.

Consider three possible values for Σ, given by

$$\Sigma_1 = \left[\begin{array}{ccc:ccc} 1 & a & 0 & c_1 & 0 & 0 \\ a & 1 & 0 & 0 & c_2 & 0 \\ 0 & 0 & 1 & 0 & 0 & c_3 \\ \hdashline c_1 & 0 & 0 & 1 & 0 & 0 \\ 0 & c_2 & 0 & 0 & 1 & a \\ 0 & 0 & c_3 & 0 & a & 1 \end{array}\right] \qquad \Sigma_2 = \left[\begin{array}{ccc:ccc} 1 & a & 0 & c_1 & 0 & 0 \\ a & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & c_3 \\ \hdashline c_1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & a \\ 0 & 0 & c_3 & 0 & a & 1 \end{array}\right]$$

$$\Sigma_3 = \left[\begin{array}{ccc:ccc} 1 & a & b & 0 & 0 & 0 \\ a & 1 & a & 0 & 0 & c \\ b & a & 1 & 0 & c & 0 \\ \hdashline 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & c & 0 & 1 & 0 \\ 0 & c & 0 & 0 & 0 & 1 \end{array}\right].$$

It can be verified that the edge sets in the examples satisfy $E^* \subsetneqq E$ (Σ_1 with nonzero a, c_i s.t. $|a| + |c_i| < 1$ for $i = 1, \ldots, 3$), $E = E^*$ (Σ_2 with same conditions as Σ_1), and $E \subsetneqq E^*$ (Σ_3 with nonzero a, b, c and $a = b = (1 - c^2)$). Hence, different violations of partial separability can lead to different types of discrepancies between the partially separable and true edge sets, or none at all. More details regarding the formulas for the precision matrices in each case can be found in Section A.9 of the Appendix.

3.3 Graph Estimation and Theory

3.3.1 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 (2.4), 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 3.2.1. 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 2.3.2. 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 2.3.1, $\mathrm{tr}(\Sigma_l) = \lambda_l \downarrow 0$ 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.

First, the estimation procedure is defined 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

$$\hat{\mu}_j = \frac{1}{n}\sum_{i=1}^{n} X_{ij}, \quad \hat{\mathcal{G}}_{jk} = \frac{1}{n}\sum_{i=1}^{n}(X_{ij} - \hat{\mu}_j) \otimes (X_{ik} - \hat{\mu}_k), \tag{3.5}$$

can be used. For practical observational designs, smoothing can be applied to the pooled data to estimate these quantities [37, 38]. 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 $\hat{\varphi}_l$. These quantities produce estimates of $\sigma_{ljk} = \langle \mathcal{G}_{jk}(\varphi_l), \varphi_l\rangle$ by plugin, as

$$s_{ljk} = (S_l)_{jk} = \langle \hat{\mathcal{G}}_{jk}(\hat{\varphi}_l), \hat{\varphi}_l\rangle. \tag{3.6}$$

A group graphical lasso approach [39] 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 estimation targets the first L inverse correlation matrices Ξ_l by

$$(\hat{\Xi}_1, \ldots, \hat{\Xi}_L) = \arg \min_{\Upsilon_l \succ 0, \Upsilon_l = \Upsilon_l^T} \sum_{l=1}^{L}\left\{\mathrm{tr}(\hat{R}_l\Upsilon_l) - \log(|\Upsilon_l|)\right\} + P(\Upsilon_1, \ldots, \Upsilon_L), \tag{3.7}$$

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

$$P(\Upsilon_1, \ldots, \Upsilon_L) = \gamma\left\{\alpha\sum_{l=1}^{L}\sum_{j\neq k}|\upsilon_{ljk}| + (1-\alpha)\sum_{j\neq k}\left(\sum_{l=1}^{L}\upsilon_{ljk}^2\right)^{1/2}\right\}, \quad (\Upsilon_l)_{jk} = \upsilon_{ljk}. \tag{3.8}$$

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 $\hat{\Xi}_{ljk} \neq 0$. The joint graphical lasso was chosen to borrow structural information across multiple bases instead of multiple classes as was done in [39]. 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 3.3.2, 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 (3.7) for out-of-sample data. Specifically, the procedure would select γ and α that minimize the average of (3.7) 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 [17]. This latter approach is implemented in the data example of Section 3.5.

3.3.2 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 proofs for this section can be found in Section A.7 in the Appendix. The approach follows that of [40], 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 [17], the results are derived under the setting of fully observed functional data, so that $\hat{\mu}$ and $\hat{\mathcal{G}}_{jk}$ are as in (3.5). An additional proof for edge selection consistency that is not restrictive on the value of the tuning parameter α can be found in the Appendix in Section A.8. As a preliminary result, we first derive a concentration inequality for the estimated covariances s_{ljk} in (3.6), requiring the following mild assumptions.

Assumption 1. *The eigenvalues λ_l of $\mathcal{H}$ are distinct, and thus strictly decreasing.*

Assumption 2. *There exists $\varsigma^2 > 0$ such that, for all $l \in \mathbb{N}$ and all $j \in V$, 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 3.3.1. *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 $l \in \mathbb{N}$ and $j, k \in V$,*

$$\mathrm{pr}\left(|s_{ljk} - \sigma_{ljk}| \geq \delta\right) \leq C_2 \exp\left(-C_1 \tau_l^{-2} n \delta^2\right). \tag{3.9}$$

Concentrations inequalities such as (3.9) 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 the truncation variable L to diverge with n. Furthermore, in contrast to standard multivariate scenarios, the bound in Theorem 3.3.1 contains the additional factor τ_l^{-2}. Since $\lambda_l \downarrow 0$, τ_l diverges to infinity with l, so that (3.9) reflects the increased difficulty of estimating covariances corresponding to eigenfunctions with smaller eigenvalue gaps.

Remark. A similar result to Theorem 3.3.1 was obtained by [17] 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 $L = n^{\beta_2}$, (3.9) 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 [17]. In addition to establishing the concentration inequality for a general eigenvalue decay rate, our proof is greatly simplified by using the inequality

$$|s_{ljk} - \sigma_{ljk}| \leq 2\tau_l \|\mathcal{G}_{jk}\|_{\mathrm{HS}} \|\hat{\mathcal{H}} - \mathcal{H}\|_{\mathrm{HS}} + \|\hat{\mathcal{G}}_{jk} - \mathcal{G}_{jk}\|_{\mathrm{HS}}, \tag{3.10}$$

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

Remark. The bound in (3.10) utilizes a basic eigenfunction inequality found, for example, in Lemma 4.3 of [41]; see also [42]. However, using expansions instead of geometric inequalities, [43] 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 (3.9) is suboptimal, although improvements along the lines of [43] would require further challenging work in order to establish the required exponential tail bounds.

As the objective (3.7) utilizes the correlations $\hat{r}_{ljk}$, the following corollary is needed.

Corollary 4. *Under the assumptions of Theorem 3.3.1, there exists constants $D_1, D_2, D_3 > 0$ such that, for any $0 < \delta \leq D_3$ and for all $l \in \mathbb{N}$ and $j, k \in V$,*

$$\mathrm{pr}\left(|\hat{r}_{ljk} - r_{ljk}| \geq \delta\right) \leq D_2 \exp\left(-D_1 n m_l^2 \delta^2\right), \quad m_l = \tau_l^{-1}\pi_l,\ \pi_l = \min_{j \in V} \sigma_{ljj}. \tag{3.11}$$

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 [40, 1].

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

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

For fixed l, Assumption 3 was employed by [40] as sufficient for model selection consistency. As Theorem 3.3.2 below implies simultaneous consistency of the first L edge sets, we require the assumption for each l. Weakening of this condition may be possible for graphs of specific structures; see Section 3.1 of [40].

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 Assumption 3 holds, 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, set

$$\begin{aligned}
\mathfrak{a}_L &= D_3 \min_{l=1,\ldots,L} m_l,\\
\mathfrak{b}_L &= \min_{l=1,\ldots,L}\left\{6y_l m_l \max(\kappa_{\Psi_l}^2\kappa_{R_l}^3, \kappa_{\Psi_l}\kappa_{R_l})(m_l^{-1} + 8\epsilon_L^{-1})^2\right\}^{-1},\\
\mathfrak{c}_L &= \min_{l=1,\ldots,L}\xi_{\min,l}\left\{4\kappa_{\Psi_l}(m_l^{-1} + 8\epsilon_L^{-1})\right\}^{-1}.
\end{aligned}$$

Tracking $\mathfrak{a}_L$ and $\mathfrak{b}_L$, including the maximal degrees y_l, allows one to obtain uniform consistency of the matrix estimates $\hat{\Xi}_l$ in (3.7), and to conclude that $\hat{E}_l \subset E_l$ with high probability; see Lemma 2 in Section A.7.1 of the supplementary material. The quantity $\mathfrak{c}_L$ involves the weakest nonzero signal $\xi_{\min,l}$ of each graph, with weaker signals requiring larger sample sizes for recovery. We then require the following for the divergence of the number of basis functions L.

Assumption 4. *$L \to \infty$ as $n \to \infty$, $L \leq np$, and* $\min(\mathfrak{a}_L, \mathfrak{b}_L, \mathfrak{c}_L)\{n/\log(n)\}^{1/2} \to \infty$.

Theorem 3.3.2. *Suppose Assumptions 1–4 hold, where $0 \leq (1-\alpha) \leq \min_{l=1,\ldots,L} \eta_l$, and that, for some $\varrho > 2$, $\gamma = 8\epsilon_L^{-1}\left\{(D_1 n)^{-1}\log\left(D_2 L^{\varrho-1}p^{\varrho}\right)\right\}^{1/2}$. If the sample size n satisfies*

$$n\min(\mathfrak{a}_L, \mathfrak{b}_L, \mathfrak{c}_L)^2 > D_1^{-1}\left\{\log(D_2) + (\varrho - 1)\log(n) + (2\varrho - 1)\log(p)\right\}, \tag{3.12}$$

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

Remark. 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 3.3.2 can lead to edge selection consistency in the functional graphical model. For any fixed p, there exists a finite L_p^* such that $E = \bigcup_{l=1}^{L_p^*} E_l$. If it is possible to choose a sequence L satisfying Assumption 4 and $L \geq L_p^*$ for large n, then one will have $E = \bigcup_{l=1}^{L} \hat{E}_l$ with probability at least $1 - (Lp)^{2-\varrho}$ under the assumptions of Theorem 3.3.2 with this choice of L. This will automatically be the case if p remains bounded as n grows, but can also hold in the high-dimensional setting.

Remark. Under Assumption 4, (3.12) becomes

$$n\min(\mathfrak{a}_L, \mathfrak{b}_L, \mathfrak{c}_L)^2 \gtrsim \varrho\log(p)$$

for large n, with $\gtrsim$ denoting inequality up to a multiplicative constant. Hence, if L grows sufficiently slowly, the conclusion of Theorem 3.3.2 will hold asymptotically as long as $\log(p) = o(n)$.

Remark. 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 (3.12) 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 [40], where the maxima over l reflect the need to satisfy the bound for the edge set E_l that is most difficult to estimate.

3.4 Numerical Experiments

3.4.1 Simulation Settings

The simulations in this section compare the proposed method for partially separable functional Gaussian graphical models, with that of [17]. Throughout this section we denote these methods as FGMParty and FGGM, respectively. Other potentially competing non-functional based approaches are not included since they are clearly outperformed by the latter (see [17]). 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 [44]. A fully detailed description of this step is included in the Section A.10 in 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

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

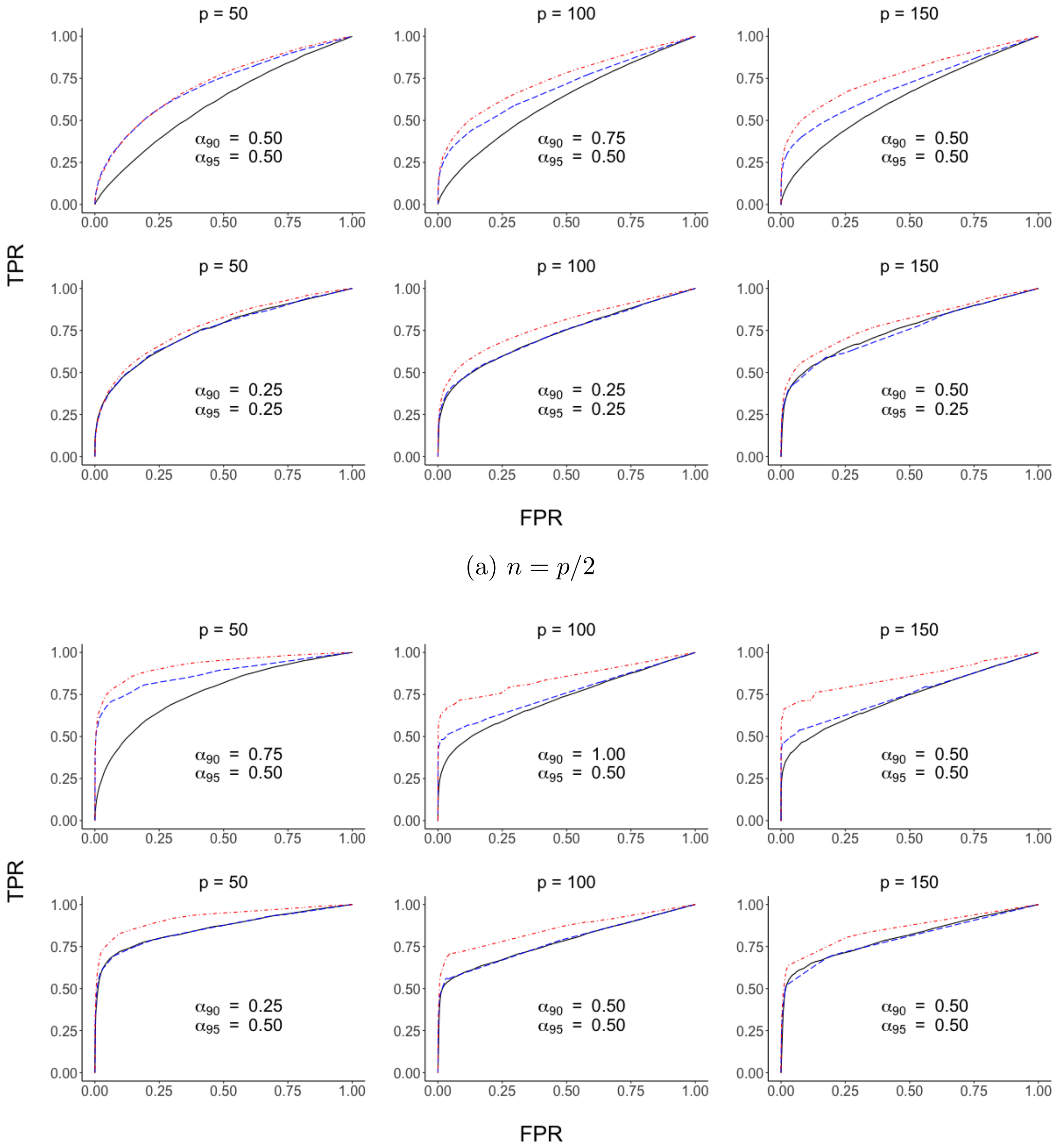
Here, $\sigma_\epsilon^2 = 0.05 \sum_{l=1}^{M} \mathrm{tr}(\Sigma_l)/p$ and $X_{ij}(t_k) = \sum_{l=1}^{M} \theta_{ilj} \varphi_l(t_k)$ according to the partially separable Karhunen-Loève expansion in (2.3). Fourier basis functions $\varphi_1, \ldots, \varphi_M$ evaluated on an equally spaced time grid 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 3.5 below, we set $T = 30$, $M = 20$ and $\pi = 5\%$ for a sparse graph.

Two models are considerd 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 $\text{tr}(\Sigma_l)$ 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 $\Omega_l^* = \Omega_l - \text{diag}(\Omega_l)$. Then, the non-partially separable covariance is computed as $\Sigma_{\text{non-ps}} = \text{diag}(\Sigma_{\text{ps}})^{1/2}\Omega^{-1}\text{diag}(\Sigma_{\text{ps}})^{1/2}$.

3.4.2 Comparison of Results

Comparisons between the proposed method and that of [17], implemented using code provided by the authors, are presented here. Additional comparisons obtained by thresholding correlations are provided in Section 3.4.3. 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 [17] 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.



(a) $n = p/2$

(b) $n = 1.5p$

Figure 3.2: Mean receiver operating characteristic curves for the proposed method (FGMParty) and that of [17] (FGGM). In subfigures (a) and (b), Σ_{ps} (top) and $\Sigma_{\text{non-ps}}$ (bottom) were used for $p = 50, 100, 150$, $\pi = 0.05$ and $\tau = 0$. Curves are coded as FGMParty (- - - -) and FGGM (——) at 90% of variance and FGMParty (—·—) at 95% of variance explained. For FGMParty, the values of α used to compute the curve values are printed in each panel.

Figure 3.2a 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 [45], a variable bandwidth smoother that uses cross-validation to find the best bandwidth. Table 3.1 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 3.2*b* and Table 3.1 summarize results for the large sample case $n = 1.5p$ with similar conclusions. Comparisons under additional simulation settings can be found in the Appendix.

Table 3.1: Mean area under the curve (and standard error) values for Figures 3.2*a* and 3.2*b*

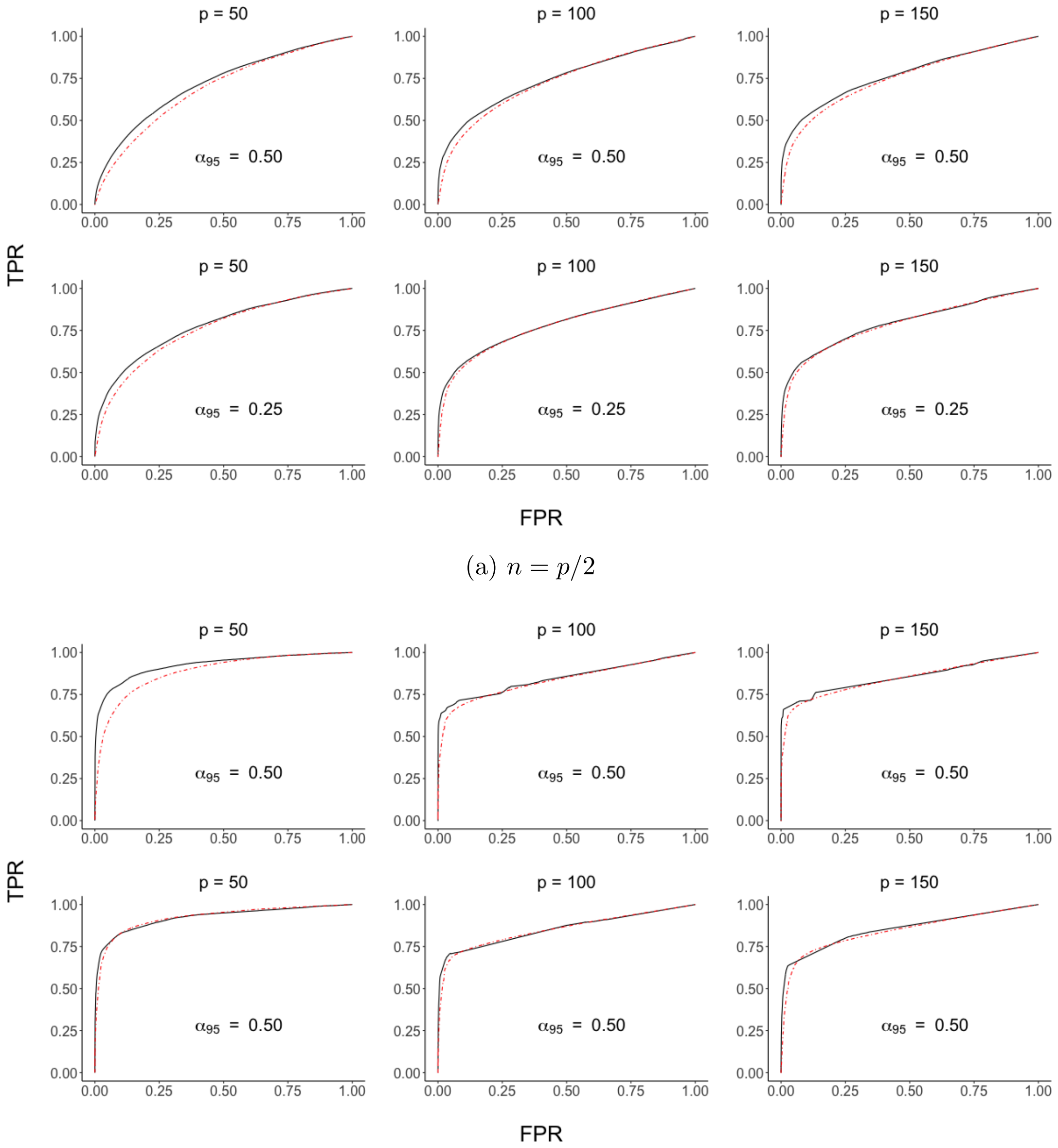
			$\Sigma = \Sigma_{\text{ps}}$			$\Sigma = \Sigma_{\text{non-ps}}$		
			$p = 50$	$p = 100$	$p = 150$	$p = 50$	$p = 100$	$p = 150$
$n = p/2$	AUC	FGGM$_{90\%}$	0.60(0.03)	0.62(0.02)	0.63(0.01)	**0.75(0.03)**	0.72(0.02)	**0.75(0.02)**
		FGMParty$_{90\%}$	**0.71(0.04)**	**0.69(0.02)**	**0.70(0.01)**	**0.75(0.03)**	**0.73(0.02)**	0.74(0.03)
		FGMParty$_{95\%}$	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)**
		FGMParty$_{90\%}$	**0.30(0.05)**	**0.35(0.02)**	**0.37(0.02)**	**0.39(0.04)**	**0.42(0.03)**	0.44(0.04)
		FGMParty$_{95\%}$	0.29(0.05)	0.40(0.03)	0.46(0.03)	0.41(0.05)	0.48(0.03)	0.51(0.03)
$n = 1.5p$	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)**
		FGMParty$_{90\%}$	**0.87(0.03)**	**0.75(0.02)**	**0.75(0.01)**	0.85(0.02)	**0.78(0.02)**	0.79(0.03)
		FGMParty$_{95\%}$	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)**
		FGMParty$_{90\%}$	**0.69(0.04)**	**0.52(0.02)**	**0.52(0.02)**	0.65(0.04)	**0.56(0.04)**	0.55(0.05)
		FGMParty$_{95\%}$	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.

3.4.3 Comparison With An Independence Screening Procedure

This section explores further practical aspect of the estimation method by comparing it with covariance thresholding approach. The main motivation is that partial separability entails a concrete relationship between (some) zeroes in the covariance and corresponding zeroes in the precision matrices. Thus, it is of interest to see how much a very naive graphical model can be improved upon based on the partial separability principle.

The FGMParty method is compared 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 [46]. It assumes partial separability just like FGMParty, 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 3.3a and 3.3b follow the sparse case settings of Section 3.4 with $\pi = 0.05$ and $\tau = 0$. Comparisons under additional simulation settings can be found in the Appendix.



(a) $n = p/2$

(b) $n = 1.5p$

Figure 3.3: Mean receiver operating characteristic curves for the proposed method (FGMParty) and the independence screening procedure (psSCREEN) under Σ_{ps} (top) and $\Sigma_{\text{non-ps}}$ (bottom) for $p = 50, 100, 150$, $\pi = 0.05$ and $\tau = 0$. We see FGMParty (——) and psSCREEN(—·—) both at 95% of variance explained for the sparse case.

Table 3.2: Mean area under the curve (and standard error) values for Figures 3.3*a* and 3.3*b*

			$\Sigma = \Sigma_{\text{ps}}$			$\Sigma = \Sigma_{\text{non-ps}}$		
			$p = 50$	$p = 100$	$p = 150$	$p = 50$	$p = 100$	$p = 150$
$n = p/2$	AUC	FGMParty$_{95\%}$	**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†	FGMParty$_{95\%}$	**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	FGMParty$_{95\%}$	**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†	FGMParty$_{95\%}$	**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.

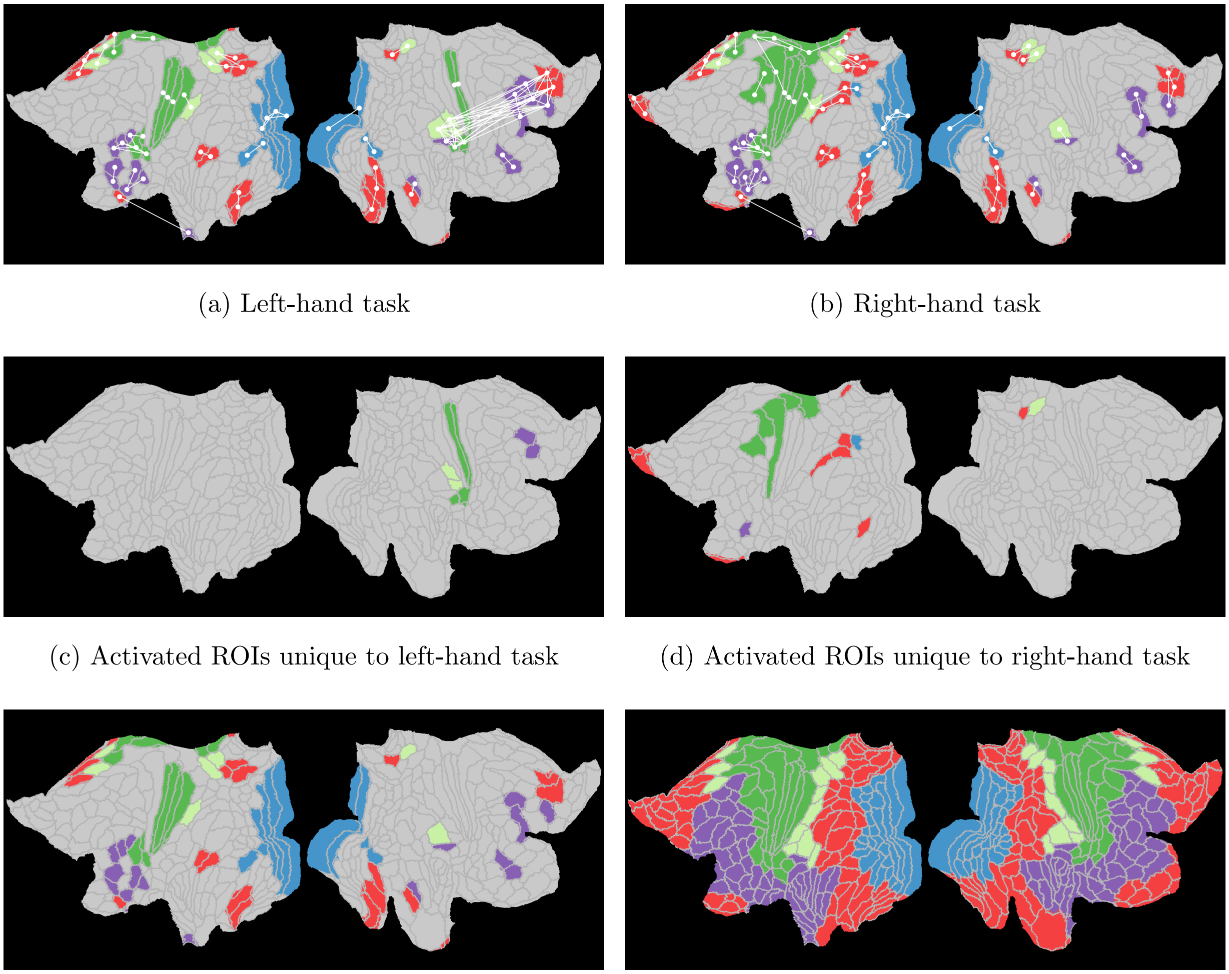
3.5 Application to Functional Brain Connectivity

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 [12]. In particular, we use the motor task fMRI dataset[1] that consists of fMRI scans of individuals performing basic body movements. During each scan, a three-second visual cue 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) [14]. After removing cool down and ramp up observations,

[1] The 1200 Subjects 3T MR imaging data available at `https://db.humanconnectome.org`

$T = 16$ time points of pure movement tasks remained. As seen in Chapter 1, the plausibility of the partial separability assumption for this dataset was discussed in detail with no indications to the contrary. Penalty parameters $\gamma = 0.91$ and $\alpha = 0.95$ were used to estimate very sparse graphs in both tasks.



(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 [14]

Figure 3.4: FGMParty 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 [14]: visual (**blue**), motor (**green**), mixed motor (**light green**), mixed other (**red**) and other (**purple**).

Figure 3.4 shows comparison of activation patterns from left and right-hand task datasets. Figures 3.4*a* and 3.4*b* show the recovered ROI graph on a flat brain map, and only those ROIs with positive degree of connectivity are colored. Figures 3.4*c* and 3.4*d* show connected ROIs that are unique to each task, whereas Figure 3.4*e* 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 (touch and motor sensory inputs) and intraparietal sulcus (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 3.4*c* and 3.4*d* 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 [47].

3.6 Conclusions

The estimation method presented in this chapter is a useful tool to infer graphical models from complex functional data. Indeed, the partial separability assumption reduces the number of free parameters substantially, especially when a large number of functional principal components is needed to explain a significant amount of variation. In the numerical experiments, this also translated in faster convergence times.

An important feature of 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 3.3.1 of this chapter or Theorem 1 of [17] have been derived under the setting of fully observed functional data, so that future work will include similar derivations under more general observation schemes.

In the light of the findings of this work there are several potential extensions to neuroimaging. First, [32] formulates a dynamic graphical model for multivariate functional data. Their estimation method uses truncated univariate Karhunen-Loève expansions as in [17] and analyses electroencephalography (EEG) data. Second, multivariate count processes are ubiquitous in neural spike data where the goal is to understand the electrical activity of the neurons [24]. In this setting, the applicability of partial separability for Poisson graphical models could be studied. And finally, multilevel Gaussian graphical models has become an active area of research with application in genetics [48] and neuroscience [49]. In the particular case of neuroimaging, voxel-level signals are aggregated into ROIs at different levels of coarseness to estimate a joint graphical models. A functional extension of this problem would greatly benefit from a joint partial separability principle to connect graphs at different levels.

Chapter 4

Functional Differential Graphical Models for Partially Separable Multivariate Gaussian Processes

This chapter focuses on estimating the difference between two functional undirected graphical models. The motivating example are different dependencies between functional magnetic resonance imaging (fMRI) signals for a large number of regions across the brain during two motor task experiments.

The methodology proposed in this chapter is within the setting of multivariate Gaussian processes as in [50], and exploits a notion of joint partial separability for multivariate functional data to estimate differences between inverse covariance objects. And included is a novel estimation method that assumes sparsity only on the graphs difference while allowing the individual graphical models to be denser.

Empirical performance of the proposed method is then compared to that of [50] 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 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.

4.1 Preliminaries

4.1.1 Differential Gaussian Graphical Model

Consider a p-variate random variables $\vartheta^X = (\vartheta_1^X, \ldots, \vartheta_p^X)^T$, and $\vartheta^Y = (\vartheta_1^Y, \ldots, \vartheta_p^Y)^T$, $p > 2$. For simplicity, the notation in this chapter is only defined for X although the definitions for Y are analogous. For any distinct indices $j, k = 1, \ldots, p$, let $\vartheta^X_{-(j,k)} \in \mathcal{R}^{p-2}$ denote the subvector of ϑ^X obtained by removing its jth and kth entries. A graphical model [35] for ϑ^X is an undirected graph $G^X = (V, E^X)$, where $V = \{1, \ldots, p\}$ is the node set and $E^X \subset V \times V \setminus \{(j,j) : j \in V\}$ is called the edge set. The edges in E^X encode the presence or absence of conditional independencies amongst the distinct components of ϑ by excluding (j,k) from E_X if and only if $\vartheta_j \perp\!\!\!\perp \vartheta_k \mid \vartheta_{-(j,k)}$. If $\vartheta \sim \mathcal{N}_p(0, \Sigma^X)$, the edges of the resulting Gaussian graphical model correspond to the non-zero entries in the precision matrix $\Omega^X = (\Sigma^X)^{-1}$. More details about this connection can be found in Chapter 3.

It is well-known that the interactions in many types of networks can change under different classes or experimental conditions. In such case, the object of study may be a differential graphical model defined as follows. For the differential matrix $\Delta = \Omega^X - \Omega^Y$, the corresponding edge set E^Δ can be defined as $\{(j,k) \in V \times V : \Delta_{jk} \neq 0\}$. And the differential graphical model is $G^\Delta = (V, E^\Delta)$. However, the edges in E^Δ should not be interpreted as encoding differences in the partial covariances between the two classes. As seen in [35], for any fixed $j, k \in V$ the partial covariance between ϑ_j^X and ϑ_k^X can be defined in terms of the precision matrix entries as $\tilde{\sigma}_{jk}^X = -\Omega_{jk}^X(\Omega_{jj}^X\Omega_{kk}^X - (\Omega_{jk}^X)^2)^{-1}$, and the partial the partial covariance difference is

$$\tilde{\sigma}_{jk}^X - \tilde{\sigma}_{jk}^Y = \frac{-\Delta_{jk}(\Omega_{jk}^X\Omega_{jk}^Y + \Omega_{jj}^X\Omega_{kk}^X) + \Omega_{jk}^X(\Omega_{jj}^X\Omega_{kk}^X - \Omega_{jj}^Y\Omega_{kk}^Y))}{(\Omega_{jj}^X\Omega_{kk}^X - (\Omega_{jk}^X)^2)(\Omega_{jj}^Y\Omega_{kk}^Y - (\Omega_{jk}^Y)^2)}.$$

So, in general, Δ_{jk} equal to zero is not a sufficient condition for $\tilde{\sigma}^X_{jk} = \tilde{\sigma}^Y_{jk}$.

Several methods have been proposed to estimate differential networks. The first group estimates both Ω_X and Ω_Y assuming some common sparsity structure. For example, [39] solves for a penalized joint log-likelihood of the ϑ^X and ϑ^Y using group lasso penalties. Other similar approaches include [51] and [52] where the common structure assumption yields hub nodes in the networks.

Other methods focus on a direct estimation of Δ, without estimating neither Ω^X nor Ω^Y. The literature on this group includes [53], [54] and [55]. In this case, sparsity assumptions are only imposed on Δ whereas Ω^X and Ω^Y are allowed to be dense. The matrix Δ is estimated by minimizing a trace loss function of the form:

$$L(\Delta|\Sigma^X, \Sigma^Y) = \text{tr}\left\{\frac{1}{2}\Sigma^Y\Delta^T\Sigma^X\Delta - \Delta^T(\Sigma^Y - \Sigma^X)\right\}$$

By denoting $\otimes$ as the Kronecker product between two matrices, the second derivative of this function in terms of Δ is

$$\frac{\partial^2 L(\Delta|\Sigma^X, \Sigma^Y)}{\partial\Delta^2} = \Sigma^Y \otimes \Sigma^X$$

which is positive definite if and only if both Σ_X and Σ_Y are positive definite. Thus, $L(\Delta|\Sigma^X, \Sigma^Y)$ is convex and the unique minimizer of this problem can be obtained by imposing a first order condition on its first derivative:

$$\frac{\partial L(\Delta, \Sigma^X, \Sigma^Y)}{\partial\Delta} = \Sigma^X\Delta\Sigma^Y - (\Sigma^Y - \Sigma^X),$$

resulting in $\Delta = (\Sigma^X)^{-1} - (\Sigma^Y)^{-1}$. Similarly, the estimator $\hat{\Delta}$ can be computed by plugin of S^X and S^Y corresponding to the sample estimates of Σ^X and Σ^Y if these sample covariances are nonsingular. For $p < n$, the non-singularity of S^X and S^Y is guaranteed but for high-dimensional cases with $p \geq n$, the estimator of $\hat{\Delta}$ needs regularization. Similar to the graphical lasso [2], this limitation can be handled by adding a penalty term that induces sparsity on the solution. Thus, the penalized trace loss minimization problem targets Δ by

$$\hat{\Delta} = \arg\min_{\Delta} \mathrm{tr}\big\{\frac{1}{2}S^Y \Delta^T S^X \Delta - \Delta^T(S^Y - S^X)\big\} + \lambda \|\Delta\|_1$$

where $\lambda > 0$ is a penalty parameter and $\|\Delta\|_1 = \sum_{i \neq j}^{p} |\Delta_{ij}|$. This penalty is similar to the standard lasso penalty and induces zeros in the off-diagonal entries of $\hat{\Delta}$.

4.1.2 Functional Differential Gaussian Graphical Model

Using the notation for functional data adopted in Chapter 2, this section introduces differential graphical models for functional data. As for differential Gaussian graphical models, differential networks are also of great interest to compare two classes of multivariate functional data in terms of the conditional independencies. Consider two multivariate process X and Y with covariance operators $\mathcal{G}^X$ and $\mathcal{G}^Y$. As discussed in Chapter 3, the covariance operators $\mathcal{G}^X$ and $\mathcal{G}^Y$ are compact and thus not invertible so the conditional independence cannot be defined. This issue is commonly addressed by performing dimensionality reduction. Specifically, one chooses orthonormal functional basis $\{\phi_{jl}^X\}_{l=1}^{\infty}$ and $\{\phi_{jl}^Y\}_{l=1}^{\infty}$ of $L^2[0,1]$ for each j, and expresses each component of X and Y as

$$\begin{aligned} X_j(t) &= \sum_{l=1}^{\infty} \xi_{jl}^X \phi_{jl}^X(t), \qquad \xi_{jl}^X = \int_0^1 X_j(t)\phi_{jl}^X(t)\mathrm{d}t \\ Y_j(t) &= \sum_{l=1}^{\infty} \xi_{jl}^Y \phi_{jl}^Y(t), \qquad \xi_{jl}^Y = \int_0^1 Y_j(t)\phi_{jl}^Y(t)\mathrm{d}t \end{aligned} \tag{4.1}$$

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.

The literature on this topic is in its very early stages, with only one preprint to date [50]. The authors of [50] truncated (4.1) at L terms using the functional principal component basis [10], and set $\xi_j^X = (\xi_{j1}^X, \ldots, \xi_{jL}^X)^T$ $(j \in V)$. Then, they define a define a $pL \times pL$ covariance matrix Γ^X blockwise for the concatenated vector $\xi^X = ((\xi_1^X)^T, \ldots, (\xi_p^X)^T)^T$, as $\Gamma^X = (\Gamma_{jk}^X)_{j,k=1}^p, (\Gamma_{jk}^X)_{lm} = \text{cov}(\xi_{jl}^X, \xi_{km}^X),\ (l, m = 1, \ldots, L)$.

Finally, a functional differential Gaussian graphical model is defined. For $\Delta = (\Gamma^X)^{-1} - (\Gamma^Y)^{-1}$, the differential graph is denoted as $G^\Delta = (V, E^\Delta)$ where the edge set can be recovered through the relation $(j, k) \in E^\Delta$ if and only if the (j, k) block in the matrix Δ is not zero. This matrix is estimated with a Joint Functional Graphical Lasso method, denoted as FuDGE, assuming sparse off-diagonal blocks in order to estimate the edge set. At the moment, this is the only work related to functional differential Gaussian graphical models available and will serve as a comparison benchmark for the simulation studies.

4.2 Partial Separability and Functional Differential Gaussian Graphical Models

This section is based on the joint partial separability assumption for multivariate functional data of two different classes. A detailed discussion and definitions of this assumption can be found in Chapter 2.

4.2.1 Consequences for Functional Differential Gaussian Graphical Models

As a starting point for this section, consider the Karhunen-Loeve expansion for joint partially separable Gaussian processes X and Y as introduced in Corollary 2 point 2 of Chapter 2:

$$X = \sum_{l=1}^{\infty} \vartheta_l^X \psi_l, \quad \vartheta_l^X = (\langle X_1, \psi_l \rangle, \dots, \langle X_p, \psi_l \rangle)^T$$
$$Y = \sum_{l=1}^{\infty} \vartheta_l^Y \psi_l, \quad \vartheta_l^Y = (\langle Y_1, \psi_l \rangle, \dots, \langle Y_p, \psi_l \rangle)^T$$

Assume that $\mathcal{G}^X$ and $\mathcal{G}^Y$ are joint partially separable covariance operators according to Definition 2.4.1, so that the joint partially separable Karhunen-Loève expansion in (2.6) holds. If we further assume that X is Gaussian, then $\vartheta_l^X \sim \mathcal{N}(0, \Sigma_l^X)$, $l \in \mathbb{N}$, are independent, where Σ_l^X is positive definite for each l. Thus, the inverse covariance matrix $\Omega_l^X = (\Sigma_l^X)^{-1}$ is defined by for each l, and so $\Delta_l = \Omega_l^X - \Omega_l^Y$

Under these assumptions and motivated by the findings in Chapter 2, the functional differential Gaussian graphical model defined by [50] corresponds to $(j, k) \in E^{\Delta}$ if and only if $(\Delta_l)_{jk} \neq 0$ for any l. Due to the above result, the edge set E^{Δ} is connected to the sequence of edge sets $\{E_l^{\Delta}\}_{l=1}^{\infty}$, for which $(j, k) \notin E_l^{\Delta}$ if and only if $(\Delta_l)_{jk} = 0$, corresponding to the sequence of differential Gaussian graphical models (V, E_l^{Δ}) for each pair of random scores vectors ϑ_l^X and ϑ_l^Y.

Corollary 5. *Under the setting of Theorem 2.3.1, the functional differential graph edge set E^{Δ} is related to the sequence of edge sets E_l^{Δ} by $E^{\Delta} = \bigcup_{l=1}^{\infty} E_l^{\Delta}$.*

This result established that, under joint partial separability, the problem of functional differential graphical model estimation can be simplified to estimation of a sequence of decoupled differential graphical models.

4.3 Graph Estimation

4.3.1 Joint Trace Loss Estimator

Consider p-variate processes X and Y. Let the components of X has mean functions $\mu_j^X(t) = E\{X_j(t)\}$ and covariance operator $\mathcal{G}^X$. Let $\{\psi_l\}_{l=1}^\infty$ be an orthonormal eigenbasis of $\mathcal{H} = (2p)^{-1}\sum_{j=1}^p \mathcal{G}_{jj}^X + \mathcal{G}_{jj}^Y$, and set $\vartheta_{lj}^X = \langle X_j, \psi_l \rangle$, $\Sigma_l^X = \mathrm{var}(\vartheta_l^X)$, and $\Delta_l = (\Sigma_l^X)^{-1} - (\Sigma_l^Y)^{-1}$. The targets are the edge sets E_l^Δ, where $(j,k) \in E_l^\Delta$ if and only if $(\Delta_l)_{jk} \neq 0$, as motivated by the developments of Section 4.2.1.

Consider a random sample $X_1, \ldots, X_n$, each distributed as X. X is not required to be Gaussian, nor $\mathcal{G}^X$ to be joint 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 $\hat{\mu}_j^X$ and $\hat{\mathcal{G}}_{jk}^X$, $j,k = 1, \ldots, p$, have been computed for each component. As an example, if the X_i are fully observed, cross-sectional estimates

$$\hat{\mu}_j^X = \frac{1}{n}\sum_{i=1}^n X_{ij}, \quad \hat{\mathcal{G}}_{jk}^X = \frac{1}{n}\sum_{i=1}^n (X_{ij} - \hat{\mu}_j^X) \otimes (X_{ik} - \hat{\mu}_k^X), \tag{4.2}$$

can be used. For practical observational designs, smoothing can be applied to the pooled data to estimate these quantities [37, 38]. And analogous procedure for sample $Y_1, \ldots, Y_n$ yields estimates $\hat{\mu}_j^Y$ and $\hat{\mathcal{G}}_{jk}^Y$, $j,k = 1, \ldots, p$. Given such preliminary estimates, the estimate of $\mathcal{H}$ is $\hat{\mathcal{H}} = (2p)^{-1}\sum_{j=1}^p \hat{\mathcal{G}}_{jj}^X + \hat{\mathcal{G}}_{jj}^Y$, leading to empirical eigenfunctions $\hat{\psi}_l$. These quantities produce estimates of $\sigma^X_{ljk} = \langle \mathcal{G}^X_{jk}(\psi_l), \psi_l \rangle$ and $\sigma^Y_{ljk} = \langle \mathcal{G}^Y_{jk}(\psi_l), \psi_l \rangle$ by plugin, as

$$\begin{aligned} s^X_{ljk} &= \left(S^X_l\right)_{jk} = \langle \hat{\mathcal{G}^X}_{jk}(\hat{\psi}_l), \hat{\psi}_l \rangle \\ s^Y_{ljk} &= \left(S^Y_l\right)_{jk} = \langle \hat{\mathcal{G}^Y}_{jk}(\hat{\psi}_l), \hat{\psi}_l \rangle. \end{aligned} \tag{4.3}$$

By Theorem 2.3.1, $\mathrm{tr}(\Sigma^X_l) \downarrow 0$ and $\mathrm{tr}(\Sigma^Y_l) \downarrow 0$ as $l \to \infty$. As a practical consideration, this makes estimators of Σ^X_l and Σ^Y_l progressively more unstable to work with as l increases. To address this issue the penalty approach of [39] is extended to a weighted group lasso estimator of Δ_l. The estimation targets the first L matrices Δ_l by

$$(\hat{\Delta}_1, \ldots, \hat{\Delta}_L) = \arg \min_{\Upsilon_l = \Upsilon_l^T} \sum_{l=1}^{L} \left[\mathrm{tr} \left\{ \frac{1}{2} S^Y_l \Upsilon^T S^X_l \Upsilon - \Upsilon^T (S^Y_l - S^X_l) \right\} \right] + P(\Upsilon_1, \ldots, \Upsilon_L), \tag{4.4}$$

By letting $\upsilon_{ljk} = (\Upsilon_l)_{jk}$, the penalty is

$$P(\Upsilon_1, \ldots, \Upsilon_L) = \gamma \left\{ \alpha \sum_{l=1}^{L} \sum_{j \neq k} w_{1,ljk} |\upsilon_{ljk}| + (1-\alpha) \sum_{j \neq k} \left(\sum_{l=1}^{L} (w_{2,ljk} \upsilon_{lij})^2 \right)^{1/2} \right\} \tag{4.5}$$

where $w_{1,ljk}, w_{2,ljk}$ are the non-negative weighting scheme parameters. 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 $(j, k) \in \hat{E}^\Delta_l$ if and only if $\hat{\Delta}_{ljk} \neq 0$. We refer to the solution of problem (4.4) as the Joint Trace Loss (JTL) estimator.

The JTL estimator can borrow structural information across multiple bases. If $\alpha = 1$, the first penalty will encourage sparsity in each $\hat{\Delta}_l$ and the corresponding edge set $\hat{E}^\Delta_l$, but the overall estimate $\hat{E}^\Delta = \bigcup_{l=1}^{L} \hat{E}^\Delta_l$ may not be sparse. The influence of the second penalty term when $\alpha < 1$ ensures that the overall differential graph estimate is sparse, enhancing interpretation.

In practice, tuning parameters γ and α can be chosen similarly to the case of a single

FGGM. In the data example of Section 4.6 the parameters are chosen to yield a desired level of sparsity.

4.3.2 Two Useful Penalty Weighting Schemes

This subsection introduces two particular choices for the weight parameters of the convex penalty function P in problem 4.4 that lead to useful functional differential graphical model estimates. Denote by $\tilde{\Omega}_l^X$ and $\tilde{\Omega}_l^Y$ the Moore–Penrose pseudoinverses of S_l^X and S_l^Y respectively for $l = 1, \ldots, L$, and let $\tilde{\Delta}_l = \tilde{\Omega}_l^X - \tilde{\Omega}_l^Y$. For weights $w_{1,ljk} = (\|\tilde{\Delta}_l\|_1)^{-1}$ and $w_{2,ljk} = 1$ the *unweighted group* estimates are the solution of problem (4.4), whose penalty becomes:

$$P(\Upsilon_1, \ldots, \Upsilon_L) = \gamma \left\{ \alpha \sum_{l=1}^{L} \frac{1}{\|\tilde{\Delta}_l\|_1} \sum_{j \neq k} |\upsilon_{ljk}| + (1-\alpha) \sum_{j \neq k} \left(\sum_{l=1}^{L} \upsilon_{lij}^2 \right)^{1/2} \right\}. \tag{4.6}$$

The weighting choice for the first term is an adaptive-lasso type penalty [56] to induce sparse solutions. On the other hand, the second term consist of a standard group lasso penalty to borrow information across the different estimates.

In principle the unweighted group estimates may exhibit some difficulties. First of all, the estimators of Σ_l^X and Σ_l^Y are progressively more unstable to work with as l increases. Second, the norm of Ω_l increases with l which could make the second term summand in (4.6) highly dependent on the entries of Δ_L. For this reason, a second estimator is considered as follows.

Let $\tilde{d}_{lj} = (\tilde{\Omega}_{ljj}^X \tilde{\Omega}_{ljj}^Y)^{1/2}$ for $l = 1, \ldots, L$ and $j = 1, \ldots, p$. For weights $w_{1,ljk} = (\|\tilde{\Delta}_l\|_1)^{-1}$ and $w_{2,ljk} = (\tilde{d}_{lj}\tilde{d}_{lk})^{-1}$ the *weighted group* estimates are the solution of problem (4.4), whose penalty becomes:

$$P(\Upsilon_1, \ldots, \Upsilon_L) = \gamma \left\{ \alpha \sum_{l=1}^{L} \frac{1}{\|\tilde{\Delta}_l\|_1} \sum_{j \neq k} |\upsilon_{ljk}| + (1-\alpha) \sum_{j \neq k} \left(\sum_{l=1}^{L} \frac{1}{\tilde{d}_{lj}\tilde{d}_{lk}} \upsilon_{ljk}^2 \right)^{1/2} \right\}. \quad (4.7)$$

The weighting choice for the first term is the same as before. The difference lies in the second term consisting of a weighted-group lasso penalty that borrows information more evenly across the different estimates.

To understand the motivation behind the weighted group estimator consider problem (4.4) at the partial correlation scale. In such case the object of interest would be: $\Omega^X_{ljk}(\Omega^X_{ljj}\Omega^X_{lkk})^{-1/2} - \Omega^Y_{ljk}(\Omega^Y_{ljj}\Omega^Y_{lkk})^{-1/2}$. However, as in problem (4.4) one cannot identify Ω^X_{ljk} and Ω^Y_{ljk}, the weighting factor $w_{2,ljk}$ cannot work at the partial correlation scale unless $\Omega^X_{ljj} = \Omega^Y_{ljj}$ for all $j = 1, \ldots, p$. Under this strong assumption one can see that:

$$\frac{\Delta_{ljk}}{(\Omega^X_{ljj}\Omega^X_{lkk}\Omega^Y_{ljj}\Omega^Y_{lkk})^{1/4}} = \frac{\Omega^X_{ljk}}{(\Omega^X_{ljj}\Omega^X_{lkk})^{1/2}} - \frac{\Omega^Y_{ljk}}{(\Omega^Y_{ljj}\Omega^Y_{lkk})^{1/2}}$$

Again, this is indeed a strong assumption on the covariance structure. But as illustrated in Section 4.5 this choice of weight exhibits a good performance, especially in the high-dimensional case.

4.4 Algorithm for the Joint Trace Loss Problem

The optimization problem in (4.4) is convex and can be solved using an alternating directions method of multipliers (ADMM) algorithm. A detailed discussion of the ADMM algorithms and their convergence guarantees can be found in [57]. First of all, problem

(4.4) can be written as

$$\min_{\Upsilon_l = \Upsilon_l^T} \sum_{l=1}^{L} L(\Upsilon_l | S_l^X, S_l^Y) + P(Z_1, \ldots, Z_L). \tag{4.8}$$

subject to the constraint $Z_l = \Upsilon_l$ for auxiliary variables Z_l and $l = 1, \ldots, L$. For simplicity, denote $\{Z\}_{l=1}^L = Z_1, \ldots, Z_L$. Second, the scaled augmented Lagrangian problem (4.8) (see [57]) is:

$$\mathcal{L}_\rho(\{\Upsilon\}_{l=1}^L, \{Z\}_{l=1}^L, \{U\}_{l=1}^L) = \sum_{l=1}^{L} L(\Upsilon_l | S_l^X, S_l^Y) + P(Z) + \frac{\rho}{2} \sum_{l=1}^{L} \|\Upsilon_l - Z_l + U_l\|_F^2 - \frac{\rho}{2} \sum_{l=1}^{L} \|U_l\|_F^2 \tag{4.9}$$

where $\{U\}_{l=1}^L = U_1, \ldots, U_l$ are dual variables and ρ is a penalty parameter of the ADMM algorithm. The goal of the ADMM algorithm is to solve (4.9) so that $\|Z_l - \Upsilon_l\|$ decreases to zero.

```
Algorithm 1: ADMM for JTL in (4.9)
  input : $\{S_l^X\}, \{S_l^Y\}$
  Initialize $\Upsilon_l = 0_{p \times p}$, $Z_l = 0_{p \times p}$, $U_l = 0_{p \times p}$ for $l = 1, \ldots, L$;
  repeat
    a) $\{\Upsilon^{(k)}\}_{l=1}^L = \text{argmin}_{\{\Upsilon\}_{l=1}^L} \mathcal{L}_\rho(\{\Upsilon\}_{l=1}^L, \{Z^{(k-1)}\}_{l=1}^L, \{U^{(k-1)}\}_{l=1}^L)$ ;
    b) $\{Z^{(k)}\}_{l=1}^L = \text{argmin}_{\{Z\}_{l=1}^L} \mathcal{L}_\rho(\{\Upsilon^{(k)}\}_{l=1}^L, \{Z\}_{l=1}^L, \{U^{(k-1)}\}_{l=1}^L)$ ;
    c) $U_l^{(k)} = U_l^{(k-1)} + \rho(\Upsilon_l^{(k)} - Z_l^{(k)})$ for $l = 1, \ldots, L$ ;
  until convergence ;
```

Finally, the $\{Z\}_{l=1}^L$ resulting from Algorithm 1 correspond to the JTL estimates of $\Delta_1, \ldots, \Delta_L$. The convergence criterion is set as

$$\frac{\sum_{l=1}^L \|Z_l^{(k)} - Z_l^{(k-1)}\|_F}{\sum_{l=1}^L \|Z_l^{(k-1)}\|_F + \epsilon} < 10^{-5}$$

with $\epsilon = 10^{-}5$. Notice that the first and second steps of Algorithm 1 correspond to finding dense and regularized estimates, respectively in an iterative way. The ADMM algorithm guarantees that $\|Z_l^{(k)} - \Upsilon_l^{(k)}\|_F \to 0$ as $k \to \infty$ (see [57]). A detailed discussion of steps a) and b) can be found in Sections 4.4.1 and 4.4.2.

4.4.1 Dense Estimation Subproblem of ADMM Algorithm 1

The first step of Algorithm 1 is solved as follows. The scaled augmented Lagrangian in (4.9) reduces to

$$\begin{aligned}\{\Upsilon^{(k)}\}_{l=1}^{L} &= \underset{\{\Upsilon\}_{l=1}^{L}}{\operatorname{argmin}} \mathcal{L}_\rho(\{\Upsilon\}_{l=1}^{L}, \{Z^{(k-1)}\}_{l=1}^{L}, \{U^{(k-1)}\}_{l=1}^{L}) \\ &= \underset{\{\Upsilon_1,\ldots,\Upsilon_L\}}{\operatorname{argmin}} \sum_{l=1}^{L} L(\Upsilon_l | S_l^X, S_l^Y) + \frac{\rho}{2}\sum_{l=1}^{L} \|\Upsilon_l - Z_l^{(k-1)} + U_l^{(k-1)}\|_F^2\end{aligned} \tag{4.10}$$

The additive structure of the problem in (4.10) allows to solve for each Υ_l separately as:

$$\hat{\Upsilon}_l = \underset{\Upsilon_l}{\operatorname{argmin}} L(\Upsilon_l | S_l^X, S_l^Y) + \frac{\rho}{2}\|\Upsilon_l - Z_l^{(k-1)} + U_l^{(k-1)}\|_F^2 \tag{4.11}$$

for $l = 1, \ldots, L$. The problem in (4.11) has closed form solution (see [58]):

$$\hat{\Upsilon}_l = V_l\big[B \circ (V_l^T C_l^{(k)} W_l)\big] W_l^T \tag{4.12}$$

where $V_l D_l^Y V_l^T$ are $W_l D_l^X W_l^T$ are the singular value decompositions of S_l^Y and S_l^X respectively, $C_l^{(k)} = (S_l^Y - S_l^X) - U_l^{(k-1)} + \rho Z_l^{(k-1)}$ for $l = 1, \ldots, L$, $B_{jk} = 1/(D_{lj}^Y D_{lk}^X + \rho)$ for $j, k = 1, \ldots, p$, and $\circ$ denotes the Hadamard matrix product. The solution is not guaranteed to be symmetric, but the final results can be symmetrized as $\hat{\Upsilon}_l = 0.5(\hat{\Upsilon}_l + \hat{\Upsilon}_l^T)$. Finally, we set $\Upsilon_l^{(k)} = \hat{\Upsilon}_l$ for $l = 1, \ldots, L$.

4.4.2 Regularized Estimation Subproblem of ADMM Algorithm 1

In the second step of Algorithm 1 the scaled augmented Lagrangian in (4.9) reduces to:

$$
\begin{aligned}
\{Z^{(k)}\}_{l=1}^{L} &= \underset{\{Z\}_{l=1}^{L}}{\operatorname{argmin}} \mathcal{L}_{\rho}(\{\Upsilon^{(k)}\}_{l=1}^{L}, \{Z\}_{l=1}^{L}, \{U^{(k-1)}\}_{l=1}^{L}) \\
&= \underset{Z_1,\dots,Z_L}{\operatorname{argmin}} \frac{\rho}{2} \sum_{l=1}^{L} \|\Upsilon_l^{(k)} - Z_l + U_l^{(k-1)}\|_F^2 + P(\{Z\}_{l=1}^{L}) \\
&= \underset{Z_1,\dots,Z_L}{\operatorname{argmin}} \frac{\rho}{2} \sum_{l=1}^{L} \|\Upsilon_l^{(k)} - Z_l + U_l^{(k-1)}\|_F^2 + \\
&\gamma \left\{ \alpha \sum_{l=1}^{L} \sum_{j \neq k} w_{1,ljk} |z_{ljk}| + (1-\alpha) \sum_{j \neq k} \left(\sum_{l=1}^{L} (w_{2,ljk} z_{lij})^2 \right)^{1/2} \right\}
\end{aligned}
\tag{4.13}
$$

with $z_{lij} = (Z_l)_{jk}$, $\lambda_1 = \gamma\alpha$ and $\lambda_2 = \gamma(1-\alpha)$. Denote $A_l = \Upsilon_l^{(k)} + U_l^{(k-1)}$ for $l = 1, \dots, L$, $\tilde{\lambda}_1 = \lambda_1/\rho$, $\tilde{\lambda}_2 = \lambda_2/\rho$ and $a_{ljk} = (A_l)_{jk}$. If $w_{2,ljk} = 1$ for $i, j = 1, \dots, p$ then problem (4.13) has closed form solution as shown in [39]:

$$
\hat{z}_{ljk} = s(a_{ljk}, \tilde{\lambda}_1 w_{1,ljk}) \max \left(0, 1 - \tilde{\lambda}_2 \left\{ \sum_{l'=1}^{L} s(a_{l'jk}, \tilde{\lambda}_1 w_{1,l'jk})^2 \right\}^{-1/2} \right) \tag{4.14}
$$

where $s(x, c) = sign(x) \max(0, |x| - c)$. For a general weighting scheme, $\hat{z}_{lij}$ can be found as the solution from the following system of equations:

$$
z_{ljk} = \frac{s(a_{ljk}, \tilde{\lambda}_1 w_{1,ljk})}{1 + \tilde{\lambda}_2 w_{2,ljk}^2 \left\{ \sum_{l'=1}^{L} z_{l'jk}^2 w_{2,l'jk}^2 \right\}^{-1/2}} \tag{4.15}
$$

for $l = 1, \ldots, L$ and $i, j = 1, \ldots, p$. The nonlinear system of equations in (4.15) can be solved very fast using the following starting point:

$$\hat{z}_{ljk}^{\text{start}} = s(a_{ljk}, \tilde{\lambda}_1 w_{1,ljk}) \max\left(0, 1 - \tilde{\lambda}_2 \left\{ \sum_{l'=1}^{L} \frac{s(a_{l'jk}, \tilde{\lambda}_1 w_{1,l'jk})^2}{w_{2,l'jk}^2} \right\}^{-1/2} \right).$$

Finally, the solution matrix $\hat{Z}_l$ has entries $\left(\hat{Z}_l\right)_{jk} = \hat{z}_{ljk}$, and set $Z_l^{(k+1)} = \hat{Z}_l$.

4.5 Numerical Experiments

The simulations in this section compare the proposed method for joint partially separable functional differential Gaussian graphical models, with that of [50]. Throughout this section we denote these methods as DFGMParty and FuDGE, respectively. Other potentially competing approaches are not included since they are clearly outperformed by the latter (see [50]). They include non-functional methods as well as estimation of two separate functional graphical models to estimate Δ.

4.5.1 Simulation Settings for Model 1

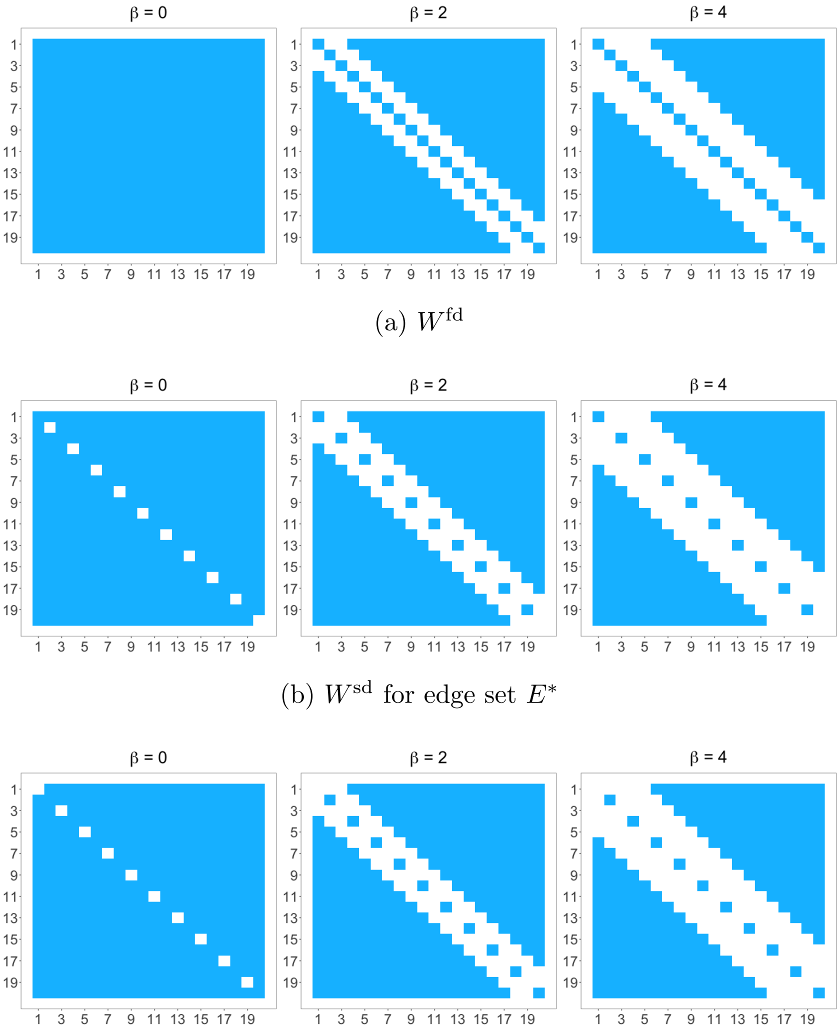
This section is based on the simulation settings of Model 1 in [50] with the purpose of comparing both estimation methods with a data-generating mechanism where partial separability does not hold. First of all, consider a graph $G^X = (V, E^X)$ with $p(p-1)/10$ edges. This graph is generated with a power-law distribution with expected power parameter equals to 2 which exhibits as hub-node structure. For more details please refer to [59].

Second, the graph G^X is used to construct a $Mp \times Mp$ matrix Ω^X as follows. Unless specified otherwise, the rows and columns of Ω^X are sorted in features-first ordering as in [50]. For $(j, k) \in E^X$ set the (j, k)-th block in Ω^X as $\delta_{jk} I_L$ with δ_{jk} a random variable sampled from a uniform distribution on $[-0.5\lambda, -0.2\lambda] \cup [0.2\lambda, 0.5\lambda]$. Next, the diagonal entries of Ω^X are set to 1 and the matrix is averaged with its transpose to guarantee symmetry. We set $p = 60$ and the positive definiteness of Ω^X is guaranteed for $\lambda = 1/4$.

Third, a differential edge set E^Δ is formed based on the the hub nodes from G^X. More specifically, the nodes in G^X are sorted in decreasing order in terms of their degrees and the first two nodes are selected. Next, for each selected hub node $j \in V$ its edges are sorted in decreasing order of edge magnitude measured as $|\delta_{jk}|$ for $(j, k) \in E^X$. And the top 20% of such edges are included in the set E^Δ.

Fourth, for a given $M \times M$ matrix W compute Ω^Y as $\Omega^Y = \Omega^X + \Delta$ where if $(j, k) \in E^\Delta$ then the (j, k)-th block of Δ is set as W and $\mathbf{0}$ otherwise. Notice that the DFGMParty method can only estimate diagonal entries of W whereas FuDGE estimates the entire matrix allowing for deviations of the partial separability principle.

With this in mind, two models for W are considered. The first one, denoted as W^{fd}, is formed by setting $W_{rs} = w$ if either $|r - s| \geq \beta$ or $r = s$ and 0 otherwise for $r, s = 1, \ldots, M$. And the second one, W^{sd}, is the same except for the diagonal entries as follows. The edge set E^Δ is partitioned into two halves denoted as E^* and E^{**} where if $(j, k) \in E^*$ set $W_{ss} = w$ for odd s and 0 otherwise. And if $(j, k) \in E^{**}$ set $W_{ss} = w$ for even s and 0 otherwise. The value of w is sampled from the same distribution as δ_{jk}, and β is a non-negative integer characterizing deviations from partial separability. The smaller the value of β, the higher the deviation. For clarity of exposition Figure 4.1 illustrates the support of matrices W^{fd} and W^{sd}.



(a) W^{fd}

(b) W^{sd} for edge set E^*

(c) W^{sd} for edge set E^{**}

Figure 4.1: Sparsity patterns for matrices W^{fd} and W^{sd}. The colored cell indicates entries with value w. As the the value of β decreases, the deviations from partial separability increases as DFGMParty can only estimate the diagonal entries. Rows and column numbers are indicated in each figure.

Finally, random vectors $\vartheta_i^X \in \mathbb{R}^{Mp}$ are generated from a mean zero multivariate normal distribution with covariance matrix $(\Omega^X)^{-1}$ yielding discrete and noisy functional data

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

Here, $\sigma_\epsilon^2 = 0.5^2$ and $X_{ij}(t_k) = \sum_{l=1}^{M} \vartheta_{ilj}^X \psi_l(t_k)$ according to the joint partially separable Karhunen-Loève expansion in (1). For brevity, only notation for X was defined although the notation for Y is defined analogously. By setting $M = 20$, the basis functions $\psi_1, \dots, \psi_{20}$ are defined as:

$$\psi_l(t) = \begin{cases} cos\left(40\pi\{x - \frac{2l-1}{40}\}\right) + 1, & \text{if } \frac{l-1}{20} \leq x < \frac{l}{20} \\ 0, & \text{otherwise} \end{cases}$$

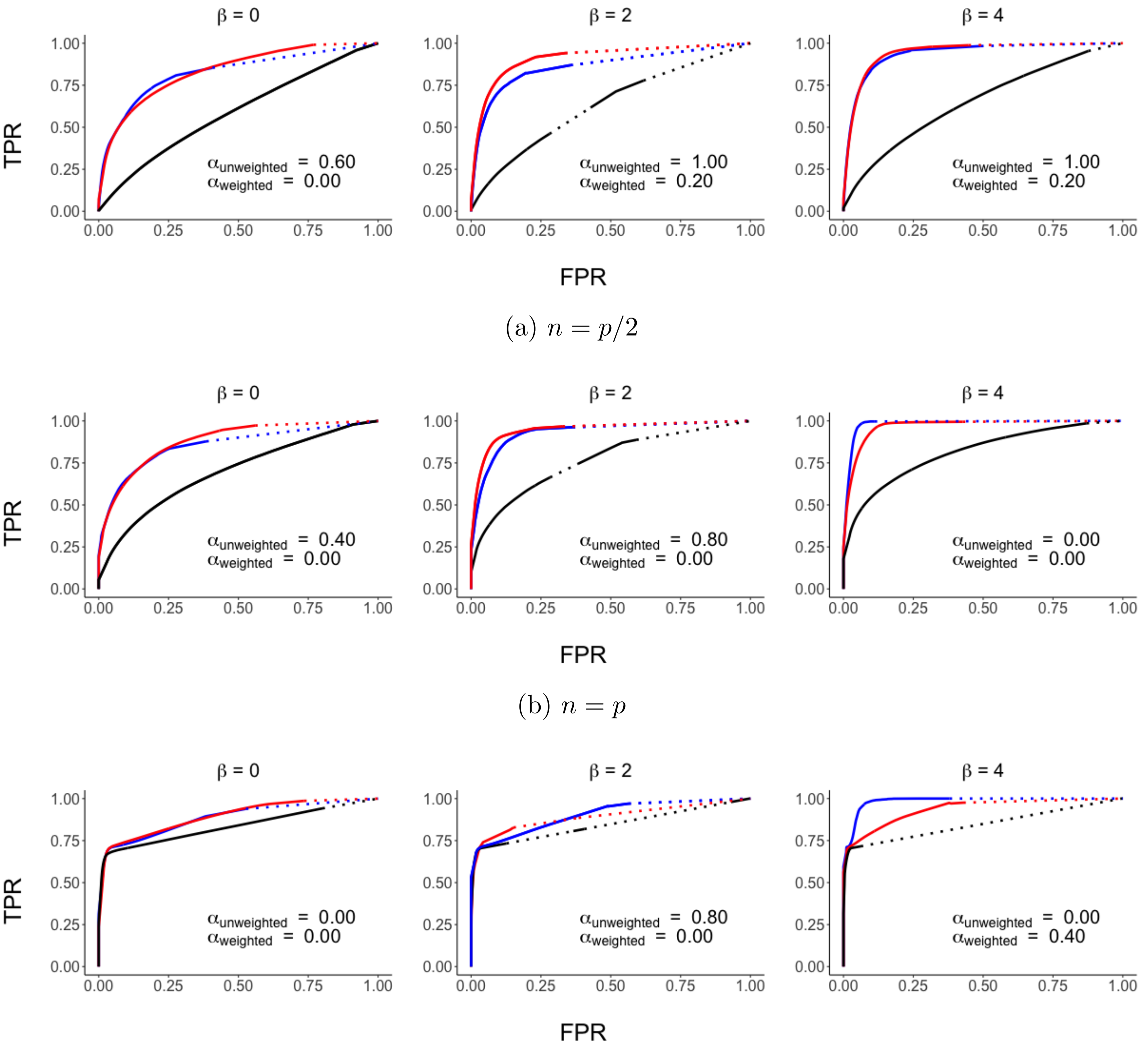
Thus, the orthogonality between these basis functions is given by their disjoint support. These functions are evaluated on an equally spaced time grid of $t_1, \dots, t_T$, with $t_1 = 0$ and $t_T = 1$. In all settings, 100 simulations were conducted with $T = 50$, $M = 20$, $p = 60$, $n \in \{30, 60, 90\}$ and $\beta \in \{0, 2, 4\}$.

4.5.2 Comparison of Results for Model 1

This section compares the proposed method and that of [50]. Figure 4.2 and Table 4.1 show average true/false positive rate curves for model W^{fd} under different deviations of the partial separability assumption for low and high-dimensional cases. The proposed method exhibits uniformly higher true positive rates across the full range of false positive rates. In addition, the weighted group estimator tends to outperform the unweighted group estimator in the high-dimensional case with $n = p/2$, whereas the reverse holds for low-dimensional cases with $n \geq p$. Moreover, in the high-dimensional case higher values of β correspond to higher levels of sparsity in the best performing DFGMParty estimators for both weighting schemes.

Finally notice that as β increases, deviations from partial separability in W decrease, and both methods FuDGE and DFGMParty exhibit higher area-under-the-curve estimates. This is an novel result for FuDGE that was not explored in [50] since the authors set the diagonal of W to zero.

Similar conclusions can be obtained from Figure 4.3 which illustrates ROC curves for model W^{sd} under different deviations of the partial separability assumption for low and high-dimensional cases. However, the performance of the methods in terms of estimated area under the curve increases for DFGMParty and decreases for FuDGE.

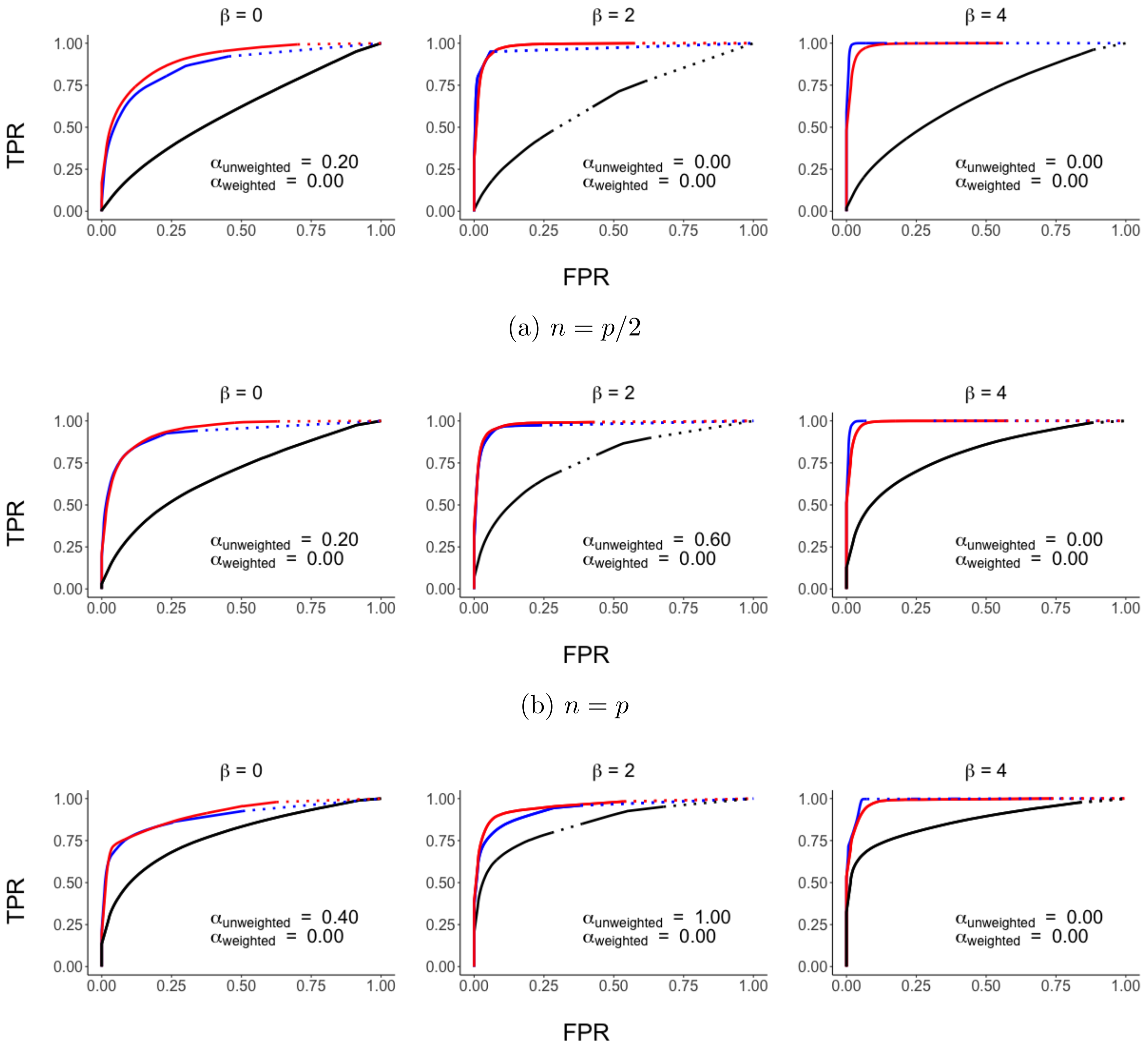


(a) $n = p/2$

(b) $n = p$

(c) $n = 1.5p$

Figure 4.2: Mean receiver operating characteristic curves for the proposed method (DFGMParty) and that of [50] (FuDGE). For $p = 60$, $W = W^{\text{fd}}$ and $\beta \in \{0, 2, 4\}$, subfigures (a), (b) and (c) correspond to values of $n \in \{30, 60, 90\}$ respectively. Curves are coded as unweighted group DFGMParty (——), weighted group DFGMParty (——) and FuDGE (——) at 95% of variance explained. In each curve adjacent points with FPR difference less or equal than 0.10 are interpolated with a solid line. Otherwise, a dashed line is used. For DFGMParty, the values of α used to compute the curve values are printed in each panel.



(a) $n = p/2$

(b) $n = p$

(c) $n = 1.5p$

Figure 4.3: Mean receiver operating characteristic curves for the proposed method (DFGMParty) and that of [50] (FuDGE). For $p = 60$, $W = W^{\text{sd}}$ and $\beta \in \{0, 2, 4\}$, subfigures (a), (b) and (c) correspond to values of $n \in \{30, 60, 90\}$ respectively. Curves are coded as unweighted group DFGMParty (——), weighted group DFGMParty (——) and FuDGE (——) at 95% of variance explained. In each curve adjacent points with FPR difference less or equal than 0.10 are interpolated with a solid line. Otherwise, a dashed line is used. For DFGMParty, the values of α used to compute the curve values are printed in each panel.

Table 4.1: Mean area under the curve (and standard error) values for Figures 4.2 and 4.3.

			W^{fd}			W^{sd}		
			$\beta = 0$	$\beta = 2$	$\beta = 4$	$\beta = 0$	$\beta = 2$	$\beta = 4$
$n = p/2$	AUC	FuDGE	0.60(0.09)	0.72(0.08)	0.86(0.03)	0.59(0.09)	0.67(0.11)	0.75(0.13)
		DFGMParty$_{\text{unweighted}}$	0.75(0.11)	0.81(0.08)	0.87(0.04)	0.79(0.11)	0.88(0.10)	0.87(0.04)
		DFGMParty$_{\text{weighted}}$	**0.84(0.09)**	**0.88(0.07)**	**0.90(0.04)**	**0.91(0.07)**	**0.92(0.06)**	**0.91(0.04)**
	AUC15†	FuDGE	0.15(0.09)	0.29(0.12)	**0.71(0.04)**	0.15(0.09)	0.23(0.13)	0.39(0.20)
		DFGMParty$_{\text{unweighted}}$	0.40(0.16)	0.53(0.12)	0.69(0.02)	0.47(0.16)	0.66(0.19)	0.69(0.07)
		DFGMParty$_{\text{weighted}}$	**0.47(0.16)**	**0.57(0.14)**	0.70(0.03)	**0.62(0.16)**	**0.69(0.14)**	**0.71(0.06)**
$n = p$	AUC	FuDGE	0.64(0.09)	0.77(0.07)	0.85(0.02)	0.62(0.09)	0.74(0.11)	0.79(0.15)
		DFGMParty$_{\text{unweighted}}$	0.83(0.09)	0.91(0.06)	0.90(0.04)	0.93(0.15)	0.97(0.03)	0.94(0.03)
		DFGMParty$_{\text{weighted}}$	**0.89(0.10)**	**0.95(0.05)**	**0.92(0.06)**	**0.97(0.04)**	**0.98(0.02)**	**0.95(0.04)**
	AUC15†	FuDGE	0.19(0.10)	0.40(0.11)	0.71(0.02)	0.18(0.10)	0.34(0.15)	0.50(0.24)
		DFGMParty$_{\text{unweighted}}$	0.55(0.13)	0.68(0.11)	0.72(0.02)	0.85(0.30)	0.88(0.07)	0.78(0.06)
		DFGMParty$_{\text{weighted}}$	**0.68(0.11)**	**0.78(0.12)**	**0.75(0.04)**	**0.88(0.08)**	**0.90(0.07)**	**0.84(0.06)**
$n = 1.5p$	AUC	FuDGE	0.67(0.07)	0.83(0.05)	0.85(0.00)	0.65(0.10)	0.79(0.13)	0.82(0.15)
		DFGMParty$_{\text{unweighted}}$	0.93(0.04)	**0.98(0.01)**	**0.98(0.01)**	**1.00(0.00)**	**1.00(0.00)**	**0.99(0.00)**
		DFGMParty$_{\text{weighted}}$	**0.94(0.03)**	0.97(0.01)	0.96(0.01)	0.99(0.01)	0.99(0.01)	0.98(0.01)
	AUC15†	FuDGE	0.22(0.09)	0.49(0.10)	0.72(0.01)	0.19(0.10)	0.42(0.17)	0.58(0.24)
		DFGMParty$_{\text{unweighted}}$	0.69(0.09)	**0.90(0.04)**	**0.88(0.04)**	**0.99(0.01)**	**0.98(0.01)**	**0.94(0.02)**
		DFGMParty$_{\text{weighted}}$	**0.73(0.09)**	0.82(0.08)	0.77(0.03)	0.94(0.04)	0.94(0.04)	0.89(0.05)

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

4.5.3 Simulation Settings for Model 2

An initial conditional independence graph $G^{\Delta} = (V, E^{\Delta})$ 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 $E_1^{\Delta}, \dots, E_M^{\Delta}$ is generated so that $E^{\Delta} = \bigcup_{l=1}^{M} E_l^{\Delta}$. A set of common edges to all edge sets is computed for a given proportion of common edges $\tau \in [0, 1]$.

Next, $p \times p$ dense precision matrices Ω_l^X and Ω_l^Y are generated and set $\Delta_l = \Omega_l^X - \Omega_l^Y$

where the set $\{(i,j):(\Delta_l)_{ij}\neq 0\}$ is the same as the edge set E_l^{Δ}. More specifically, a graph $G^X=(V,E^X)$ is generated according to a power law with parameter $\tilde{\pi}>\pi$ and precision matrices Ω_l^X are obtained by following the steps in Section 3.4 in the Appendix. Then a $p\times p$ matrix Δ_l is computed with entries $(\Delta_l)_{ij}=c\left(\Omega_l^X\right)_{ij}$ if $(i,j)\in E_l^{\Delta}$, and 0 otherwise for $c\in[0,1]$. Finally, set $\Omega_l^Y=\Omega_l^X-\Delta_l$. Thus, the parameter c characterizes the magnitude of the off-diagonal entries in Δ_l. A fully detailed description of this step is included in the Section A.13 in the Appendix. Then, following the steps in 3.4, a graph $G^X=(V,E^X)$ with parameter $\tilde{\pi}>\pi$ and a $p\times p$ precision matrices Ω_l^X are obtained. Then a $p\times p$ matrix Δ_l is computed with entries $(\Delta_l)_{ij}=c\left(\Omega_l^X\right)_{ij}$ if $(i,j)\in E_l^{\Delta}$, and 0 otherwise. Finally, set $\Omega_l^Y=\Omega_l^X-\Delta_l$.

Denote by Σ^X a block diagonal covariance matrix with $p\times p$ diagonal blocks $\Sigma_l^X=a_l(\Omega_l^X)^{-1}$. The decaying factors $a_l=3l^{-1.8}$ guarantee that $\mathrm{tr}(\Sigma_l^X)$ decreases monotonically in l. Then, random vectors $\vartheta_i^X\in\mathbb{R}^{Mp}$ are generated from a mean zero multivariate normal distribution with covariance matrix Σ^X yielding discrete and noisy functional data

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

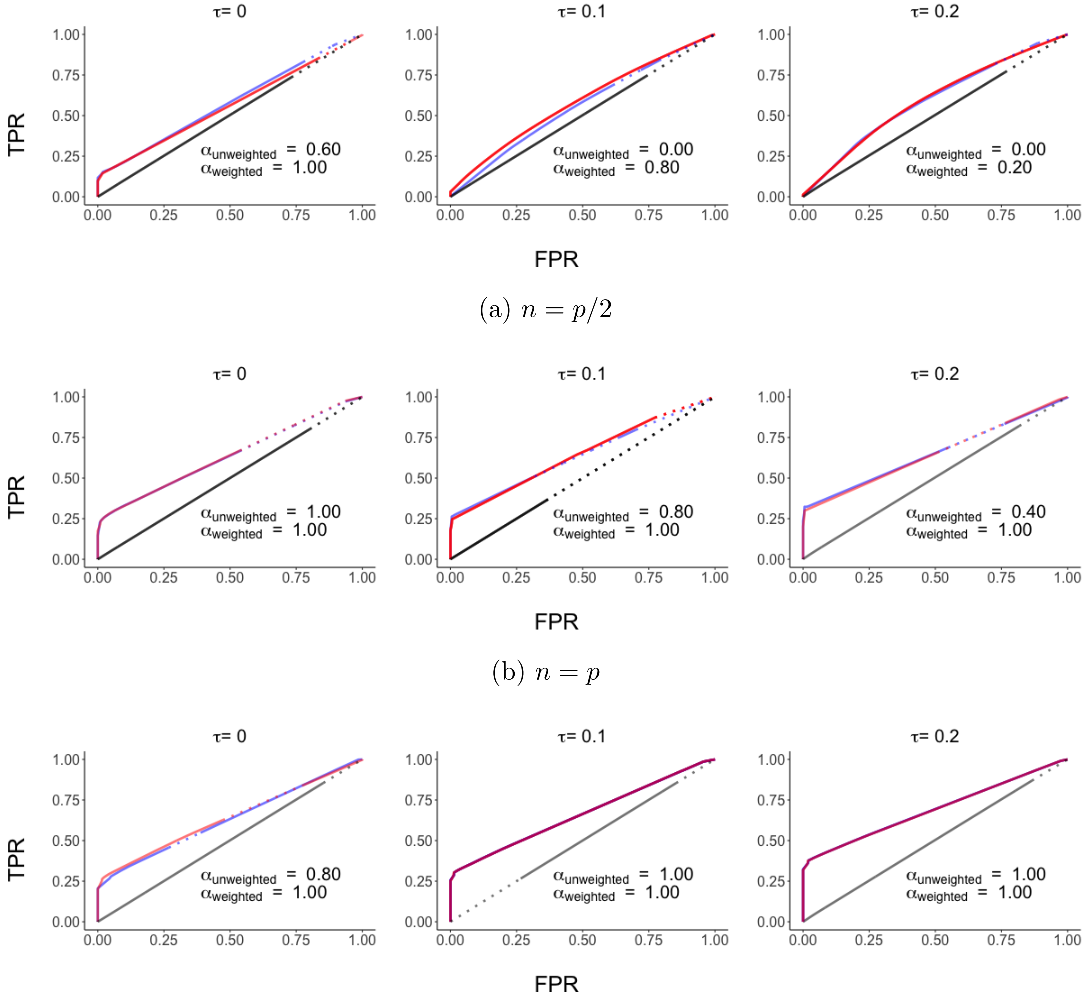
Here, $\sigma_\epsilon^2=0.05$ and $X_{ij}(t_k)=\sum_{l=1}^{M}\vartheta_{ilj}^X\psi_l(t_k)$ according to the joint partially separable Karhunen-Loève expansion in (1). For brevity, only notation for X was defined although the notation for Y is defined analogously. Fourier basis functions $\psi_1,\dots,\psi_M$ evaluated on an equally spaced time grid of $t_1,\dots,t_T$, 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 4.6 below, we set $T=50$, $M=20$, $\tilde{\pi}=0.30$, $c=0.20$ and $\pi=10\%$ for a sparse graph.

4.5.4 Comparison of Results for Model 2

This section compares the proposed method and that of [50] with a data-generating mechanism where partial separability holds. In particular, FuDGE is implemented using code provided by the authors.

As performance metrics, the true and false positive rates of correctly identifying edges in graph G^{Δ} 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 $\pi = 0.10$ and $c = 0.2$. The two methods are compared using L principal components explaining at least 95% of the variance. For all simulations and both methods, this threshold results in the choice of $L = 9$ or $L = 10$ components.

Figure 4.4a and Table 4.2 show average true/false positive rate curves for different proportion of common edges across basis for the high-dimensional case $n = p/2$. The smoothed curves are computed using the `supsmu` R package that implements Super-Smoother [45], a variable bandwidth smoother that uses cross-validation to find the best bandwidth. Table 4.2 shows the mean and standard deviation of area under the curve estimates for various settings. The proposed method exhibits uniformly higher true positive rates across the full range of false positive rates. In particular, the weighted group estimator tends to outperform the unweighted group estimator. Moreover, as the proportion of common edges increases, the best performing value of α in terms of area under the curve decreases. On the other hand, Figures 4.4b and 4.4c summarize results for the large sample case $n \geq p$. In both cases, and contrary to the high-dimensional case, the unweighted estimator exhibits the highest area under the curve. Similar conclusions can be obtained for $c = 0.4$ as seen in Section A.14 in the Appendix.



(a) $n = p/2$

(b) $n = p$

(c) $n = 1.5p$

Figure 4.4: Mean receiver operating characteristic curves for the proposed method (DFGMParty) and that of [50] (FuDGE). For $p = 60$, $\tau \in \{0, 0.1, 0.2\}$, $\pi = 0.10$ and $c = 0.2$ subfigures (a), (b) and (c) correspond to values of $n \in \{30, 60, 90\}$ respectively. Curves are coded as unweighted group DFGMParty (——), weighted group DFGMParty (——) and FuDGE (——) at 95% of variance explained. In each curve adjacent points with FPR difference less or equal than 0.10 are interpolated with a solid line. Otherwise, a dashed line is used. For DFGMParty, the values of α used to compute the curve values are printed in each panel.

Table 4.2: Mean area under the curve (and standard error) values for Figure 4.4.

			τ		
			0	0.1	0.2
$p/2$	AUC	FuDGE	0.51(0.02)	0.50(0.02)	0.51(0.02)
		DFGMParty$_{\text{unweighted}}$	**0.57**(**0.02**)	0.55(0.02)	0.59(0.02)
		DFGMParty$_{\text{weighted}}$	0.56(0.02)	**0.59**(**0.02**)	**0.60**(**0.02**)
	AUC15†	FuDGE	0.08(0.01)	0.08(0.01)	0.08(0.02)
		DFGMParty$_{\text{unweighted}}$	**0.21**(**0.04**)	0.10(0.01)	0.12(0.01)
		DFGMParty$_{\text{weighted}}$	0.20(0.05)	**0.15**(**0.03**)	**0.13**(**0.02**)
$n = p$	AUC	FuDGE	0.50(0.02)	0.50(0.02)	0.51(0.02)
		DFGMParty$_{\text{unweighted}}$	**0.63**(**0.02**)	**0.64**(**0.01**)	**0.66**(**0.02**)
		DFGMParty$_{\text{weighted}}$	**0.63**(**0.02**)	**0.64**(**0.02**)	0.65(0.03)
	AUC15†	FuDGE	0.08(0.02)	0.08(0.02)	0.08(0.02)
		DFGMParty$_{\text{unweighted}}$	0.26(0.06)	0.29(0.06)	**0.31**(**0.08**)
		DFGMParty$_{\text{weighted}}$	**0.28**(**0.05**)	**0.31**(**0.07**)	**0.31**(**0.07**)
$n = 1.5p$	AUC	FuDGE	0.50(0.02)	0.50(0.02)	0.51(0.02)
		DFGMParty$_{\text{unweighted}}$	**0.64**(**0.02**)	**0.67**(**0.02**)	**0.70**(**0.03**)
		DFGMParty$_{\text{weighted}}$	**0.64**(**0.02**)	**0.67**(**0.02**)	**0.70**(**0.03**)
	AUC15†	FuDGE	0.08(0.02)	0.07(0.02)	0.08(0.02)
		DFGMParty$_{\text{unweighted}}$	**0.33**(**0.03**)	**0.36**(**0.03**)	**0.43**(**0.06**)
		DFGMParty$_{\text{weighted}}$	0.32(0.04)	0.36(0.03)	**0.43**(**0.06**)

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

4.6 Application to Functional Brain Connectivity

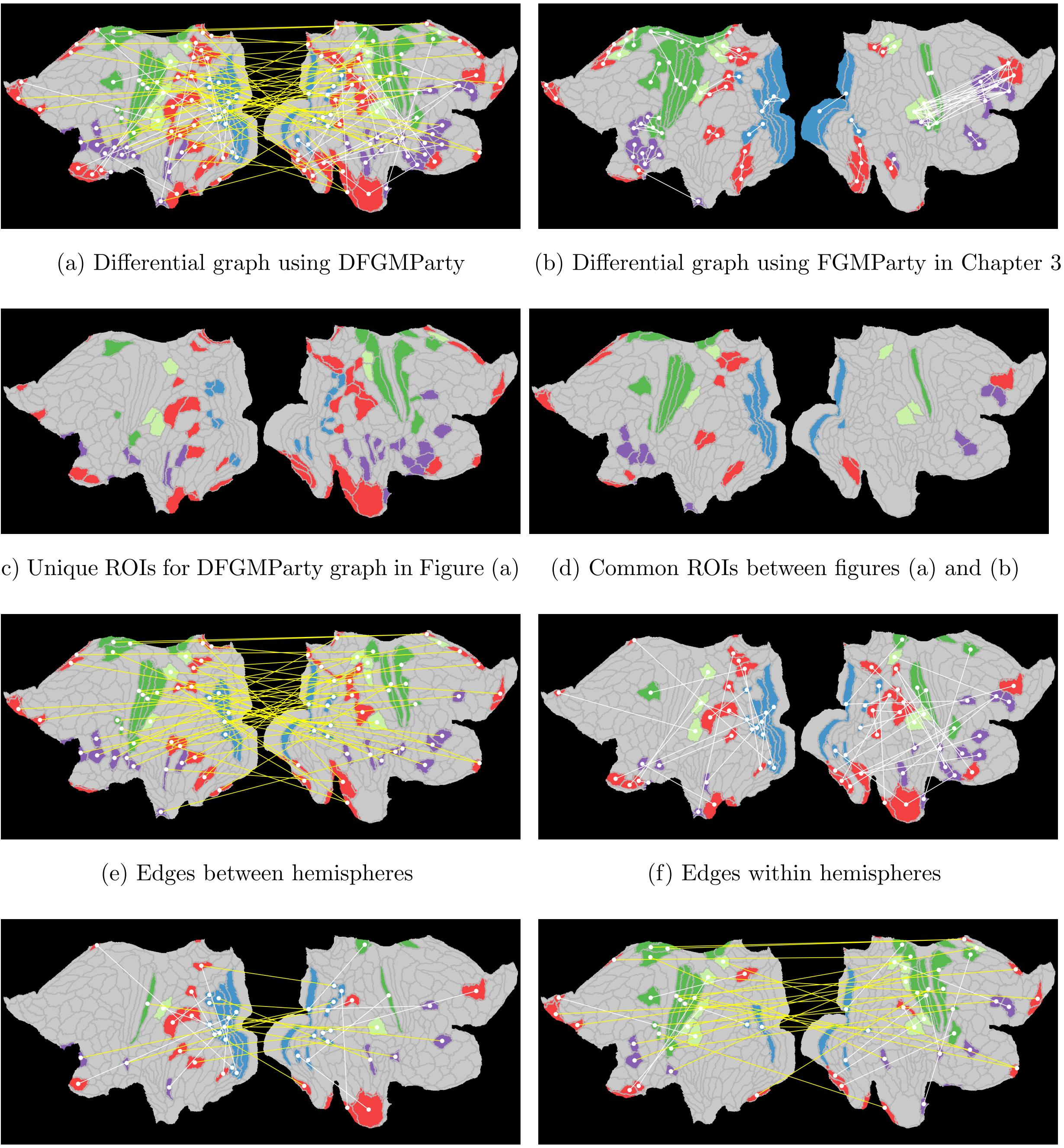
In this section, the proposed method is used to reconstruct the differential brain connectivity structure using functional magnetic resonance imaging (fMRI) data from the Human Connectome Project. Specifically, the data from left- and right-hand finger movements are considered. A full description of the data can be found in Application section in Chapter 3.

Figure 4.5a shows the differential graph estimated using the DFGMParty method. The unweighted group penalty is used as the data consists of $n = 1054$ subjects and $p = 360$ regions of interest (ROIs). The choice of penalty parameters $\gamma = 2.7 \times 10^{-3}$ and $\alpha = 0.95$ were used to estimate a very sparse differential graph exhibiting a similar number of edges to a differential graph obtained from the psFGGM method in Figure 4.5b. More details regarding the application of the psFGGM method to the HCP data can be found in Section 3.5 in Chapter 3.

The graphs in figures 4.5a and 4.5b have several differences. First, the two graphs do not have any common edge. The DFGMParty differential graph exhibits a similar number of edges between hemispheres relative to the within-hemisphere edges as seen in figures 4.5e and 4.5f. This is a contrasting feature as the psFGGM differential graph only presents within-hemishphere edges. Indeed, this result holds even for higher values of the γ penalty parameter in the DFGMParty method.

Second, the DFGMParty differential graph contains almost all the connected ROIs in the psFGGM differential graph as seen in 4.5d. However, the DFGMParty differential graph connects many more ROIs as seen in Figure 4.5c.

Finally, almost all the edges in the psFUDGE differential graph connect either non-adjacent ROIs or ROIs with different functionalities. This is another main difference, as the psFGGM differential graph connects mostly adjacent ROIs mostly with same functionalities. More specifically, Figures 4.5g and 4.5h show the connectivity patterns for the visual and motor ROIs in the DFGMParty differential graph. The former exhibit mostly within-hemisphere edges, and the latter between-hemisphere edges. And there are five motor ROIS connected to five visual ROIs.



(a) Differential graph using DFGMParty

(b) Differential graph using FGMParty in Chapter 3

(c) Unique ROIs for DFGMParty graph in Figure (a)

(d) Common ROIs between figures (a) and (b)

(e) Edges between hemispheres

(f) Edges within hemispheres

(g) Edges for visual ROIs

(h) Edges for motor ROIs

Figure 4.5: DFGMParty 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 [14]: visual (**blue**), motor (**green**), mixed motor (**light green**), mixed other (**red**) and other (**purple**). Edges within and between hemispheres are colored in white and yellow, respectively.

4.7 Conclusions

The estimation method presented in this chapter is a useful tool to infer differential graphical models from complex functional data. Indeed, the joint partial separability assumption reduces the number of free parameters substantially, especially when a large number of functional principal components is needed to explain a significant amount of variation. This also translated in faster convergence times. Moreover, the empirical findings in the application section highlights the importance of direct estimation methods for differential graphs rather than computing differences on separate estimators as the resulting graph can change substantially.

The proposed method for functional graphical model estimation is equally applicable to dense or sparse functional data, observed with or without noise. Moreover, the JTL estimator provides a very efficient estimation method that can work with high dimensional functional data.

In the light of the findings of this work there are several potential extensions. First of all, the JTL estimator could be used in other problems. For instance, multi-class extensions of both the differential network estimator of [53] and the sparse quadratic discriminant analysis of [58] would benefit from the JTL estimator. Second, joint partial separability provide a useful framework to extend the differential latent variable graphical models of [60] to the functional setting.

Appendix A

Appendix

A.1 Computation of Karhunen-Loève Expansion

This section illustrates the computation of the eigenvalues and eigenfunctions of the covariance kernel function $G_{jj}(s,t) = \text{cov}\{X_j(s), X_j(t)\}$ to obtain the univariate Karhunen-Loève expansion of $X_j(j = 1, \dots, p)$ in equation (2.2). More details can be found in Chapter 8 of [61].

Consider functional data $\{X_{ij}(t) \in \mathbb{R}^p : t \in \tau\}$ where subjects are indexed by $i = 1, \dots, n$, ROIs by $j = 1, \dots, p$, and measurements are taken on a grid of equally spaced time points $\tau = \{t_1 = 0, \dots, t_K = 1\}$. Denote $x_{ijk} = X_{ij}(t_k)$, and consider noisy measurements $y_{ijk} = x_{ijk} + \epsilon_{ijk}$ where ϵ_{ijk} are i.i.d. with zero mean and variance σ_ϵ.

First of all, the covariance kernel function is estimated. Let the sample mean function be computed as $\hat{\mu}_j(t_k) = n^{-1}\sum_{i=1}^n y_{ijk}$ and let the $n \times K$ matrix $\tilde{\mathbf{Y}}$ with entries $\left(\tilde{\mathbf{Y}}\right)_{jk} = y_{ijk} - \hat{\mu}_j(t_k)$. Then, the sample covariance matrix is computed as $\mathbf{G}_j = n^{-1}\tilde{\mathbf{Y}}\tilde{\mathbf{Y}}^T$. Second, the eigenfunctions of $\mathbf{G}_j$ are computed. By denoting the singular value decomposition of $\tilde{\mathbf{Y}}$ as $\mathbf{UDW}^T$ then $\mathbf{G}_j = n^{-1}\mathbf{WD}^2\mathbf{W}^T$. Thus, the eigenfunctions $\{\phi_{jl}\}_{l\geq 1}$ are estimated by interpolating the k-dimensional column vector $n^{-1/2}\mathbf{W}_{\cdot l}$ with $\hat{\phi}_{jl}(t_k) = n^{-1/2}\mathbf{W}_{kl}$. And finally, the random scores estimators $\hat{\xi}_{ij}$ are computed as the inner product between

$\hat{\phi}_{jl}$ and X_j. This inner product equals $\int_0^1 \hat{\phi}_{jl}(s)X_j(s)ds$ which is estimated by trapezoidal rule.

A.2 Proofs for Section 2.3

Proof of Theorem 2.3.1.

($1 \Leftrightarrow 2$) 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 $\Sigma_l = \sum_{j=1}^p \lambda_{jl}^{\mathcal{G}} e_{lj} e_{lj}^T$. Since $(e_{lj}\varphi_l) \otimes_p (e_{lj}\varphi_l) = (e_{lj}e_{lj}^T)\varphi_l \otimes \varphi_l$,

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

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

$$\begin{aligned} \mathcal{G}(e_{lj}\varphi_l) &= \sum_{l'=1}^{\infty} \varphi_{l'} \otimes \varphi_{l'} \left\{ \Sigma_{l'} \left(e_{lj}\varphi_l \right) \right\} = \sum_{l'=1}^{\infty} \left(\Sigma_{l'} e_{lj} \right) \varphi_{l'} \otimes \varphi_{l'}(\varphi_l) \\ &= (\Sigma_l e_{lj})\varphi_l = \lambda_j^{\Sigma_l} e_{lj}\varphi_l, \end{aligned}$$

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

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

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

so 3 holds. If 3 holds, then define $\sigma_{ljk} = \operatorname{cov}(\langle X_j, \varphi_l \rangle, \langle X_k, \varphi_l \rangle)$ 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 [10], and since $\sum_{j=1}^p e_{lj}e_{lj}^T$ is the identity matrix,

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

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 eigenbasis for Σ_l. Then

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

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

Proof of Theorem 2.3.2.

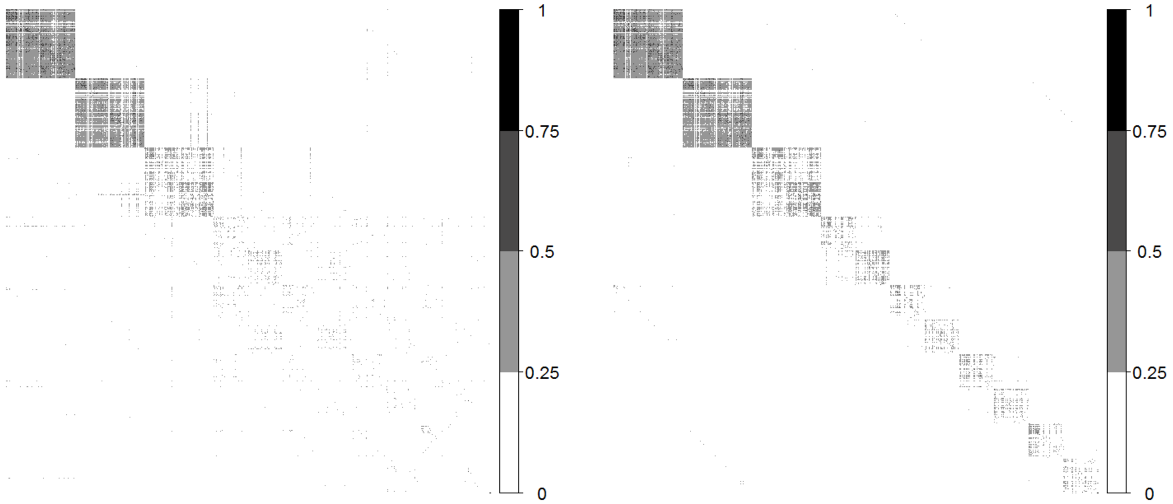
To prove 1, for any orthonormal basis $\{\tilde{\varphi}_l\}_{l=1}^{\infty}$ of $L^2[0,1]$,

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

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 2.3.1, we have $\mathcal{H} = p^{-1}\sum_{l=1}^{\infty}\operatorname{tr}(\Sigma_l)\varphi_l \otimes \varphi_l$, 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 $\operatorname{tr}(\Sigma_l)$ is nonincreasing, this yields the spectral decomposition of $\mathcal{H}$.

A.3 Left-Hand Task Partial Separability Results for Sections 2.3 and 2.4.2



(a) (b)

Figure A.1: Estimated correlation structures of $\mathbb{R}^{Lp}$-valued random coefficients from different L-truncated Karhunen-Loève type expansions for the left-hand task dataset. 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 equation 2.2 as in [17], and (b) random coefficients $(\theta_1^T, \ldots, \theta_L^T)^T$ under partial separability in (2.3)

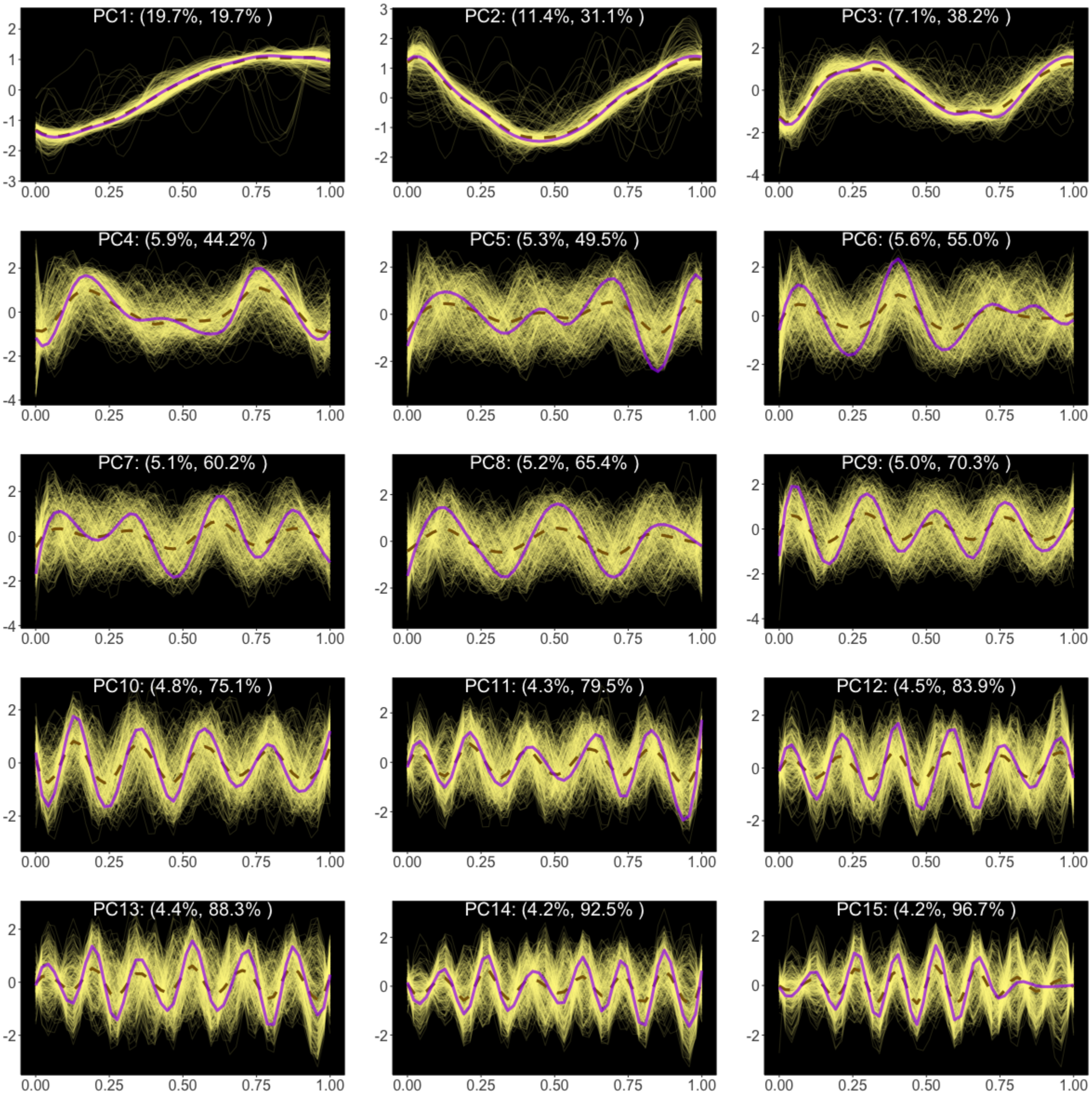
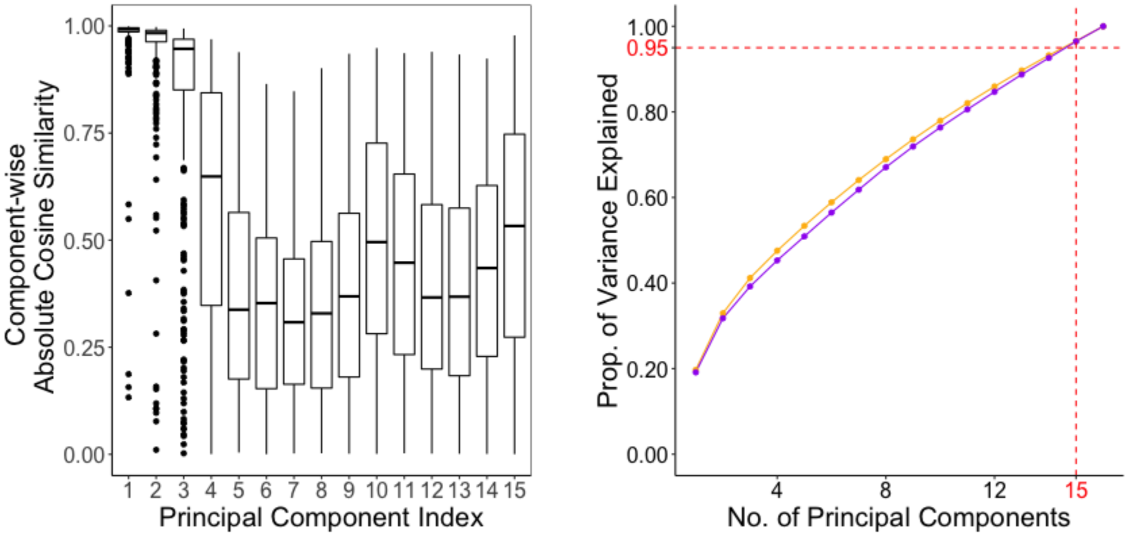


Figure A.2: Estimated functional principal components from L-truncated Karhunen-Loève (KL) type expansions for the left-hand task dataset. Curves are coded as: univariate KL eigenfunctions $\{\phi_{jl}\}_{j=1}^{p}$ (- - -) and their average $p^{-1}\sum_{j=1}^{p}\phi_{jl}$ (- - -), and partially separable eigenfunctions φ_l (—). Values on the top of each figure indicate the marginal and cumulative proportion of variance explained under partial separability.



(a) (b)

Figure A.3: Comparison between functional principal components from L-truncated Karhunen-Loève (KL) type expansions for the left-hand dataset. (a): Boxplots for component-wise absolute cosine similarity between φ_l and ϕ_{jl} for $j = 1, \ldots, p$ for every principal component. (b): Proportion of variance explained for each expansion different number of principal components. Curves are coded as: univariate KL eigenfunctions ϕ_{jl} (—) and partially separable eigenfunctions φ_l (—).

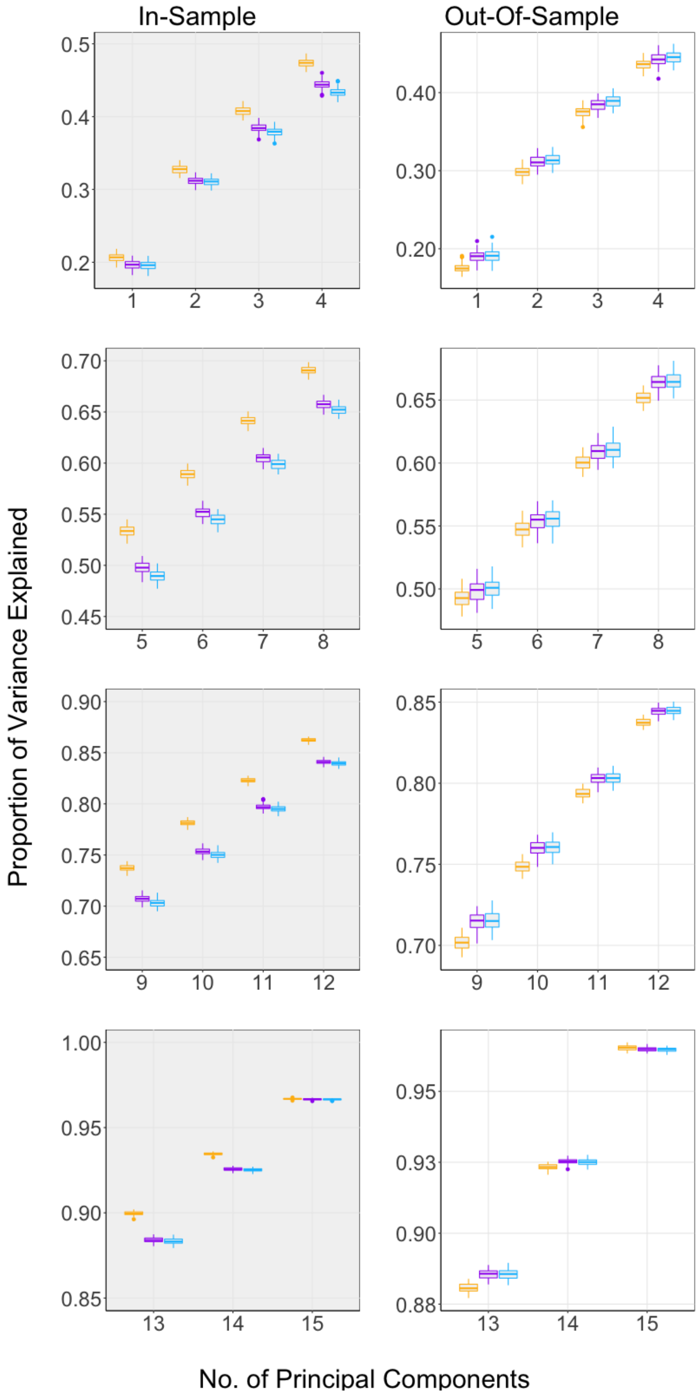


Figure A.4: Estimated variance explained for different L-truncated Karhunen-Loève type expansions for left-hand task fMRI curves. Left: In-Sample. Right: Out-of-Sample. Boxplots are coded as: functional principal components (—) in (2.2), partially separable expansion (—) in (2.3), and joint partially separable expansion (—) in (2.6).

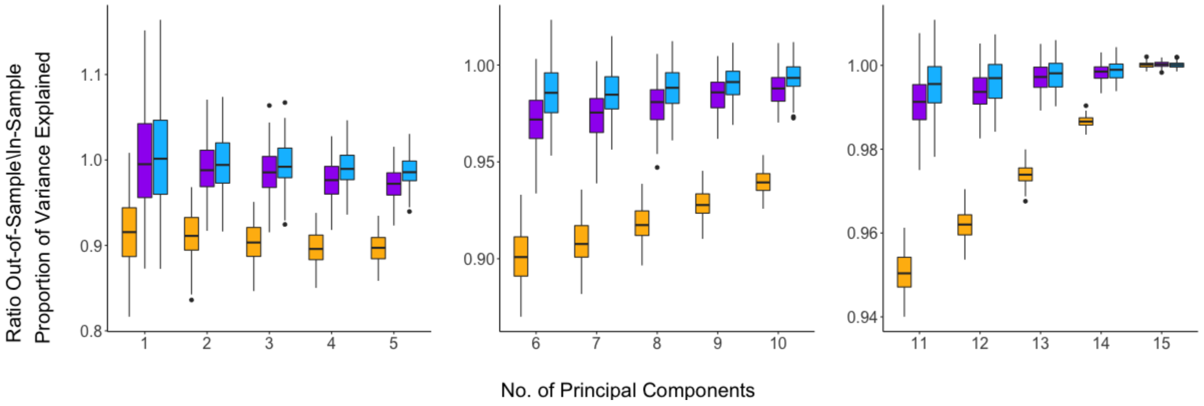


Figure A.5: Estimated variance explained for different L-truncated Karhunen-Loève type expansions for left-hand task fMRI curves. The figure shows boxplots for the ratio out-of-sample over in-sample variance explained. Boxplots are coded as: functional principal components (—) in (2.2), partially separable expansion (—) in (2.3), and joint partially separable expansion (—) in (2.6).

A.4 Block Correlation Computation in Sections 2.2 and 2.3.3

In this section I discussed how to compute the sample correlation matrices for the random score of the Karhunen-Loève type expansions. Throughout this section, the data consists of curves $\{X_{ij}(\cdot)\}$ for $i = 1, \dots, n$ and $j = 1, \dots, p$.

A.4.1 Univariate Karhunen-Loève Expansion

Consider a one dimensional Karhunen-Loève decomposition for each $j = 1, \dots, p$ and $l \geq 1$ according to the model:

$$X_j(t) = \sum_{l=1}^{\infty} \xi_{jl}\phi_{jl}(t)$$

First, I compute estimates $\hat{\phi}_{jl}(\cdot)$,$j = 1, \dots, p$, $l = 1, \dots, L$ from the full data set. Second, I compute (since the curves are already centered) the scores $\hat{\xi}_{ilj} = \int_0^1 \hat{\phi}_{jl}(t)X_{ij}(t)dt$ for $i = 1, \dots, n$. These integrals are estimated numerically. Then, a $L \times L$ block covariance matrix $\hat{\Gamma}$, where the (l, m)-th block $\hat{\Gamma}_{lm}$ has (j, k)-th element

$$\frac{1}{n}\sum_{i=1}^{n} \hat{\xi}_{ilj}\hat{\xi}_{imk}.$$

Finally, a block correlation matrix is computed as:

$$\text{diag}(\hat{\Gamma})^{-1/2}\hat{\Gamma}\text{diag}(\hat{\Gamma})^{-1/2}$$

.

A.4.2 Partially Separable Karhunen-Loève Expansion

Under partial separability we consider a multivariate Karhunen-Loève expansion for $l \geq 1$ according to the model:

$$X_j(t) = \sum_{l=1}^{\infty} \theta_{lj} \varphi_l(t)$$

First, I compute estimates $\hat{\varphi}_l(\cdot)$, $l = 1, \ldots, L$ from the full data set. Second, I compute (since the curves are already centered) the scores $\hat{\theta}_{ilj} = \int_0^1 \hat{\varphi}_l(t) X_{ij}(t) dt$ for $i = 1, \ldots, n$. These integrals are estimated numerically. Then, a $L \times L$ block covariance matrix $\hat{\Sigma}$, where the (l, m)-th block $\hat{\Sigma}_{lm}$ has (j, k)-th element

$$\frac{1}{n} \sum_{i=1}^{n} \hat{\theta}_{ilj} \hat{\theta}_{imk}.$$

Finally, a block correlation matrix is computed as:

$$\text{diag}(\hat{\Sigma})^{-1/2} \hat{\Sigma} \text{diag}(\hat{\Sigma})^{-1/2}$$

.

A.5 Out-of-Sample Performance Analysis In Section 2.3.4

In this section I discuss in detail the in- and out-of-sample analysis to compare the univariate and partially separable Karhunen-Loève expansions. Let $\{X_{ij}(t) : t \in [0,1]\}$ be the featured-centered observed fMRI curves for $i = 1, \dots, n$ and $j = 1, \dots, p$ so that $n^{-1} \sum_{i=1}^{n} X_{ij} = 0$.

A.5.1 Univariate Karhunen-Loève Expansion

Based on the work of [17], I consider a one dimensional Karhunen-Loève decomposition for each $j = 1, \dots, p$ and $l \geq 1$ according to the model:

$$X_j(t) = \sum_{l=1}^{\infty} \xi_{jl} \phi_{jl}(t)$$

I repeatedly split the data into $\{X_{\cdot j}(\cdot)\}_{\text{train}}$ and $\{X_{\cdot j}(\cdot)\}_{\text{test}}$ each time. Let $\overline{X}_j^{\text{train}}$ and $\overline{X}_j^{\text{test}}$ be the testing and training means, which will not be 0. First, I compute estimators $\hat{\phi}_{jl}(\cdot)$ for $j = 1, \dots, p$ and $l = 1, \dots, L$ from a training set $\{X_{\cdot j}(\cdot)\}_{\text{train}}$.

Next, for curves in the testing set, I compute projected curves as:

$$\hat{X}_{ij}(t) = \overline{X}_j^{\text{test}}(t) + \sum_{l=1}^{L} \hat{\xi}_{jl} \hat{\phi}_{jl}(t), \quad \hat{\xi}_{ijl} = \int_0^1 \hat{\phi}_{jl}(t) [X_{ij}(t) - \overline{X}_j^{\text{test}}(t)] dt.$$

Finally, the variance explained in the testing set is computed as:

$$1 - \frac{\sum_i \sum_{j=1}^{p} \int_0^1 \left[X_{ij}(t) - \hat{X}_{ij}(t) \right]^2 dt}{\sum_i \sum_{j=1}^{p} \left[X_{ij}(t) - \overline{X}_j^{\text{test}} \right]^2 dt},$$

where the index i ranges over the testing set. Since the curves can only be computed on a discrete grid, the integrals are computed using numerical integration.

A.5.2 Partially Separable Karhunen-Loève Expansion

Under partial separability we consider a multivariate Karhunen-Loève expansion for $l \geq 1$ according to the model:

$$X_j(t) = \sum_{l=1}^{\infty} \theta_{lj} \varphi_l(t)$$

Using the same splits as for fgm, I first compute estimators $\hat{\varphi}_l(\cdot)$ for $l = 1, \ldots, L$ from a training set $\{X_{\cdot j}(\cdot)\}_{\text{train}}$, where the training curves must be centered as part of this step with respect to $\overline{X}_j^{\text{train}}$. Next, I compute:

$$\hat{X}_{ij}(t) = \overline{X}_j^{\text{test}}(t) + \sum_{l=1}^{L} \hat{\theta}_{ilj} \hat{\varphi}_l(t), \quad \hat{\theta}_{ilj} = \int_0^1 \hat{\varphi}_l(t)[X_{ij}(t) - \overline{X}_j^{\text{test}}]dt.$$

Finally, the fraction of variance explained by the partially separable basis on the testing set is computed as

$$1 - \frac{\sum_i \sum_{j=1}^{p} \int_0^1 \left[X_{ij}(t) - \hat{X}_{ij}(t)\right]^2 dt}{\sum_i \sum_{j=1}^{p} \left[X_{ij}(t) - \overline{X}_j^{\text{test}}\right]^2 dt},$$

A.6 Proofs For Section 3.2.1

Proof of Theorem 3.2.1.

We have

$$\begin{aligned}
&\operatorname{cov}\left\{X_j(s), X_k(t) \mid X_{-(j,k)}\right\} \\
&\qquad = \operatorname{cov}\left\{\sum_{l=1}^{\infty} \theta_{lj}\varphi_l(s), \sum_{l'=1}^{\infty} \theta_{l'j}\varphi_{l'}(s) \mid X_{-(j,k)}\right\} \\
&\qquad = \sum_{l,l'=1}^{\infty} \operatorname{cov}\left\{\theta_{lj}, \theta_{l'k} \mid X_{-(j,k)}\right\} \varphi_l(s)\varphi_{l'}(t) \\
&\qquad = \sum_{l=1}^{\infty} \operatorname{cov}\left\{\theta_{lj}, \theta_{lk} \mid \theta_{l,-(j,k)}\right\} \varphi_l(s)\varphi_l(t) \\
&\qquad = \sum_{l=1}^{\infty} \tilde{\sigma}_{ljk}\varphi_l(s)\varphi_l(t)
\end{aligned}$$

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

Proof of Corollary 3.

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* $\operatorname{cov}(\theta_{lj}, \theta_{l'k}|X_{-(j,k)}) = 0$.

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,

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

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 $\operatorname{cov}\{\theta_{lj}, \theta_{lk} \mid \theta_{l,-(j,k)}\} = 0$ 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 $\operatorname{cov}\{\theta_{lj}, \theta_{l'k} \mid X_{-(j,k)}\} = 0$ for all $l \neq l'$, whence all of the coefficients in the above display are zero, and $(j,k) \notin E$.

A.7 Proof For Section 3.3

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* (**??**) *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 $0 < \delta \leq \tilde{C}_3$ and for all $j, k \in V$,*

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

Proof of Lemma 1.

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

$$\operatorname{pr}\left(\|\hat{\mathcal{G}}_{jk} - \mathcal{G}_{jk}\|_{\mathrm{HS}} \geq \delta\right) \leq \operatorname{pr}\left(\|\overline{Y}_{jk} - \mathcal{G}_{jk}\|_{\mathrm{HS}} \geq \frac{\delta}{2}\right) + 2\max_{j\in V}\operatorname{pr}\left(\|\hat{\mu}_j\|^2 \geq \frac{\delta}{2}\right). \quad \text{(A.1)}$$

We begin with the first term on the right-hand side of (A.1), and will apply Theorem 2.5 of [41]. Specifically, we need to find $L_1, L_2 > 0$ such that

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

which will then imply that

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

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

$$\begin{aligned}\|Y_{ijk} - \mathcal{G}_{jk}\|_{\text{HS}}^b &= \left\{ \sum_{l,l'=1}^{\infty} \sigma_{ljj}\sigma_{l'kk} \left(\xi_{ilj}\xi_{il'k} - \delta_{ll'}r_{ljk}\right)^2 \right\}^{b/2} \\ &= (M_j M_k)^{b/2} \left\{ \sum_{l,l'=1}^{\infty} \frac{\sigma_{ljj}\sigma_{l'kk}}{M_j M_k} \left(\xi_{ilj}\xi_{il'k} - \delta_{ll'}r_{ljk}\right)^2 \right\}^{b/2} \\ &\le (M_j M_k)^{b/2-1} \sum_{l,l'=1}^{\infty} \sigma_{ljj}\sigma_{l'kk} \left|\xi_{ilj}\xi_{il'k} - \delta_{ll'}r_{ljk}\right|^b,\end{aligned}$$

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

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

Thus,

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

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

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

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

whence

$$\text{pr}\left(\|\hat{\mu}_j\| \ge \frac{\delta}{2}\right) \le 2\exp\left(-\frac{n\delta^2}{8\tilde{L}_1 + 4\tilde{L}_2\delta}\right) \tag{A.3}$$

Now, setting $\tilde{C}_3 \le 2$ and $0 < \delta < \tilde{C}_3$, we find that $\text{pr}(\|\hat{\mu}_j\|^2 \ge \delta/2)$ is also bounded by the right hand side of (A.3), 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 $\tilde{C}_2 = 6$, (A.1)–(A.3) together imply the result.

Proof of Theorem 3.3.1.

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 [41] and Assumption 2,

$$\begin{aligned} |s_{ljk} - \sigma_{ljk}| &\leq |\langle \mathcal{G}_{jk}(\varphi_l), \varphi_l - \hat{\varphi}_l \rangle| + |\langle \mathcal{G}_{jk}(\varphi_l - \hat{\varphi}_l), \hat{\varphi}_l \rangle| + |\langle [\mathcal{G}_{jk} - \hat{\mathcal{G}}_{jk}](\hat{\varphi}_l), \hat{\varphi}_l \rangle| \\ &\leq \|\mathcal{G}_{jk}(\varphi_l)\| \|\varphi_l - \hat{\varphi}_l\| + \|\mathcal{G}_{jk}(\varphi_l - \hat{\varphi}_l)\| \|\hat{\varphi}_l\| + \|[\mathcal{G}_{jk} - \hat{\mathcal{G}}_{jk}](\hat{\varphi}_l)\| \|\hat{\varphi}_l\| \\ &\leq 2M\tau_l \|\hat{\mathcal{H}} - \mathcal{H}\|_{\mathrm{HS}} + \|\hat{\mathcal{G}}_{jk} - \mathcal{G}_{jk}\|_{\mathrm{HS}}. \end{aligned} \tag{A.4}$$

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 (A.4), when $0 < \delta < C_3$, for any $l \in \mathbb{N}$ and any $j, k \in V$,

$$\begin{aligned} \mathrm{pr}\left(|s_{ljk} - \sigma_{ljk}| > \delta \right) &\leq \mathrm{pr}\left(\|\hat{\mathcal{H}} - \mathcal{H}\|_{\mathrm{HS}} > \frac{\delta}{4M\tau_l} \right) + \mathrm{pr}\left(\|\hat{\mathcal{G}_{jk}} - \mathcal{G}_{jk}\|_{\mathrm{HS}} > \frac{\delta}{2} \right) \\ &\leq C_2^* \exp\{-C_1^* n \delta^2 (4M\tau_l)^{-2}\} + \tilde{C}_2 \exp\{-\tilde{C}_1 n (\delta/2)^2\}. \end{aligned}$$

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 4.

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

$$\begin{aligned}\mathrm{pr}\left(|\hat{r}_{ljk} - r_{ljk}| \geq 2\delta\right) &\leq 2\max_{j\in V}\mathrm{pr}\{A_{lj}(\epsilon)^c\} + \mathrm{pr}\left[A_{lj}(\epsilon)\cap A_{lk}(\epsilon)\cap\{|\hat{r}_{ljk} - r_{ljk}| \geq 2\delta\}\right] \\ &\leq 2\max_{j\in V}\mathrm{pr}\{A_{lj}(\epsilon)^c\} + \mathrm{pr}\left\{|s_{ljk} - \sigma_{ljk}| \geq \delta(1-\epsilon)^2\pi_l\right\} \\ &\quad + \mathrm{pr}\left\{|1 - \hat{c}_{lj}\hat{c}_{lk}| \geq \delta(1-\epsilon)^2\right\}. \qquad (A.5)\end{aligned}$$

We next obtain bounds for the first and last terms of the last line above. Let C_j be as in Theorem 3.3.1, and $\overline{\pi} = \max_{l\in\mathbb{N}}\pi_l$. If $D_3 = \min(\epsilon, C_3\overline{\pi}^{-1})$ and $0 < \delta \leq D_3$, then

$$\begin{aligned}\mathrm{pr}\{A_{lj}(\epsilon)^c\} &\leq \mathrm{pr}(|1 - \hat{c}_{lj}^2| > \epsilon) = \mathrm{pr}(|s_{ljj} - \sigma_{ljj}| > \epsilon\sigma_{ljj}) \leq \mathrm{pr}(|s_{ljj} - \sigma_{ljj}| > \delta\pi_l) \\ &\leq C_2\exp\{-C_1 n\tau_l^{-2}\pi_l^2\delta^2\}.\end{aligned}$$

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 3.3.1,

$$\begin{aligned}\mathrm{pr}\left\{|1 - \hat{c}_{lj}| > \frac{\delta(1-\epsilon)^2}{3}\right\} &\leq \mathrm{pr}\left\{|s_{ljj} - \sigma_{ljk}| > \frac{\delta(1-\epsilon)^2\pi_l}{3}\right\} \\ &\leq C_2\exp\left\{-C_1 n\pi_l^2\delta^2\tau_l^{-2}(1-\epsilon)^4/9\right\}.\end{aligned}$$

Putting these facts together, (A.5) becomes

$$\begin{aligned}\mathrm{pr}\left(|\hat{r}_{ljk} - r_{ljk}| > 2\delta\right) &\leq 2C_2\exp\{-C_1 n\tau_l^{-2}\pi_l^2\delta^2\} + C_2\exp\{-C_1(1-\epsilon)^4 n\tau_l^{-2}\pi_l^2\delta^2\} \\ &\quad + 2C_2\exp\left\{-C_1 n\pi_l^2\delta^2\tau_l^{-2}(1-\epsilon)^4/9\right\}.\end{aligned}$$

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

A.7.1 Lemma 2 and Proof of Theorem 3.3.2

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, $|||\Delta|||_\infty = \max_{j=1,\ldots,p} \sum_{k=1}^p |\Delta_{jk}|$, $|||\Delta|||_1 = \left|\left|\left|\Delta^T\right|\right|\right|_\infty$, 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 4, define functions

$$\overline{n}(\delta; c) = \frac{\log(D_2 c)}{D_1 \delta^2}, \quad \overline{\delta}(n; c) = \left\{\frac{\log(D_2 c)}{D_1 n}\right\}^{1/2} \quad (c, \delta > 0, n \in \mathbb{N}). \tag{A.6}$$

Before proceeding to the results, we describe our primal-dual witness approach as a modification of that of [40] to account for the presence of the group Lasso penalty in (3.8). Of importance are the sub-differentials of each of the penalty terms in (3.8), omitting the tuning parameter factor, evaluated at a generic set of inputs $(\Upsilon_1, \ldots, \Upsilon_L)$. Let $\upsilon_{ljk} = (\Upsilon_l)_{jk}$. 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

$$(Z_l)_{jk} = \begin{cases} 0 & \text{if } j = k \\ \operatorname{sgn}(\upsilon_{ljk}) & \text{if } j \neq k,\ \upsilon_{ljk} \neq 0 \\ \in [-1, 1] & \text{if } j \neq k,\ \upsilon_{ljk} = 0. \end{cases} \tag{A.7}$$

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

$$z_{\cdot jk} = \begin{cases} 0 & \text{if } j = k \\ \frac{\upsilon_{\cdot jk}}{\|\upsilon_{\cdot jk}\|_E} & \text{if } j \neq k,\ \|\upsilon_{\cdot jk}\|_E \neq 0 \\ \in \{y \in \mathcal{R}^L : \|y\|_E \leq 1\} & \text{if } j \neq k, \|\upsilon_{\cdot jk}\|_E = 0. \end{cases} \tag{A.8}$$

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

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

$$(\tilde{\Xi}_1, \ldots, \tilde{\Xi}_L) = \arg \min_{\Upsilon_l \succ 0, \Upsilon_l = \Upsilon_l^T, \Upsilon_{l,\overline{E}_l^c}=0} \sum_{l=1}^{L} \left\{ \mathrm{tr}(\hat{R}_l \Upsilon_L) - \log(|\Upsilon_l|) \right\}$$
$$+ \gamma \left\{ \alpha \sum_{l=1}^{L} \sum_{j \neq k} |\upsilon_{ljk}| + (1-\alpha) \sum_{j \neq k} \left(\sum_{l=1}^{L} \upsilon_{ljk}^2 \right)^{1/2} \right\} \qquad \text{(A.9)}$$

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

$$\left[\hat{R}_l - \tilde{\Xi}_l^{-1} + \gamma \left\{ \alpha \tilde{Z}_{1l} + (1-\alpha) \tilde{Z}_{2l} \right\} \right]_{\overline{E}_l} \quad (l = 1, \ldots, L). \qquad \text{(A.10)}$$

3. Update

$$\left(\tilde{Z}_{1,l} \right)_{jk} = \frac{1}{\gamma \alpha} \left\{ \left(\tilde{\Xi}_l^{-1} \right)_{jk} - \hat{r}_{ljk} \right\}, \quad \left(\tilde{Z}_{2,l} \right)_{jk} = 0, \qquad \{(j,k) \in \overline{E}_l^c, \, l = 1, \ldots, L\}. \qquad \text{(A.11)}$$

4. Verify strict dual feasibility condition

$$\left| \left(\tilde{Z}_{1,l} \right)_{jk} \right| < 1, \quad \{(j,k) \in \overline{E}_l^c, \, l = 1, \ldots, L\}. \qquad \text{(A.12)}$$

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*

$$n > \overline{n} \left(\min \{ \mathfrak{a}_L, \mathfrak{b}_L \} ; L^{\varrho-1} p^{\varrho} \right), \qquad \text{(A.13)}$$

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

$$\|\hat{\Xi}_l - \Xi_l\|_\infty \leq 2\kappa_{\Psi_l}\left(m_l^{-1} + 8\epsilon_L^{-1}\right)\overline{\delta}(n; L^{\varrho-1}p^{\varrho}) \tag{A.14}$$

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

Proof of Lemma 2.

Define D_j and m_l as in Corollary 4. Observe that $n > \overline{n}(\delta; c)$ implies $\overline{\delta}(n; c) < \delta$. Since (A.13) implies that $n > \overline{n}(D_3 m_l; L^{\varrho-1}p^{\varrho})$ for each l, $1 \leq l \leq L$, we have $m_l^{-1}\overline{\delta}(n; L^{\varrho-1}p^{\varrho}) \leq D_3$, for $1 \leq l \leq L$. Define

$$\mathcal{A}_l = \left\{\|R_l - \hat{R}_l\|_\infty \leq m_l^{-1}\overline{\delta}(n; L^{\varrho-1}p^{\varrho})\right\}. \tag{A.15}$$

Applying (3.11) with $\delta = m_l^{-1}\overline{\delta}(n; L^{\varrho-1}p^{\varrho})$ together with the union bound, we obtain $\mathrm{pr}(\mathcal{A}_l^c) \leq L^{1-\varrho}p^{2-\varrho}$, $1 \leq l \leq L$, so that $\mathrm{pr}\left(\bigcap_{l=1}^L \mathcal{A}_l\right) \geq 1-(Lp)^{2-\varrho}$. Let $\{(\tilde{\Xi}_l, \tilde{Z}_l),\ l = 1, \ldots, L\}$ be the primal-dual witness solutions constructed in steps 1–4 preceding the lemma statement. The result in (A.14) 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 (A.14) 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

$$\|\hat{R}_l - R_l\|_\infty \leq m_l^{-1}\overline{\delta}(n; L^{\varrho-1}p^{\varrho}) = \left(\frac{\epsilon_L}{m_l\eta_l'}\right)\left\{8\epsilon_L^{-1}\overline{\delta}(n; L^{\varrho-1}p^{\varrho})\right\}\frac{\eta_l'}{8} \leq \frac{\gamma\eta_l'}{8}, \tag{A.16}$$

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

$$\begin{aligned} 2\kappa_{\Psi_l}\left(\|\hat{R}_l - R_l\|_\infty + \gamma\right) &\leq 2\kappa_{\Psi_l}(m_l^{-1} + 8\epsilon_L^{-1})\overline{\delta}(n; L^{\varrho-1}p^{\varrho}) \\ &\leq \frac{2\kappa_{\Psi_l}(m_l^{-1} + 8\epsilon_L^{-1})}{6y_l m_l \max(\kappa_{\Psi_l}^2\kappa_{R_l}^3, \kappa_{\Psi_l}\kappa_{R_l})(m_l^{-1} + 8\epsilon_L^{-1})^2} \\ &\leq \min\left(\frac{1}{3y_l\kappa_{\Psi_l}\kappa_{R_l}^3}, \frac{1}{3y_l\kappa_{R_l}}\right) \end{aligned} \tag{A.17}$$

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 [40] are satisfied whenever $\mathcal{A}_l$ holds, so that

$$\|\tilde{\Xi}_l - \Xi_l\|_\infty \leq 2\kappa_{\Psi_l}(\|\hat{R}_l - R_l\|_\infty + \gamma) \leq 2\kappa_{\Psi_l}(m_l^{-1} + 8\epsilon_L^{-1})\overline{\delta}(n; L^{\varrho-1}p^{\varrho}). \tag{A.18}$$

Define $\mathcal{W}_l = \tilde{\Xi}_l^{-1} - R_l + R_l(\tilde{\Xi}_l - \Xi_l)R_l$. Having established (A.17) and (A.18), we apply Lemma 5 of [40] to conclude that

$$\begin{aligned}\|\mathcal{W}_l\|_\infty &\leq \frac{3}{2}y_l\|\tilde{\Xi}_l - \Xi_l\|_\infty^2\kappa_{R_l}^3 \\ &\leq 6\kappa_{R_l}^3\kappa_{\Psi_l}^2 y_l(m_l^{-1} + 8\epsilon_L^{-1})^2\left\{\overline{\delta}(n; L^{\varrho-1}p^{\varrho})\right\}^2 \\ &= \left\{6\kappa_{R_l}^3\kappa_{\Psi_l}^2 y_l(m_l^{-1} + 8\epsilon_L^{-1})^2\overline{\delta}(n; L^{\varrho-1}p^{\varrho})\right\}\left(\frac{\epsilon_L}{\eta_l'}\right)\frac{\gamma\eta_l'}{8} \\ &\leq \left\{6\kappa_{R_l}^3\kappa_{\Psi_l}^2 y_l m_l(m_l^{-1} + 8\epsilon_L^{-1})^2\overline{\delta}(n; L^{\varrho-1}p^{\varrho})\right\}\left(\frac{\epsilon_L}{m_l\eta_l'}\right)\frac{\gamma\eta_l'}{8} \\ &\leq \frac{\gamma\eta_l'}{8}.\end{aligned} \tag{A.19}$$

The last line follows by (A.13) and because $\epsilon_L \leq m_l\eta_l'$.

Together, (A.16) and (A.19) imply that, when $\bigcap_{l=1}^L \mathcal{A}_l$ holds,

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

Following similar derivations to those of Lemma 4 of [40], for any $l = 1, \ldots, L$ and $(j,k) \in \overline{E}_l^c$,

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

Using Assumption 3, the definitions of the sub-differentials in (A.7) and (A.8), 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, $\tilde{\hat{\Xi}}_l = \hat{\Xi}_l$ for each l when $\bigcap_{l=1}^L \mathcal{A}_l$ holds. Together with (A.18), this completes the proof.

Proof of Theorem 3.3.2.

By construction of the primal witness in (A.9), it is clear that $(j,k) \notin E_l$ implies $\tilde{\hat{\Xi}}_{ljk} = 0$. Under the given constraint on the sample size and using Assumption 4, we have $\tilde{\hat{\Xi}}_l = \hat{\Xi}_l$ 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, Assumption 4 and (3.12) imply $n > \overline{n}(\mathfrak{c}_l; L^{\varrho-1}p^{\varrho})$, 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.

A.8 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 3.3.2, 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 3.3.2. 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 $\left\{\Psi_{1,(j,k),(j',k')}, \ldots, \Psi_{L,(j,k),(j',k')}\right\}$. 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 $\tilde{\Psi}_{S^cS}$, with row and column blocks indexed by S^c and S, respectively. Similarly, define Ψ_{SS}.

We next define an alternative operator norm on $\tilde{\Psi}_{S^cS}\tilde{\Psi}_{SS}^{-1}$ tailored to the group Lasso sub-differential defined in (A.8). 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

$$|\!|\!| A |\!|\!|_{\infty,2} = \max_{e \in S^c} \max_{\|z_{e'}\|_E \leq 1} \left\| \sum_{e' \in S} A_{ee'} z_{e'} \right\|_E = \max_{e \in S^c} \left\{ \sum_{l=1}^{L} \left(\sum_{e' \in S} |A_{ee'll}| \right)^2 \right\}^{1/2}. \tag{A.20}$$

We require the following group irrepresentability condition.

Assumption 5. *For some* $\eta \in (0, 1]$, $\left|\!\left|\!\left| \tilde{\Psi}_{S^cS}\tilde{\Psi}_{SS}^{-1} \right|\!\right|\!\right|_{\infty,2} \leq 1 - \eta$.

Next, define $\tilde{\kappa}_{\Psi_l} = \left|\!\left|\!\left| (\Psi_{l,SS})^{-1} \right|\!\right|\!\right|_\infty$, $y = \max_{j \in V} |\{k \in V : \sum_{l=1}^{L} \Xi_{ljk}^2 \neq 0\}$, and $\tilde{\xi}_{\min} = \min_{(j,k) \in \bigcup_{l=1}^{L} E_l} \{\max_{l=1,\ldots,L} |\Xi_{ljk}|\}$. With D_j as in Corollary 4 and $\tilde{\epsilon}_L = \eta \min_{l=1,\ldots,L} m_l$, set

$$\begin{aligned}
\tilde{\mathfrak{a}}_L &= D_3 \min_{l=1,\ldots,L} m_l, \\
\tilde{\mathfrak{b}}_L &= (6y)^{-1} \min_{l=1,\ldots,L} \left\{ m_l(m_l^{-1} + 8\tilde{\epsilon}_L^{-1})^2 \max\left(\tilde{\kappa}_{\Psi_l}^2 \kappa_{R_l}^2, \tilde{\kappa}_{\Psi_l}\kappa_{R_l}\right)\right\}^{-1} \qquad (A.21) \\
\tilde{\mathfrak{c}}_L &= \tilde{\xi}_{\min}\{4\tilde{\kappa}_{\Psi_l}(m_l^{-1} + 8\tilde{\epsilon}_L^{-1})\}^{-1}.
\end{aligned}$$

As a direct analog of Assumption 4, we require the following.

Assumption 6. *$L \to \infty$ as $n \to \infty$, $L \leq np$, and $\min(\tilde{\mathfrak{a}}_L, \tilde{\mathfrak{b}}_L, \tilde{\mathfrak{c}}_L) \{n/\log(n)\}^{1/2} \to \infty$.*

Theorem A.8.1. *Suppose Assumptions 1–2 and 5–6 hold and that, for some $\varrho > 2$, $\gamma = 8\tilde{\epsilon}_L^{-1} \{(D_1 n)^{-1} \log(D_2 L^{\varrho-1} p^{\varrho})\}^{1/2}$ and that the sample size satisfies the lower bound*

$$n \min(\tilde{\mathfrak{a}}_L, \tilde{\mathfrak{b}}_L, \tilde{\mathfrak{c}}_L)^2 > D_1^{-1} \{\log(D_2) + (\varrho - 1)\log(n) + (2\varrho - 1)\log(p)\}. \qquad (A.22)$$

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 (A.22) 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 $\min_{l=1,\ldots,L} m_l \gtrsim n^{-d}$ for $0 < d < 1/4$, (A.22) 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 3.3.2 under analagous settings, it is weaker in the first term since $\tilde{\xi}_{\min} > \min_{l=1,\ldots,L} \xi_{\min,l}$. 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, similar to Remark 3.3.2, one can deduce edge selection consistency if L is capable of growing faster than $\tilde{L}_n^* = \min\left\{L : \bigcup_{l=1}^{L} E_l = \bigcup_{l=1}^{\infty} E_l\right\}$ while still satisfying Assumption 6.

Proof of Theorem A.8.1.

As the proof follows the same logical flow as that of Theorem 3.3.2, we will sketch the proof while outlining major differences. First of all, steps 1–3 from Section A.7.1 that detail the construction of the primal/dual witness pairs $(\tilde{\Xi}_l, \tilde{Z}_l)$ is modified as follows. In the first step, one computes the penalized estimator as in (A.9) 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 (A.10), 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

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

With these amendments, one proceeds by showing that, when

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

one has $\|\tilde{\Xi}_l - \Xi_l\|_\infty \le b_l$. This result can be proven using the same logic used in the proofs of Lemmas 5 and 6 of [40] using the sub-differential properties in (A.7) and (A.8). 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 (A.15). 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

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

so that

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

Then, using arguments similar to Lemma 1 of [40], the bound

$$\begin{aligned} \|\{(\tilde{Z}_1)_{jk}, \ldots, (\tilde{Z}_L)_{jk}\}^T\|_E &\le \frac{2}{\gamma}\left(\left|\!\left|\!\left| \tilde{\Psi}_{S^cS}\tilde{\Psi}_{SS}^{-1} \right|\!\right|\!\right|_{2,\infty} + 1\right)\left(\frac{\gamma\eta}{8}\right) + \left|\!\left|\!\left| \tilde{\Psi}_{S^cS}\tilde{\Psi}_{SS}^{-1} \right|\!\right|\!\right|_{2,\infty} \\ &\le \frac{\eta(2-\eta)}{4} + 1 - \eta < 1 \end{aligned}$$

by Assumption 5. 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 $(j,k) \in \bigcup_{l=1}^{L} E_l$, one has

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

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

A.9 Partial Separability Violation Examples of Section 3.2.2

The examples in this section analyze how different violations of partial separability affect the true edge sets.

For simplicity, consider a generic example. Let X has three components with each one lying on a common two-dimensional space with probability one, and let $\mathcal{H} = 3^{-1}(\mathcal{G}_{11} + \mathcal{G}_{22} + \mathcal{G}_{33})$ have eigenfunctions (φ_1, φ_2), where $\mathcal{G}_{jj}$ is the covariance operator of X_j. Then one has $X_j = \theta_{1j}\varphi_1 + \theta_{2j}\varphi_2$ for $j = 1, 2, 3$. If X is Gaussian, the conditional dependence is completely determined by the block covariance matrix $\Sigma = \{\Sigma_{ll'}\}_{l,l'=1}^{2}$, $\Sigma_{ll'} = \{\sigma_{ll'jk}\}_{j,k=1}^{3}$, where $\Sigma_{ll'} = 0$ for $l \neq l'$ if and only if X is partially separable. Thus, let the partially separable edge set be $E^* = E_1 \cup E_2$, where $(j, k) \in E_l$ if and only if $(\Sigma_{ll}^{-1})_{jk} \neq 0$. On the other hand, the true edge set E has $(j, k) \in E$ if and only if the (j, k)-th element in at least one of the blocks in Σ^{-1} is nonzero.

Consider three possible values for Σ, given by

$$\Sigma_1 = \left[\begin{array}{ccc|ccc} 1 & a & 0 & c_1 & 0 & 0 \\ a & 1 & 0 & 0 & c_2 & 0 \\ 0 & 0 & 1 & 0 & 0 & c_3 \\ \hline c_1 & 0 & 0 & 1 & 0 & 0 \\ 0 & c_2 & 0 & 0 & 1 & a \\ 0 & 0 & c_3 & 0 & a & 1 \end{array}\right] \qquad \Sigma_2 = \left[\begin{array}{ccc|ccc} 1 & a & 0 & c_1 & 0 & 0 \\ a & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & c_3 \\ \hline c_1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & a \\ 0 & 0 & c_3 & 0 & a & 1 \end{array}\right]$$

$$\Sigma_3 = \left[\begin{array}{ccc|ccc} 1 & a & b & 0 & 0 & 0 \\ a & 1 & a & 0 & 0 & c \\ b & a & 1 & 0 & c & 0 \\ \hline 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & c & 0 & 1 & 0 \\ 0 & c & 0 & 0 & 0 & 1 \end{array}\right].$$

A.9.1 Example 1

Consider the following block covariance matrix and its inverse:

$$\Sigma = \left(\begin{array}{ccc|ccc} 1 & a & 0 & c_1 & 0 & 0 \\ a & 1 & 0 & 0 & c_2 & 0 \\ 0 & 0 & 1 & 0 & 0 & c_3 \\ \hline c_1 & 0 & 0 & 1 & 0 & 0 \\ 0 & c_2 & 0 & 0 & 1 & a \\ 0 & 0 & c_3 & 0 & a & 1 \end{array}\right) \qquad \Omega = \Sigma^{-1} = \gamma^{-1} \left(\begin{array}{ccc|ccc} \tilde{\omega}_{11} & \tilde{\omega}_{12} & \tilde{\omega}_{13} & \tilde{\omega}_{14} & \tilde{\omega}_{15} & \tilde{\omega}_{16} \\ \tilde{\omega}_{12} & \tilde{\omega}_{22} & \tilde{\omega}_{23} & \tilde{\omega}_{24} & \tilde{\omega}_{25} & \tilde{\omega}_{26} \\ \tilde{\omega}_{13} & \tilde{\omega}_{23} & \tilde{\omega}_{33} & \tilde{\omega}_{34} & \tilde{\omega}_{35} & \tilde{\omega}_{36} \\ \hline \tilde{\omega}_{14} & \tilde{\omega}_{24} & \tilde{\omega}_{34} & \tilde{\omega}_{44} & \tilde{\omega}_{45} & \tilde{\omega}_{46} \\ \tilde{\omega}_{15} & \tilde{\omega}_{25} & \tilde{\omega}_{35} & \tilde{\omega}_{45} & \tilde{\omega}_{55} & \tilde{\omega}_{56} \\ \tilde{\omega}_{16} & \tilde{\omega}_{26} & \tilde{\omega}_{36} & \tilde{\omega}_{46} & \tilde{\omega}_{56} & \tilde{\omega}_{66} \end{array}\right)$$

where:

- $\gamma = a^2c_1^2 + a^2c_3^2 + a^4 - 2a^2 - c_1^2 + c_1^2c_2^2 - c_2^2 + c_1^2c_3^2 - c_1^2c_2^2c_3^2 + c_2^2c_3^2 - c_3^2 + 1$
- $\tilde{\omega}_{11} = -a^2 - c_2^2 + c_2^2c_3^2 - c_3^2 + 1$
- $\tilde{\omega}_{12} = a^3 + ac_3^2 - a$
- $\tilde{\omega}_{13} = a^2c_2c_3$
- $\tilde{\omega}_{14} = a^2c_1 + c_1c_2^2 - c_1c_2^2c_3^2 + c_1c_3^2 - c_1$
- $\tilde{\omega}_{15} = ac_2 - ac_2c_3^2$
- $\tilde{\omega}_{16} = -a^2c_2$
- $\tilde{\omega}_{22} = a^2c_1^2 - a^2 - c_1^2 + c_1^2c_3^2 - c_3^2 + 1$
- $\tilde{\omega}_{23} = ac_1^2c_2c_3 - ac_2c_3$
- $\tilde{\omega}_{24} = -a^3c_1 - ac_1c_3^2 + ac_1$
- $\tilde{\omega}_{25} = -c_2c_3^2c_1^2 + c_2c_1^2 + c_2c_3^2 - c_2$

- $\tilde{\omega}_{26} = ac_2 - ac_1^2c_2$
- $\tilde{\omega}_{33} = a^2c_1^2 + a^4 - 2a^2 - c_1^2 + c_1^2c_2^2 - c_2^2 + 1$
- $\tilde{\omega}_{34} = -a^2c_1c_2c_3$
- $\tilde{\omega}_{35} = -a^3c_3 - ac_1^2c_3 + ac_3$
- $\tilde{\omega}_{36} = a^2c_3 + c_1^2c_3 - c_1^2c_2^2c_3 + c_2^2c_3 - c_3$
- $\tilde{\omega}_{44} = a^2c_3^2 + a^4 - 2a^2 - c_2^2 + c_2^2c_3^2 - c_3^2 + 1$
- $\tilde{\omega}_{45} = ac_1c_2c_3^2 - ac_1c_2$
- $\tilde{\omega}_{46} = a^2c_1c_2$
- $\tilde{\omega}_{55} = a^2c_3^2 - a^2 - c_1^2 + c_1^2c_3^2 - c_3^2 + 1$
- $\tilde{\omega}_{56} = a^3 + ac_1^2 - a$
- $\tilde{\omega}_{66} = -a^2 - c_1^2 + c_1^2c_2^2 - c_2^2 + 1$

Claim 1. *$E^* \subsetneqq E$ if a, c_i are not zero and so that $|a| + |c_i| < 1$ for $i = 1, \ldots, 3$*

Proof. First, if $c_i = 0$ for $i = 1, 2, 3$ then partial separability holds. Thus, $E_1 = \{(1,2)\}, E_2 = \{(2,3)\}$ and $E^* = E_1 \cup E_2 = \{(1,2),(2,3)\}$.

On the other hand, if $c_i \neq 0$ for $i = 1, 2, 3$ then partial separability does not hold. Sort the rows and columns of Ω in a features-first ordering as:

$$\gamma^{-1} \left(\begin{array}{cc:cc:cc}
\tilde{\omega}_{11} & \tilde{\omega}_{14} & \tilde{\omega}_{12} & \tilde{\omega}_{15} & \tilde{\omega}_{13} & \tilde{\omega}_{16} \\
\tilde{\omega}_{14} & \tilde{\omega}_{44} & \tilde{\omega}_{24} & \tilde{\omega}_{45} & \tilde{\omega}_{34} & \tilde{\omega}_{46} \\
\hdashline
\tilde{\omega}_{12} & \tilde{\omega}_{24} & \tilde{\omega}_{22} & \tilde{\omega}_{25} & \tilde{\omega}_{23} & \tilde{\omega}_{26} \\
\tilde{\omega}_{15} & \tilde{\omega}_{45} & \tilde{\omega}_{25} & \tilde{\omega}_{55} & \tilde{\omega}_{35} & \tilde{\omega}_{56} \\
\hdashline
\tilde{\omega}_{13} & \tilde{\omega}_{34} & \tilde{\omega}_{23} & \tilde{\omega}_{35} & \tilde{\omega}_{33} & \tilde{\omega}_{36} \\
\tilde{\omega}_{16} & \tilde{\omega}_{46} & \tilde{\omega}_{26} & \tilde{\omega}_{56} & \tilde{\omega}_{36} & \tilde{\omega}_{66}
\end{array}\right)$$

We first prove that $(1,2) \in E$. If $|a| + |c_i| < 1$ then $a^2 + c_i^2 < 1$ for $i = 1, 2, 3$ and so $\tilde{\omega}_{12} \neq 0$. Thus, the $(1,2)$-block is not zero, so $(1,2) \in E$.

Next, we prove that $(1,3) \in E$. $\tilde{\omega}_{34} \neq 0$ since a, c_i are not zero for $j = 1, 2, 3$. Thus, the $(1,3)$-block is not zero, so $(2,3) \in E$.

Finally, we prove that $(2,3) \in E$. If $|a| + |c_i| < 1$ then $a^2 + c_i^2 < 1$ for $i = 1, 2, 3$ and so $\tilde{\omega}_{12} \neq 0$. Thus, the $(2,3)$-block is not zero, so $(2,3) \in E$.

□

A.9.2 Example 2

Consider the following block covariance matrix and its inverse:

$$\Sigma = \left(\begin{array}{ccc:ccc} 1 & a & 0 & c_1 & 0 & 0 \\ a & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & c_3 \\ \hdashline c_1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & a \\ 0 & 0 & c_3 & 0 & a & 1 \end{array}\right) \qquad \Omega = \Sigma^{-1} = \gamma^{-1} \left(\begin{array}{ccc:ccc} \tilde{\omega}_{11} & \tilde{\omega}_{12} & 0 & \tilde{\omega}_{14} & 0 & 0 \\ \tilde{\omega}_{12} & \tilde{\omega}_{22} & 0 & \tilde{\omega}_{24} & 0 & 0 \\ 0 & 0 & \tilde{\omega}_{33} & 0 & \tilde{\omega}_{35} & \tilde{\omega}_{36} \\ \hdashline \tilde{\omega}_{14} & \tilde{\omega}_{24} & 0 & \tilde{\omega}_{44} & 0 & 0 \\ 0 & 0 & \tilde{\omega}_{35} & 0 & \tilde{\omega}_{55} & \tilde{\omega}_{56} \\ 0 & 0 & \tilde{\omega}_{36} & 0 & \tilde{\omega}_{56} & \tilde{\omega}_{66} \end{array}\right)$$

with:

- $\gamma = a^2c_1^2 + a^2c_3^2 + a^4 - 2a^2 - c_1^2 + c_1^2c_3^2 - c_3^2 + 1$
- $\tilde{\omega}_{11} = -a^2 - c_3^2 + 1$
- $\tilde{\omega}_{12} = a^3 + ac_3^2 - a$
- $\tilde{\omega}_{14} = a^2c_1 + c_1c_3^2 - c_1$
- $\tilde{\omega}_{22} = a^2c_1^2 - a^2 - c_1^2 + c_1^2c_3^2 - c_3^2 + 1$
- $\tilde{\omega}_{24} = -a^3c_1 - ac_1c_3^2 + ac_1$
- $\tilde{\omega}_{33} = a^2c_1^2 + a^4 - 2a^2 - c_1^2 + 1$
- $\tilde{\omega}_{35} = -a^3c_3 - ac_1^2c_3 + ac_3$
- $\tilde{\omega}_{36} = a^2c_3 + c_1^2c_3 - c_3$
- $\tilde{\omega}_{44} = a^2c_3^2 + a^4 - 2a^2 - c_3^2 + 1$
- $\tilde{\omega}_{55} = a^2c_3^2 - a^2 - c_1^2 + c_1^2c_3^2 - c_3^2 + 1$
- $\tilde{\omega}_{56} = a^3 + ac_1^2 - a$
- $\tilde{\omega}_{66} = -a^2 - c_1^2 + 1$

Claim 2. *$E^* = E$ if a, c_i are not zero and so that $|a| + |c_i| < 1$ for $i = 1, \ldots, 3$*

Proof. First, if $c_i = 0$ for $i = 1, 3$ then partial separability holds. Thus, for $|a| < 1$ we have $E_1 = \{(1, 2)\}, E_2 = \{(2, 3)\}$ and $E^* = E_1 \cup E_2 = \{(1, 2), (2, 3)\}$.

On the other hand, if $c_i \neq 0$ for $i = 1, 3$ then partial separability does not hold. Sort the rows and columns of Ω in a features-first ordering as:

$$\gamma^{-1} \left(\begin{array}{cc|cc|cc} \tilde{\omega}_{11} & \tilde{\omega}_{14} & \tilde{\omega}_{12} & 0 & 0 & 0 \\ \tilde{\omega}_{14} & \tilde{\omega}_{44} & \tilde{\omega}_{24} & 0 & 0 & 0 \\ \hline \tilde{\omega}_{12} & \tilde{\omega}_{24} & \tilde{\omega}_{22} & 0 & 0 & 0 \\ 0 & 0 & 0 & \tilde{\omega}_{55} & \tilde{\omega}_{35} & \tilde{\omega}_{56} \\ \hline 0 & 0 & 0 & \tilde{\omega}_{35} & \tilde{\omega}_{33} & \tilde{\omega}_{36} \\ 0 & 0 & 0 & \tilde{\omega}_{56} & \tilde{\omega}_{36} & \tilde{\omega}_{66} \end{array} \right)$$

We first prove that $(1, 2) \in E$. If $|a| + |c_i| < 1$ then $a^2 + c_i^2 < 1$ for $i = 1, 3$ and so $\tilde{\omega}_{12} \neq 0$. Thus, the $(1, 2)$-block is not zero, so $(1, 2) \in E$.

Next, we prove that $(2, 3) \in E$. $\tilde{\omega}_{34} \neq 0$ since a, c_i are not zero for $j = 1, 2, 3$. And $\tilde{\omega}_{56} \neq 0$ since $|a| + |c_1| < 1$. Thus, the $(2, 3)$-block is not zero, so $(2, 3) \in E$.

We prove that $E^* \subset E$. $\tilde{\omega}_{16} \neq 0$ since a, c_2 are not zero. Thus, the $(1, 3)$-block is not zero, so $(1, 3) \in E$ but $(1, 3) \notin E$.

And since the $(1, 3)$-block is zero, then $E^* = E$.

□

A.9.3 Example 3

Consider the following block covariance matrix and its inverse:

$$\Sigma = \left(\begin{array}{ccc:ccc} 1 & a & b & 0 & 0 & 0 \\ a & 1 & a & 0 & 0 & c \\ b & a & 1 & 0 & c & 0 \\ \hdashline 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & c & 0 & 1 & 0 \\ 0 & c & 0 & 0 & 0 & 1 \end{array}\right) \qquad \Omega = \Sigma^{-1} = \gamma^{-1} \left(\begin{array}{ccc:ccc} \tilde{\omega}_{11} & \tilde{\omega}_{12} & \tilde{\omega}_{13} & 0 & \tilde{\omega}_{15} & \tilde{\omega}_{16} \\ \tilde{\omega}_{12} & \tilde{\omega}_{22} & \tilde{\omega}_{23} & 0 & \tilde{\omega}_{25} & \tilde{\omega}_{26} \\ \tilde{\omega}_{13} & \tilde{\omega}_{23} & \tilde{\omega}_{33} & 0 & \tilde{\omega}_{35} & \tilde{\omega}_{36} \\ \hdashline 0 & 0 & 0 & \tilde{\omega}_{44} & 0 & 0 \\ \tilde{\omega}_{15} & \tilde{\omega}_{25} & \tilde{\omega}_{35} & 0 & \tilde{\omega}_{55} & \tilde{\omega}_{56} \\ \tilde{\omega}_{16} & \tilde{\omega}_{26} & \tilde{\omega}_{36} & 0 & \tilde{\omega}_{56} & \tilde{\omega}_{66} \end{array}\right)$$

with:

- $\gamma = 2a^2b + a^2c^2 - 2a^2 + b^2c^2 - b^2 + c^4 - 2c^2 + 1$
- $\tilde{\omega}_{11} = -a^2 + c^4 - 2c^2 + 1$
- $\tilde{\omega}_{12} = ab + ac^2 - a$
- $\tilde{\omega}_{13} = a^2 + bc^2 - b$
- $\tilde{\omega}_{15} = -a^2c - bc^3 + bc$
- $\tilde{\omega}_{16} = -abc - ac^3 + ac$
- $\tilde{\omega}_{22} = -b^2 - c^2 + 1$
- $\tilde{\omega}_{23} = ab - a$
- $\tilde{\omega}_{25} = ac - abc$
- $\tilde{\omega}_{26} = b^2c + c^3 - c$
- $\tilde{\omega}_{33} = -a^2 - c^2 + 1$
- $\tilde{\omega}_{35} = a^2c + c^3 - c$
- $\tilde{\omega}_{36} = ac - abc$
- $\tilde{\omega}_{44} = 1$
- $\tilde{\omega}_{55} = 2a^2b - 2a^2 + b^2c^2 - b^2 - c^2 + 1$
- $\tilde{\omega}_{56} = abc^2 - ac^2$
- $\tilde{\omega}_{66} = 2a^2b + a^2c^2 - 2a^2 - b^2 - c^2 + 1$

Claim 3. *$E \subsetneq E^*$ if a, b, c are not zero and $a = b = (1 - c^2)$*

Proof. First, if $c = 0$ we can find nonzero a and b so that partial separability holds. Thus,

$E_1 = \{(1,2),(1,3),(2,3)\}$ and $E^* = E_1 \cup E_2 = \{(1,2),(1,3),(2,3)\}$.

On the other hand, if $c \neq 0$ then partial separability does not hold. Sort the rows and columns of Ω in a features-first ordering as:

$$\gamma^{-1} \left(\begin{array}{cc|cc|cc} \tilde{\omega}_{11} & 0 & \tilde{\omega}_{12} & \tilde{\omega}_{15} & \tilde{\omega}_{13} & \tilde{\omega}_{16} \\ 0 & \tilde{\omega}_{44} & 0 & 0 & 0 & 0 \\ \hline \tilde{\omega}_{12} & 0 & \tilde{\omega}_{22} & \tilde{\omega}_{25} & \tilde{\omega}_{23} & \tilde{\omega}_{26} \\ \tilde{\omega}_{15} & 0 & \tilde{\omega}_{25} & \tilde{\omega}_{55} & \tilde{\omega}_{35} & \tilde{\omega}_{56} \\ \hline \tilde{\omega}_{13} & 0 & \tilde{\omega}_{23} & \tilde{\omega}_{35} & \tilde{\omega}_{33} & \tilde{\omega}_{36} \\ \tilde{\omega}_{16} & 0 & \tilde{\omega}_{26} & \tilde{\omega}_{56} & \tilde{\omega}_{36} & \tilde{\omega}_{66} \end{array} \right)$$

We prove that $(1,2) \in E$. If $b = (1 - c^2)$ then $\tilde{\omega}_{12} = 0$. In addition, if $b = a$ then $\tilde{\omega}_{125} = 0$. Thus, the $(1,2)$-block is not zero, so $(1,2) \in E$.

□

A.10 Simulation Details and Additional Figures for Section 3.4

A.10.1 Simulation Details for Section 3.4

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 3.4. 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 $E = \bigcup_{l=1}^{M} E_l$. 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 2.

Next, $p \times p$ precision matrices Ω_l are generated for each E_l based on the algorithm of [44]. Let $\tilde{\Omega}_l$ be a $p \times p$ matrix with entries

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

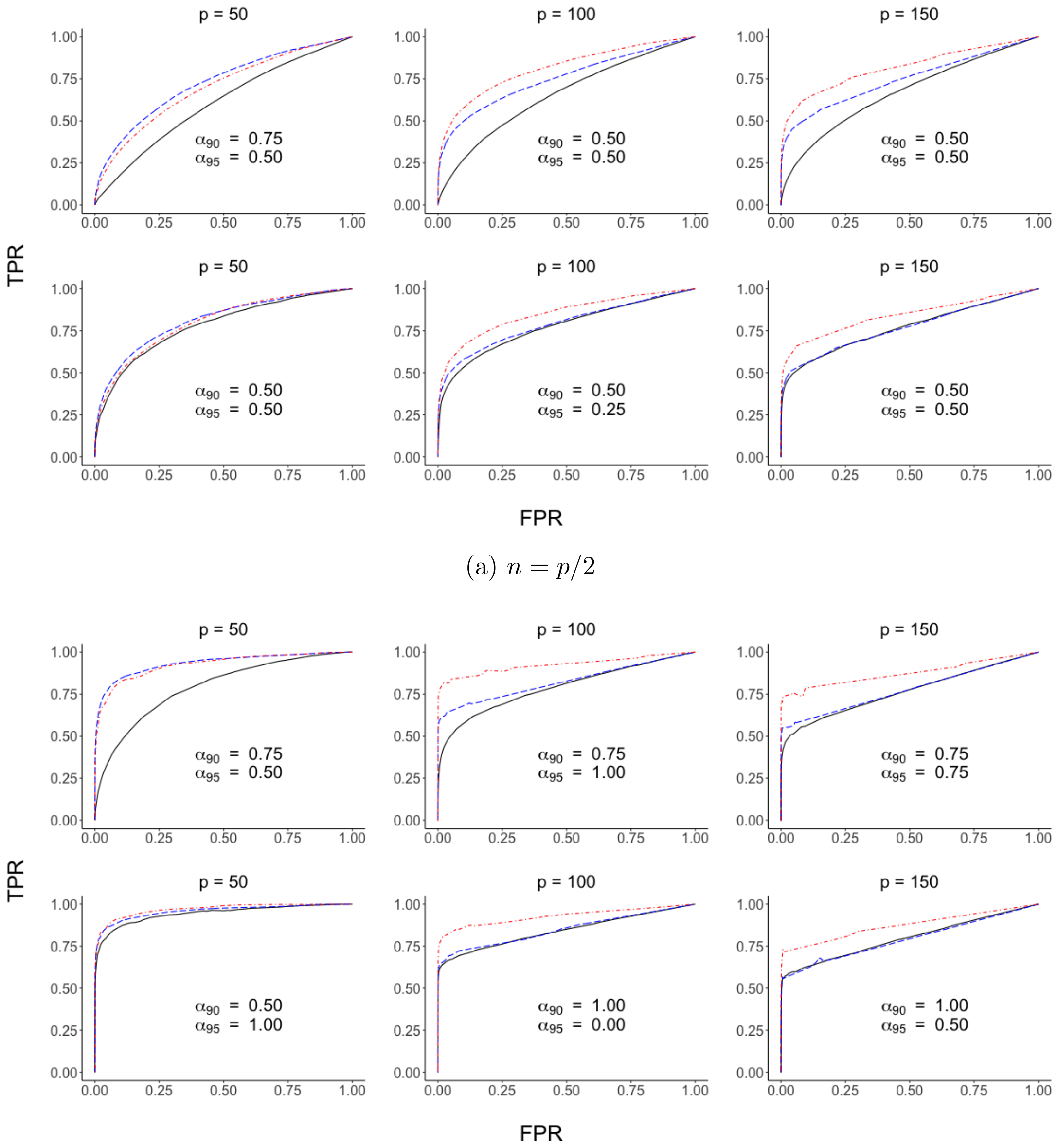
where $\mathcal{D} = [-2/3, -1/3] \cup [1/3, 2/3]$. Then, rowwise, we sum the absolute value of the off-diagonal entries and divide each row by 1.5 times this sum componentwise. Finally, $\tilde{\Omega}_l$ is averaged with its transpose and has its diagonal entries set to one. The output is a precision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

Algorithm 2: Pseudocode to create the edge sets $E_1, \ldots, E_M$

```
Input: graph $G = (V, E)$ with nodes $V = \{1, \ldots, p\}$ and edge set $E$
       number of basis $M$
       proportion of common edges $\tau$
Result: Edge sets $E_1, \ldots, E_M$
$E_c \leftarrow$ random subset of size $\tau|E|$ from $E$;
$E_l \leftarrow E_c$ for $l = 1, \ldots, M$;
$l \leftarrow 1$;
$B \leftarrow 1$;
for $e \in E \setminus E_c$ do
    $E_l \leftarrow E_l \bigcup e$;
    $l \leftarrow l + 1$;
    if $l > B$ then
        $l \leftarrow 1$;
        $B \leftarrow (B + 1) \mod M$;
```

A.11 Additional Results for Section 3.4.2

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.



(a) $n = p/2$

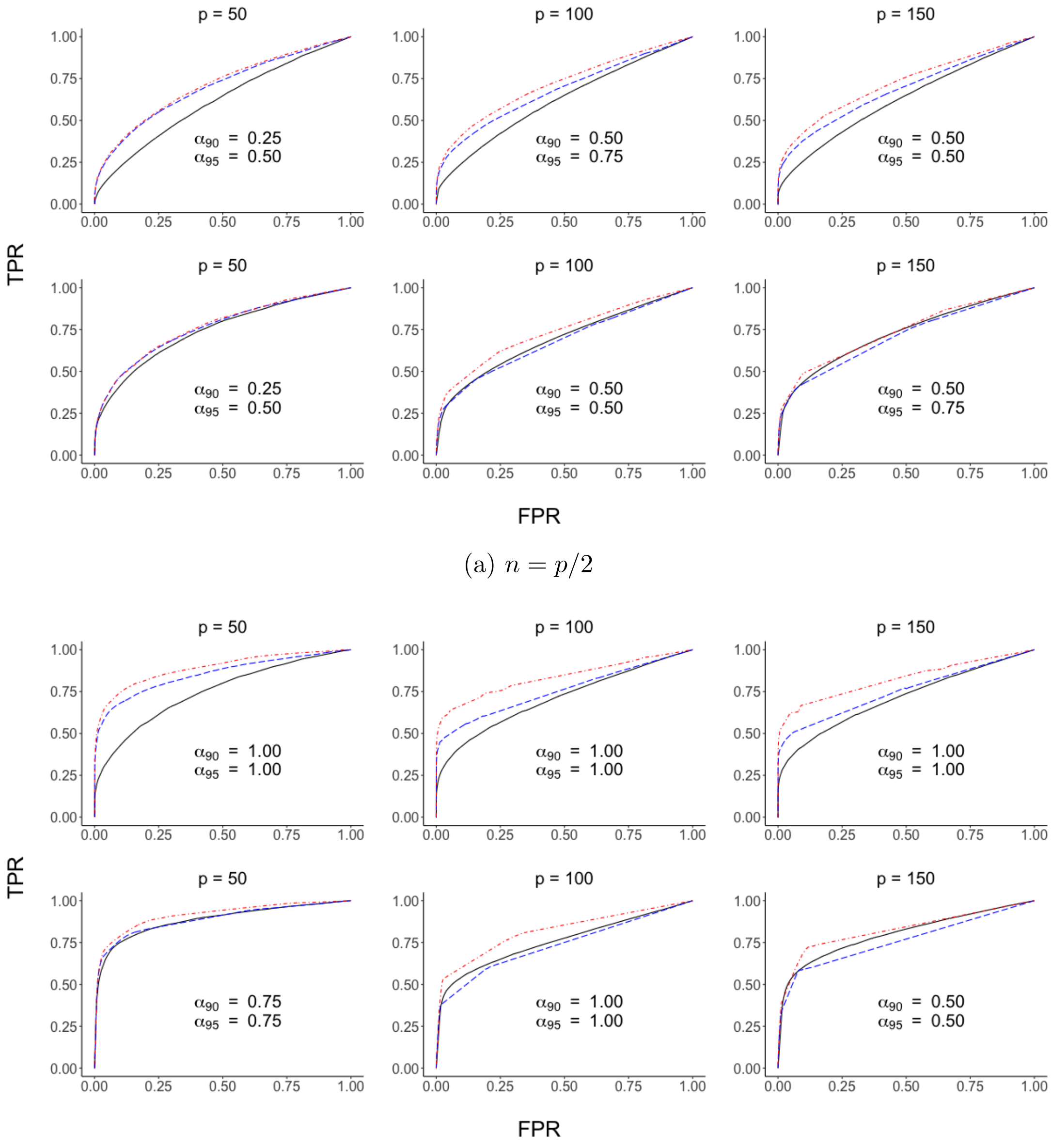
(b) $n = 1.5p$

Figure A.6: Mean receiver operating characteristic curves for the proposed method (FGMParty) with that of [17] (FGGM) under Σ_{ps} (top) and $\Sigma_{\text{non-ps}}$ (bottom) for $p = 50, 100, 150$, $\pi = 0.025$ and $\tau = 0$. We see FGMParty (- - - -) and FGGM (——) at 90% of variance and FGMParty (—·—) at 95% of variance explained.

Table A.1: Mean area under the curve (and standard error) values for Figures *A*.6*a* and *A*.6*b*

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

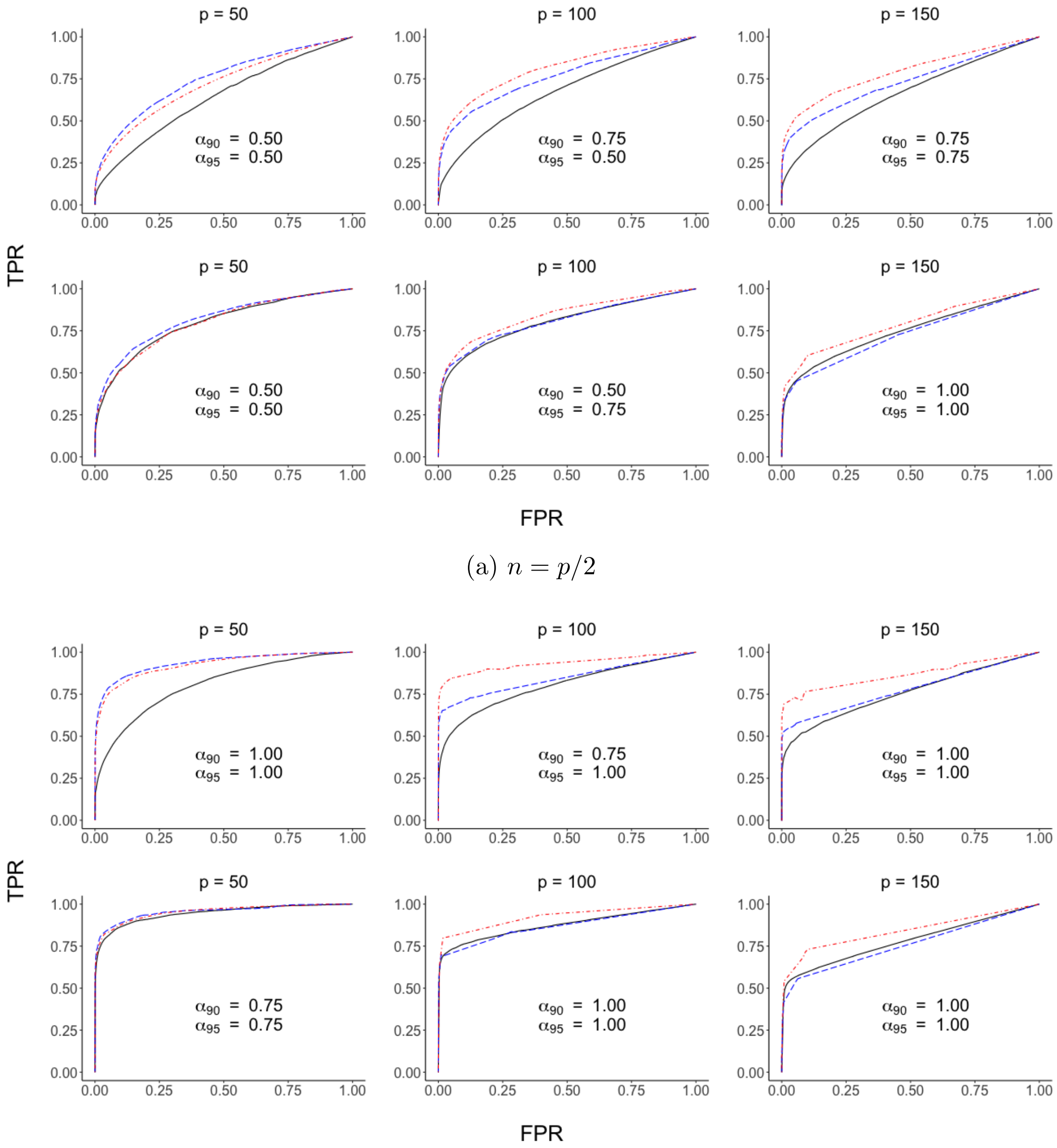
(b) $n = 1.5p$

Figure A.7: Mean receiver operating characteristic curves for the proposed method (FGMParty) with that of [17] (FGGM) under Σ_{ps} (top) and $\Sigma_{\text{non-ps}}$ (bottom) for $p = 50, 100, 150$, $\pi = 0.05$ and $\tau = 0.1$. We see FGMParty (- - - -) and FGGM (——) at 90% of variance and FGMParty (—·—) at 95% of variance explained.

Table A.2: Mean area under the curve (and standard error) values for Figures *A.7a* and *A.7b*

			$\Sigma = \Sigma_{\text{ps}}$			$\Sigma = \Sigma_{\text{non-ps}}$		
			$p = 50$	$p = 100$	$p = 150$	$p = 50$	$p = 100$	$p = 150$
$n = p/2$	AUC	FGGM$_{90\%}$	0.61(0.03)	0.62(0.02)	0.62(0.01)	0.74(0.03)	**0.69(0.02)**	**0.72(0.01)**
		FGMParty$_{90\%}$	**0.70(0.03)**	**0.68(0.02)**	**0.68(0.01)**	**0.76(0.03)**	0.68(0.02)	0.70(0.02)
		FGMParty$_{95\%}$	0.71(0.03)	0.71(0.02)	0.72(0.01)	0.76(0.03)	0.73(0.02)	0.73(0.02)
	AUC15†	FGGM$_{90\%}$	0.18(0.04)	0.20(0.02)	0.22(0.01)	0.35(0.04)	0.32(0.02)	0.35(0.02)
		FGMParty$_{90\%}$	**0.30(0.04)**	**0.32(0.02)**	**0.33(0.01)**	**0.39(0.05)**	**0.33(0.02)**	**0.36(0.02)**
		FGMParty$_{95\%}$	0.31(0.04)	0.35(0.03)	0.37(0.02)	0.39(0.05)	0.39(0.03)	0.40(0.03)
$n = 1.5p$	AUC	FGGM$_{90\%}$	0.75(0.02)	0.71(0.02)	0.71(0.01)	**0.88(0.02)**	**0.76(0.01)**	**0.80(0.01)**
		FGMParty$_{90\%}$	**0.85(0.03)**	**0.75(0.02)**	**0.75(0.01)**	**0.88(0.02)**	0.73(0.01)	0.76(0.01)
		FGMParty$_{95\%}$	0.89(0.02)	0.83(0.02)	0.82(0.01)	0.91(0.02)	0.81(0.02)	0.81(0.02)
	AUC15†	FGGM$_{90\%}$	0.37(0.05)	0.37(0.02)	0.38(0.01)	0.67(0.04)	**0.48(0.02)**	**0.52(0.01)**
		FGMParty$_{90\%}$	**0.62(0.04)**	**0.50(0.02)**	**0.50(0.02)**	**0.68(0.04)**	0.43(0.02)	0.49(0.02)
		FGMParty$_{95\%}$	0.68(0.04)	0.63(0.03)	0.61(0.02)	0.70(0.04)	0.54(0.03)	0.56(0.03)

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



(a) $n = p/2$

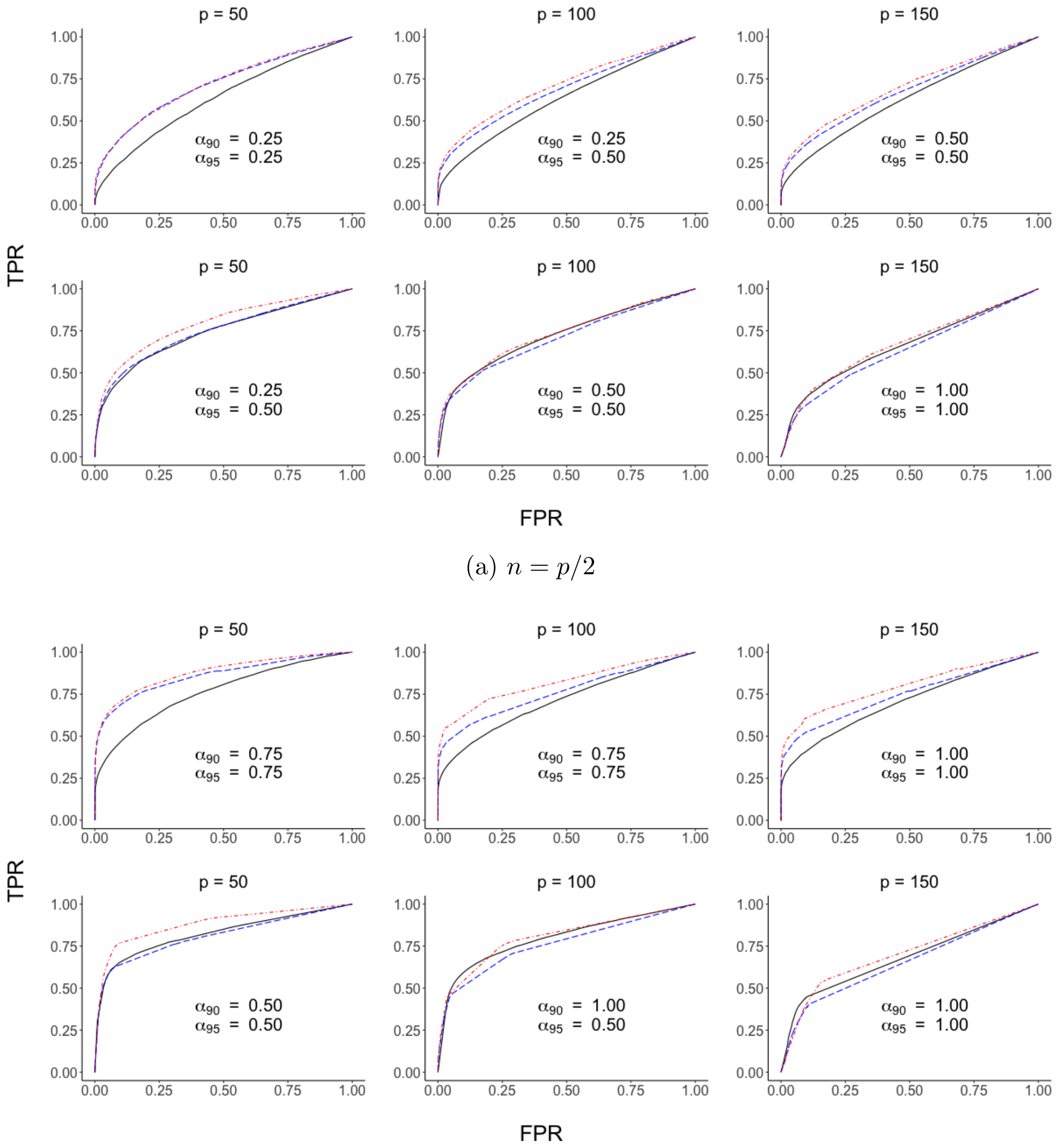
(b) $n = 1.5p$

Figure A.8: Mean receiver operating characteristic curves for the proposed method (FGMParty) with that of [17] (FGGM) under Σ_{ps} (top) and $\Sigma_{\text{non-ps}}$ (bottom) for $p = 50, 100, 150$, $\pi = 0.025$ and $\tau = 0.1$. We see FGMParty (- - - -) and FGGM (——) at 90% of variance and FGMParty (—·—) at 95% of variance explained.

Table A.3: Mean area under the curve (and standard error) values for Figures $A.8a$ and $A.8b$

			$\Sigma = \Sigma_{\text{ps}}$			$\Sigma = \Sigma_{\text{non-ps}}$		
			$p = 50$	$p = 100$	$p = 150$	$p = 50$	$p = 100$	$p = 150$
$n = p/2$	AUC	$\text{FGGM}_{90\%}$	0.64(0.05)	0.67(0.02)	0.67(0.02)	0.80(0.05)	0.80(0.02)	**0.75(0.01)**
		$\text{FGMParty}_{90\%}$	**0.75(0.05)**	**0.76(0.03)**	**0.73(0.02)**	**0.82(0.04)**	**0.81(0.02)**	0.73(0.02)
		$\text{FGMParty}_{95\%}$	0.72(0.05)	0.81(0.02)	0.79(0.02)	0.79(0.05)	0.83(0.02)	0.78(0.01)
	AUC15†	$\text{FGGM}_{90\%}$	0.22(0.06)	0.26(0.03)	0.28(0.02)	0.44(0.07)	0.51(0.03)	**0.46(0.02)**
		$\text{FGMParty}_{90\%}$	**0.37(0.07)**	**0.45(0.04)**	**0.44(0.02)**	**0.49(0.08)**	**0.54(0.04)**	0.44(0.02)
		$\text{FGMParty}_{95\%}$	0.33(0.07)	0.50(0.04)	0.52(0.03)	0.45(0.07)	0.57(0.04)	0.52(0.03)
$n = 1.5p$	AUC	$\text{FGGM}_{90\%}$	0.81(0.04)	0.80(0.02)	0.76(0.02)	0.94(0.02)	**0.87(0.02)**	**0.78(0.01)**
		$\text{FGMParty}_{90\%}$	**0.93(0.02)**	**0.84(0.02)**	**0.78(0.01)**	**0.95(0.02)**	**0.87(0.02)**	0.76(0.01)
		$\text{FGMParty}_{95\%}$	0.92(0.03)	0.93(0.02)	0.86(0.02)	0.95(0.02)	0.92(0.02)	0.83(0.02)
	AUC15†	$\text{FGGM}_{90\%}$	0.45(0.07)	0.53(0.03)	0.49(0.02)	0.82(0.05)	**0.73(0.03)**	**0.56(0.02)**
		$\text{FGMParty}_{90\%}$	**0.78(0.06)**	**0.68(0.03)**	**0.57(0.02)**	**0.84(0.04)**	0.72(0.04)	0.53(0.02)
		$\text{FGMParty}_{95\%}$	0.74(0.06)	0.83(0.03)	0.72(0.03)	0.82(0.05)	0.79(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) $n = p/2$

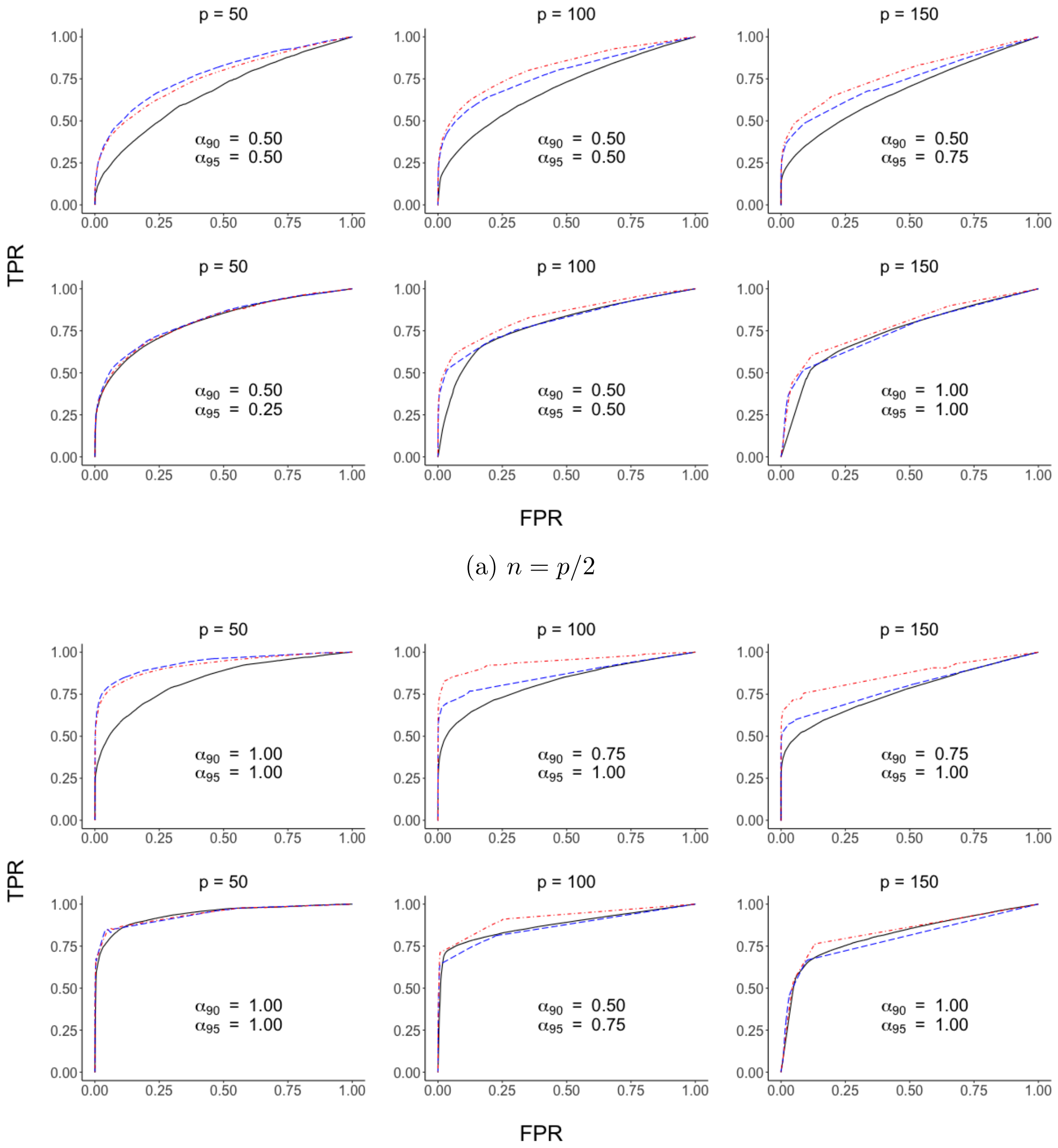
(b) $n = 1.5p$

Figure A.9: Mean receiver operating characteristic curves for the proposed method (FGMParty) with that of [17] (FGGM) under Σ_{ps} (top) and $\Sigma_{\text{non-ps}}$ (bottom) for $p = 50, 100, 150$, $\pi = 0.05$ and $\tau = 0.2$. We see FGMParty (- - - -) and FGGM (——) at 90% of variance and FGMParty (—·—) at 95% of variance explained.

Table A.4: Mean area under the curve (and standard error) values for Figures *A.9a* and *A.9b*

			$\Sigma = \Sigma_{\text{ps}}$			$\Sigma = \Sigma_{\text{non-ps}}$		
			$p = 50$	$p = 100$	$p = 150$	$p = 50$	$p = 100$	$p = 150$
$n = p/2$	AUC	$\text{FGGM}_{90\%}$	0.63(0.04)	0.63(0.02)	0.62(0.01)	0.74(0.03)	**0.72(0.02)**	**0.66(0.01)**
		$\text{FGMParty}_{90\%}$	**0.71(0.04)**	**0.68(0.02)**	**0.67(0.01)**	**0.75(0.03)**	0.70(0.02)	0.63(0.01)
		$\text{FGMParty}_{95\%}$	0.72(0.03)	0.71(0.02)	0.69(0.01)	0.79(0.03)	0.73(0.02)	0.67(0.01)
	AUC15†	$\text{FGGM}_{90\%}$	0.21(0.04)	0.22(0.03)	0.23(0.01)	0.39(0.04)	0.35(0.02)	**0.28(0.01)**
		$\text{FGMParty}_{90\%}$	**0.34(0.05)**	**0.32(0.02)**	**0.31(0.01)**	**0.41(0.03)**	**0.36(0.02)**	0.24(0.01)
		$\text{FGMParty}_{95\%}$	0.34(0.04)	0.35(0.02)	0.34(0.02)	0.45(0.04)	0.38(0.02)	0.27(0.02)
$n = 1.5p$	AUC	$\text{FGGM}_{90\%}$	0.76(0.03)	0.71(0.02)	0.70(0.01)	**0.82(0.02)**	**0.79(0.02)**	**0.67(0.02)**
		$\text{FGMParty}_{90\%}$	**0.86(0.02)**	**0.76(0.02)**	**0.74(0.01)**	0.81(0.02)	0.76(0.02)	0.65(0.01)
		$\text{FGMParty}_{95\%}$	0.87(0.02)	0.82(0.02)	0.79(0.01)	0.88(0.02)	0.80(0.02)	0.70(0.01)
	AUC15†	$\text{FGGM}_{90\%}$	0.41(0.04)	0.38(0.02)	0.37(0.01)	**0.57(0.03)**	**0.48(0.02)**	**0.34(0.02)**
		$\text{FGMParty}_{90\%}$	**0.63(0.04)**	**0.49(0.02)**	**0.47(0.02)**	0.56(0.03)	0.44(0.02)	0.29(0.01)
		$\text{FGMParty}_{95\%}$	0.64(0.04)	0.58(0.03)	0.54(0.02)	0.64(0.04)	0.47(0.02)	0.30(0.02)

†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 A.10: Mean receiver operating characteristic curves for the proposed method (FGMParty) with that of [17] (FGGM) under Σ_{ps} (top) and $\Sigma_{\text{non-ps}}$ (bottom) for $p = 50, 100, 150$, $\pi = 0.025$ and $\tau = 0.2$. We see FGMParty (- - - -) and FGGM (——) at 90% of variance and FGMParty (—·—) at 95% of variance explained.

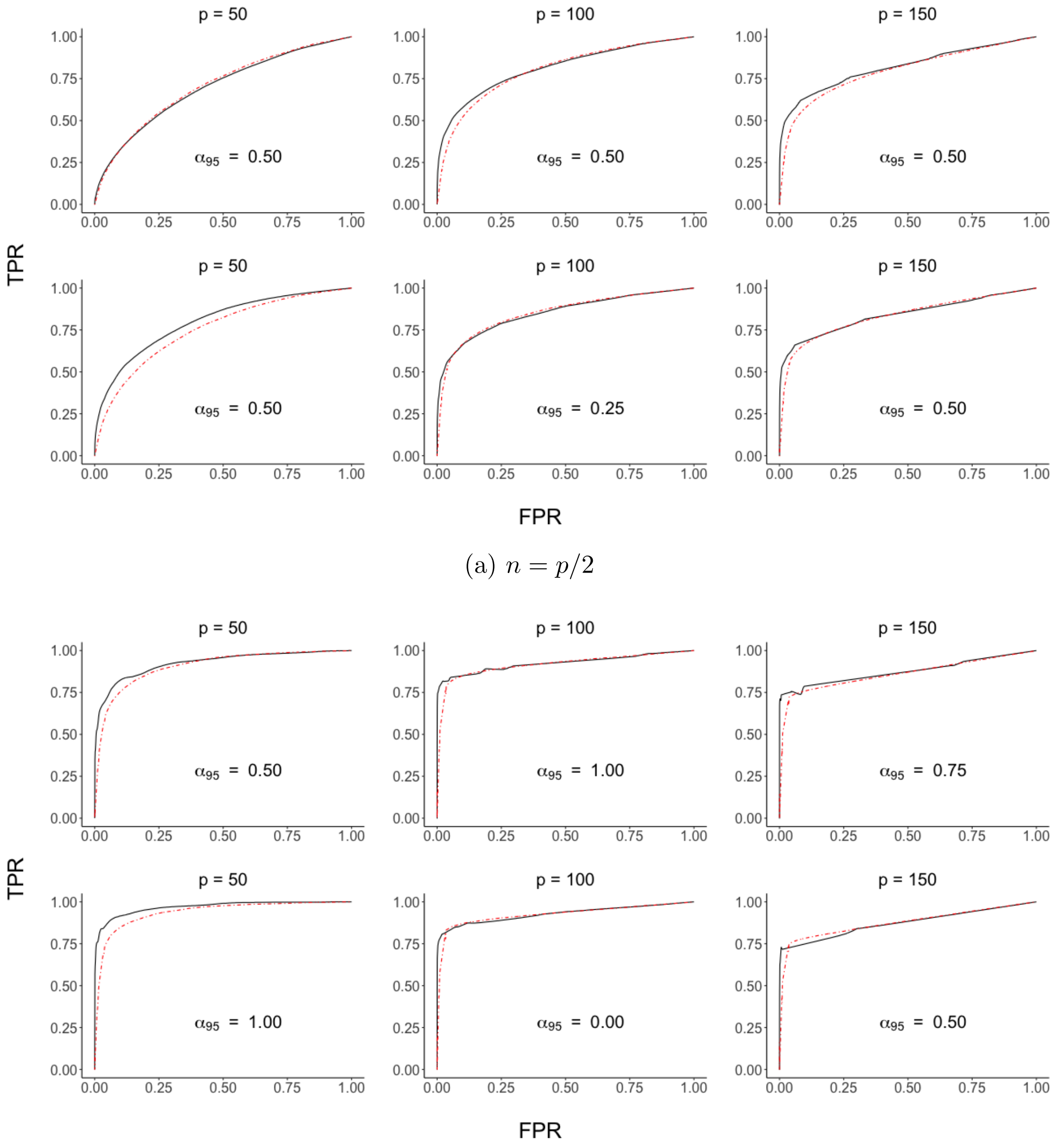
Table A.5: Mean area under the curve (and standard error) values for Figures $A.10a$ and $A.10b$

			$\Sigma = \Sigma_{\text{ps}}$			$\Sigma = \Sigma_{\text{non-ps}}$		
			$p = 50$	$p = 100$	$p = 150$	$p = 50$	$p = 100$	$p = 150$
$n = p/2$	AUC	$\text{FGGM}_{90\%}$	0.67(0.05)	0.69(0.02)	0.68(0.02)	0.81(0.04)	0.77(0.02)	0.69(0.06)
		$\text{FGMParty}_{90\%}$	**0.78(0.05)**	**0.78(0.02)**	**0.74(0.02)**	**0.81(0.04)**	**0.80(0.02)**	**0.75(0.02)**
		$\text{FGMParty}_{95\%}$	0.76(0.05)	0.81(0.02)	0.78(0.02)	0.80(0.04)	0.83(0.02)	0.78(0.01)
	AUC15†	$\text{FGGM}_{90\%}$	0.25(0.06)	0.30(0.03)	0.31(0.02)	0.48(0.07)	0.31(0.07)	0.22(0.09)
		$\text{FGMParty}_{90\%}$	**0.43(0.07)**	**0.48(0.03)**	**0.44(0.02)**	**0.49(0.06)**	**0.53(0.04)**	**0.43(0.02)**
		$\text{FGMParty}_{95\%}$	0.41(0.06)	0.51(0.03)	0.49(0.03)	0.47(0.06)	0.57(0.03)	0.46(0.02)
$n = 1.5p$	AUC	$\text{FGGM}_{90\%}$	0.83(0.04)	0.82(0.02)	0.76(0.01)	0.94(0.02)	**0.87(0.02)**	0.78(0.05)
		$\text{FGMParty}_{90\%}$	**0.93(0.03)**	**0.86(0.02)**	**0.79(0.01)**	**0.94(0.03)**	0.86(0.02)	**0.79(0.02)**
		$\text{FGMParty}_{95\%}$	0.92(0.03)	0.94(0.02)	0.86(0.02)	0.93(0.02)	0.91(0.01)	0.84(0.01)
	AUC15†	$\text{FGGM}_{90\%}$	0.52(0.06)	0.56(0.02)	0.50(0.02)	0.79(0.05)	**0.69(0.02)**	0.40(0.12)
		$\text{FGMParty}_{90\%}$	**0.79(0.06)**	**0.71(0.03)**	**0.59(0.02)**	**0.81(0.06)**	0.68(0.03)	**0.53(0.02)**
		$\text{FGMParty}_{95\%}$	0.77(0.06)	0.84(0.03)	0.71(0.02)	0.80(0.06)	0.75(0.03)	0.55(0.02)

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

A.12 Additional Results for Section 3.4.3

This section provides additional results comparing the FGMParty method against an independence screening procedure psSCREEN as described in Section 3.4.3. Figures A.11a and A.11b follow the very sparse case settings with $\pi = 0.025$ and $\tau = 0$.



(a) $n = p/2$

(b) $n = 1.5p$

Figure A.11: Mean receiver operating characteristic curves for the proposed method (FGMParty) and the independence screening procedure (psSCREEN) under Σ_{ps} (top) and $\Sigma_{\text{non-ps}}$ (bottom) for $p = 50, 100, 150$, $\pi = 0.025$ and $\tau = 0$. We see FGMParty (——) and psSCREEN(—·—) both at 95% of variance explained for the very sparse case.

Table A.6: Mean area under the curve (and standard error) values for Figures $A.11a$ and $A.11b$

			$\Sigma = \Sigma_{\text{ps}}$			$\Sigma = \Sigma_{\text{non-ps}}$		
			$p = 50$	$p = 100$	$p = 150$	$p = 50$	$p = 100$	$p = 150$
$n = p/2$	AUC	FGMParty$_{95\%}$	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†	FGMParty$_{95\%}$	**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)
$n = 1.5p$	AUC	FGMParty$_{95\%}$	**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†	FGMParty$_{95\%}$	**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.

A.13 Simulation Details and Additional Figures for Section 4.5.3

A.13.1 Simulation Details for Section 4.5.3

This section describes the generation of edge sets $E_1^\Delta, \ldots, E_M^\Delta$, and precision matrices $\Omega_1^X, \ldots, \Omega_M^X$ and $\Omega_1^Y, \ldots, \Omega_M^Y$ for the simulation settings in section 4.5.3. An initial differential graph $G^\Delta = (V, E^\Delta)$ is generated from a power law distribution with parameter $\pi = \text{pr}\{(j,k) \in E^\Delta\}$. Then, for a fixed M, a sequence of edge sets $E_1^\Delta, \ldots, E_M^\Delta$ is generated so that $E^\Delta = \bigcup_{l=1}^M E_l^\Delta$. This process has two main steps. First, a set of common edges to all edge sets is computed and denoted as E_c^Δ for a given proportion of common edges $\tau \in [0,1]$. Next, the set of edges $E^\Delta \setminus E_c^\Delta$ is partitioned into $\tilde{E}_1^\Delta, \ldots, \tilde{E}_M^\Delta$ where $|\tilde{E}_l^\Delta| \geq |\tilde{E}_{l'}^\Delta|$ for $l < l'$ and set $E_l^\Delta = E_c^\Delta \cup \tilde{E}_l^\Delta$. The details for this process are described in Algorithm 3.

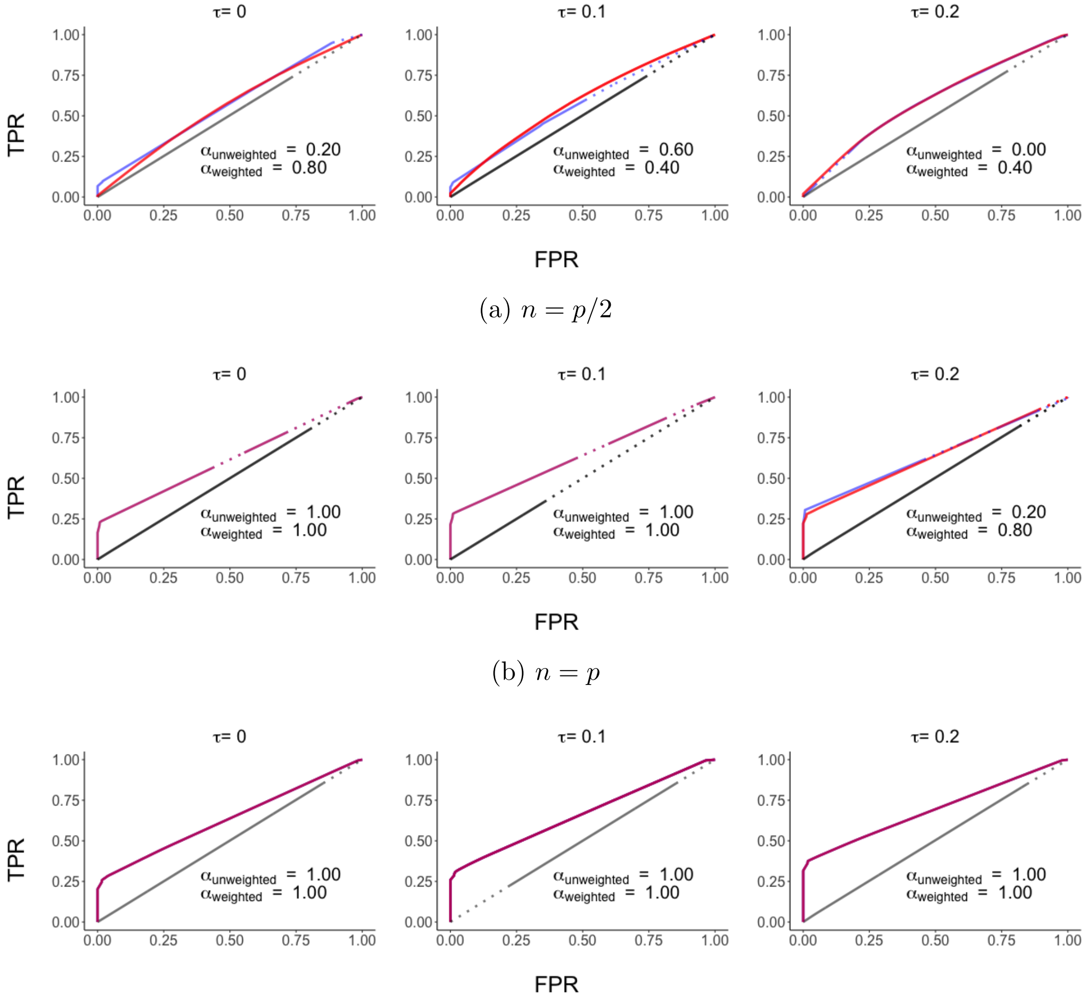
Next, following the steps in A.10, a graph $G^X = (V, E^X)$ with parameter $\tilde{\pi} > \pi$ and a $p \times p$ precision matrices Ω_l^X are obtained. Then a $p \times p$ matrix Δ_l is computed with entries $(\Delta_l)_{ij} = c\left(\Omega_l^X\right)_{ij}$ if $(i, j) \in E_l^\Delta$, and 0 otherwise. Finally, set $\Omega_l^Y = \Omega_l^X - \Delta_l$.

Algorithm 3: Pseudocode to create the edge sets $E_1^\Delta, \ldots, E_M^\Delta$

```
input: graph $G^\Delta = (V, E^\Delta)$ with nodes $V = \{1, \ldots, p\}$ and edge set $E^\Delta$
       number of basis $M$
       proportion of common edges $\tau$
Result: Edge sets $E_1^\Delta, \ldots, E_M^\Delta$
$E_c^\Delta \leftarrow$ random subset of size $\tau|E^\Delta|$ from $E^\Delta$;
$E_l^\Delta \leftarrow E_c^\Delta$ for $l = 1, \ldots, M$;
$l \leftarrow 1$;
$B \leftarrow 1$;
for $e \in E^\Delta \setminus E_c^\Delta$ do
    $E_l^\Delta \leftarrow E_l^\Delta \bigcup e$;
    $l \leftarrow l + 1$;
    if $l > B$ then
        $l \leftarrow 1$;
        $B \leftarrow (B + 1) \mod M$;
```

A.14 Additional Results for Section 4.5.3

This section includes results for the simulation analysis in Section 4.5.4 with parameters $T = 50$, $M = 20$, $\tilde{\pi} = 0.30$, $c = 0.4$ and $\pi = 10\%$ for a sparse graph.



(a) $n = p/2$

(b) $n = p$

(c) $n = 1.5p$

Figure A.12: Mean receiver operating characteristic curves for the proposed method (psFuDGE) and that of [50] (FuDGE). For $p = 60$, $n \in \{30, 60, 90\}$, $\pi = 0.10$ and $c = 0.4$ subfigures (a), (b) and (c) correspond to values of $n \in \{30, 60, 90\}$ respectively. Curves are coded as unweighted group psFuDGE (——), weighted group psFuDGE (——) and FuDGE (——) at 95% of variance explained. In each curve adjacent points with FPR difference less or equal than 0.10 are interpolated with a solid line. Otherwise, a dashed line is used. For psFuDGE, the values of α used to compute the curve values are printed in each panel.

Table A.7: Mean area under the curve (and standard error) values for Figure A.12.

			$\tau = 0$	$\tau = 0.10$	$\tau = 0.20$
$n = p/2$	AUC	FuDGE	0.51(0.02)	0.50(0.02)	0.51(0.02)
		psFuDGE$_{\text{unweighted}}$	**0.57**(**0.02**)	0.56(0.03)	0.59(0.03)
		psFuDGE$_{\text{weighted}}$	0.56(0.02)	**0.58**(**0.02**)	**0.60**(**0.03**)
	AUC15†	FuDGE	0.08(0.01)	0.08(0.01)	0.08(0.02)
		psFuDGE$_{\text{unweighted}}$	**0.19**(**0.04**)	**0.20**(**0.04**)	0.12(0.01)
		psFuDGE$_{\text{weighted}}$	0.10(0.01)	0.13(0.02)	**0.14**(**0.03**)
$n = p$	AUC	FuDGE	0.50(0.02)	0.50(0.02)	0.51(0.02)
		psFuDGE$_{\text{unweighted}}$	**0.62**(**0.02**)	**0.65**(**0.03**)	**0.65**(**0.02**)
		psFuDGE$_{\text{weighted}}$	**0.62**(**0.02**)	**0.65**(**0.03**)	0.64(0.02)
	AUC15†	FuDGE	0.08(0.02)	0.08(0.01)	0.08(0.02)
		psFuDGE$_{\text{unweighted}}$	**0.25**(**0.07**)	0.30(0.07)	0.30(0.07)
		psFuDGE$_{\text{weighted}}$	**0.25**(**0.07**)	**0.30**(**0.07**)	**0.31**(**0.06**)
$n = 1.5p$	AUC	FuDGE	0.50(0.02)	0.50(0.02)	0.51(0.02)
		psFuDGE$_{\text{unweighted}}$	**0.64**(**0.02**)	**0.67**(**0.02**)	**0.70**(**0.03**)
		psFuDGE$_{\text{weighted}}$	**0.64**(**0.02**)	**0.67**(**0.02**)	**0.70**(**0.03**)
	AUC15†	FuDGE	0.07(0.02)	0.07(0.02)	0.08(0.02)
		psFuDGE$_{\text{unweighted}}$	**0.32**(**0.04**)	**0.36**(**0.04**)	**0.42**(**0.07**)
		psFuDGE$_{\text{weighted}}$	**0.32**(**0.04**)	**0.36**(**0.04**)	**0.42**(**0.07**)

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

Bibliography

[1] N. Meinshausen, P. Bühlmann, *et. al.*, *High-dimensional graphs and variable selection with the lasso*, *Annals of statistics* **34** (2006), no. 3 1436–1462.

[2] M. Yuan and Y. Lin, *Model selection and estimation in the gaussian graphical model*, *Biometrika* **94** (2007), no. 1 19–35.

[3] N. Friedman, *Inferring cellular networks using probabilistic graphical models*, *Science* **303** (Feb., 2004) 799–805.

[4] S. L. Lauritzen and N. A. Sheehan, *Graphical Models for Genetic Analyses*, *Statistical Science* **18** (2003), no. 4 489 – 514.

[5] S. M. Smith, K. L. Miller, G. Salimi-Khorshidi, M. Webster, C. F. Beckmann, T. E. Nichols, J. D. Ramsey, and M. W. Woolrich, *Network modelling methods for fmri*, *NeuroImage* **54** (2011), no. 2 875–891.

[6] S. Ullah and C. Finch, *Applications of functional data analysis: A systematic review*, *BMC Medical Research Methodology* **13** (2013), no. 43.

[7] W. Jank and G. Shmueli, *Functional Data Analysis in Electronic Commerce Research*, *Statistical Science* **21** (2006), no. 2 155 – 166.

[8] A. Sood, G. M. James, and G. J. Tellis, *Functional regression: A new model for predicting market penetration of new products*, *Marketing Science* **28** (2009), no. 1 36–51.

[9] P. Hall, *Principal component analysis for functional data: Methodology, theory, and discussion*, in *The Oxford Handbook of Functional Data Analysis.* 2011.

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

[11] L. N. K., *The neural basis of the blood-oxygen-level-dependent functional magnetic resonance imaging signal*, *. Philosophical transactions of the Royal Society of London. Series B, Biological sciences* **357** (2002), no. 1424 1003–1037.

[12] M. F. Glasser, S. N. Sotiropoulos, J. A. Wilson, T. S. Coalson, B. Fischl, J. L. Andersson, J. Xu, S. Jbabdi, M. Webster, J. R. Polimeni, D. C. V. Essen, and M. Jenkinson, *The minimal preprocessing pipelines for the human connectome project*, *NeuroImage* **80** (2013) 105–124.

[13] *Human brain mapping: A systematic comparison of parcellation methods for the human cerebral cortex*, *NeuroImage* **170** (2018) 5–30. Segmenting the Brain.

[14] M. F. Glasser, T. S. Coalson, E. C. Robinson, C. D. Hacker, J. Harwell, E. Yacoub, K. Ugurbil, J. Andersson, C. F. Beckmann, M. Jenkinson, S. M. Smith, and D. C. Van Essen, *A multi-modal parcellation of human cerebral cortex*, *Nature* **536** (2016) 171–178.

[15] N. López-López, A. Vázquez, J. Houenou, C. Poupon, J.-F. Mangin, S. Ladra, and P. Guevara, *From coarse to fine-grained parcellation of the cortical surface using a fiber-bundle atlas*, *Frontiers in Neuroinformatics* **14** (2020) 32.

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

[17] X. Qiao, S. Guo, and G. M. James, *Functional graphical models*, *Journal of the American Statistical Association* **114** (2019), no. 525 211–222.

[18] T. Gneiting, M. G. Genton, and P. Guttorp, *Geostatistical space-time models, stationarity, separability, and full symmetry*, *Monographs On Statistics and Applied Probability* **107** (2006) 151.

[19] M. G. Genton, *Separable approximations of space-time covariance matrices*, *Environmetrics: The official journal of the International Environmetrics Society* **18** (2007), no. 7 681–695.

[20] J. A. Aston, D. Pigoli, S. Tavakoli, *et. al.*, *Tests for separability in nonparametric covariance operators of random surfaces*, *The Annals of Statistics* **45** (2017), no. 4 1431–1461.

[21] B. Lynch and K. Chen, *A test of weak separability for multi-way functional data, with application to brain connectivity studies*, *Biometrika* **105** (2018), no. 4 815–831.

[22] F. X. Castellanos, A. Di Martino, R. C. Craddock, A. D. Mehta, and M. P. Milham, *Continuing progress of spike sorting in the era of big data*, *Neuroimage* **80** (2013) 527–40. Machine Learning, Big Data, and Neuroscience.

[23] Y. Du, Z. Fu, and V. D. Calhoun, *Classification and prediction of brain disorders using functional connectivity: Promising but challenging*, *Frontiers in Neuroscience* **12** (2018) 525.

[24] L. Buesing, T. A. Machado, J. P. Cunningham, and L. Paninski, *Clustered factor analysis of multineuronal spike data*, in *Advances in Neural Information Processing Systems* (Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Q. Weinberger, eds.), vol. 27, Curran Associates, Inc., 2014.

[25] D. Carlson and L. Carin, *Continuing progress of spike sorting in the era of big data*, *Current Opinion in Neurobiology* **55** (2019) 90–96. Machine Learning, Big Data, and Neuroscience.

[26] C.-Z. Di, C. M. Crainiceanu, B. S. Caffo, and N. M. Punjabi, *Multilevel functional principal component analysis*, *The Annals of Applied Statistics)* **3** (2009), no. 1 458–488.

[27] Z. Liu and W. Guo, *Functional mixed effects models*, *WIREs Computational Statistics* **4** (2012), no. 6 527–534.

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

[29] S. Zhou, J. Lafferty, and L. Wasserman, *Time varying undirected graphs*, *Machine Learning* **80** (2010), no. 2-3 295–319.

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

[31] H. Qiu, F. Han, H. Liu, and B. Caffo, *Joint estimation of multiple graphical models from high dimensional time series*, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **78** (2016), no. 2 487–504.

[32] X. Qiao, C. Qian, G. M. James, and S. Guo, *Doubly functional graphical models in high dimensions*, *Biometrika* **107** (2020), no. 2 415–431.

[33] H. Zhu, N. Strawn, and D. B. Dunson, *Bayesian graphical models for multivariate functional data*, *Journal of Machine Learning Research* **17** (2016), no. 204 1–27.

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

[35] S. L. Lauritzen, *Graphical models*, vol. 17. Clarendon Press, 1996.

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

[37] F. Yao, H.-G. Müller, and J.-L. Wang, *Functional data analysis for sparse longitudinal data*, *Journal of the American Statistical Association* **100** (2005), no. 470 577–590.

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

[39] P. Danaher, P. Wang, and D. M. Witten, *The joint graphical lasso for inverse covariance estimation across multiple classes*, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **76** (2014), no. 2 373–397.

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

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

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

[43] M. Jirak, *Optimal eigen expansions and uniform bounds*, *Probability Theory and Related Fields* **166** (2016), no. 3-4 753–799.

[44] J. Peng, P. Wang, N. Zhou, and J. Zhu, *Partial correlation estimation by joint sparse regression models*, *Journal of the American Statistical Association* **104** (2009), no. 486 735–746.

[45] J. H. Friedman, *A variable span scatterplot smoother*, Tech. Rep. 5, Stanford University, 1984.

[46] J. Fan and J. Lv, *Sure independence screening for ultrahigh dimensional feature space*, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **70** (2008), no. 5 849–911.

[47] D. M. Barch, G. C. Burgess, M. P. Harms, S. E. Petersen, B. L. Schlaggar, M. Corbetta, M. F. Glasser, S. Curtiss, S. Dixit, C. Feldt, D. Nolan, E. Bryant, T. Hartley, O. Footer, J. M. Bjork, R. Poldrack, S. Smith, H. Johansen-Berg, A. Z. Snyder, and D. C. V. Essen, *Function in the human connectome: Task-fMRI and individual differences in behavior*, *NeuroImage* **80** (2013) 169 – 189.

[48] L. Cheng, L. Shan, and I. Kim, *Multilevel gaussian graphical model for multilevel networks*, *Journal of Statistical Planning and Inference* **190** (2017) 1–14.

[49] E. Pircalabelu, G. Claeskens, and L. J. Waldorp, *Zoom-in–out joint graphical lasso for different coarseness scales*, *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **69** (2020), no. 1 47–67.

[50] B. Zhao, Y. S. Wang, and M. Kolar, *Fudge: Functional differential graph estimation with fully and discretely observed curves*, 2020.

[51] J. Chiquet, Y. Grandvalet, and C. Ambroise, *Inferring multiple graphical structures*, *Statistics and Computing* **21** (2011) 537–553.

[52] K. M. Tan, P. London, K. Mohan, S.-I. Lee, M. Fazel, and D. Witten, *Learning graphical models with hubs*, *J. Mach. Learn. Res.* **15** (Jan., 2014) 3297–3331.

[53] S. D. Zhao, T. T. Cai, and H. Li, *Direct estimation of differential networks.*, *Biometrika* **101** (2014), no. 2 253–268.

[54] P. Xu and Q. Gu, *Semiparametric differential graph models*, in *Advances in Neural Information Processing Systems* (D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, eds.), vol. 29, Curran Associates, Inc., 2016.

[55] H. Yuan, R. Xi, C. Chen, and M. Deng, *Differential network analysis via lasso penalized D-trace loss*, *Biometrika* **104** (10, 2017) 755–770.

[56] H. Zou, *The adaptive lasso and its oracle properties*, *Journal of the American Statistical Association* **101** (2006), no. 476 1418–1429.

[57] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, *Distributed optimization and statistical learning via the alternating direction method of multipliers*, *Found. Trends Mach. Learn.* **3** (Jan., 2011) 1–122.

[58] B. Jiang, X. Wang, and C. Leng, *A direct approach for sparse quadratic discriminant analysis*, *Journal of Machine Learning Research* **19** (2018), no. 31 1–37.

[59] M. E. J. Newman, *The structure and function of complex networks*, *SIAM Rev.* **45** (Jan., 2003) 167–256.

[60] S. Na, M. Kolar, and O. Koyejo, *Estimating differential latent variable graphical models with applications to brain connectivity*, *Biometrika* (Sep, 2020).

[61] J. Ramsay and B. W. Silverman, *Functional Data Analysis.* Springer-Verlag New York, second ed., 2005.