The Pennsylvania State University

The Graduate School

JOINT ESTIMATION OF MULTIPLE GRAPHICAL MODELS FROM HETEROGENEOUS SUBPOPULATIONS

A Dissertation in

Statistics

by

Ilias Moysidis

Submitted in Partial Fulfillment

of the Requirements

for the Degree of

Doctor of Philosophy

August 2021

The dissertation of Ilias Moysidis was reviewed and approved by the following:

Bing Li
Verne M. Willaman Professor of Statistics
Dissertation Advisor
Committe Chair

Runze Li
Eberly Family Chair in Statistics, Associate Department Head

Bharath K. Sriperumbudur
Associate Professor of Statistics

Alexei Novikov
Professor of Mathematics

Ephraim Hanks
Associate Professor of Statistics
Chair of Graduate Studies

Abstract

It is a frequent occurrence in the field of graphical models to estimate graphs from different subpopulations that share behavioral or genetic similarities. For example, such subpopulations can be alcoholic and non-alcoholic groups, cancer and non-cancer patients, students and faculty. We can expect these similarities to translate into a common graph structure. Compiling all the data together to estimate a single graph would ignore the differences among subpopulations. Dividing the data to estimate a graph for each subpopulation separately does not make efficient use of the common structure in the data. It is therefore crucial to develop methods for combining all the data to estimate the graphs of the subpopulations simultaneously.

My research with Dr.Bing Li is focused on two important areas of simultaneous graph estimation: **(1)** the simultaneous estimation of functional graphical models by graphical lasso, **(2)** the simultaneous estimation of high-dimensional graphical models by neighborhood selection. Concerning **(1)**, functional graphical models explore dependence relationships of random processes. This is achieved through estimating the precision matrix of the coefficients from the Karhunen-Loeve expansion. Our research deals with the problem of estimating functional graphs that consist of the same random processes and share some of the dependence structure. Concerning **(2)**, we develop a new method for simultaneous estimation of multiple graphical models by estimating the topological neighborhoods of the involved variables under a sparse inducing penalty that takes into account the common structure in the subpopulations. Unlike the existing methods for joint graphical models, our method does not rely on spectral decomposition of large matrices, and is therefore more computationally attractive for estimating large networks.

Table of Contents

List of Figures

List of Tables

Acknowledgments

I cannot believe this is over, but unfortunately all the good things come to an end. It has been a wonderful journey that helped me mature as a person intellectually. I would never have thought that my ideas would have been challenged so fundamentally. Again and again. Each failure made me go back to the roots and rethink the entire problem.

I would like to express my gratitude to my friends Adam and Emily, that made America a second home for me. All the Italian gang with which we shared some amazing experiences: Tobia, Umberto, Martina, Roberto and Luca. Many thanks to my adviser Bing who even though I repeatedly and aggressively chased for multiple meetings per week, he never complained. I would also like to thank my friends in Greece for the great moments we have every time I take a break and go back: Vaggelis, Dimitris, Tolis, Thanasis, Saro, Psilos, Tasos. I could not forget of course my friend Stavros, who like me emigrated to foreign lands to find his fortune. Finally, I would like to dedicate this entire body of research to my parents Antonia and Vasili who have given me everything a child would ever want and more.

Chapter 1

Joint Functional Graphical Models

1.1 Introduction

Functional graphical models provide insight on the conditional dependence structure among the components of a multivariate random function. Datasets such as those arising from functional magnetic resonance imaging (fMRI) and electroencephalography (EEG) motivate research in this area. In particular, for the EEG dataset, a curve is recorded at each location of the brain and a network is constructed based on a sample of subjects each having a vector of curves. However, these datasets often consist of samples originating from different subpopulations that share behavioral or genetic characteristics. Such subpopulations could be ADHD and non-ADHD patients or alcoholic and non-alcoholic subjects. We can assume that the similarities between the groups translate into a common graph structure. Merging the data together to estimate a single graph would ignore the differences among subpopulations. Dividing the data to estimate a graph individually for each subpopulation would waste potential information in the common structure that lies across the different subpopulations. The goal of this paper is to develop a model that uses all the data to estimate the common structure, but also leaves room for differences between the graphs. We call this model the joint functional graphical model.

Consider a vector of random functions $\mathbf{g} = (g_1, \ldots, g_p)^\intercal$. The graphical model of $\mathbf{g}$ is represented by an undirected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V} = \{1, \ldots, p\}$ is the set of nodes and $\mathcal{E} \subseteq \{(i,j) : i, j \in \mathcal{V}, i \neq j\}$ is the set of edges. For convenience we assume $i < j$ for $(i,j) \in \mathcal{E}$, because for an undirected graph (i,j) and (j,i) represent the same edge. The set of edges $\mathcal{E}$ is defined by the relation

$$(i,j) \notin \mathcal{E} \quad \Leftrightarrow \quad g_i \perp\!\!\!\perp g_j | \mathbf{g}_{\setminus\{i,j\}} \tag{1.1}$$

where $\mathbf{g}_{\setminus\{i,j\}}$ represents $\mathbf{g}$ with its i-th and j-th components removed.

Functional graphical models have undergone dynamic development in the recent years. Zhu et al. (2016) extended the notions of Markov distribution and hyper Markov laws to the functional case and developed a Bayesian approach for functional graphical models by proposing a hyper-inverse Wishart process prior for the covariance operator, and assuming the random elements are multivariate Gaussian processes. Qiao et al. (2019) introduced the functional Gaussian graphical model (FGGM) where the random functions are assumed to be Gaussian random elements in a Hilbert space, and the network is constructed using an association of the relation (1.1) with the coefficients of their Karhunen-Loeve expansion. Solea and Li (2019) proposed a semiparametric functional copula Gaussian graphical model for random functions that are not originally Gaussian processes, but for which there exist one-to-one transformations of their Karhunen-Loeve expansion coefficients that makes them Gaussian. Li and Solea (2018) bypasses the Gaussian assumption altogether by replacing the conditional independence relationship with additive conditional independence. This approach allows for nonlinear or heteroscedastic relations between the random processes.

In the multivariate case, the research on covariance selection dates back to Dempster (1972) who proposed backward selection, in which we start with a fully connected graph and at each step remove edges that are not significant according to a partial

correlation test until all remaining edges are significant. Another approach to constructing graphical models is via sparse inducing penalty which has become popular because it can be applied to the case where the dimension of the graph is much bigger than the sample size. Meinshausen et al. (2006) reformulated the covariance selection problem to an ℓ_1 penalized variable selection problem where each variable is regressed against the others to estimate its set of effective predictors. Yuan and Lin (2007) propose an ℓ_1 penalized loglikelihood method which has the advantage of merging variable selection and precision matrix estimation into one problem.

The first paper that deals with the joint estimation problem in the multivariate case is by Guo et al. (2011). They decompose each precision matrix into a shared component and a unique component, and proposed a nonconvex hierarchical penalty that utilizes the information from all the data by penalizing first on a common and then on an individual level. Danaher et al. (2014) developed the computationally attractive methods of Fused Graphical Lasso and Group Graphical Lasso that allow joint estimation while also have the advantage of employing convex penalties. The first forces a common structure using a penalty that promotes identical edge values across subpopulations, while the latter only propels shrinkage on a common level. A general framework for joint precision matrix estimation is provided in Saegusa and Shojaie (2016). They use a graph Laplacian penalty that allows for different levels of similarity between subpopulations. Motivated by the high dimensionality of the datasets, Cai et al. (2016) propose a computationally fast method for jointly estimating precision matrices, following the steps of the Dantzig Selector (Candes et al., 2007).

In this paper we focus on the joint functional graphical model under the Gaussian assumption, and we call our method the joint functional Gaussian graphical model (JFGGM). We propose a nonconvex objective function that achieves regularization which encourages both a common graph structure and individual sparsity. To estimate

the global minimizer we are using the local linear approximation (LLA) method, which is an algorithm that at each step proposes a convex, non-smooth objective function. To calculate the global minimizer at each step of the LLA algorithm we employ the alternating direction method of multipliers (ADMM) algorithm. Furthermore, we establish the asymptotic consistency of our method with overwhelming probability. In addition, we do a simulation study, where we compare our method with the FGGM applied on each subpopulation individually. Finally, we estimate the graph of the EEG dataset with the use of our method.

The rest of the paper is organized as follows. In section 2 we give an overview of the functional graphical lasso (fglasso) method, and introduce the JFGGM. In section 3 we develop the algorithmic procedure for estimating the JFGGM. In section 4 we establish the asymptotic consistency of our estimator. In section 5 we compare by simulation our method with the FGGM. In section 6 we apply JFGGM to the EEG dataset.

1.2 Methodology

In this section we provide the theoretical background for the functional graphical model presented in Qiao et al. (2019). We, then, propose our method, the JFGGM, for joint estimation of graphical models that come from different subpopulations.

1.2.1 Functional Graphical Models

Let $(\mathbf{\Omega}, \mathcal{F}, P)$ be a probability space, let T be an interval in $\mathbb{R}$, let $\mathbb{H}_1, \ldots, \mathbb{H}_p$ be Hilbert spaces of real-valued functions on T. Let $g_i : \Omega \to \mathbb{H}_i$ be a random element in $\mathbb{H}_i$, so that $(g_1, \ldots, g_p)$ is a random element in the direct sum $\oplus_{i=1}^{p} \mathbb{H}_i$, which is the Cartesian product $\mathbb{H}_1 \times \cdots \times \mathbb{H}_p$ together with the inner product $\langle f_1, g_1 \rangle_{\mathbb{H}_1} + \ldots + \langle f_p, g_p \rangle_{\mathbb{H}_p}$.

The functional Gaussian graphical model (FGGM) developed by Qiao et al. (2019)

is under the assumption that $\mathbf{g} = (g_1, \ldots, g_p)$ is a multivariate Gaussian random element in $\mathbb{H}_1 \times \cdots \times \mathbb{H}_p$. Each $\mathbb{H}_i$ is an M-dimensional subspace of the space of square integrable functions $\mathbb{L}_2(T, \mathcal{B}(T), \mu)$, where $\mathcal{B}(T)$ is the Borel σ-field generated by the open sets in T and μ is the Lebesgue measure. Under this assumption, relation (1.1) reduces to

$$(i,j) \notin \mathcal{E} \quad \Leftrightarrow \quad \operatorname{cov}[g_i(s), g_j(t) \,|\, g_{\setminus\{i,j\}}] = 0 \ \ \forall\, s,t \in T. \tag{1.2}$$

Associated with the stochastic process g_j is the mean function $h_j(t) = \mathrm{E}(g_j(t))$ and the covariance function $K_j(s,t) = \operatorname{cov}(g_j(s), g_j(t))$. Processes with well defined mean and covariance functions are called second-order processes. If such a process is mean-square continuous, then the Karhunen-Loeve expansion theorem applies. To move forward, we need the following assumption.

Assumption 1. *Each g_j is jointly measurable with respect to the product σ-field $\mathcal{B}(T) \times \mathcal{F}$ and $g_j(\cdot, \omega) \in \mathbb{H}_j$ for each $\omega \in \mathbf{\Omega}$. Furthermore, we have $\mathrm{E}\|\mathbf{g}\|^2 < \infty$.*

This assumption implies the existence of the covariance operator of the j-th process

$$\mathscr{K}_j : \mathbb{H}_j \to \mathbb{H}_j, \quad f \mapsto \int_T K_j(\cdot, t) f(t) dt,$$

and its spectral decomposition. That is, there exist pairs $\{(\lambda_{jm}, \phi_{jm})\}_{m=1}^{M}$, with $\lambda_{j1} \geq \ldots \geq \lambda_{jM}$, such that

$$\int_T K_j(s,t)\phi_{jm}(t)dt = \lambda_{jm}\phi_{jm}(s), \quad s \in T, \quad m = 1, \ldots, M,$$

and

$$\int_T \phi_{jm}(t)\phi_{jm'}(t)dt = \delta_{mm'}, \quad m, m' = 1, \ldots, M,$$

where $\delta_{mm'}$ is the Kronecker δ-function. To link the spectral decomposition of $\mathscr{K}_j$ with the Karhunen-Loeve expansion of g_j we need the following assumption.

Assumption 2. *For each j, the mean and covariance functions of g_j are continuous.*

Without loss of generality we assume that $h_j = \mathbf{0}$. Under Assumptions 1 and 2, we have the following Karhunen-Loeve expansion for g_j:

$$g_j(t) = \sum_{m=1}^{M} a_{jm}\phi_{jm}(t),$$

where $a_{jm} = \int_T g_j(t)\phi_{jm}(t)dt$ is distributed as $\mathcal{N}(0, \lambda_{jm})$, and $a_{jm} \perp\!\!\!\perp a_{jm'}$ for $m \neq m'$. A more detailed analysis of these matters can be found in Hsing and Eubank (2015).

Let $\mathbf{a}_j = (a_{j1}, \ldots, a_{jM})^\intercal$. Since $\mathbf{g}$ is a mean zero Gaussian random element, the vector $\mathbf{a} = (\mathbf{a}_1^\intercal, \ldots, \mathbf{a}_p^\intercal)^\intercal$ follows a pM-dimensional normal distribution with mean $\mathbf{0}$ and covariance matrix $\mathbf{\Sigma} = (\mathbf{\Sigma}_{jl})_{p\times p} \in \mathbb{R}^{pM\times pM}$ where $\mathbf{\Sigma}_{jl} = \mathrm{cov}(\mathbf{a}_j, \mathbf{a}_l) \in \mathbb{R}^{M\times M}$. Let $\mathbf{\Omega} = \mathbf{\Sigma}^{-1}$ be the precision matrix. The following theorem (Qiao et al., 2019) links the conditional independence of random elements with the zero entries of $\mathbf{\Omega}$.

Theorem 1. *Let $i, j \in \mathcal{V}$, $i \neq j$, and let $\mathbf{\Omega}_{ij}$ be the $M \times M$ matrix corresponding to the (i, j)-th block matrix of $\mathbf{\Omega}$. Then,*

$$g_i \perp\!\!\!\perp g_j | \mathbf{g}_{\setminus\{i,j\}} \quad \Leftrightarrow \quad \mathbf{\Omega}_{ij} = \mathbf{0}.$$

The defining relation (1.2) of the functional graph is difficult to work with. Theorem 1 provides an equivalent condition relating it to the coefficients of the Karhunen-Loeve expansion, thus, reducing the functional setting to the multivariate setting, which has been extensively studied.

Denote the estimator of $\mathbf{\Sigma}$ by $\hat{\mathbf{\Sigma}}$, whose precise definition is given in Section 1.3, where we give detailed analysis of $\hat{\mathbf{\Sigma}}$ at the sample level. Based on Theorem 1, Qiao

et al. (2019) introduce the fglasso to estimate $\mathbf{\Omega}$ by

$$\hat{\mathbf{\Omega}} = \underset{\mathbf{\Omega} \succ \mathbf{0}}{\operatorname{argmin}} \left\{ \operatorname{trace}(\hat{\mathbf{\Sigma}}\mathbf{\Omega}) - \log\det(\mathbf{\Omega}) + \lambda \sum_{i \neq j} \|\mathbf{\Omega}_{ij}\|_F \right\},$$

where $\|\cdot\|_F$ is the Frobenius norm. The loss function $\operatorname{trace}(\hat{\mathbf{\Sigma}}\mathbf{\Omega}) - \log\det(\mathbf{\Omega})$ is the negative loglikelihood and is responsible for producing a precision matrix that is going to make the data most likely to be observed given the assumptions of the model, and the penalty function is responsible for enforcing sparsity on the estimator. Notice that to encourage blockwise sparsity for $\mathbf{\Omega}$, they use a groupwise penalty induced by the Frobenius norm.

1.2.2 Joint Functional Graphical Models

We now relax the assumption that all of the observed functional data $\mathbf{g}_i = (g_{i1}, \ldots, g_{ip})^\intercal$ are realizations of the same Gaussian process. Instead, we assume that there are K different subpopulations whose graphs, though not identical, share a set of common edges. Suppose, for each $k = 1, \ldots, K$, we observe an i.i.d. sample $\mathbf{g}_1^{(k)}, \ldots, \mathbf{g}_n^{(k)}$ of random elements in $\oplus_{j=1}^p \mathbb{H}_j$, where $\mathbf{g}_i^{(k)} = (g_{i1}^{(k)}, \ldots, g_{ip}^{(k)})^\intercal$. We assume that $\mathbf{g}_i^{(k)}$ is a zero mean Gaussian random element. By performing Karhunen-Loeve expansion on each $g_{ij}^{(k)}$ we obtain the vector of coefficients $\mathbf{a}_{ij}^{(k)} = (a_{ij1}^{(k)}, \ldots, a_{ijM}^{(k)})^\intercal$. Since each $\mathbf{g}_i^{(k)}$ is a mean zero Gaussian random element, the vector $\mathbf{a}_i^{(k)} = (\mathbf{a}_{i1}^{(k)\intercal}, \ldots, \mathbf{a}_{ip}^{(k)\intercal})^\intercal$ is distributed as $\mathcal{N}_{pM}(\mathbf{0}, \mathbf{\Sigma}^{(k)})$, for $i = 1, \ldots, n, k = 1, \ldots, K$. Let $\mathbf{\Omega}^{(k)} = \mathbf{\Sigma}^{(k)-1}$ be the precision matrix corresponding to the k-th subpopulation.

To take advantage of the information across the subpopulations we reparameterize each off-diagonal block $\mathbf{\Omega}_{jl}^{(k)}$ of $\mathbf{\Omega}^{(k)}$ as $\theta_{jl}\mathbf{\Gamma}_{jl}^{(k)}$, where $\theta_{jl} \in \mathbb{R}$ is common to all subpopulations and $\mathbf{\Gamma}_{jl}^{(k)} \in \mathbb{R}^{M \times M}$ is specific to subpopulation k. For identifiability, we assume $\theta_{jl} \geq 0$. To preserve symmetry we require $\theta_{lj} = \theta_{jl}$ and $\mathbf{\Gamma}_{lj}^{(k)\intercal} = \mathbf{\Gamma}_{jl}^{(k)}$. For the diagonal-block matrices we require $\theta_{jj} = 1$ and $\mathbf{\Gamma}_{jj}^{(k)} = \mathbf{\Omega}_{jj}^{(k)}$. In this reparameter-

ization the common factor θ_{jl} controls the presence of the common edge between the nodes j and l in all of the graphs, and $\mathbf{\Gamma}_{jl}^{(k)}$ accommodates the differences between individual graphs. This reparameterization is similar to that of Guo et al. (2011) with the difference that in our case the individual element is a matrix rather than a number. To estimate this model we propose to minimize

$$\sum_{k=1}^{K}\left[\text{trace}\left(\hat{\mathbf{\Sigma}}^{(k)}\mathbf{\Omega}^{(k)}\right)-\log\det\left(\mathbf{\Omega}^{(k)}\right)\right]+\lambda_1\sum_{j\neq l}\theta_{jl}+\lambda_2\sum_{j\neq l}\sum_{k=1}^{K}\|\mathbf{\Gamma}_{jl}^{(k)}\|_F \qquad (1.3)$$

over all θ_{jl} and $\mathbf{\Gamma}_{jl}^{(k)}$ specified above. The first penalty function penalizes the common factors θ_{jl} and is responsible for identifying the zeros across all precision matrices. That is, if θ_{jl} is zero then there is no edge between the nodes j and l in all K graphs. The second penalty function penalizes the individual factors $\mathbf{\Gamma}_{jl}^{(k)}$ and is responsible for identifying the zeros that are specific to each graph. That is, for a nonzero θ_{jl} some of the matrices $\mathbf{\Gamma}_{jl}^{(1)},\ldots,\mathbf{\Gamma}_{jl}^{(K)}$ can be zero, which means that the edge connecting the nodes j and l may be absent in some of the K graphs but present in others.

However, the objective function (1.3) is hard to minimize because of its complexity. It includes two groups of variables over which we have to optimize, and two parameters that we have to tune. Additionally, due to the restrictions on the variables that ensure identifiability and positive definiteness, the domain of the objective function is not directly intuitive. As will be shown in Theorem 2, (1.3) is equivalent to the much simpler form

$$\sum_{k=1}^{K}\left[\text{trace}\left(\hat{\mathbf{\Sigma}}^{(k)}\mathbf{\Omega}^{(k)}\right)-\log\det\left(\mathbf{\Omega}^{(k)}\right)\right]+2(\lambda_1\lambda_2)^{1/2}\sum_{j\neq l}\left(\sum_{k=1}^{K}\|\mathbf{\Omega}_{jl}^{(k)}\|_F\right)^{1/2}, \qquad (1.4)$$

where $\mathbf{\Omega}^{(1)},\ldots,\mathbf{\Omega}^{(K)}$ are positive definite matrices.

Let $\mathbf{\Theta}=(\theta_{jl})$, $\mathbf{\Gamma}^{(k)}=(\mathbf{\Gamma}_{jl}^{(k)})$, $\mathbf{\Omega}^{(k)}=(\mathbf{\Omega}_{jl}^{(k)})$ for $\theta_{jl},\mathbf{\Gamma}_{jl}^{(k)},\mathbf{\Omega}_{jl}^{(k)}$ defined above, and define the lists $\mathbf{\Gamma}=(\mathbf{\Gamma}^{(1)},\ldots,\mathbf{\Gamma}^{(K)})$, $\mathbf{\Omega}=(\mathbf{\Omega}^{(1)},\ldots,\mathbf{\Omega}^{(K)})$, $\hat{\mathbf{\Gamma}}=(\hat{\mathbf{\Gamma}}^{(1)},\ldots,\hat{\mathbf{\Gamma}}^{(K)})$, $\hat{\mathbf{\Omega}}=$

$(\hat{\mathbf{\Omega}}^{(1)}, \ldots, \hat{\mathbf{\Omega}}^{(K)})$. The proof of the following theorem can be found in the appendix.

Theorem 2. *Let* $(\hat{\mathbf{\Theta}}, \hat{\mathbf{\Gamma}})$ *be a local minimizer of* (1.3). *Then, there exists a local minimizer* $\hat{\mathbf{\Omega}}$ *of* (1.4) *such that* $\hat{\mathbf{\Omega}}_{jl}^{(k)} = \hat{\theta}_{jl}\hat{\mathbf{\Gamma}}_{jl}^{(k)}$ *for all* j, l, k. *Conversely, let* $\hat{\mathbf{\Omega}}$ *be a local minimizer of* (1.4). *Then, there exists a local minimizer* $(\hat{\mathbf{\Theta}}, \hat{\mathbf{\Gamma}})$ *of* (1.3) *such that* $\hat{\theta}_{jl}\hat{\mathbf{\Gamma}}_{jl}^{(k)} = \hat{\mathbf{\Omega}}_{jl}^{(k)}$ *for all* j, l, k.

1.3 Sample-Level Implementation

The goal of this section is to develop a procedure to calculate the minimizer of (1.4). We begin by providing a formula for the computation of the sample covariance matrix $\hat{\mathbf{\Sigma}}$ mentioned at the end of subsection 1.2.1. Furthermore, because the objective function (1.4) is nonconvex, we instead optimize an approximate version of it. Finally, to calculate the minimizer of the approximate objective function we employ the ADMM algorithm because it provides closed form solutions for the updates, makes good use of R's vectorized operations, and has shown to outperform other popular algorithms on similar problems (Scheinberg et al., 2010).

1.3.1 Sample Covariance Matrix Estimation

Since we will be concerned with a fixed k in this subsection, we drop the superscript (k) for simplicity of notation. First, we need to estimate the eigenpairs $(\lambda_{jm}, \phi_{jm})$. To do that we follow the procedure described in Ramsay and Silverman (2001). Assume that we observe each curve g_{ij} at equally-spaced time points $t_1 < \ldots < t_\nu$, where t_1 and t_ν are the endpoints of the interval T. Without loss of generality, we assume that the data are centered. That is $\sum_{i=1}^{n} g_{ij}(t_q) = 0$, for all j, q.

As we will show, the procedure for estimating the eigenpairs of the covariance operator is very similar to the multivariate case. We start by providing a matrix approximation of the covariance operator at the population level. Then, we show how to

perform eigenvalue decomposition on the matrix approximation. Finally, by replacing the population-level quantities with sample-level quantities in the approximation of the covariance operator, we obtain the estimators of the eigenpairs.

Define

$$\mathbf{K}_j = \begin{pmatrix} K_j(t_1,t_1) & \dots & K_j(t_1,t_\nu) \\ \vdots & \ddots & \vdots \\ K_j(t_\nu,t_1) & \dots & K_j(t_\nu,t_\nu) \end{pmatrix}.$$

Let also $\boldsymbol{\phi}_{jm} = (\phi_{jm}(t_1), \dots, \phi_{jm}(t_\nu))^\intercal$, and $w = (t_\nu - t_1)/(\nu - 1)$ be the gap between two adjacent time points. Then, for large ν,

$$\int_T K_j(t_r,t)\phi_{jm}(t)dt \approx w \sum_{q=1}^{\nu} K_j(t_r,t_q)\phi_{jm}(t_q).$$

Therefore, the integral equation

$$\int_T K_j(s,t)\phi_{jm}(t)dt = \lambda_{jm}\phi_{jm}(s)$$

can be approximated by

$$w\mathbf{K}_j\boldsymbol{\phi}_{jm} = \lambda_{jm}\boldsymbol{\phi}_{jm}.$$

Then $(\lambda_{jm}, \boldsymbol{\phi}_{jm}) = (w^{-1}\mathbf{u}_m^\intercal\mathbf{K}_j\mathbf{u}_m, \mathbf{u}_m)$, where $\mathbf{u}_m$ is the solution to the eigenvalue problem:

$$\begin{aligned} &\text{maximize} \quad \mathbf{u}_m^\intercal\mathbf{K}_j\mathbf{u}_m \\ &\text{subject to} \quad \|\mathbf{u}_m\|_2 = 1 \quad \text{and} \quad \mathbf{u}_l^\intercal\mathbf{u}_m = 0 \text{ for } l < m. \end{aligned} \tag{1.5}$$

Let $\mathbf{a}_i = (\mathbf{a}_{i1}^\intercal, \dots, \mathbf{a}_{ip}^\intercal)^\intercal$ denote the sample principal component score vector, where

$\mathbf{a}_{ij} = (a_{ij1}, \ldots, a_{ijM})^\intercal$. Similarly, for large ν,

$$a_{ijm} = \int_T g_{ij}(t)\phi_{jm}(t)dt \approx w \sum_{q=1}^{\nu} g_{ij}(t_q)\phi_{jm}(t_q).$$

Let $\hat{K}_j(s,t) = n^{-1}\sum_{i=1}^n g_{ij}(s)g_{ij}(t)$ be the estimator of the covariance function $K_j(s,t)$, and let $\hat{\mathbf{K}}_j = (\hat{K}_j(t_r, t_q))$. Define

$$\mathbf{G}_j = \begin{pmatrix} g_{1j}(t_1) & \ldots & g_{1j}(t_\nu) \\ \vdots & \ddots & \vdots \\ g_{nj}(t_1) & \ldots & g_{nj}(t_\nu) \end{pmatrix}.$$

Then $\hat{\mathbf{K}}_j = n^{-1}\mathbf{G}_j^\intercal\mathbf{G}_j$. The estimator $(\hat{\lambda}_{jm}, \hat{\boldsymbol{\phi}}_{jm})$ of the eigenpair $(\lambda_{jm}, \boldsymbol{\phi}_{jm})$ is the solution to (1.5) with $\mathbf{K}_j = \hat{\mathbf{K}}_j$. Finally, the estimated principal component scores are given by

$$\hat{a}_{ijm} = \int_T g_{ij}(t)\hat{\phi}_{jm}(t)dt \approx w \sum_{q=1}^{\nu} g_{ij}(t_q)\hat{\phi}_{jm}(t_q)$$

with which we calculate the sample covariance matrix $\hat{\boldsymbol{\Sigma}} = n^{-1}\sum_{i=1}^n \hat{\mathbf{a}}_i\hat{\mathbf{a}}_i^\intercal$.

1.3.2 Penalty linearization

We now resume the use of the superscript (k) as we are concerned with calculations across the K subpopulations in this subsection. Writing $2(\lambda_1\lambda_2)^{1/2}$ as λ, the objective function (1.4) becomes

$$\sum_{k=1}^{K}\left[\text{trace}\left(\hat{\boldsymbol{\Sigma}}^{(k)}\boldsymbol{\Omega}^{(k)}\right) - \log\det\left(\boldsymbol{\Omega}^{(k)}\right)\right] + \lambda\sum_{j\neq l}\left(\sum_{k=1}^{K}\|\boldsymbol{\Omega}_{jl}^{(k)}\|_F\right)^{1/2}. \tag{1.6}$$

Notice that because of the square root in the penalty function, (1.6) is not convex. To tackle this issue we use the Local Linear Approximation (LLA) method

developed in Zou and Li (2008). Suppose that we are given an initial value $\hat{\boldsymbol{\Omega}}_{(0)} = (\hat{\boldsymbol{\Omega}}_{(0)}^{(1)}, \ldots, \hat{\boldsymbol{\Omega}}_{(0)}^{(K)})$ that is close to the true value. They propose locally approximating the penalty function by a linear function

$$\sum_{j\neq l}\left(\sum_{k=1}^{K}\|\boldsymbol{\Omega}_{jl}^{(k)}\|_F\right)^{1/2} \approx \sum_{j\neq l}\left(\sum_{k=1}^{K}\|\hat{\boldsymbol{\Omega}}_{(0),jl}^{(k)}\|_F\right)^{1/2} + \sum_{j\neq l}\tau_{(0),jl}\left(\sum_{k=1}^{K}\|\boldsymbol{\Omega}_{jl}^{(k)}\|_F - \sum_{k=1}^{K}\|\hat{\boldsymbol{\Omega}}_{(0),jl}^{(k)}\|_F\right)$$

where $\tau_{(0),jl} = 2^{-1}\left(\sum_{k=1}^{K}\|\hat{\boldsymbol{\Omega}}_{(0),jl}^{(k)}\|_F\right)^{-1/2}$. Thus, at the t-th iteration, problem (1.6) is decomposed into K individual optimization problems

$$\hat{\boldsymbol{\Omega}}_{(t)}^{(k)} = \underset{\boldsymbol{\Omega}\succ\mathbf{0}}{\operatorname{argmin}}\left\{\operatorname{trace}\left(\hat{\boldsymbol{\Sigma}}^{(k)}\boldsymbol{\Omega}\right) - \log\det(\boldsymbol{\Omega}) + \lambda\sum_{j\neq l}\tau_{(t-1),jl}\|\boldsymbol{\Omega}_{jl}\|_F\right\}, \tag{1.7}$$

where $k = 1, \ldots, K$. It is shown in the same paper, that if the initial value at iteration $t = 0$ is reasonably good, then with only one iteration ($t = 1$) we can get a good sparse estimate. In our paper, for the initial value we use the precision matrices estimated separately using the functional graphical lasso of Qiao et al. (2019) for each subpopulation.

1.3.3 ADMM algorithm for optimization

To solve (1.7) we employ the ADMM algorithm as described in Boyd et al. (2011). The primal problem is given by

$$\begin{aligned}&\text{minimize} \quad \operatorname{trace}(\mathbf{S}\boldsymbol{\Omega}) - \log\det(\boldsymbol{\Omega}) + \lambda\sum_{j\neq l}\tau_{jl}\|\boldsymbol{\Omega}_{jl}\|_F\\ &\text{subject to} \quad \boldsymbol{\Omega}\succ\mathbf{0},\end{aligned}$$

while the dual problem is

$$\begin{aligned} &\text{minimize} \quad \operatorname{trace}(\mathbf{S}\boldsymbol{\Omega}) - \log\det(\boldsymbol{\Omega}) + \lambda \sum_{j \neq l} \tau_{jl} \|\mathbf{Z}_{jl}\|_F \\ &\text{subject to } \boldsymbol{\Omega} \succ \mathbf{0},\ \mathbf{Z} \text{ is symmetric},\ \boldsymbol{\Omega} - \mathbf{Z} = \mathbf{0}. \end{aligned} \tag{1.8}$$

Under certain conditions, the solutions of the primal and dual problems coincide.

Define the augmented Lagrangian function

$$L_b(\boldsymbol{\Omega}, \mathbf{Z}, V) = \operatorname{trace}(\mathbf{S}\boldsymbol{\Omega}) - \log\det(\boldsymbol{\Omega}) + \lambda \sum_{j \neq l} \tau_{jl} \mathbf{Z}_{jl} + \operatorname{trace}[\mathbf{V}^\intercal(\boldsymbol{\Omega} - \mathbf{Z})] + \frac{b}{2}\|\boldsymbol{\Omega} - \mathbf{Z}\|_F^2.$$

The ADMM algorithm provides iterative formulas that approach the solution of the dual problem. The iteration formulas for (1.8) are given by

$$\begin{aligned} \boldsymbol{\Omega}^{(t+1)} &= \underset{\boldsymbol{\Omega}}{\operatorname{argmin}} \frac{\partial L_b(\boldsymbol{\Omega}, \mathbf{Z}^{(t)}, \mathbf{V}^{(t)})}{\partial \boldsymbol{\Omega}} = \mathbf{Y} \left\{ \frac{1}{2} \left[\boldsymbol{\Lambda} + \left(\boldsymbol{\Lambda}^2 + \frac{4}{b} \mathbf{I}_{pM} \right)^{1/2} \right] \right\} \mathbf{Y}^\intercal \\ \mathbf{Z}_{jl}^{(t+1)} &= \underset{\mathbf{Z}}{\operatorname{argmin}} \frac{\partial L_b(\boldsymbol{\Omega}^{(t+1)}, \mathbf{Z}, \mathbf{V}^{(t)})}{\partial \mathbf{Z}_{jl}} \\ &= \begin{cases} \boldsymbol{\Omega}_{jj}^{(t+1)} + \frac{1}{b}\mathbf{V}_{jj}^{(t)}, & \text{if } j = l \\ \max\left(\mathbf{0}, 1 - \frac{\lambda \tau_{jl}}{b\|\boldsymbol{\Omega}_{jl}^{(t+1)} + \frac{1}{b}\mathbf{V}_{jl}^{(t)}\|_F} \right) (\boldsymbol{\Omega}_{jl}^{(t+1)} + \frac{1}{b}\mathbf{V}_{jl}^{(t)}), & \text{if } j \neq l \end{cases} \\ \mathbf{V}^{(t+1)} &= \mathbf{V}^{(t)} + b(\boldsymbol{\Omega}^{(t+1)} - \mathbf{Z}^{(t+1)}), \end{aligned}$$

where $\mathbf{Y}$ is a matrix whose columns are the eigenvectors, $\boldsymbol{\Lambda}$ is a diagonal matrix of the eigenvalues, obtained by performing eigenvalue decomposition on

$$\mathbf{Z}^{(t)} - \frac{1}{b}(\mathbf{S} + \mathbf{V}^{(t)}),$$

and b is a positive constant that affects the convergence speed and accuracy of the algorithm. The initial values for $\mathbf{Z}$ and $\mathbf{V}$ are $\mathbf{1}_{pM}\mathbf{1}_{pM}^\intercal$ and $\mathbf{0}_{pM}$, respectively. From

the formulas of the updates, we can see that the computational complexity for the ADMM algorithm at each iteration is that of an eigenvalue decomposition of a matrix in $\mathbb{R}^{pM\times pM}$, which is $\mathcal{O}(p^3M^3)$.

When it comes to estimating the graph, we do not use a threshold to determine the zero components of $\mathbf{\Omega}$ produced by the ADMM. Instead, we use the zero entries of the dual variable $\mathbf{Z}$ produced by the ADMM, for two reasons. First, as it can be seen by the iteration formula, $\mathbf{Z}$ is a thresholding rule. Second, the ADMM theory states that as the number of iterations increase, the updates of $\mathbf{\Omega}$ and $\mathbf{Z}$ converge to the same point.

Algorithm 1: ADMM algorithm for solving (1.8) in R

Input : Sample Covariance Matrix $\mathbf{S}$, Matrix of weights $\mathbf{T} = (\tau_{jl})$, Tuning parameter λ, Step size b, Maximum number of iterations max_iter

Output: $(\mathbf{\Omega}, \mathbf{Z}, \mathbf{V})$

begin

  $\mathbf{Z} \leftarrow \mathbf{1}_{pM}\mathbf{1}_{pM}^{\intercal}$, $\mathbf{V} \leftarrow \mathbf{0}_{pM}$, iter $\leftarrow 1$,

  $\mathbf{X} \leftarrow \begin{pmatrix} \mathbf{1}_M & \dots & \mathbf{0}_M \\ \vdots & \ddots & \vdots \\ \mathbf{0}_M & \dots & \mathbf{1}_M \end{pmatrix}$

  while $iter \leq max\_iter$ **do**

    temp $\leftarrow$ eigen $\left(\mathbf{Z} - \frac{1}{b}(\mathbf{S} + \mathbf{V})\right)$

    $\mathbf{Y} \leftarrow$ temp\$vectors; $\mathbf{\Lambda} \leftarrow \text{diag}(temp\$values, nrow = pM)$

    $\mathbf{\Omega} \leftarrow \mathbf{Y}\left\{\frac{1}{2}\left[\mathbf{\Lambda} + \left(\mathbf{\Lambda}^2 + \frac{4}{b}\mathbf{I}_{pM}\right)^{1/2}\right]\right\}\mathbf{Y}^{\intercal}$

    $\mathbf{Z} \leftarrow \left\{\left[\text{pmax}\left(\mathbf{0}_p, \mathbf{1}_p - \frac{\lambda\mathbf{T}}{[\mathbf{X}^{\intercal}(b\mathbf{\Omega}+\mathbf{V})^2\mathbf{X}]^{1/2}}\right)\right] \otimes \mathbf{1}_M\mathbf{1}_M^{\intercal}\right\} \circ \left(\mathbf{\Omega} + \frac{1}{b}\mathbf{V}\right)$

    $\mathbf{V} \leftarrow \mathbf{V} + b(\mathbf{\Omega} - \mathbf{Z})$

    iter $\leftarrow$ iter $+ 1$

  end

  return($\mathbf{\Omega}$,$\mathbf{Z}$,$\mathbf{V}$)

end

1.4 Asymptotics

In this section we prove the asymptotic consistency of the one-step version of (1.7), that is, the estimator produced by the first step of the LLA algorithm uncovers the true graph with probability tending to 1.

The asymptotic consistency of the fglasso estimator was established in Qiao et al. (2019). The main difference of our setting is twofold. First, we do not make a distinction between the truncated and true process, as we assume that the dimension of the Hilbert space is known for a given sample size n. Second, we have to take into account the weights τ_{jl} that accommodate the common structure, which adds an extra layer of complexity.

We denote the true precision matrix by $\mathbf{\Omega}_0^{(k)} = (\mathbf{\Sigma}_0^{(k)})^{-1}$, the number of principal components by M_n, the number of stochastic processes by p_n, and define the degree of the graph

$$d_n^{(k)} = \max_{j=1,\ldots,p_n} \operatorname{card}\left(\left\{l : l \neq j, \mathbf{\Omega}_{0,jl}^{(k)} \neq \mathbf{0}\right\}\right),$$

where card denotes the cardinality of the set. In our framework we assume that all three quantities $M_n, p_n, d_n^{(k)}$ diverge to infinity.

Let $\mathbf{A}$ be the block-matrix $(\mathbf{A}_{jl})$, with $\mathbf{A}_{jl} \in \mathbb{R}^{M\times M}$. Define the M-block versions of the ℓ_∞-matrix norm, the ℓ_∞-vector norm, and the ℓ_1-matrix norm to be

$$\begin{aligned}
\|\mathbf{A}\|_\infty^{(M)} &= \max_{j=1,\ldots,p} \sum_{l=1}^{p} \|\mathbf{A}_{jl}\|_F, \\
\|\mathbf{A}\|_{\max}^{(M)} &= \max_{1\leq j,l\leq p} \|\mathbf{A}_{jl}\|_F, \\
\|\mathbf{A}\|_1^{(M)} &= \max_{l=1,\ldots,p} \sum_{j=1}^{p} \|\mathbf{A}_{jl}\|_F,
\end{aligned}$$

respectively. Similar block versions of these norms are also going to be used for various

block matrices and vectors, and their definition will be implied. For two sequence of real numbers a_n, b_n we denote $a_n \asymp b_n$ if there exist positive constants c_1, c_2 such that $c_1 \leq |a_n|/|b_n| \leq c_2$ for all n.

Let $\mathcal{B} = \{(i_1, j_{11}), \ldots, (i_1, j_{1r_1}), \ldots, (i_q, j_{q1}), \ldots, (i_1, j_{qr_q})\}$, such that $i_1 < \ldots < i_q$, and $j_{sm} < j_{sl}$ for all s and $m < l$. We define

$$\mathbf{A}_{\mathcal{B}} = \left(\text{vec}(\mathbf{A}_{i_1 j_{11}})^\intercal, \ldots, \text{vec}(\mathbf{A}_{i_1 j_{1r_1}})^\intercal, \ldots, \text{vec}(\mathbf{A}_{i_q j_{q1}})^\intercal, \ldots, \text{vec}(\mathbf{A}_{i_q j_{qr_q}})^\intercal\right)^\intercal.$$

Let $\mathbf{\Gamma}^{(k)} = (\mathbf{\Omega}_0^{(k)})^{-1} \otimes (\mathbf{\Omega}_0^{(k)})^{-1}$. We use $\mathbf{\Gamma}^{(k)}_{\mathcal{BC}} \in \mathbb{R}^{M^2|\mathcal{B}| \times M^2|\mathcal{C}|}$ to denote the submatrix of $\mathbf{\Gamma}^{(k)}$ with blocks $\mathbf{\Gamma}^{(k)}_{(i,j),(m,l)} \in \mathbb{R}^{M^2 \times M^2}$, where $(i,j) \in \mathcal{B}$ and $(m,l) \in \mathcal{C}$. To construct the matrix $\mathbf{\Gamma}^{(k)}_{\mathcal{BC}}$ we first fix the coordinates (i,j) to locate the block $\text{cov}(\mathbf{a}_i^{(k)}, \mathbf{a}_j^{(k)}) \otimes \mathbf{\Sigma}_0^{(k)}$ and then we fix the coordinates (m,l) to locate the block $\text{cov}(\mathbf{a}_i^{(k)}, \mathbf{a}_j^{(k)}) \otimes \text{cov}(\mathbf{a}_m^{(k)}, \mathbf{a}_l^{(k)})$. For a set D, we denote by D^c its complement. Let $\mathcal{S}^{(k)} = \mathcal{E}^{(k)} \cup \{(1,1), \ldots, (p,p)\}$, where $\mathcal{E}^{(k)} = \{(j,l) : \mathbf{\Omega}_{0,jl}^{(k)} \neq \mathbf{0}\}$. Define the quantities

$$C_{\mathbf{\Sigma}}^{(k)} = \|(\mathbf{\Omega}_0^{(k)})^{-1}\|_\infty^{(M)}, \quad C_{\mathbf{\Gamma}}^{(k)} = \|(\mathbf{\Gamma}_{\mathcal{SS}}^{(k)})^{-1}\|_\infty^{(M)}, \quad C_{\mathbf{\Gamma}^2}^{(k)} = \|(\mathbf{\Gamma}_{\mathcal{S}^c\mathcal{S}}^{(k)} \mathbf{\Gamma}_{\mathcal{SS}}^{(k)})^{-1}\|_\infty^{(M^2)}. \quad (1.9)$$

We first need to find conditions to establish concentration bounds for all entries of $\hat{\mathbf{\Sigma}}^{(k)} - \mathbf{\Sigma}_0^{(k)}$. To do so we adopt the same conditions as in Qiao et al. (2019).

Condition 1. *(i) The number of principal components, M_n, satisfies $M_n \asymp n^\alpha$ with some constant $\alpha \geq 0$; (ii) For each $j \in \mathcal{V}$, the eigenvalue sequence $\{\lambda_{jm}^{(k)}\}_{m=1}^{M_n}$ is decreasing; (iii) There exists some constant $\beta > 1$ with $\alpha\beta < 1/4$ such that $\lambda_{jm}^{(k)} \asymp m^{-\beta}$ and $d_{jm}^{(k)} \lambda_{jm}^{(k)} = \mathcal{O}(m)$ for each $m = 1, \ldots, M_n$ and $j \in \mathcal{V}$, where $d_{jm}^{(k)} = 2\sqrt{2}\{(\lambda_{j(m-1)}^{(k)} - \lambda_{jm}^{(k)})^{-1}, (\lambda_{jm}^{(k)} - \lambda_{j(m+1)}^{(k)})^{-1}\}$.*

Parameter α controls the dimension of the Hilbert spaces, while parameter β determines how fast the eigenvalues $\lambda_{jm}^{(k)}$ and eigengaps $\lambda_{jm}^{(k)} - \lambda_{j(m-1)}^{(k)}$ converge to zero. We also need a condition for the weights τ_{jl} in order to establish the optimality

of our estimator.

Condition 2. *For any $\gamma > 2$, there exist positive numbers $a_1^{(k)}, a_2^{(k)}$ such that $a_1^{(k)} > a_2^{(k)} C_{\mathbf{\Gamma}^2}^{(k)}$ and*

$$\min_{(j,l)\in\mathcal{S}^{(k)c}} \tau_{jl} > a_1^{(k)}, \qquad \max_{(j,l)\in\mathcal{S}^{(k)}} \tau_{jl} < a_2^{(k)},$$

with probability greater than $1 - (M_n p_n)^{2-\gamma}$ each.

To gain intuition about this condition, let us assume that the initial estimator $\hat{\mathbf{\Omega}}_{(0)}$ is not far from the truth. In the best case scenario, where the edge sets $\mathcal{S}^{(k)}$ of all the subpopulations are identical and equal to $\mathcal{A}$, Condition 2 can be seen as a minimum and maximum signal strength for the existent and absent edges respectively.

$$\begin{aligned}
(j,l) \notin \mathcal{A} \quad &\Rightarrow \quad \sum_{k=1}^{K} \|\hat{\mathbf{\Omega}}_{(0),jl}^{(k)}\|_F \to 0 \quad \Rightarrow \quad \tau_{jl} \to \infty, \\
(j,l) \in \mathcal{A} \quad &\Rightarrow \quad \sum_{k=1}^{K} \|\hat{\mathbf{\Omega}}_{(0),jl}^{(k)}\|_F \nrightarrow 0 \quad \Rightarrow \quad \tau_{jl} \nrightarrow \infty.
\end{aligned}$$

The richer the diversity of the graphs, the harder it is for this condition to hold.

We are now ready to prove graph selection consistency for each subpopulation $k = 1, \ldots, K$.

Theorem 3. *Suppose conditions 1 and 2 hold, $\gamma > 2$,*

$$\lambda_n = \frac{2(1 + C_{\mathbf{\Gamma}^2}^{(k)}) M_n}{a_1 - a_2 C_{\mathbf{\Gamma}^2}^{(k)}} \sqrt{\frac{\log C_2 + \gamma \log(M_n p_n)}{C_1 n^{1-2\alpha(1+\beta)}}},$$

and

$$\min_{(j,l)\in\mathcal{S}^{(k)}} \|\mathbf{\Omega}_{0,jl}^{(k)}\|_F > \min\left\{ \frac{1}{3C_{\mathbf{\Sigma}}^{(k)} d_n^{(k)}}, \frac{1}{3(C_{\mathbf{\Sigma}}^{(k)})^3 C_{\mathbf{\Gamma}}^{(k)} d_n^{(k)}} \right\}.$$

Then, for all n satisfying the lower bound

$$n^{1-2\alpha(1+\beta)} > \frac{\log[C_2(M_n p_n)^\gamma]}{C_1} \max\left\{ \frac{1}{C_1}, \frac{2M_n C_{\mathbf{\Gamma}}^{(k)} \left[1 + \frac{2a_2(1+C_{\mathbf{\Gamma}^2}^{(k)})}{a_1 - a_2 C_{\mathbf{\Gamma}^2}^{(k)}}\right]}{\min\left\{\frac{1}{3C_{\mathbf{\Sigma}}^{(k)} d_n^{(k)}}, \frac{1}{3(C_{\mathbf{\Sigma}}^{(k)})^3 C_{\mathbf{\Gamma}}^{(k)} d_n^{(k)}}\right\}}, 6M_n d_n^{(k)} (C_{\mathbf{\Gamma}}^{(k)})^2 C_{\mathbf{\Gamma}^2}^{(k)} \left[1 + \frac{2a_2(1+C_{\mathbf{\Gamma}^2}^{(k)})}{a_1 - a_2 C_{\mathbf{\Gamma}^2}^{(k)}}\right]^2 \right\}^2,$$

we have $\hat{\mathcal{S}}^{(k)} = \mathcal{S}^{(k)}$ with probability greater than $1 - 3(M_n p_n)^{2-\gamma}$.

1.5 Simulations

In this section we use simulation to compare the performance of the JFGGM with the FGGM applied separately to each subpopulation. To generate the data, we first construct the edge sets for all subpopulations, and then form the precision matrices described in section 2.

To form the edge sets $\mathcal{E}^{(1)}, \ldots, \mathcal{E}^{(K)}$, we follow two steps:

1. Randomly choose a set of pairs $(j, l) \in \mathcal{V} \times \mathcal{V}$, $j < l$, as a percentage s of the total number of edges $\binom{p}{2}$. This set constitutes the common graphical structure of the K subpopulations and is denoted by $\mathcal{A}$.

2. For each subpopulation k, randomly choose a set of pairs as a percentage ρ of the number of common edges and denote this set by $\mathcal{B}^{(k)}$. The sets $\mathcal{B}^{(1)}, \ldots, \mathcal{B}^{(K)}$ must satisfy

$$\bigcup_{k=1}^{K} \mathcal{B}^{(k)} \cap \mathcal{A} = \varnothing,$$

$$\text{and} \quad \bigcap_{k=1}^{K} \mathcal{B}^{(k)} = \varnothing.$$

   These sets are the individual edge structure of each subpopulation. Combining the above, we define $\mathcal{E}^{(k)} = \mathcal{A} \cup \mathcal{B}^{(k)}$, $k = 1, \ldots, K$.

To form the precision matrices $\mathbf{\Omega}^{(1)}, \ldots, \mathbf{\Omega}^{(K)}$, we follow three steps:

1. Generate $a_{jl}^{(k)}$ independently from $\mathcal{U}(0,1)$ and form $\mathbf{A}^{(k)} = (\mathbf{A}_{jl}^{(k)})$, $k = 1, \ldots, K$, by

$$\mathbf{A}_{jl}^{(k)} = \begin{cases} a_{jl}^{(k)}\mathbf{I}_M, & (j,l) \in \mathcal{E}^{(k)} \\ \mathbf{I}_M, & j = l \\ \mathbf{0}, & \text{otherwise} \end{cases},$$

2. To ensure symmetry, let

$$\mathbf{B}^{(k)} = \frac{\mathbf{A}^{(k)} + \mathbf{A}^{(k)\intercal}}{2}.$$

   Note that by construction, the diagonal elements of $\mathbf{B}^{(k)}$ are 1.

3. Let $b_{rs}^{(k)}$, $r = 1, \ldots, pM$, $s = 1, \ldots, pM$, be the (r,s)-th element of $\mathbf{B}^{(k)}$. To ensure positive definiteness, we use Gershgorin's Circle Theorem (Bell, 1965) to define the precision matrices $\mathbf{\Omega}^{(k)} = (\omega_{rs}^{(k)})$, $k = 1, \ldots, K$, such that

$$\omega_{rs}^{(k)} = \begin{cases} \frac{b_{rs}^{(k)}}{\sum_{q \neq r} |b_{rq}^{(k)}|}, & r \neq s \quad \text{and} \quad \sum_{q \neq r} |b_{rq}^{(k)}| > 0 \\ 0, & r \neq s \quad \text{and} \quad \sum_{q \neq r} |b_{rq}^{(k)}| = 0 \cdot \\ 1, & r = s \end{cases}$$

With $\mathbf{\Omega}^{(1)}, \ldots, \mathbf{\Omega}^{(K)}$ thus constructed, we are now ready to generate the observed data for each subpopulation. To do so, we follow three steps:

1. Choose a basis $\boldsymbol{\phi}_j^{(k)} = (\phi_{j1}^{(k)}, \ldots, \phi_{jM}^{(k)})^\intercal$, for all j, k.

2. Generate $\mathbf{a}_i^{(k)} = (\mathbf{a}_{i1}^{(k)\intercal}, \ldots, \mathbf{a}_{ip}^{(k)\intercal})^\intercal$ from a $\mathcal{N}_{pM}(\mathbf{0}, (\mathbf{\Omega}^{(k)})^{-1})$ distribution, for $i = 1, \ldots, n$ and all $k = 1, \ldots, K$.

3. Create the observed data $\mathbf{h}_i^{(k)} = (\mathbf{h}_{i1}^{(k)\intercal}, \ldots, \mathbf{h}_{ip}^{(k)\intercal})^\intercal$, where

$$\mathbf{h}_{ij}^{(k)}(t) = \mathbf{a}_{ij}^{(k)\intercal} \boldsymbol{\phi}_j^{(k)}(t) + \epsilon_{ijt}^{(k)},$$

and $\epsilon_{ijt}^{(k)}$ are i.i.d. $\mathcal{N}(0, \sigma^2)$, for all i, j, k, t.

We compare the JFGGM with the separate estimation with the FGGM on 12 different scenarios, which consist of all combinations of the variables $n = 100$ and 200, $p = 80$ and 100, $\rho = 0, 0.5$ and 1. In all settings, the common structure of the $K = 3$ subpopulations consists of $s = 5\%$ of all possible edges $\binom{p}{2}$. The basis for the functional data, for each population, is $\{1, \sin t, \cos t\}$. Thus $M = 3$. The variance of the error is $\sigma^2 = 0.05$, the number of time points is $\nu = 100$, starting from 0 and ending at 1.

Receiver operating characteristic (ROC) curves are used to evaluate the performance of the two competing methods. For these curves we plot the average proportion of correctly detected links (ATPR) against the average proportion of falsely detected links (AFPR), over a range of values of λ. In particular,

$$\mathrm{AFPR}(\lambda) = \frac{1}{K} \sum_{k=1}^{K} \frac{\sum_{1 \le j < l \le p} \mathbb{1}\left(\boldsymbol{\Omega}_{0,jl}^{(k)} = \mathbf{0}\,,\, \hat{\boldsymbol{\Omega}}_{jl}^{(k)}(\lambda) \neq \mathbf{0}\right)}{\sum_{1 \le j < l \le p} \mathbb{1}\left(\boldsymbol{\Omega}_{0,jl}^{(k)} = \mathbf{0}\right)}$$

$$\mathrm{ATPR}(\lambda) = \frac{1}{K} \sum_{k=1}^{K} \frac{\sum_{1 \le j < l \le p} \mathbb{1}\left(\boldsymbol{\Omega}_{0,jl}^{(k)} \neq \mathbf{0}\,,\, \hat{\boldsymbol{\Omega}}_{jl}^{(k)}(\lambda) \neq \mathbf{0}\right)}{\sum_{1 \le j < l \le p} \mathbb{1}\left(\boldsymbol{\Omega}_{0,jl}^{(k)} \neq \mathbf{0}\right)},$$

where $\mathbb{1}$ is the indicator function, $\boldsymbol{\Omega}_0^{(k)}$ is the true precision matrix and $\hat{\boldsymbol{\Omega}}^{(k)}(\lambda)$ is the estimated precision matrix using tuning parameter λ. The above quantities are calculated for 100 values of λ, where 90 of them are in $[0, 0.67]$ and 10 of them are in $[0.6784, 1.5]$. All of them are equally spaced in their respective intervals and start

$\big((p,n),\rho\big)$	Method	1	0.5	0
$(80,100)$	JFGGM	0.68	0.76	0.90
	FGGM	0.64	0.72	0.83
$(80,200)$	JFGGM	0.76	0.83	0.96
	FGGM	0.73	0.80	0.91
$(100,100)$	JFGGM	0.65	0.70	0.83
	FGGM	0.61	0.66	0.75
$(100,200)$	JFGGM	0.71	0.79	0.91
	FGGM	0.68	0.74	0.85

Table 1.1. Table with the area under the ROC curves from Figure 2.1

and end at the boundaries of their respective intervals. Each scenario is simulated 5 times, and the final ROC curve is the average of them.

Figure 1.1 shows the ROC curves by the JFGGM and those by separate estimation with FGGM. Overall, our method outperforms separate estimation with the FGGM. When the number of individual edges is the same with the number of common edges the two methods are close. As the number of individual edges decreases, JFGGM significantly outperforms separate estimation. In addition, the JFGGM performs well on high dimensions with a relatively small sample size. Table 1.1 provides the area under the curve for each scenario, verifying the visual results. In all scenarios, the ADMM algorithm produced accurate estimators after no more than 100 iterations.

1.6 Application

In this section we apply the JFGGM to the EEG dataset mentioned in the Introduction. The dataset consists of two groups of subjects: alcoholic and non-alcoholic. The first group is comprised of 77 subjects and the second of 45. Sixty four electrodes were strategically placed on each subject's scalp, which measured their brain activity while they were shown pictures of a variety of objects. Measurements of brain activity were sampled at 256 Hz for 1 second. The purpose of this study is to uncover genetic predisposition to alcoholism.

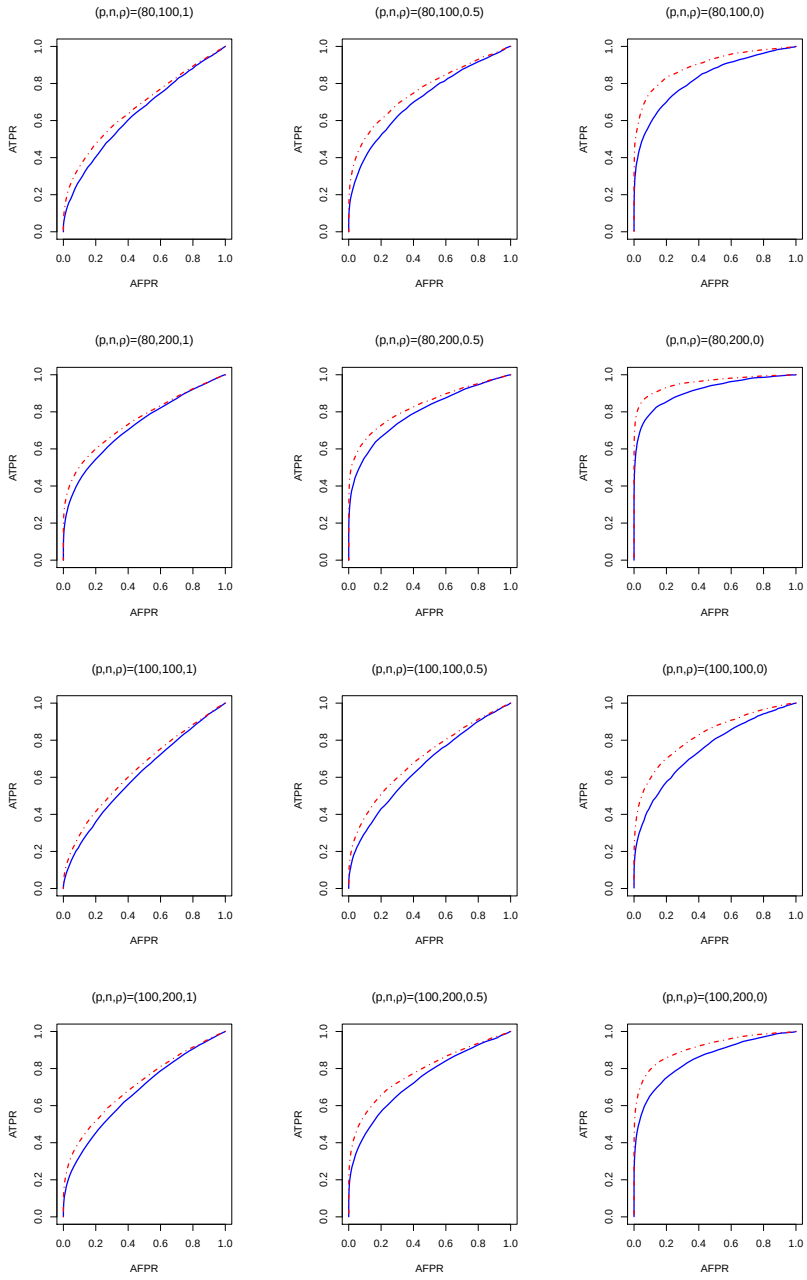

Figure 1.1. ROC curves by JFGGM (red dashed line) and by separate estimation with FGGM (solid blue line) for different combinations of (p, n, ρ).

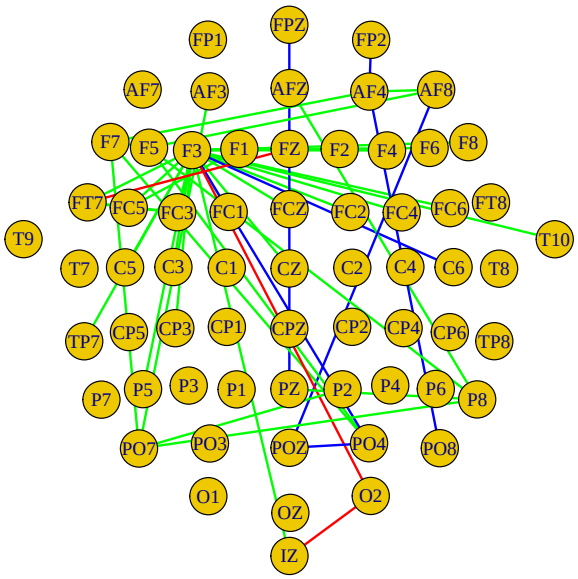

Figure 1.2. Graph of the 64 electrodes, produced by the JFGGM. Green represents the common edges, red the edges unique to the alcoholic group, and blue the edges unique to the non-alcoholic group.

Li et al. (2010) applied a dimension folding method to the EEG dataset where brain activity recorded at each electrode was treated as a multivariate random vector of 256 entries. Qiao et al. (2019) and Solea and Li (2019) treated the same quantities as stochastic processes. Both of them, however, apply their methods separately to the alcoholic and the control group, losing the joint information for prediction. In contrast, with the JFGGM we can exploit the information that exists across these

groups with joint estimation of their graphs. It is also computationally efficient, since we would have to choose two λ's for the separate estimation, one for each group. Finally, JFGGM makes comparison of the two graphs easier, since the level of sparsity for both is controlled by the same λ.

Figure 1.2 shows the graph estimated by the JFGGM among the 64 stochastic processes describing the brain activity at the electrodes. The layout of the vertices represents the position of the electrodes on the scalp, with the top side being the front of the skull. We chose $\lambda = 3.5$ so that the sparsity level of the graph is at 2.5%. For every stochastic process of each group, we estimated a Karhunen-Loeve expansion of $M = 5$ eigenfunctions.

From Figure 1.2 we see that the graphs of the two groups have a rich common structure, which indicates that our joint estimation procedure has a significant advantage. The majority of the common structure is located at the front side of the left hemisphere of the scalp. Furthermore, the two groups exhibit important differences: the edges unique to the alcoholic group are observed on an acute diagonal strip near the center of the scalp, whereas the edges unique to the non-alcoholic group occupy the right hemisphere of the scalp. Finally, the non-alcoholic group has seven extra individual edges, while the alcoholic group has only three, indicating heightened brain activity in the control group.

1.7 Discussion

In this paper, we develop a method for jointly estimating functional graphical models. The assumption is that these graphs share a significant common structure. We can see the common structure in the distribution of the Karhunen-Loeve expansion coefficients $\mathcal{N}(\mathbf{0}, \mathbf{\Omega}^{(k)})$, for each subpopulation k. Each precision matrix $\mathbf{\Omega}^{(k)}$ can be decomposed into the Hadamard product of a matrix $\mathbf{\Theta}$, that is common for all subpopulations and expresses the common structure, and a subpopulation specific matrix

$\mathbf{\Gamma}^{(k)}$, that expresses the individual structure of the k-th subpopulation. By estimating a single graph for all the data, we would be ignoring the individuality of each subpopulation. On the other hand, by estimating a single graph for each subpopulation, we would not be using the existence of the common structure to our advantage. We accommodate our method with two optimization algorithms. The first dealing with the nonconvex nature of the objective function, and the second dealing with the nonsmooth nature of the first algorithm. To complete the theoretical novelty of our model, we establish the asymptotic consistency of our estimator. The theoretical accuracy of the JFGGM is demonstrated in a simulation experiment against a separate estimation of the graphs with the FGGM method developed in Qiao et al. (2019).

To conclude our work, we would like to point three possible extensions of this article. First, we have assumed that the gaussian processes associated with each subpopulation are realizations of Hilbert spaces with the same dimension M. Hence, the first possibility could be to extended this method to the case where the dimension of the Hilbert spaces is subpopulation specific. Second, the assumption that the Karhunen-Loeve expansion coefficients follow directly a $\mathcal{N}(\mathbf{0}, \mathbf{\Omega}^{(k)})$ for each subpopulation k, can be further relaxed by assuming instead that the coefficients may not follow initially a multivariate normal, but there exists a transformation such that the transformed coefficients follow it, as described in Solea and Li (2019). A third possible extension would be to get rid of the multivariate normal distributional assumption altogether for all subpopulations by replacing the conditional independence relationship which defines the graphs with the additive conditional independence relationship studied in Li and Solea (2018).

Chapter 2

Simultaneous Estimation of Graphical Models by Neighborhood Selection

2.1 Introduction

Graphical models are a useful tool for constructing statistical networks for a wide range of applications such as speech recognition, computer vision and genomics. One recent focus in this research is the simultaneous estimation of multiple graphs from several subpopulations. The current approach to this problem is by modifying the graphical lasso (Yuan and Lin, 2007) to take into account of the common structure in the subpopulations (Guo et al., 2011; Danaher et al., 2014). However, this involves the spectral decomposition of large matrices and is computationally infeasible for some biological applications. In this paper we propose to generalize the neighborhood selection (Meinshausen et al., 2006) to the simultaneous estimation problem, which saves substantial amount of computing time and can be applied to large genetic networks.

Consider a p-dimensional random vector $\mathbf{X} = (X_1, \ldots, X_p)^\intercal$. The graphical model

of $\mathbf{X}$ is represented by an undirected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V} = \{1, \ldots, p\}$ is the set of vertices and $\mathcal{E} = \{(i,j) : i,j \in \mathcal{V}, i \neq j\}$ is the set of edges. Because for an undirected graph (i,j) and (j,i) represent the same edge, we assume $i < j$ for $(i,j) \in \mathcal{E}$ without loss of generality. In a statistical graphical model, the set of edges $\mathcal{E}$ is defined by the relation

$$(i,j) \notin \mathcal{E} \quad \Leftrightarrow \quad X_i \perp\!\!\!\perp X_j | \mathbf{X}_{-\{i,j\}}, \tag{2.1}$$

where $\mathbf{X}_{-\{i,j\}}$ is the vector $\mathbf{X}$ with its i-th and j-th components removed, that is $\mathbf{X}_{-\{i,j\}} = \{X_k : k \in \mathcal{V} \backslash \{i,j\}\}$, and the notation $A \perp\!\!\!\perp B | C$ means A and B are conditionally independent given C. Of special interest is the case where $\mathbf{X}$ follows a multivariate Gaussian distribution $\mathcal{N}_p(\mathbf{0}, \mathbf{\Sigma})$. Let $\mathbf{\Omega} = \mathbf{\Sigma}^{-1}$ be the precision matrix and ω_{ij} the (i,j)-th component of $\mathbf{\Omega}$. Under the Gaussian assumption, because of the relation

$$X_i \perp\!\!\!\perp X_j | \mathbf{X}_{-\{i,j\}} \quad \Leftrightarrow \quad \omega_{ij} = 0,$$

the estimation of the edge set $\mathcal{E}$ is equivalent to the estimation of the positions of the zero components of the precision matrix. Lauritzen (1996) and Sinoquet (2014) give excellent reviews of the theory of statistical graphical models.

The problem of estimating the zero components of the precision matrix dates back to Dempster (1972). He proposed a forward selection procedure in which he starts with the sample covariance matrix and gradually chooses elements of the inverse of the sample covariance matrix to go to zero. The procedure is stopped when adding a new zero component does not change the fit significantly. However, the overall error properties of the step-wise procedure are hard to study. Drton and Perlman (2004) proposed a joint hypothesis testing under the same significance level for all the pairs of nodes.

Meinshausen et al. (2006) developed the neighborhood selection method that uses an equivalent relationship to (2.1) to define the edges of the graph. They applied penalized regressions to find the neighborhood of each node and assemble these neighborhoods together to recover the entire graph. Up until that point, graph and precision matrix estimation were separate problems. Yuan and Lin (2007) proposed a penalized likelihood method, called the graphical lasso, to estimate the precision matrix that forces its small components to zero. Rocha et al. (2008) developed a method that allows for joint estimation of all the neighborhoods in the graph that is based on a minimization of a penalized pseudo-likelihood. Other important recent advances include Liu et al. (2009, 2012) and Xue et al. (2012), which considered the non-Gaussian case.

The simultaneous estimation problem was first considered by Guo et al. (2011), which assumes that the data come from different subpopulations that share a common graph structure but also have significant differences. Under such circumstances, to estimate each graph separately would be to waste information in uncovering the common structure. On the other hand, estimating a single graph by merging data from all the subpopulations together would ignore their differences. Guo et al. (2011) proposed a simultaneous estimation procedure by decomposing each precision matrix into two components: one common to all subpopulations and another specific to each subpopulation. More recently, Danaher et al. (2014) introduced two different simultaneous estimation methods: the fused graphical lasso (FGL) and the group graphical lasso (GGL). By making some additional assumptions for the common structure, they overcame the problem of the non-convex penalty, resulting in a more computationally reliable method. Other important contributions in the simultaneous estimation of multiple precision matrices include Cai et al. (2016) and Saegusa and Shojaie (2016). The former developed a weighted ℓ_∞/ℓ_1 constrained estimator, following the steps of the Dantzig Selector (Candes et al., 2007), which can efficiently estimate high-

dimensional precision matrices. The latter introduced a general framework for the joint estimation of precision matrices with the use of a Laplacian graph penalty.

Our proposed method, which we call Simultaneous Neighborhood Selection (SNS), has the following advantages over the above methods. Compared with Guo et al. (2011) and Saegusa and Shojaie (2016), instead of trying to adapt graphical lasso to simultaneous estimation, we adapt neighborhood selection to simultaneous estimation by introducing an extra penalty term that enforces the common structure among the subpopulations. We further simplify the penalty with local linear approximation (Zou and Li, 2008) which reduces the procedure to that of adaptive lasso. This enables us to bypass the large eigenvalue decomposition. Compared with the FGL in Danaher et al. (2014), event though it is a computationally fast method, the penalty is designed to enforce similarity not only in the edge sets but also in the edge values, whereas SNS only enforces similarity among the edge sets, which is closer to our goal of simultaneous estimation. Finally, compared with the GGL in Danaher et al. (2014) and the method in Cai et al. (2016), since these methods use the ℓ_2 penalty, they do not induce sparsity on a common level, while the SNS does.

The rest of the paper is organized as follows. In section 2 we give an overview of the penalized neighborhood selection method. In section 3 we propose our new method for simultaneous estimation via neighborhood selection. In section 4 we develop the algorithm for finding the minimum of the objective function and demonstrate its computational complexity. In section 5 we compare the performance of SNS with existing methods on simulated datasets, both in terms of ROC curves and CPU time. In section 5 we apply SNS on a lung cancer dataset.

2.2 Methodology

2.2.1 Neighborhood Selection Method

We first give an overview of the neighborhood selection method for estimating a single graph. Suppose $\mathbf{X} = (X_1, \ldots, X_p)^\intercal$ is a random vector that follows a multivariate Gaussian distribution $\mathcal{N}_p(\mathbf{0}, \mathbf{\Sigma})$, with precision matrix $\mathbf{\Omega} = \mathbf{\Sigma}^{-1}$. The neighborhood of a vertex $i \in \mathcal{V}$ is the smallest set $\mathcal{A} \subset \mathcal{V}$ such that $X_i \perp\!\!\!\perp \mathbf{X}_{-\mathcal{A}\cup\{i\}} | \mathbf{X}_{\mathcal{A}}$, and is denoted by $\mathrm{ne}(i)$. Let $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ be the undirected graph with edge set $\mathcal{E}$ defined by the relation

$$(i,j) \in \mathcal{E} \quad \Leftrightarrow \quad j \in \mathrm{ne}(i). \tag{2.2}$$

It can be shown that the edge sets determined by the relations (2.1) and (2.2) are identical (Lauritzen, 1996, Proposition C.5). Since $\mathbf{X} \sim \mathcal{N}(\mathbf{0}, \mathbf{\Omega}^{-1})$, where $\mathbf{\Omega} = (\omega_{ij})$, then for each $j \in \mathcal{V}$, the conditional distribution of $X_j | \mathbf{X}_{-j}$ is $\mathcal{N}\left(\sum_{i \neq j} \theta_{ij} X_i, \sigma^2\right)$, where

$$\theta_{ij} = -\frac{\omega_{ij}}{\omega_{jj}} \quad , \quad \sigma^2 = \mathrm{Var}\left(X_j\right) - \mathrm{cov}\left(X_j, \mathbf{X}_{-j}\right) \mathrm{Var}\left(\mathbf{X}_{-j}\right)^{-1} \mathrm{cov}\left(\mathbf{X}_{-j}, X_j\right).$$

From this we can see that, under the assumption $\mathbf{X} \sim \mathcal{N}(\mathbf{0}, \mathbf{\Omega}^{-1})$,

$$\theta_{ij} = 0 \quad \Leftrightarrow \quad \omega_{ij} = 0 \quad \Leftrightarrow \quad X_i \perp\!\!\!\perp X_j | \mathbf{X}_{-\{i,j\}} \quad \Leftrightarrow \quad (i,j) \notin \mathcal{E}.$$

Thus, the edge set $\{(i,j) : i < j,\, \omega_{ij} \neq 0\}$ is the same as the set $\{(i,j) : i < j,\, \theta_{ij} \neq 0\}$, which is completely determined by the system of neighborhoods $\{\mathrm{ne}(i) : i \in \mathcal{V}\}$. In this way, estimating the graph reduces to neighborhood selection.

Another way to represent the statement $X_j | \mathbf{X}_{-j} \sim \mathcal{N}\left(\sum_{i \neq j} \theta_{ij} X_i, \sigma^2\right)$ is by the

linear regression model

$$X_j = \sum_{i \neq j} \theta_{ij} X_i + \epsilon,$$

where the error $\epsilon = X_j - \sum_{i \neq j} \theta_{ij} X_i$ follows a $\mathcal{N}(0, \sigma^2)$ distribution and is independent of $\sum_{i \neq j} \theta_{ij} X_i$. Therefore, finding the graph $\mathcal{G}$ boils down to regressing each variable X_j against the remaining $\mathbf{X}_{-j}$ and finding the zero coefficients θ_{ij}.

The neighborhood selection method for estimating $\mathcal{G}$ proceeds as follows. Define the matrix $\boldsymbol{\Theta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_p)$, where $\boldsymbol{\theta}_j = (\theta_{1j}, \ldots, \theta_{pj})^\intercal$ such that $\theta_{jj} = 0$ and θ_{lj} is the regression coefficient defined above for $l \neq j$. Suppose we observe an i.i.d. sample $\mathbf{X}_1, \ldots, \mathbf{X}_n$, where $\mathbf{X}_i = (X_{i1}, \ldots, X_{ip})^\intercal$. Let $\mathbb{X}$ denote the matrix $(\mathbf{X}_1 \ldots \mathbf{X}_n)^\intercal$ and $\mathbb{X}_j$ the j-th column of $\mathbb{X}$. We perform linear regression of $\mathbb{X}_j$ on $\{\mathbb{X}_1, \ldots, \mathbb{X}_p\} \backslash \{\mathbb{X}_j\}$ by minimizing the least squares criterion

$$\frac{1}{2n} ||\mathbb{X}_j - \mathbb{X}\boldsymbol{\theta}_j||_2^2,$$

where $\|\cdot\|_2$ denotes the ℓ_2 norm. To induce sparsity so that we can cover the $p > n$ case, we add an ℓ_1-penalty to produce the penalized estimator

$$\hat{\boldsymbol{\theta}}_j = \underset{\boldsymbol{\theta} \in \mathbb{R}^p, \theta_{jj}=0}{\operatorname{argmin}} \left(\frac{1}{2n} ||\mathbb{X}_j - \mathbb{X}\boldsymbol{\theta}||_2^2 + \lambda \|\boldsymbol{\theta}\|_1 \right), \quad j = 1, \ldots, p, \tag{2.3}$$

where λ is a nonnegative tuning parameter that controls the degree of sparsity. We rewrite the p equations in (2.3) into a matrix form as

$$\hat{\boldsymbol{\Theta}} = \underset{\boldsymbol{\Theta} \in \mathbb{R}^{p \times p}, \operatorname{diag}(\boldsymbol{\Theta}) = \mathbf{0}}{\operatorname{argmin}} \left(\frac{1}{2n} \|\mathbb{X}(\mathbf{I} - \boldsymbol{\Theta})\|_F^2 + \lambda \|\boldsymbol{\Theta}\|_1 \right),$$

where $\|\cdot\|_F$ and $\|\cdot\|_1$ denote the Frobenius norm and the elementwise matrix ℓ_1 norm, respectively, and $\operatorname{diag}(\boldsymbol{\Theta}) = (\theta_{11}, \ldots, \theta_{pp})^\intercal$.

Let $\hat{\boldsymbol{\theta}}_j$ be the solution of (2.3) and let $\hat{\text{ne}}(i) = \{i : i \neq j, \hat{\theta}_{ij} \neq 0\}$. Because $j \in \hat{\text{ne}}(i)$ does not imply $i \in \hat{\text{ne}}(j)$ and vice versa, Meinshausen et al. (2006) proposed two estimators for the edge set $\mathcal{E}$: a conservative estimator $\hat{\mathcal{E}} = \{(i,j) : j \in \hat{\text{ne}}(i) \text{ and } i \in \hat{\text{ne}}(j)\}$, and a liberal estimator $\hat{\mathcal{E}} = \{(i,j) : j \in \hat{\text{ne}}(i) \text{ or } i \in \hat{\text{ne}}(j)\}$.

2.2.2 Simultaneous Neighborhood Selection

We now consider simultaneous estimation of multiple graphs from several subpopulations. We assume that there are K different subpopulations whose graphs, though different, share a set of common edges. For each $k = 1, \ldots, K$, suppose we observe an i.i.d. sample $\mathbf{X}_1^{(k)}, \ldots, \mathbf{X}_{n_k}^{(k)}$, where $\mathbf{X}_i^{(k)} = \left(X_{i1}^{(k)}, \ldots, X_{ip}^{(k)}\right)^\intercal$. We assume that $\mathbf{X}_i^{(k)}$ is distributed as $\mathcal{N}_p(\mathbf{0}, \boldsymbol{\Sigma}^{(k)})$. Let $\mathbb{X}^{(k)}$ denote the matrix $\left(\mathbf{X}_1^{(k)}, \ldots, \mathbf{X}_{n_k}^{(k)}\right)^\intercal$ and $\boldsymbol{\Theta}^{(k)}$ the matrix of coefficients defined in section 2.2 for each subpopulation k.

To take advantage of the information across the subpopulations we reparameterize each $\boldsymbol{\Theta}^{(k)}$ as $\mathbf{H} \circ \boldsymbol{\Gamma}^{(k)}$, where $\circ$ is the Hadamard matrix product, $\mathbf{H} = (\eta_{lj})$ is a matrix common to all subpopulations and $\boldsymbol{\Gamma}^{(k)} = (\gamma_{lj}^{(k)})$ is a matrix specific to subpopulation k. To eliminate sign ambiguity we assume $\eta_{lj} \geq 0$ for all l and j, and to be consistent with the fact that $\theta_{jj}^{(k)} = 0$ we set $\eta_{jj} = \gamma_{jj}^{(k)} = 0$ for all j, k. In this reparameterization, the common factor η_{lj} controls the presence of vertex l in the neighborhood of j in all of the graphs, and $\gamma_{lj}^{(k)}$ accommodates the differences in the neighborhood of j between individual graphs. For the simultaneous estimation we propose to minimize

$$\frac{1}{2n} \sum_{k=1}^{K} \|\mathbb{X}^{(k)}(\mathbf{I} - \boldsymbol{\Theta}^{(k)})\|_F^2 + \lambda_1 \|\mathbf{H}\|_1 + \lambda_2 \sum_{k=1}^{K} \|\boldsymbol{\Gamma}^{(k)}\|_1 \tag{2.4}$$

over all $\mathbf{H}$ and $\boldsymbol{\Gamma}^{(k)}$ specified above, where $n = max\{n_k : k = 1, \ldots, K\}$. The first penalty function penalizes the common factors η_{lj} and is responsible for identifying the zeros across all coefficient vectors $\boldsymbol{\theta}_j^{(1)}, \ldots, \boldsymbol{\theta}_j^{(K)}$. That is, if η_{lj} is zero then vertex l is not in the neighborhood of j in all K graphs. The second penalty function penalizes

the individual factors $\gamma_{lj}^{(k)}$ and is responsible for identifying the zeros in $\boldsymbol{\theta}_j^{(k)}$ specific to each individual graph. That is, for a non-zero η_{lj} some of the coefficients $\gamma_{lj}^{(1)}, \ldots, \gamma_{lj}^{(K)}$ can be zero, which means that l may be absent from the neighborhood of j in some of the K graphs but present in others.

However, the objective function (2.4) is difficult to minimize because of its complexity. It involves two groups of variables over which we have to optimize, and two parameters that we have to tune. As will be shown in the Theorem 4, (2.4) is equivalent to the much simpler form

$$\frac{1}{2n}\sum_{k=1}^{K}\|\mathbb{X}^{(k)}(\mathbf{I}-\boldsymbol{\Theta}^{(k)})\|_F^2+2(\lambda_1\lambda_2)^{1/2}\sum_{l\neq j}\left(\sum_{k=1}^{K}|\theta_{lj}^{(k)}|\right)^{1/2}. \tag{2.5}$$

Let $\boldsymbol{\Theta}=(\boldsymbol{\Theta}^{(1)},\ldots,\boldsymbol{\Theta}^{(K)})$ and $\boldsymbol{\Gamma}=(\boldsymbol{\Gamma}^{(1)},\ldots,\boldsymbol{\Gamma}^{(K)})$. The proof of the following theorem can be found in the supplementary material.

Theorem 4. *If* $(\hat{\mathbf{H}},\hat{\boldsymbol{\Gamma}})$ *is a local minimizer of* (2.4)*, then there exists a local minimizer* $\hat{\boldsymbol{\Theta}}$ *of* (2.5) *such that* $\hat{\boldsymbol{\Theta}}^{(k)}=\hat{\mathbf{H}}\circ\hat{\boldsymbol{\Gamma}}^{(k)}$ *for all* k*. Conversely, if* $\hat{\boldsymbol{\Theta}}$ *is a local minimizer of* (2.5)*, then there exists a local minimizer* $(\hat{\mathbf{H}},\hat{\boldsymbol{\Gamma}})$ *of* (2.4) *such that* $\hat{\mathbf{H}}\circ\hat{\boldsymbol{\Gamma}}^{(k)}=\hat{\boldsymbol{\Theta}}^{(k)}$ *for all* k*.*

2.3 Computation

2.3.1 Penalty linearization

Writing $2(\lambda_1\lambda_2)^{1/2}$ as λ, the objective function (2.5) becomes

$$\frac{1}{2n}\sum_{k=1}^{K}\|\mathbb{X}^{(k)}(\mathbf{I}-\boldsymbol{\Theta}^{(k)})\|_F^2+\lambda\sum_{l\neq j}\left(\sum_{k=1}^{K}|\theta_{lj}^{(k)}|\right)^{1/2}. \tag{2.6}$$

Due to the presence of the square root in the penalty function, (2.6) is not convex. To tackle this issue we approximate (2.6) by using the Local Linear Approximation

(LLA) method developed in Zou and Li (2008), which proceeds as follows. Given an initial value $\hat{\boldsymbol{\Theta}}_{(0)} = \left(\hat{\boldsymbol{\Theta}}_{(0)}^{(1)}, \ldots, \hat{\boldsymbol{\Theta}}_{(0)}^{(K)}\right)$ that is close to the true value, we locally approximate the penalty function by a linear function

$$\left(\sum_{k=1}^{K} |\theta_{lj}^{(k)}|\right)^{1/2} \approx \left(\sum_{k=1}^{K} |\hat{\theta}_{(0),lj}^{(k)}|\right)^{1/2} + \tau_{(0),lj}\left(\sum_{k=1}^{k} |\theta_{lj}^{(k)}| - \sum_{k=1}^{k} |\hat{\theta}_{(0),lj}^{(k)}|\right),$$

where $\tau_{(0),lj} = 2^{-1}\left(\sum_{k=1}^{K} |\hat{\theta}_{(0),lj}^{(k)}|\right)^{-1/2}$. Then, at the t-th iteration of the LLA algorithm, problem (2.6) is decomposed into K individual optimization problems

$$\hat{\boldsymbol{\Theta}}_{(t)}^{(k)} = \underset{\boldsymbol{\Theta}\in\mathbb{R}^{p\times p}, \mathrm{diag}(\boldsymbol{\Theta})=\mathbf{0}}{\operatorname{argmin}} \frac{1}{2n}\|\mathbb{X}^{(k)}(\mathbf{I}-\boldsymbol{\Theta})\|_F^2 + \lambda\sum_{l\neq j}\tau_{(t-1),lj}|\theta_{lj}|, \tag{2.7}$$

where $\tau_{(t-1),lj}$ is defined as above for $\hat{\boldsymbol{\Theta}}_{(t-1)}$ and $l \neq j$. In the asymptotics sections, it will be shown that if the initial estimator and the data satisfy certain conditions, then with only one iteration we can get an estimator with the oracle property.

2.3.2 ADMM algorithm for optimization

We employ the Alternating Direction Method of Multiplies (ADMM; Boyd et al., 2011) to solve the optimization problem (2.7) for each subpopulation. Since this procedure is the same for all subpopulations k and all LLA iterations t, we omit the subscript (t) and the superscript (k) in this section. The primal problem is given by

$$\begin{aligned} &\text{minimize} \quad \frac{1}{2n}\|\mathbb{X}(\mathbf{I}-\boldsymbol{\Theta})\|_F^2 + \lambda\sum_{l\neq j}\tau_{lj}|\theta_{lj}| \\ &\text{subject to} \quad \boldsymbol{\Theta}\in\mathbb{R}^{p\times p}, \quad \mathrm{diag}(\boldsymbol{\Theta}) = \mathbf{0}. \end{aligned} \tag{2.8}$$

To free ourselves of the zero diagonal constraint we reformulate the problem into an equivalent form. Let $\mathbb{X}_{-j}$ denote the matrix $\mathbb{X}$ with its j-th column removed, $\boldsymbol{\tau}_j$ denote the vector of weights $(\tau_{1j}, \ldots, \tau_{pj})^\intercal$, and $\boldsymbol{\theta}_{j,-j}, \boldsymbol{\tau}_{j,-j}$ the vectors $\boldsymbol{\theta}_j, \boldsymbol{\tau}_j$ with

their j-th elements removed. Define

$$\mathbb{Y} = \begin{pmatrix} \mathbb{X}_1 \\ \vdots \\ \mathbb{X}_p \end{pmatrix}, \quad \mathbb{Z} = \begin{pmatrix} \mathbb{X}_{-1} & \dots & \mathbf{0} \\ \vdots & \ddots & \vdots \\ \mathbf{0} & \dots & \mathbb{X}_{-p} \end{pmatrix}, \quad \mathbf{v} = \begin{pmatrix} \boldsymbol{\theta}_{1,-1} \\ \vdots \\ \boldsymbol{\theta}_{p,-p} \end{pmatrix}, \quad \mathbf{w} = \begin{pmatrix} \boldsymbol{\tau}_{1,-1} \\ \vdots \\ \boldsymbol{\tau}_{p,-p} \end{pmatrix}.$$

Then, the primal problem in (2.8) is equivalent to

$$\begin{aligned} &\text{minimize} \quad \frac{1}{2n}\|\mathbb{Y} - \mathbb{Z}\mathbf{v}\|_2^2 + \lambda \sum_{l=1}^{p(p-1)} w_l |v_l| \\ &\text{subject to} \quad \mathbf{v} \in \mathbb{R}^{p(p-1)}. \end{aligned} \tag{2.9}$$

The dual problem of (2.9) is given by

$$\begin{aligned} &\text{minimize} \quad \frac{1}{2n}\|\mathbb{Y} - \mathbb{Z}\mathbf{v}\|_2^2 + \lambda \sum_{l=1}^{p(p-1)} w_l |r_l| \\ &\text{subject to} \quad \mathbf{v}, \mathbf{r} \in \mathbb{R}^{p(p-1)}, \quad \mathbf{v} - \mathbf{r} = \mathbf{0}. \end{aligned} \tag{2.10}$$

The iteration formulas for the ADMM are given by

$$\begin{aligned} \mathbf{v}^{t+1} &= (\mathbb{Z}^\intercal \mathbb{Z} + nb\mathbf{I}_{p(p-1)})^{-1}[\mathbb{Z}^\intercal \mathbb{Y} + nb(\mathbf{r}^t - \mathbf{u}^t)], \\ \mathbf{r}^{t+1} &= \max\left(\mathbf{v}^{t+1} + \mathbf{u}^t - \frac{\lambda}{b}\mathbf{w}, \mathbf{0}\right) - \max\left(-\mathbf{v}^{t+1} - \mathbf{u}^t - \frac{\lambda}{b}\mathbf{w}, \mathbf{0}\right), \\ \mathbf{u}^{t+1} &= \mathbf{u}^t + \mathbf{v}^{t+1} - \mathbf{r}^{t+1}, \end{aligned} \tag{2.11}$$

where the max is taken for each component separately and b is the step size. In practice, because of the size of $\mathbf{v}^{t+1}$, we compute it iteratively. This is possible as will be shown below.

2.3.3 Computational complexity

To show that our approach is computationally faster than the joint graphical lasso of Guo et al. (2011) we first illustrate the ADMM for the method developed therein. The primal problem for the joint graphical lasso is

$$\begin{aligned} &\text{minimize} \quad \operatorname{trace}(\mathbf{S\Omega}) - \log\det(\mathbf{\Omega}) + \lambda \sum_{l<j} q_{lj}|\omega_{lj}| \\ &\text{subject to} \quad \mathbf{\Omega} \succ \mathbf{0}, \end{aligned}$$

and the dual is

$$\begin{aligned} &\text{minimize} \quad \operatorname{trace}(\mathbf{S\Omega}) - \log\det(\mathbf{\Omega}) + \lambda \sum_{l<j} q_{lj}|z_{lj}| \\ &\text{subject to} \quad \mathbf{\Omega} \succ \mathbf{0}, \quad \mathbf{\Omega} - \mathbf{Z} = \mathbf{0}. \end{aligned}$$

The iteration formulas are given by

$$\begin{aligned} \mathbf{\Omega}^{t+1} &= \mathbf{Y}\left\{\frac{1}{2b}\left[\mathbf{\Lambda} + \left(\mathbf{\Lambda}^2 + 4b\mathbf{I}_p\right)^{1/2}\right]\right\}\mathbf{Y}^\intercal, \\ \mathbf{Z}^{t+1} &= \max\left(\mathbf{\Omega}^{t+1} + \mathbf{U}^t - \frac{\lambda}{b}\mathbf{Q}, \mathbf{0}\right) - \max\left(-\mathbf{\Omega}^{t+1} - \mathbf{U}^t - \frac{\lambda}{b}\mathbf{Q}, \mathbf{0}\right), \\ \mathbf{U}^{t+1} &= \mathbf{U}^t + \mathbf{\Omega}^{t+1} - \mathbf{Z}^{t+1}, \end{aligned} \tag{2.12}$$

where the max is performed componentwise, $\mathbf{Q} = (q_{lj})$ is the symmetric matrix of weights, $\mathbf{Y}$ is a matrix whose columns are the eigenvectors, $\mathbf{\Lambda}$ is a diagonal matrix of the eigenvalues obtained by performing spectral decomposition on the symmetric matrix

$$b(\mathbf{Z}^t - \mathbf{U}^t) - \mathbf{S}.$$

Thus, the most time consuming part of (2.12) is that of performing spectral decomposition on $b(\mathbf{Z}^t - \mathbf{U}^t) - \mathbf{S}$, which is of computational complexity $\mathcal{O}(p^3)$.

On the other hand, the most time consuming part of (2.11) is that of inverting the matrix $\mathbb{Z}^\intercal\mathbb{Z} + nb\mathbf{I}_{p(p-1)}$, which is of computational complexity $\mathcal{O}(p^6)$. Because of its block-diagonal form, it is equivalent to inverting the matrices $\mathbb{X}_{-j}^\intercal\mathbb{X}_{-j} + nb\mathbf{I}_{p-1}$, $j = 1, \ldots, p$, which can be done swiftly, as the next proposition shows.

Proposition 1. *Let* $\mathbf{M}$ *denote the matrix* $\mathbb{X}\mathbb{X}^\intercal + nb\mathbf{I}_n$. *The following identity holds.*

$$(\mathbb{X}_{-j}^\intercal\mathbb{X}_{-j} + nb\mathbb{I}_{p-1})^{-1} = \frac{1}{nb}\left(\mathbf{I}_{p-1} - \mathbb{X}_{-j}^\intercal\mathbf{M}^{-1}\mathbb{X}_{-j} + \frac{\mathbb{X}_{-j}^\intercal\mathbf{M}^{-1}\mathbb{X}_j\mathbb{X}_j^\intercal\mathbf{M}^{-1}\mathbb{X}_{-j}}{1 - \mathbb{X}_j^\intercal\mathbf{M}^{-1}\mathbb{X}_j}\right).$$

Proposition 1 helps us reduce the computational complexity of (2.11), as is shown in the following proposition.

Proposition 2. *The computational complexity of* (2.11) *is* $\mathcal{O}(np^2)$.

2.4 Asymptotics

In this section we prove the asymptotic consistency of the one-step version of (2.7), that is, the estimator produced by the first step of the LLA algorithm uncovers the true graph with probability tending to 1. Note that (2.7) is essentially the summation of p adaptive lasso problems (Zou, 2006). In this section, since k does not depend on n, we assume that we have the same number of samples n for each subpopulation.

The asymptotic consistency of the adaptive lasso has been established by Huang et al. (2008), where they proved that if the initial estimate satisfies certain conditions, then we choose the right variables with probability tending to 1. The difference of our case is that we are dealing with p_n regressions each having $p_n - 1$ variables. Thus, if we wish to prove graph consistency, we need to establish the asymptotic consistency uniformly for all regressions with diverging p_n.

To do so, we begin by defining some useful notation. Let $\boldsymbol{\theta}_j^{(k)}$ denote the vector of regression coefficients defined in section 2.2. We assume that $\boldsymbol{\theta}_j^{(k)}$ has $q_{nj}^{(k)}$ nonzero elements and $s_{nj}^{(k)}$ zero elements, so that $q_{nj}^{(k)} + s_{nj}^{(k)} = p_n - 1$. Without loss of generality, we assume the first $q_{nj}^{(k)}$ elements of $\boldsymbol{\theta}_j^{(k)}$ to be nonzero, and the next $s_{nj}^{(k)}$ elements to be 0. We denote the l-th element of $\boldsymbol{\theta}_j^{(k)}$ by $\theta_{lj}^{(k)}$ and define

$$
\begin{aligned}
b_n &= \min\left\{|\theta_{lj}^{(k)}| : j = 1, \ldots, p_n,\, l = 1, \ldots, q_{nj}^{(k)}\right\} \\
q_n &= \max\{q_{nj}^{(k)} : j = 1, \ldots, p_n\} \\
s_n &= \max\{s_{nj}^{(k)} : j = 1, \ldots, p_n\}
\end{aligned}
$$

Notice that b_n, q_n and s_n depend on k. However, because k does not depend on n, for simplicity, we write b_n, q_n, s_n instead of $b_n^{(k)}, q_n^{(k)}, s_n^{(k)}$.

Without loss of generality, we assume that the matrices $\mathbb{X}^{(1)}, \ldots, \mathbb{X}^{(K)}$ are centered and scaled, that is

$$
\sum_{i=1}^{n} x_{ij}^{(k)} = 0 \quad \text{and} \quad \frac{1}{n}\sum_{i=1}^{n}(x_{ij}^{(k)})^2 = 1, \tag{2.13}
$$

for all j, k. Furthermore, for each j, k, we compartmentalize $\mathbb{X}_{-j}^{(k)}$ into $\left[\mathbb{X}_{-j;1}^{(k)}, \mathbb{X}_{-j;2}^{(k)}\right]$, where $\mathbb{X}_{-j;1}^{(k)} \in \mathbb{R}^{n \times q_{nj}^{(k)}}$ and $\mathbb{X}_{-j;2}^{(k)} \in \mathbb{R}^{n \times s_{nj}^{(k)}}$, and define $\hat{\boldsymbol{\Sigma}}_{jj}^{(k)}$ to be the matrix

$$
\frac{1}{n}\mathbb{X}_{-j;1}^{(k)\intercal}\mathbb{X}_{-j;1}^{(k)}.
$$

We denote by $\upsilon_{nj}^{(k)}$ its smallest eigenvalue.

Assumption 3. *There exists a positive number ξ such that $\upsilon_{nj}^{(k)} > \xi$ for all n, j.*

Let $\tilde{\boldsymbol{\theta}}_j^{(k)}$ be an initial estimate of $\boldsymbol{\theta}_j^{(k)}$ for all j, k. By construction, the weight τ_{lj}

is equal to

$$\frac{1}{2}\left(\sum_{k=1}^{K}|\tilde{\theta}_{lj}^{(k)}|\right)^{-1/2}, \quad l=1,\ldots,p_n-1.$$

Let $\text{sgn}:\mathbb{R}\to\mathbb{R}$ denote the sign function, such that

$$\text{sgn}(t)=\begin{cases}-1 & \text{,if} \quad \text{t}<0\\ 0 & \text{,if} \quad \text{t}=0\\ 1 & \text{,if} \quad \text{t}>0\end{cases}.$$

Define $\mathbf{s}_j^{(k)}=\left(s_{1j}^{(k)},\ldots,s_{q_{nj}^{(k)}j}^{(k)}\right)^{\intercal}$, where

$$s_{lj}^{(k)}=\tau_{lj}\,\text{sgn}\left(\theta_{lj}^{(k)}\right), \quad l=1,\ldots,q_{nj}^{(k)},$$

for all regressions $j=1,\ldots,p_n$.

To establish uniform consistency for all regressions $j=1,\ldots,p_n$, we need stronger assumptions on the initial estimators $\tilde{\boldsymbol{\theta}}_j^{(k)}$ than those made in Huang et al. (2008), particularly in the tail probability behaviors. We make the following assumptions.

Assumption 4. *For each $\tilde{\theta}_{lj}^{(k)}$ there exists a nonrandom constant $h_{lj}^{(k)}$ such that, for all $t\geq 0$,*

$$P\left(\max_{\substack{1\leq j\leq p_n\\ 1\leq l\leq p_n-1}}\left|\frac{\sum_{k=1}^{K}|h_{lj}^{(k)}|}{\sum_{k=1}^{K}|\tilde{\theta}_{lj}^{(k)}|}-1\right|\geq t\right)\leq\exp(-Ct),$$

$$P\left(\max_{\substack{1\leq j\leq p_n\\ 1\leq l\leq p_n-1}}\left|\sum_{k=1}^{K}|\tilde{\theta}_{lj}^{(k)}|-\sum_{k=1}^{K}|h_{lj}^{(k)}|\right|\geq t\right)\leq\exp(-Ct).$$

Furthermore, the constants $h_{lj}^{(k)}$ *satisfy*

$$\max_{\substack{1\le j\le p_n\\ 1\le l\le p_n-1}} \frac{1}{\sum_{k=1}^{K}|h_{lj}^{(k)}|} \le M_1, \qquad \max_{\substack{1\le j\le p_n\\ 1\le l\le p_n-1}} \sum_{k=1}^{K}|h_{lj}^{(k)}| \le M_2.$$

We denote the l-th column of $\mathbb{X}_{-j}^{(k)}$ by $(\mathbb{X}_{-j}^{(k)})_l$.

Assumption 5. *For each* $t \ge 0$ *it holds that*

$$P\left(\max_{\substack{1\le j\le p_n\\ q_{nj}^{(k)}+1\le l\le p_n-1}} \left|(\mathbb{X}_{-j}^{(k)})_l^{\intercal}\mathbb{X}_{-j;1}^{(k)}\hat{\mathbf{\Sigma}}_{jj}^{(k)-1}\mathbf{s}_j\right| \ge t\right) \le \exp(-Ct^2).$$

Assumption 6. *The sequences* p_n, b_n, q_n, λ_n *satisfy*

$$\frac{\log p_n}{nb_n^2} + \frac{\lambda_n^2 q_n \log p_n}{b_n^2} + \frac{\log p_n}{\lambda_n^2 + n} \to 0$$

as $n \to +\infty$.

We are now ready to prove the main theorem of this section. We need to introduce the notion of sign equality between vectors, which is crucial for our proof. For any vector $\mathbf{v} = (v_1, \ldots, v_t)^{\intercal}$, we denote its sign vector by $\mathrm{sgn}(\mathbf{v}) = (\mathrm{sgn}(x_1), \ldots, \mathrm{sgn}(x_t))^{\intercal}$. We say that two vectors $\boldsymbol{v}, \boldsymbol{u} \in \mathbb{R}^t$ are equal in sign if $\mathrm{sgn}(\boldsymbol{v}) = \mathrm{sgn}(\boldsymbol{u})$, and denote this by $\boldsymbol{v} =_s \boldsymbol{u}$.

Theorem 5. *Suppose that Assumptions 3, 4, 5, 6 hold. Then,*

$$P\left(\hat{\mathcal{E}}^{(k)} = \mathcal{E}^{(k)}\right) \to 1 \quad as \quad n \to +\infty.$$

The proof Theorem 5 can be found in the supplementary material.

2.5 Simulations

In this section we use simulation to compare the performance of the SNS with three existing methods: neighborhood selection (Meinshausen et al., 2006) as applied to each subpopulation, which we refer to as the individual neighborhood selection (INS), joint graphical lasso method (JGL; Guo et al., 2011), and graphical lasso method (Yuan and Lin, 2007) as applied to each subpopulation, which we refer to as the individual graphical lasso (IGL). We compare the methods both in ROC curve performance as well as CPU execution time of the optimization algorithm.

To generate the data, we first construct the edge sets for all subpopulations, and then form the precision matrices. To form the edge sets $\mathcal{E}^{(1)}, \ldots, \mathcal{E}^{(K)}$, we follow two steps:

1. Randomly choose a set of pairs $(j, l) \in \mathcal{V} \times \mathcal{V}, j < l$, as a percentage $\boldsymbol{s}$ of the total number of edges $\binom{p}{2}$. This set constitutes the common edge set of the K graphs and is denoted by $\mathcal{A}$.

2. For each subpopulation k, randomly choose a set of pairs as a percentage ρ of the number of common edges and denote this set by $\mathcal{B}^{(k)}$. The sets $\mathcal{B}^{(1)}, \ldots, \mathcal{B}^{(K)}$ must satisfy

$$\bigcup_{k=1}^{K} \mathcal{B}^{(k)} \cap \mathcal{A} = \varnothing \quad \text{and} \quad \bigcap_{k=1}^{K} \mathcal{B}^{(k)} = \varnothing.$$

These sets are the individual edge structure of each subpopulation. Combining the above, we define $\mathcal{E}^{(k)} = \mathcal{A} \cup \mathcal{B}^{(k)}, k = 1, \ldots, K$.

To form the precision matrices $\boldsymbol{\Omega}^{(1)}, \ldots, \boldsymbol{\Omega}^{(K)}$, we follow three steps:

1. Generate $a_{jl}^{(k)}$ and $b_{jl}^{(k)}$ independently from $\mathcal{U}(0.5, 1)$ and the Rademacher distri-

bution, respectively, to form the matrices

$$\mathbf{A}^{(k)} = \begin{cases} a_{jl}^{(k)} b_{jl}^{(k)} & , (j,l) \in \mathcal{E}^{(k)} \\ 0 & , \text{otherwise} \end{cases}.$$

2. To ensure symmetry, let

$$\mathbf{C}^{(K)} = \frac{\mathbf{A}^{(k)} + (\mathbf{A}^{(k)})^{\intercal}}{2}.$$

3. Let $c_{jl}^{(k)}$ be the (j,l)-th element of $\mathbf{C}^{(k)}$. To ensure positive definiteness , we use Gershgorin's Circle Theorem (Bell, 1965) to define the precision matrices $\mathbf{\Omega}^{(k)} = (\omega_{jl}^{(k)})$, with diagonal elements $\omega_{jj}^{(k)} = \sum_{q \neq j} |c_{jq}^{(k)}| + 1$, and off-diagonal elements $\omega_{jl}^{(k)} = c_{jl}^{(k)}$.

With $\mathbf{\Omega}^{(1)}, \ldots, \mathbf{\Omega}^{(K)}$ thus constructed, we are now ready to generate the observed data for each subpopulation. To do so we generate $\mathbf{X}_1^{(k)}, \ldots, \mathbf{X}_n^{(k)}$ from a $\mathcal{N}_p(\mathbf{0}, (\mathbf{\Omega}^{(k)})^{-1})$, where $\mathbf{X}_i^{(k)} = \left(X_{i1}^{(k)}, \ldots, X_{ip}^{(k)}\right)^{\intercal}$.

We compare the four methods on 12 different scenarios, which consist of the combinations of the dimensions $p = 100, 1000, 2000, 3000$ and proportions $\rho = 0, 0.5, 1$. For each of the p's mentioned above the common structure consists of $s = 5 \cdot 10^{-3}$, $5 \cdot 10^{-4}$, $15 \cdot 10^{-5}$, $75 \cdot 10^{-6}$, as proportions of $\binom{p}{2}$, to achieve high levels of sparsity. In all settings the number of samples is $n = 100$.

For the ROC curves we plot the average true positive rate (ATPR) against the average false positive rate (AFPR), over a range of values of λ. Specifically,

$$\text{ATPR}(\lambda) = \frac{1}{K} \sum_{k=1}^{K} \frac{\sum_{1 \leq j < l \leq p} \mathbb{1}\left(\theta_{jl}^{(k)} \neq 0\,,\, \hat{\theta}_{jl}^{(k)}(\lambda) \neq 0\right)}{\sum_{1 \leq j < l \leq p} \mathbb{1}\left(\theta_{jl}^{(k)} \neq 0\right)}$$

p	Method	ρ		
		1	0.5	0
100	SNS	0.86	0.91	0.94
	INS	0.84	0.88	0.91
	JGL	0.90	0.94	0.96
	IGL	0.87	0.90	0.93
1000	SNS	0.88	0.89	0.90
	INS	0.86	0.87	0.88
	JGL	0.91	0.92	0.93
	IGL	0.89	0.90	0.91
2000	SNS	0.92	0.93	0.96
	INS	0.91	0.92	0.93
3000	SNS	0.93	0.94	0.97
	INS	0.92	0.93	0.94

Table 2.1. Table of the areas under the ROC curves from Figure 2.1.

$$\text{AFPR}(\lambda) = \frac{1}{K}\sum_{k=1}^{K}\frac{\sum_{1\le j<l\le p}\mathbb{1}\left(\theta_{jl}^{(k)}=0\,,\,\hat{\theta}_{jl}^{(k)}(\lambda)\neq 0\right)}{\sum_{1\le j<l\le p}\mathbb{1}\left(\theta_{jl}^{(k)}=0\right)},$$

where $\mathbb{1}$ is the indicator function and $\hat{\theta}_{jl}^{(k)}$ is the estimate of $\theta_{jl}^{(k)}$ using tuning parameter λ. The above quantities are calculated for 100 equally spaced values of λ, starting at 10^{-5} and ending at 1. Each scenario is simulated 5 times, and the final ROC curve is the average of them. The curves are shown in Figure 2.1 and their respective areas-under-curve (AUC) in Table 2.1. We did not include the ROC results for IGL and JGL when $p = 2000$ and 3000, because computing them would take more than 48 hours which is the wall time for the open queue in the PSU super-computer. Finally, in Table 2.2, we demonstrate the CPU times of SNS against JGL for one iteration of the ADMM algorithm mentioned in section 3. All the experiments were conducted on a 2.2 GHz Intel Xeon processor.

To summarize the results, SNS performs significantly better than INS, due to the fact that the former takes advantage of the common structure. As expected, SNS does

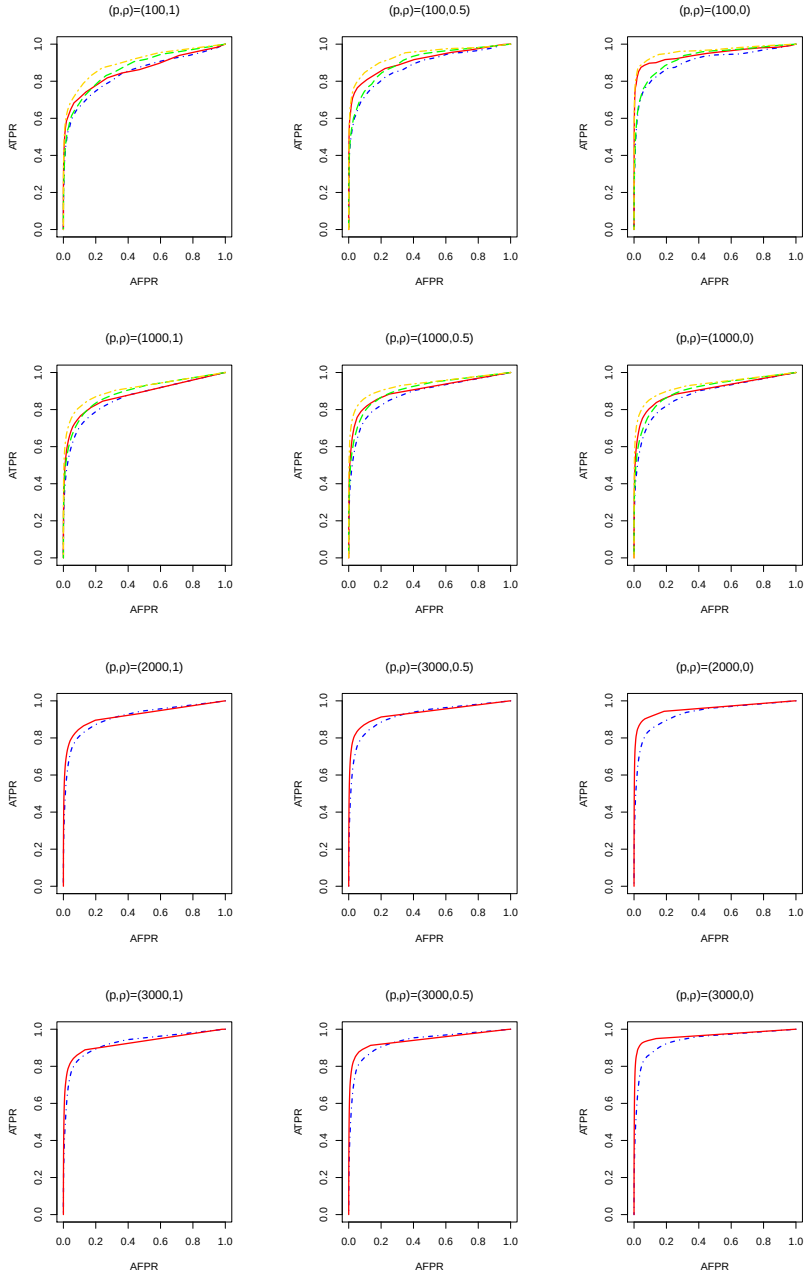

Figure 2.1. ROC curves by SNS (red solid line), INS (blue dashed line), JGL (golden dashed line) and IGL (green dashed line) for different combinations of (p, ρ). Note that for $p = 2000, 3000$, the computation of ROC results are only feasible for INS and SNS.

Method	p			
	100	1000	2000	3000
SNS	0.191	6.262	27.352	91.848
JGL	0.086	14.764	124.376	513.693

Table 2.2. CPU times for one iteration of the ADMM algorithm for the SNS and JGL for different numbers of features.

not perform as well as JGL. This is because the loss function of the SNS is a second-order approximation of the loss function of JGL. Yuan and Lin (2007) discussed this point in the context of estimating a single graph. However, SNS requires significantly less computing time than JGL or IGL as reported in Table 2.2. Note that we did not test the computing time of IGL, because it has the same ADMM algorithm as JGL with the only difference that the weights are equal to 1. The performance of SNS and IGL is comparable but as with JGL it is infeasible to estimate the graph for large number of variables, such as the dataset we analyze next.

2.6 Application

In this section we apply our method on the dataset mentioned in the introduction, which can be found in the GEO DataSet Browser under the name GDS2771. It consists of data from large airway epithelial cells. The data were collected from three subpopulations of smokers: with lung cancer, without lung cancer and with suspected lung cancer. The three subpopulations consist of 97, 90 and 5 subjects, respectively.

We removed the subjects with suspected lung cancer, since our goal is not fitting a regression model to predict which of these subjects have cancer and which do not. We also removed all the duplicate variables along with variables that contained missing values. The remaining dataset consist of 14062 variables. Finally, we centered and scaled our data to match the input of the SNS.

To get an initial estimate for $\boldsymbol{\Theta}$ in (2.7) we applied neighborhood selection separately to each subpopulation. We set a λ relatively small (0.05), so that we have a

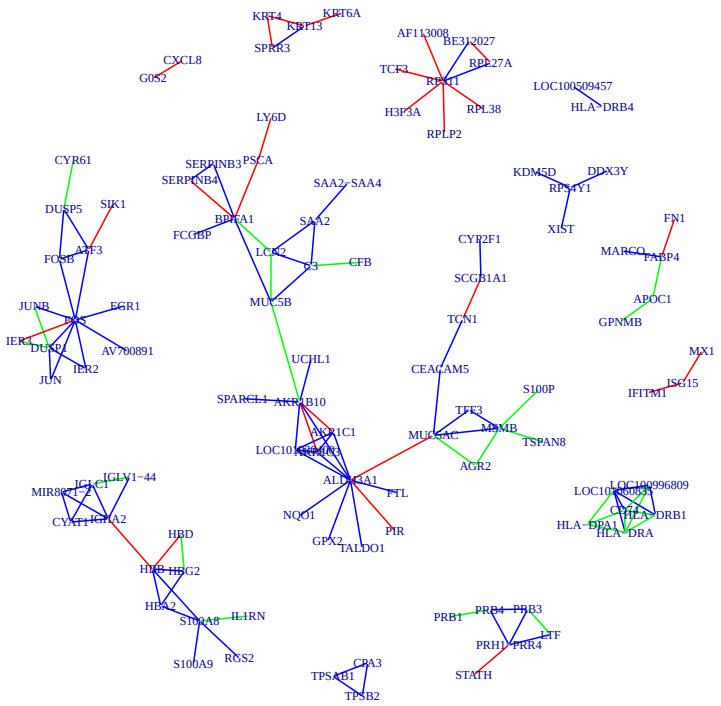

Figure 2.2. Graph of lung cancer dataset, recovered with the simultaneous neighborhood selection method. Edges with blue color are common to both subpopulations, edges with green color are exclusive to the subpopulation with lung cancer and edges with red are exclusive to the subpopulation without lung cancer.

rich structure from which to select edges with our simultaneous estimation method. After getting the initial estimate, we applied our method to the dataset with a $\lambda = 50$ to get a sparse graph. The first reason why λ has to be this high is that the sample size (187) is very small compared to the size of the graph (14062), which requires strong regularization. The second reason is that we were only interested in a good visual representation of the graph of the dataset, which favored an interpretable sparse result.

Appendix A

Proofs of Chapter 1

A.1 Proving Theorem 2

For a matrix $\mathbf{A} = (a_{jl})$, let $\|\cdot\|_1$ denote the usual ℓ_1 vector norm $\|\mathbf{A}\|_1 = \sum_{j,l} |a_{jl}|$. Define the objective function

$$\sum_{k=1}^{K} \left[\operatorname{trace}\left(\hat{\mathbf{\Sigma}}^{(k)}\mathbf{\Omega}^{(k)}\right) - \log\det\left(\mathbf{\Omega}^{(k)}\right)\right] + \sum_{j\neq l}\theta_{jl} + \lambda_1\lambda_2 \sum_{j\neq l}\sum_{k=1}^{K} \|\mathbf{\Gamma}_{jl}^{(k)}\|_F, \qquad \text{(A.1)}$$

for θ_{jl} and $\mathbf{\Gamma}_{jl}^{(k)}$ specified above. We first show that the objective functions (1.3) and (A.1) are equivalent.

Lemma A.1. *Suppose* $(\hat{\mathbf{\Theta}}, \hat{\mathbf{\Gamma}})$ *is a local minimizer of* (1.3). *Then, there exists a local minimizer* $(\tilde{\mathbf{\Theta}}, \tilde{\mathbf{\Gamma}})$ *of* (A.1) *such that* $\tilde{\theta}_{jl}\tilde{\mathbf{\Gamma}}_{jl}^{(k)} = \hat{\theta}_{jl}\hat{\mathbf{\Gamma}}_{jl}^{(k)}$ *for all* j, l, k. *Conversely, let* $(\tilde{\mathbf{\Theta}}, \tilde{\mathbf{\Gamma}})$ *be a local minimizer of* (A.1). *Then, there exists a local minimizer* $(\hat{\mathbf{\Theta}}, \hat{\mathbf{\Gamma}})$ *of* (1.3) *such that* $\hat{\theta}_{jl}\hat{\mathbf{\Gamma}}_{jl}^{(k)} = \tilde{\theta}_{jl}\tilde{\mathbf{\Gamma}}_{jl}^{(k)}$ *for all* j, l, k.

Proof. Let $Q_1(\lambda_1, \lambda_2, \mathbf{\Theta}, \mathbf{\Gamma})$ and $Q_2(\lambda_1\lambda_2, \mathbf{\Theta}, \mathbf{\Gamma})$ denote the objective functions (1.3) and (A.1), respectively. Observe that

$$Q_1(\lambda_1, \lambda_2, \mathbf{\Theta}, \mathbf{\Gamma}) = Q_2(\lambda_1\lambda_2, \lambda_1\mathbf{\Theta}, \lambda_1^{-1}\mathbf{\Gamma})$$

$$Q_2(\lambda_1\lambda_2, \boldsymbol{\Theta}, \boldsymbol{\Gamma}) = Q_1(\lambda_1, \lambda_2, \lambda_1^{-1}\boldsymbol{\Theta}, \lambda_1\boldsymbol{\Gamma})$$

Since $(\hat{\boldsymbol{\Theta}}, \hat{\boldsymbol{\Gamma}})$ is a local minimizer of $Q_1(\lambda_1, \lambda_2, \cdot, \cdot)$, there exists $\delta > 0$ such that for every $(\boldsymbol{\Theta}, \boldsymbol{\Gamma})$ with

$$\|\boldsymbol{\Theta} - \hat{\boldsymbol{\Theta}}\|_1 + \sum_{k=1}^{K} \|\boldsymbol{\Gamma}^{(k)} - \hat{\boldsymbol{\Gamma}}^{(k)}\|_1 < \delta,$$

we have

$$Q_1(\lambda_1, \lambda_2, \hat{\boldsymbol{\Theta}}, \hat{\boldsymbol{\Gamma}}) \leq Q_1(\lambda_1, \lambda_2, \boldsymbol{\Theta}, \boldsymbol{\Gamma}).$$

Let $0 < \delta^* \leq \delta \min\left(\lambda_1, \lambda_1^{-1}\right)$, and define $(\tilde{\boldsymbol{\Theta}}, \tilde{\boldsymbol{\Gamma}}) = (\lambda_1\hat{\boldsymbol{\Theta}}, \lambda_1^{-1}\hat{\boldsymbol{\Gamma}})$. Then, for any $(\boldsymbol{\Theta}, \boldsymbol{\Gamma})$ satisfying

$$\|\boldsymbol{\Theta} - \tilde{\boldsymbol{\Theta}}\|_1 + \sum_{k=1}^{K} \|\boldsymbol{\Gamma}^{(k)} - \tilde{\boldsymbol{\Gamma}}^{(k)}\|_1 < \delta^*,$$

we have

$$\|\lambda_1^{-1}\boldsymbol{\Theta} - \hat{\boldsymbol{\Theta}}\|_1 + \sum_{k=1}^{K} \|\lambda_1\boldsymbol{\Gamma}^{(k)} - \hat{\boldsymbol{\Gamma}}^{(k)}\|_1 \leq \frac{\|\boldsymbol{\Theta} - \tilde{\boldsymbol{\Theta}}\|_1 + \sum_{k=1}^{K} \|\boldsymbol{\Gamma}^{(k)} - \tilde{\boldsymbol{\Gamma}}^{(k)}\|_1}{\min\left(\lambda_1, \lambda_1^{-1}\right)} \leq \delta.$$

Thus

$$Q_1(\lambda_1, \lambda_2, \hat{\boldsymbol{\Theta}}, \hat{\boldsymbol{\Gamma}}) \leq Q_1(\lambda_1, \lambda_2, \lambda_1^{-1}\boldsymbol{\Theta}, \lambda_1\boldsymbol{\Gamma}) \Rightarrow Q_2(\lambda_1\lambda_2, \tilde{\boldsymbol{\Theta}}, \tilde{\boldsymbol{\Gamma}}) \leq Q_2(\lambda_1\lambda_2, \boldsymbol{\Theta}, \boldsymbol{\Gamma}),$$

which means that $(\tilde{\boldsymbol{\Theta}}, \tilde{\boldsymbol{\Gamma}})$ is a local minimizer of (A.1). The other direction is proven similarly. □

Lemma A.2. *Suppose* $(\hat{\boldsymbol{\Theta}}, \hat{\boldsymbol{\Gamma}})$ *is a local minimizer of* (A.1) *and* $\hat{\boldsymbol{\Omega}}_{jl}^{(k)} = \hat{\theta}_{jl}\hat{\boldsymbol{\Gamma}}_{jl}^{(k)}$ *for all* j, l, k. *Then, for* $j, l \in \mathcal{V}$, $j \neq l$, *the following are true:*

1. $\hat{\theta}_{jl} = 0$ *if and only if* $\hat{\mathbf{\Gamma}}_{jl}^{(k)} = 0$ *for all* $k = 1, \ldots, K$.

2. *If* $\hat{\theta}_{jl} \neq 0$, *then* $\hat{\theta}_{jl} = \left(\lambda \sum_{k=1}^{K} \|\hat{\mathbf{\Omega}}_{jl}^{(k)}\|_F\right)^{1/2}$, *where* $\lambda = \lambda_1\lambda_2$.

Proof. 1. If θ_{jl} is 0, then $\mathbf{\Gamma}_{jl}^{(1)}, \ldots, \mathbf{\Gamma}_{jl}^{(K)}$ only appear in the third term in (A.1). Thus, in order to minimize Q_2, we need $\mathbf{\Gamma}_{jl}^{(k)} = \mathbf{0}$, for all $k = 1, \ldots, K$. The other direction is similar.

2. Suppose $\hat{\theta}_{jl} \neq 0$ and let

$$c = \frac{\left(\lambda \sum_{k=1}^{K} \|\hat{\mathbf{\Omega}}_{jl}^{(k)}\|_F\right)^{1/2}}{\hat{\theta}_{jl}}.$$

We will show $c = 1$. By definition,

$$\hat{\mathbf{\Gamma}}_{jl}^{(k)} = c\frac{\hat{\mathbf{\Omega}}_{jl}^{(k)}}{\left(\lambda \sum_{k=1}^{K} \|\hat{\mathbf{\Omega}}_{jl}^{(k)}\|_F\right)^{1/2}}.$$

Suppose $c > 1$. Since $(\hat{\mathbf{\Theta}}, \hat{\mathbf{\Gamma}})$ is a local minimizer of $Q_2(\lambda, \cdot, \cdot)$, there exists $\delta > 0$ such that for all $(\mathbf{\Theta}, \mathbf{\Gamma})$ with

$$\|\mathbf{\Theta} - \hat{\mathbf{\Theta}}\|_1 + \sum_{k=1}^{K} \|\mathbf{\Gamma}^{(k)} - \hat{\mathbf{\Gamma}}^{(k)}\|_1 < \delta$$

we have $Q_2(\lambda, \hat{\mathbf{\Theta}}, \hat{\mathbf{\Gamma}}) \leq Q_2(\lambda, \mathbf{\Theta}, \mathbf{\Gamma})$.

Then there exists $\delta^* \in (1, c)$, slightly greater than 1, such that for $(\tilde{\mathbf{\Theta}}, \tilde{\mathbf{\Gamma}})$ defined by

$$\begin{cases} \tilde{\theta}_{j'l'} = \hat{\theta}_{j'l'} \text{ and } \tilde{\mathbf{\Gamma}}_{j'l'}^{(k)} = \hat{\mathbf{\Gamma}}_{j'l'}^{(k)}, & (j', l') \neq (j, l) \\ \tilde{\theta}_{jl} = \delta^*\hat{\theta}_{jl} \text{ and } \tilde{\mathbf{\Gamma}}_{jl}^{(k)} = \frac{1}{\delta^*}\hat{\mathbf{\Gamma}}_{jl}^{(k)} \end{cases}$$

we have

$$\|\tilde{\mathbf{\Theta}} - \hat{\mathbf{\Theta}}\|_1 + \sum_{k=1}^{K} \|\tilde{\mathbf{\Gamma}}^{(k)} - \hat{\mathbf{\Gamma}}^{(k)}\|_1 < \delta.$$

But this implies

$$\begin{aligned} & Q_2(\lambda, \hat{\mathbf{\Theta}}, \hat{\mathbf{\Gamma}}) - Q_2(\lambda, \tilde{\mathbf{\Theta}}, \tilde{\mathbf{\Gamma}}) \\ =& (1-\delta^*)\hat{\theta}_{jl} + \left(1 - \frac{1}{\delta^*}\right) \lambda \sum_{k=1}^{K} \|\hat{\mathbf{\Gamma}}_{jl}^{(k)}\|_F \\ =& \frac{1}{c}(\delta^* - 1)\left(\frac{c^2}{\delta^*} - 1\right)\left(\lambda \sum_{k=1}^{K} \|\hat{\mathbf{\Omega}}_{jl}^{(k)}\|_F\right)^{1/2} > 0, \end{aligned}$$

which is impossible because $(\hat{\mathbf{\Theta}}, \hat{\mathbf{\Gamma}})$ is a local minimizer. Hence $c \leq 1$. Following the same argument we can show $c \geq 1$. Thus $c = 1$. □

We are now ready to establish the equivalence between the objective functions (1.3) and (1.4), for which it suffices to show the equivalence of (1.4) and (A.1).

Lemma A.3. *Let $(\hat{\mathbf{\Theta}}, \hat{\mathbf{\Gamma}})$ be a local minimizer of* (A.1). *Then, there exists a local minimizer $\hat{\mathbf{\Omega}}$ of* (1.4) *such that $\hat{\mathbf{\Omega}}_{jl}^{(k)} = \hat{\theta}_{jl}\hat{\mathbf{\Gamma}}_{jl}^{(k)}$ for all j, l, k. Conversely, let $\hat{\mathbf{\Omega}}$ be a local minimizer of* (1.4). *Then, there exists a local minimizer $(\hat{\mathbf{\Theta}}, \hat{\mathbf{\Gamma}})$ of* (A.1) *such that $\hat{\theta}_{jl}\hat{\mathbf{\Gamma}}_{jl}^{(k)} = \hat{\mathbf{\Omega}}_{jl}^{(k)}$ for all j, l, k.*

Proof. Let $Q_3(\lambda_1\lambda_2, \mathbf{\Omega})$ denote the objective function (1.4). Suppose $(\hat{\mathbf{\Theta}}, \hat{\mathbf{\Gamma}})$ is a local minimizer of (A.1). Then, there exists $\delta > 0$ such that for all $(\mathbf{\Theta}, \mathbf{\Gamma})$ with

$$\|\mathbf{\Theta} - \hat{\mathbf{\Theta}}\|_1 + \sum_{k=1}^{K} \|\mathbf{\Gamma}^{(k)} - \hat{\mathbf{\Gamma}}^{(k)}\|_1 < \delta,$$

we have

$$Q_2(\lambda, \hat{\mathbf{\Theta}}, \hat{\mathbf{\Gamma}}) \leq Q_2(\lambda, \mathbf{\Theta}, \mathbf{\Gamma}).$$

Let $\hat{\mathbf{\Omega}}$ be the estimator associated with $(\hat{\mathbf{\Theta}}, \hat{\mathbf{\Gamma}})$, that is $\hat{\mathbf{\Omega}}_{jl}^{(k)} = \hat{\theta}_{jl}\hat{\mathbf{\Gamma}}_{jl}^{(k)}$ for all j, l, k. In order to find a neighborhood where $\hat{\mathbf{\Omega}}$ is a minimizer we need to define the constants which will appear in the course of the proof. Let

$$a = \min\left\{\sum_{k=1}^{K} \|\hat{\mathbf{\Omega}}_{jl}^{(k)}\|_F : j, l \in \mathcal{V}, \hat{\theta}_{jl} \neq 0\right\},$$
$$b = \max\left\{\sum_{k=1}^{K} \|\hat{\mathbf{\Omega}}_{jl}^{(k)}\|_F : j, l \in \mathcal{V}\right\},$$

and

$$c = 2p^2 \max\left\{\lambda^{1/2}, \frac{M}{\lambda^{1/2}}, \left(\frac{\lambda}{2a}\right)^{1/2}, \left(\frac{M^2}{2a}\right)^{1/2} + \left[\left(\frac{2M^2}{a\lambda}\right)^{1/2} + \left(\frac{bM^2}{a^2}\right)^{1/2}\right]\right\}.$$

Let $0 < \delta^* < \min\left(\frac{a}{2}, 1, c^{-2}\delta^2\right)$, and $\mathbf{\Delta} = (\mathbf{\Delta}^{(1)}, \ldots, \mathbf{\Delta}^{(K)})$, where $\mathbf{\Delta}^{(k)} = (\mathbf{\Delta}_{jl}^{(k)}) \in \mathbb{R}^{pM \times pM}$, $\mathbf{\Delta}_{jl}^{(k)} \in \mathbb{R}^{M \times M}$, that satisfies

$$0 < \|\mathbf{\Delta}_{jl}^{(k)}\|_F < \min\left\{\|\hat{\mathbf{\Omega}}_{jl}^{(k)}\|_F : j, l \in \mathcal{V}, \hat{\theta}_{jl} \neq 0\right\},$$

for all j, l, k, and $\sum_{k=1}^{K} \|\mathbf{\Delta}^{(k)}\|_F < \delta^*$. Let $\tilde{\mathbf{\Omega}} = \hat{\mathbf{\Omega}} + \mathbf{\Delta}$. Then

$$\sum_{k=1}^{K} \|\tilde{\mathbf{\Omega}}^{(k)} - \hat{\mathbf{\Omega}}^{(k)}\|_1 < \delta^*.$$

This means that $\tilde{\mathbf{\Omega}}$ is a generic element of the ball with radius less than δ^* and center $\hat{\mathbf{\Omega}}$.

Define $(\tilde{\mathbf{\Theta}}, \tilde{\mathbf{\Gamma}})$ by

$$\tilde{\theta}_{jl} = \left(\lambda \sum_{k=1}^{K} \|\hat{\mathbf{\Omega}}_{jl}^{(k)} + \mathbf{\Delta}_{jl}^{(k)}\|_F\right)^{1/2} \text{ and } \tilde{\mathbf{\Gamma}}_{jl}^{(k)} = \frac{\hat{\mathbf{\Omega}}_{jl}^{(k)} + \mathbf{\Delta}_{jl}^{(k)}}{\left(\lambda \sum_{k=1}^{K} \|\hat{\mathbf{\Omega}}_{jl}^{(k)} + \mathbf{\Delta}_{jl}^{(k)}\|_F\right)^{1/2}},$$

for all j, l, k. By Lemma A.2,

$$Q_2(\lambda, \hat{\mathbf{\Theta}}, \hat{\mathbf{\Gamma}}) = Q_3(\lambda, \hat{\mathbf{\Omega}}),$$

and by the definition of $(\tilde{\mathbf{\Theta}}, \tilde{\mathbf{\Gamma}})$,

$$Q_2(\lambda, \tilde{\mathbf{\Theta}}, \tilde{\mathbf{\Gamma}}) = Q_3(\lambda, \tilde{\mathbf{\Omega}}).$$

If $\hat{\theta}_{jl} = 0$, then $\hat{\mathbf{\Gamma}}_{jl}^{(k)} = \hat{\mathbf{\Omega}}_{jl}^{(k)} = \mathbf{0}$ for all k. Hence,

$$|\tilde{\theta}_{jl} - \hat{\theta}_{jl}| = \left(\lambda \sum_{k=1}^{K} \|\mathbf{\Delta}_{jl}^{(k)}\|_F\right)^{1/2} \leq \left(\lambda \sum_{k=1}^{K} \|\mathbf{\Delta}^{(k)}\|_F\right)^{1/2} < \lambda^{1/2}\delta^{*1/2},$$

and

$$\begin{aligned}
\sum_{k=1}^{K} \|\tilde{\mathbf{\Gamma}}_{jl}^{(k)} - \hat{\mathbf{\Gamma}}_{jl}^{(k)}\|_1 &= \frac{\sum_{k=1}^{K} \|\mathbf{\Delta}_{jl}^{(k)}\|_1}{\left(\lambda \sum_{k=1}^{K} \|\mathbf{\Delta}_{jl}^{(k)}\|_F\right)^{1/2}} \\
&\leq \frac{M}{\lambda^{1/2}} \frac{\sum_{k=1}^{K} \|\mathbf{\Delta}_{jl}^{(k)}\|_F}{\left(\sum_{k=1}^{K} \|\mathbf{\Delta}_{jl}^{(k)}\|_F\right)^{1/2}} \\
&\leq \frac{M}{\lambda^{1/2}} \delta^{*1/2}.
\end{aligned}$$

If $\hat{\theta}_{jl} \neq 0$, then

$$\begin{aligned}
|\tilde{\theta}_{jl} - \hat{\theta}_{jl}| &\leq \frac{\lambda^{1/2}}{2} \frac{\sum_{k=1}^{K} \|\mathbf{\Delta}_{jl}^{(k)}\|_F}{\left(\sum_{k=1}^{K} \|\hat{\mathbf{\Omega}}_{jl}^{(k)}\|_F - \sum_{k=1}^{K} \|\mathbf{\Delta}_{jl}^{(k)}\|_F\right)^{1/2}} \\
&\leq \frac{\lambda^{1/2}}{2} \frac{\delta^{*1/2}}{\sqrt{a - \frac{a}{2}}} \\
&= \left(\frac{\lambda}{2a}\right)^{1/2} \delta^{*1/2},
\end{aligned}$$

and

$$\sum_{k=1}^{K} \|\tilde{\mathbf{\Gamma}}_{jl}^{(k)} - \hat{\mathbf{\Gamma}}_{jl}^{(k)}\|_1 \leq I_1 + I_2,$$

where

$$\begin{aligned} I_1 &= \lambda^{-1/2} \frac{\sum_{k=1}^{K} \|\mathbf{\Delta}_{jl}^{(k)}\|_1}{\left(\sum_{k=1}^{K} \|\hat{\mathbf{\Omega}}_{jl}^{(k)} + \mathbf{\Delta}_{jl}^{(k)}\|_F\right)^{1/2}} \\ &\leq \frac{M}{\lambda^{1/2}} \frac{\sum_{k=1}^{K} \|\mathbf{\Delta}_{jl}^{(k)}\|_F}{\left(\sum_{k=1}^{K} \|\hat{\mathbf{\Omega}}_{jl}^{(k)}\|_F - \sum_{k=1}^{K} \|\mathbf{\Delta}_{jl}^{(k)}\|_F\right)^{1/2}} \\ &\leq \left(\frac{M^2}{2a}\right)^{1/2} \delta^{*1/2}, \end{aligned}$$

and

$$\begin{aligned} I_2 &= \sum_{k=1}^{K} \|\hat{\mathbf{\Omega}}_{jl}^{(k)}\|_1 \frac{\left| \left(\sum_{k=1}^{K} \|\hat{\mathbf{\Omega}}_{jl}^{(k)} + \mathbf{\Delta}_{jl}^{(k)}\|_F\right)^{1/2} - \left(\sum_{k=1}^{K} \|\hat{\mathbf{\Omega}}_{jl}^{(k)}\|_F\right)^{1/2} \right|}{\left(\sum_{k=1}^{K} \|\hat{\mathbf{\Omega}}_{jl}^{(k)}\|_F\right)^{1/2} \left(\sum_{k=1}^{K} \|\hat{\mathbf{\Omega}}_{jl}^{(k)} + \mathbf{\Delta}_{jl}^{(k)}\|_F\right)^{1/2}} \\ &\leq M \left(\sum_{k=1}^{K} \|\hat{\mathbf{\Omega}}_{jl}^{(k)}\|_F\right)^{1/2} \frac{|\tilde{\theta}_{jl} - \hat{\theta}_{jl}|}{\left(\sum_{k=1}^{K} \|\hat{\mathbf{\Omega}}_{jl}^{(k)}\|_F - \sum_{k=1}^{K} \|\hat{\mathbf{\Delta}}_{jl}^{(k)}\|_F\right)^{1/2}} \\ &\leq \left[\left(\frac{2M^2}{a\lambda}\right)^{1/2} + \left(\frac{bM^2}{a^2}\right)^{1/2} \right] \delta^{*1/2}. \end{aligned}$$

Therefore,

$$\begin{aligned} &\|\tilde{\mathbf{\Theta}} - \hat{\mathbf{\Theta}}\|_1 + \sum_{k=1}^{K} \|\tilde{\mathbf{\Gamma}}^{(k)} - \hat{\mathbf{\Gamma}}\|_1 \\ &< p^2 \max\left\{ \lambda^{1/2}, \left(\frac{\lambda}{2a}\right)^{1/2} \right\} \delta^{*1/2} \\ &+ p^2 \max\left\{ \frac{M}{\lambda^{1/2}}, \left(\frac{M^2}{2a}\right)^{1/2} + \left[\left(\frac{2M^2}{a\lambda}\right)^{1/2} + \left(\frac{bM^2}{a^2}\right)^{1/2} \right] \right\} \delta^{*1/2} \end{aligned}$$

$$\leq c\delta^{*1/2}$$

$$< \delta.$$

Thus

$$Q_2(\lambda, \hat{\mathbf{\Theta}}, \hat{\mathbf{\Gamma}}) \leq Q_2(\lambda, \tilde{\mathbf{\Theta}}, \tilde{\mathbf{\Gamma}}) \Rightarrow Q_3(\lambda, \hat{\mathbf{\Omega}}) \leq Q_3(\lambda, \tilde{\mathbf{\Omega}}),$$

which means that $\hat{\mathbf{\Omega}}$ is a local minimizer of (1.4). The other direction is proven similarly. □

Combining the results from Lemmas A.2 and A.3 yields Theorem 2.

A.2 Important norm inequalities

Lemma A.4. *For $1 \leq j, l \leq p$, let $\mathbf{A} = (\mathbf{A}_{jl})$ and $\mathbf{B} = (\mathbf{B}_{jl})$ be block matrices with $\mathbf{A}_{jl}, \mathbf{B}_{jl} \in \mathbb{R}^{M \times M}$, let $\mathbf{u} = (\mathbf{u}_j)$ be the block vector with $\mathbf{u}_j \in \mathbb{R}^M$, and let $\mathbf{x} = (\mathbf{x}_j)$ and $\mathbf{y} = (\mathbf{y}_j)$ be block matrices with j-th blocks $\mathbf{x}_j, \mathbf{y}_j \in \mathbb{R}^{M \times M}$. The following norm properties hold:*

$$\|\mathbf{A}\|_{\max}^{(M)} = \|\mathrm{vec}(\mathbf{A})\|_{\max}^{(M^2)} \tag{A.2a}$$

$$\|\mathbf{A}\mathbf{u}\|_{\max}^{(M)} \leq \|\mathbf{A}\|_{\infty}^{(M)} \|\mathbf{u}\|_{\max}^{(M)} \tag{A.2b}$$

$$\|\mathbf{x}^\intercal \mathbf{y}\|_F \leq \|\mathbf{x}\|_{\max}^{(M)} \|\mathbf{y}\|_1^{(M)} \tag{A.2c}$$

$$\|\mathbf{A}\mathbf{x}\|_{\max}^{(M)} \leq \|\mathbf{A}\|_{\max}^{(M)} \|\mathbf{x}\|_1^{(M)} \tag{A.2d}$$

$$\|\mathbf{A}\|_{\infty}^{(M)} = \|\mathbf{A}^\intercal\|_1^{(M)} \tag{A.2e}$$

$$\|\mathbf{A}\mathbf{B}\|_{\infty}^{(M)} \leq \|\mathbf{A}\|_{\infty}^{(M)} \|\mathbf{B}\|_{\infty}^{(M)} \tag{A.2f}$$

Proof. Proof of (A.2a):

$$\|\mathbf{A}\|_{\max}^{(M)} = \max_{1\le j,l\le p} \|\mathbf{A}_{jl}\|_F = \max_{1\le j,l\le p} \|\mathrm{vec}(\mathbf{A}_{jl})\|_2 = \|\mathrm{vec}(\mathbf{A})\|_{\max}^{(M^2)}.$$

Proof of (A.2b): Note that for a matrix $\mathbf{G} = (g_{jl})$ and a vector $\mathbf{b} = (b_j)$ of appropriate dimensions,

$$\begin{aligned}
\|\mathbf{Gb}\|_2^2 &= \sum_j \sum_l (g_{jl} b_l)^2 \\
&\le \sum_j \left(\sum_l g_{jl}^2 \right) \left(\sum_l b_l^2 \right) \\
&= \left(\sum_{j,l} g_{jl}^2 \right) \left(\sum_l b_l^2 \right) \\
&= \|\mathbf{G}\|_F^2 \, \|\mathbf{b}\|_2^2 \\
\Rightarrow \|\mathbf{Gb}\|_2 &\le \|\mathbf{G}\|_F \, \|\mathbf{b}\|_2.
\end{aligned}$$

Therefore,

$$\begin{aligned}
\|\mathbf{Au}\|_{\max}^{(M)} &= \max_{1\le j,l\le p} \|\mathbf{A}_{jl}\mathbf{u}_l\|_2 \\
&\le \max_{1\le j,l\le p} \|\mathbf{A}_{jl}\|_F \, \|\mathbf{u}_l\|_2 \\
&\le \left(\max_{1\le j,l\le p} \|\mathbf{A}_{jl}\|_F \right) \left(\max_{1\le l\le p} \|\mathbf{u}_l\|_2 \right) \\
&\le \max_{1\le j\le p} \left(\sum_{l=1}^{p} \|\mathbf{A}_{jl}\|_F \right) \left(\max_{1\le l\le p} \|\mathbf{u}_l\|_2 \right) \\
&= \|\mathbf{A}\|_{\infty}^{(M)} \, \|\mathbf{u}\|_{\max}^{(M)}.
\end{aligned}$$

Proof of (A.2c):

$$\begin{aligned}
\|\mathbf{x}^\intercal \mathbf{y}\|_F &= \left\| \sum_{j=1}^{p} \mathbf{x}_j^\intercal \mathbf{y}_j \right\|_F \\
&\leq \sum_{j=1}^{p} \|\mathbf{x}_j^\intercal \mathbf{y}_j\|_F \\
&\leq \sum_{j=1}^{p} \|\mathbf{x}_j\|_F \, \|\mathbf{y}_j\|_F \\
&\leq \left(\max_{1 \leq l \leq p} \|\mathbf{x}_l\|_F \right) \sum_{j=1}^{p} \|\mathbf{y}_j\|_F \\
&= \|\mathbf{x}\|_{\max}^{(M)} \, \|\mathbf{y}\|_1^{(M)}.
\end{aligned}$$

Proof of (A.2d):

$$\begin{aligned}
\|\mathbf{A}\mathbf{x}\|_{\max}^{(M)} &= \max_{1 \leq j \leq p} \left\| \sum_{l=1}^{p} \mathbf{A}_{jl} \mathbf{x}_l \right\|_F \\
&\leq \max_{1 \leq j \leq p} \sum_{l=1}^{p} \|\mathbf{A}_{jl} \mathbf{x}_l\|_F \\
&\leq \max_{1 \leq j \leq p} \sum_{l=1}^{p} \|\mathbf{A}_{jl}\|_F \, \|\mathbf{x}_l\|_F \\
&= \max_{1 \leq j \leq p} \sum_{l=1}^{p} \left(\max_{1 \leq m \leq p} \|\mathbf{A}_{jm}\|_F \right) \|\mathbf{x}_l\|_F \\
&= \left(\max_{1 \leq j,m \leq p} \|\mathbf{A}_{jm}\|_F \right) \sum_{l=1}^{p} \|\mathbf{x}_l\|_F \\
&= \|\mathbf{A}\|_{\max}^{(M)} \, \|\mathbf{x}\|_1^{(M)}.
\end{aligned}$$

Proof of (A.2e):

$$\|\mathbf{A}^\intercal\|_1^{(M)} = \max_{1 \leq l \leq p} \sum_{j=1}^{p} \|(\mathbf{A}^\intercal)_{jl}\|_F$$

$$
\begin{aligned}
&= \max_{1\le l\le p} \sum_{j=1}^{p} \|\mathbf{A}_{lj}\|_F \\
&= \max_{1\le j\le p} \sum_{l=1}^{p} \|\mathbf{A}_{jl}\|_F \\
&= \|\mathbf{A}\|_\infty^{(M)}.
\end{aligned}
$$

Proof of (A.2f): Note that

$$
\begin{aligned}
\|\mathbf{AB}\|_F^2 &= \sum_{i,j} \left(\sum_k a_{ik} b_{kj} \right)^2 \\
&\le \sum_{i,j} \left(\sum_k a_{ik}^2 \right) \left(\sum_k b_{kj}^2 \right) \\
&= \left(\sum_{i,k} a_{ik}^2 \right) \left(\sum_{l,j} b_{lj}^2 \right) \\
&= \|\mathbf{A}\|_F^2 \, \|\mathbf{B}\|_F^2 \\
\Rightarrow \|\mathbf{AB}\|_F &\le \|\mathbf{A}\|_F \, \|\mathbf{B}\|_F.
\end{aligned}
$$

Therefore,

$$
\begin{aligned}
\|\mathbf{AB}\|_\infty^{(M)} &= \max_{1\le j\le p} \sum_{l=1}^{p} \|(\mathbf{AB})_{jl}\|_F \\
&= \max_{1\le j\le p} \sum_{l=1}^{p} \left\| \sum_{k=1}^{p} \mathbf{A}_{jk} \mathbf{B}_{kl} \right\|_F \\
&\le \max_{1\le j\le p} \sum_{l=1}^{p} \sum_{k=1}^{p} \|\mathbf{A}_{jk} \mathbf{B}_{kl}\|_F \\
&\le \max_{1\le j\le p} \sum_{l=1}^{p} \sum_{k=1}^{p} \|\mathbf{A}_{jk}\|_F \, \|\mathbf{B}_{kl}\|_F \\
&= \max_{1\le j\le p} \left(\sum_{k=1}^{p} \|\mathbf{A}_{jk}\|_F \sum_{l=1}^{p} \|\mathbf{B}_{kl}\|_F \right)
\end{aligned}
$$

$$\leq \max_{1\leq j\leq p}\left(\sum_{k=1}^{p}\|\mathbf{A}_{jk}\|_F \max_{1\leq m\leq p}\sum_{l=1}^{p}\|\mathbf{B}_{ml}\|_F\right)$$
$$= \|\mathbf{A}\|_\infty^{(M)}\|\mathbf{B}\|_\infty^{(M)}.$$

□

A.3 Proving Theorem 3

Since the proof is the same for all K graphs, K is only present in the weights τ_{jl}, and K does not depend on n, we omit the superscript k in this section. It is therefore implied that we are working within the k-th subpopulation for proving consistency.

Let $\mathbf{Z} = (\mathbf{Z}_{jl})$ be an element of the subdifferential $\partial(\sum_{j\neq l}\|\mathbf{\Omega}_{jl}\|_F)/\partial\mathbf{\Omega}$, where

$$\mathbf{Z}_{jl} \in \begin{cases} \{\mathbf{0}\}, & \text{if} \quad j = l \\ \left\{\frac{\mathbf{\Omega}_{jl}}{\|\mathbf{\Omega}_{jl}\|_F}\right\}, & \text{if} \quad j \neq l \quad \text{and} \quad \mathbf{\Omega}_{jl} \neq \mathbf{0}\,. \\ \{\mathbf{G}\in\mathbb{R}^{M\times M} \,:\, \|\mathbf{G}\|_F \leq 1\}, & \text{if} \quad j \neq l \quad \text{and} \quad \mathbf{\Omega}_{jl} = \mathbf{0} \end{cases} \tag{A.3}$$

Define the matrix of weights $\mathbf{T} = (\tau_{jl})$ with diagonal elements equal to 0, and the vector $\mathbf{1}_M = (1,\ldots,1)^\intercal \in \mathbb{R}^M$.

In the proof to follow p_n, M_n, d_n will be abbreviated simply by p, M, d and the n subscript is going to be used only when it is meaningful. We begin by proving the existence of the solution of the adaptive fglasso problem

$$\underset{\mathbf{\Omega}\succ\mathbf{0}}{\operatorname{argmin}}\left\{\operatorname{trace}(\hat{\mathbf{\Sigma}}\mathbf{\Omega}) - \log\det(\mathbf{\Omega}) + \lambda_n\sum_{j\neq l}\tau_{jl}\|\mathbf{\Omega}_{jl}\|_F\right\} \tag{A.4}$$

and providing optimality conditions for it.

Lemma A.5. *For any $\lambda_n > 0$ and sample covariance matrix $\hat{\mathbf{\Sigma}} = (\hat{\sigma}_{jl})$ with strictly positive diagonal elements, the adaptive fglasso problem* (A.4) *has a unique solution.*

Furthermore, this solution is equal to $\hat{\mathbf{\Omega}}$ *if and only if*

$$\hat{\mathbf{\Sigma}} - \hat{\mathbf{\Omega}}^{-1} + \lambda_n(\mathbf{T} \otimes \mathbf{1}_M \mathbf{1}_M^\intercal) \circ \hat{\mathbf{Z}} = \mathbf{0}, \tag{A.5}$$

where $\circ$ *denotes the Hadamard product, and* $\hat{\mathbf{Z}}$ *is the subdifferential in* (A.3) *evaluated at* $\hat{\mathbf{\Omega}}$.

Proof. By the Lagrangian duality, for $\lambda_n > 0$, there is a constant $C(\lambda_n) > 0$ such that the problem (A.4) can be written in the equivalent constrained form

$$\underset{\mathbf{\Omega} \in \mathcal{A}}{\operatorname{argmin}} \left\{ \operatorname{trace}(\hat{\mathbf{\Sigma}}\mathbf{\Omega}) - \log\det(\mathbf{\Omega}) \right\}, \tag{A.6}$$

where $\mathcal{A} = \{\mathbf{\Omega} \succ \mathbf{0} \ : \ \sum_{j \neq l} \tau_{jl} |\omega_{jl}| \leq C(\lambda_n)\}$. It can be easily proved that the function

$$L : \mathcal{A} \to \mathbb{R}, \quad L(\mathbf{\Omega}) = \operatorname{trace}(\hat{\mathbf{\Sigma}}\mathbf{\Omega}) - \log\det(\mathbf{\Omega})$$

is convex (Boyd and Vandenberghe, 2004). If L is also bounded from below on its domain, then it has a unique minimum.

Since the off-diagonal elements are bounded within an ℓ_1-ball, the only possible issue is the behavior of the objective function on the diagonal elements. By Hadamard's inequality (Zhang, 2006, p. 35)

$$\det(\mathbf{\Omega}) \leq \prod_{i=1}^{pM} \omega_{ii} \Rightarrow \log\det(\mathbf{\Omega}) \leq \sum_{i=1}^{pM} \log(\omega_{ii}).$$

Thus,

$$\sum_{i=1}^{pM} \hat{\sigma}_{ii}\omega_{ii} - \log\det(\mathbf{\Omega}) \geq \sum_{i=1}^{pM} [\hat{\sigma}_{ii}\omega_{ii} - \log(\omega_{ii})] \geq pM[1 + \log(\min_i(\hat{\sigma}_{ii})],$$

which is bounded from below. Therefore, (A.6) has a unique solution $\hat{\mathbf{\Omega}}$.

By the interior extremum theorem (Spivak, 1980), if the global minimum of L is achieved at $\hat{\mathbf{\Omega}}$, then the first derivative of L at $\hat{\mathbf{\Omega}}$ will be zero, i.e.

$$\hat{\mathbf{\Sigma}} - \hat{\mathbf{\Omega}}^{-1} + \lambda_n(\mathbf{T} \otimes \mathbf{1}_M \mathbf{1}_M^\intercal) \circ \hat{\mathbf{Z}} = \mathbf{0}$$

The other direction is also true, since L is a convex function. □

Based on this lemma, we construct the primal-dual witness solution $(\tilde{\mathbf{\Omega}}, \tilde{\mathbf{Z}})$ as follows:

(a) We determine the matrix $\tilde{\mathbf{\Omega}}$ by solving the restricted adaptive fglasso

$$\tilde{\mathbf{\Omega}} = \underset{\mathbf{\Omega} \succ \mathbf{0}, \mathbf{\Omega}_{\mathcal{S}^c} = \mathbf{0}}{\operatorname{argmin}} \left\{ \operatorname{trace}(\hat{\mathbf{\Sigma}}\mathbf{\Omega}) - \log\det(\mathbf{\Omega}) + \lambda_n \sum_{j \neq l} \tau_{jl} \|\mathbf{\Omega}_{jl}\|_F \right\}. \tag{A.7}$$

(b) We choose $\tilde{\mathbf{Z}}_{\mathcal{S}}$ as a member of $\partial(\sum_{j \neq l} \|\tilde{\mathbf{\Omega}}_{jl}\|_F)/\partial \mathbf{\Omega}_{\mathcal{S}}$.

(c) For each $(j, l) \in \mathcal{S}^c$, we define

$$\tilde{\mathbf{Z}}_{jl} := \frac{1}{\lambda_n \tau_{jl}} \left[-\hat{\mathbf{\Sigma}}_{jl} + (\tilde{\mathbf{\Omega}}^{-1})_{jl} \right].$$

(d) We verify the strict dual feasibility condition

$$\|\mathbf{Z}_{jl}\|_F < 1 \quad \text{for all} \quad (i, j) \in \mathcal{S}^c.$$

With step (a) we ensure that

$$\hat{\mathbf{\Sigma}}_{\mathcal{S}} - (\tilde{\mathbf{\Omega}}^{-1})_{\mathcal{S}} + \lambda_n(\mathbf{T} \otimes \mathbf{1}_M \mathbf{1}_M^\intercal)_{\mathcal{S}} \circ \tilde{\mathbf{Z}}_{\mathcal{S}} = \mathbf{0},$$

which can be argued similarly as in Lemma A.5. The problem is that $\tilde{\mathbf{Z}}_{\mathcal{S}^c}$ is undefined, since in problem (A.7) we fix the elements $\mathbf{\Omega}_{\mathcal{S}^c}$ to be equal to zero. With step (c) we define $\tilde{\mathbf{Z}}_{\mathcal{S}^c}$ so that $(\tilde{\mathbf{\Omega}}, \tilde{\mathbf{Z}})$ is a solution of (A.5). The only thing that remains to show is that $\tilde{\mathbf{Z}}_{\mathcal{S}^c}$ is an element of $\partial \left(\sum_{j \neq l} \|\tilde{\mathbf{\Omega}}\|_F \right) / \partial \mathbf{\Omega}_{\mathcal{S}^c}$, which is the purpose of step (d). The result of the steps (a)-(d) is $\hat{\mathbf{\Omega}} = \tilde{\mathbf{\Omega}}$, which we need in order to show that $\hat{\mathcal{S}} \subset \mathcal{S}$.

In the analysis to follow, some additional notation is useful. We let $\mathbf{W} \in \mathbb{R}^{pM \times pM}$ denote the "effective noise" in the sample covariance matrix $\hat{\mathbf{\Sigma}}$–namely, the quantity

$$\mathbf{W} := \hat{\mathbf{\Sigma}} - \mathbf{\Omega}_0^{-1}. \tag{A.8}$$

Second, we use $\mathbf{\Delta} = \tilde{\mathbf{\Omega}} - \mathbf{\Omega}_0$ to measure the discrepancy between the primal witness matrix $\tilde{\mathbf{\Omega}}$ and the truth $\mathbf{\Omega}_0$. Note that by the definition of $\tilde{\mathbf{\Omega}}$, $\mathbf{\Delta}_{\mathcal{S}^c} = \mathbf{0}$. Finally, we let $\mathbf{R}(\mathbf{\Delta})$ denote the difference of the gradient $\nabla(\log\det(\tilde{\mathbf{\Omega}}))$ from its first-order Taylor expansion around $\mathbf{\Omega}_0$. Using known results on the first and second derivatives of the log-determinant function (Boyd and Vandenberghe, 2004, p. 641), this remainder takes the form

$$\mathbf{R}(\mathbf{\Delta}) = \tilde{\mathbf{\Omega}}^{-1} - \mathbf{\Omega}_0^{-1} + \mathbf{\Omega}_0^{-1}\mathbf{\Delta}\mathbf{\Omega}_0^{-1}. \tag{A.9}$$

We begin by stating and proving a lemma that provides sufficient conditions for strict dual feasibility to hold, so that $\|\tilde{\mathbf{Z}}_{\mathcal{S}^c}\|_{\max}^{(M^2)} < 1$.

Lemma A.6 (Strict dual feasibility)**.** *Suppose that*

$$\max\left\{\|\mathbf{W}\|_{\max}^{(M)}, \|\mathbf{R}(\mathbf{\Delta})\|_{\max}^{(M)}\right\} \leq \frac{\lambda_n \left(a_1 - a_2 C_{\mathbf{\Gamma}^2}\right)}{2\left(1 + C_{\mathbf{\Gamma}^2}\right)},$$

and

$$\min_{(j,l)\in\mathcal{S}^c} \tau_{jl} > a_1, \quad \max_{(j,l)\in\mathcal{S}} \tau_{jl} < a_2,$$

for a_1, a_2 specified in condition 2. Then, the vector $\tilde{\mathbf{Z}}_{\mathcal{S}^c}$ constructed in step (c) satisfies $\|\tilde{\mathbf{Z}}_{\mathcal{S}^c}\|_{\max}^{(M^2)} < 1$, and therefore $\tilde{\mathbf{\Omega}} = \hat{\mathbf{\Omega}}$.

Proof. The optimality condition (A.5) can be rewritten in the alternative but equivalent form

$$\mathbf{\Omega}_0^{-1}\mathbf{\Delta}\mathbf{\Omega}_0^{-1} + \mathbf{W} - \mathbf{R}(\mathbf{\Delta}) + \lambda_n(\mathbf{T} \otimes \mathbf{1}_M\mathbf{1}_M^\intercal) \circ \tilde{\mathbf{Z}} = \mathbf{0}. \tag{A.10}$$

The vectorized version of (A.10) is

$$\mathbf{\Gamma}\text{vec}(\mathbf{\Delta}) + \text{vec}(\mathbf{W}) - \text{vec}(\mathbf{R}) + \lambda_n\text{vec}(\mathbf{T} \otimes \mathbf{1}_M\mathbf{1}_M^\intercal) \circ \text{vec}(\tilde{\mathbf{Z}}) = \mathbf{0},$$

where we have abbreviated $\mathbf{R}(\mathbf{\Delta})$ by $\mathbf{R}$. Equivalently,

$$\begin{pmatrix} \mathbf{\Gamma}_{\mathcal{SS}} & \mathbf{\Gamma}_{\mathcal{SS}^c} \\ \mathbf{\Gamma}_{\mathcal{S}^c\mathcal{S}} & \mathbf{\Gamma}_{\mathcal{S}^c\mathcal{S}^c} \end{pmatrix} \begin{pmatrix} \mathbf{\Delta}_{\mathcal{S}} \\ \mathbf{\Delta}_{\mathcal{S}^c} \end{pmatrix} + \begin{pmatrix} \mathbf{W}_{\mathcal{S}} - \mathbf{R}_{\mathcal{S}} + \lambda_n(\mathbf{T} \otimes \mathbf{1}_M\mathbf{1}_M^\intercal)_{\mathcal{S}} \circ \tilde{\mathbf{Z}}_{\mathcal{S}} \\ \mathbf{W}_{\mathcal{S}^c} - \mathbf{R}_{\mathcal{S}^c} + \lambda_n(\mathbf{T} \otimes \mathbf{1}_M\mathbf{1}_M^\intercal)_{\mathcal{S}^c} \circ \tilde{\mathbf{Z}}_{\mathcal{S}^c} \end{pmatrix} = \mathbf{0}.$$

From this we get the system of equations

$$\mathbf{\Gamma}_{\mathcal{SS}}\mathbf{\Delta}_{\mathcal{S}} + \mathbf{W}_{\mathcal{S}} - \mathbf{R}_{\mathcal{S}} + \lambda_n(\mathbf{T} \otimes \mathbf{1}_M\mathbf{1}_M^\intercal)_{\mathcal{S}} \circ \tilde{\mathbf{Z}}_{\mathcal{S}} = \mathbf{0} \tag{A.11a}$$

$$\mathbf{\Gamma}_{\mathcal{S}^c\mathcal{S}}\mathbf{\Delta}_{\mathcal{S}} + \mathbf{W}_{\mathcal{S}^c} - \mathbf{R}_{\mathcal{S}^c} + \lambda_n(\mathbf{T} \otimes \mathbf{1}_M\mathbf{1}_M^\intercal)_{\mathcal{S}^c} \circ \tilde{\mathbf{Z}}_{\mathcal{S}^c} = \mathbf{0} \tag{A.11b}$$

Solving (A.11a) for $\mathbf{\Delta}_S$ and then substituting in (A.11b), we get

$$\begin{aligned} \lambda_n(\mathbf{T} \otimes \mathbf{1}_M\mathbf{1}_M^\intercal)_{\mathcal{S}^c} \circ \tilde{\mathbf{Z}}_{\mathcal{S}^c} = \mathbf{R}_{\mathcal{S}^c} - \mathbf{W}_{\mathcal{S}^c} - \mathbf{\Gamma}_{\mathcal{S}^c\mathcal{S}}\mathbf{\Gamma}_{\mathcal{SS}}^{-1}(\mathbf{R}_{\mathcal{S}} - \mathbf{W}_{\mathcal{S}}) \\ + \lambda_n\mathbf{\Gamma}_{\mathcal{S}^c\mathcal{S}}\mathbf{\Gamma}_{\mathcal{SS}}^{-1}(\mathbf{T} \otimes \mathbf{1}_M\mathbf{1}_M^\intercal)_{\mathcal{S}} \circ \tilde{\mathbf{Z}}_{\mathcal{S}}. \end{aligned}$$

Taking the M^2-block versions of the ℓ_∞ norm on both sides, we have

$$\lambda_n \left(\min_{(j,l)\in\mathcal{S}^c} \tau_{jl} \right) \|\tilde{\mathbf{Z}}_{\mathcal{S}^c}\|_{\max}^{(M^2)} \leq \|\mathbf{R}_{\mathcal{S}^c}\|_{\max}^{(M^2)} + \|\mathbf{W}_{\mathcal{S}^c}\|_{\max}^{(M^2)}$$

$$+ C_{\mathbf{\Gamma}^2} \left(\|\mathbf{R}_{\mathcal{S}}\|_{\max}^{(M^2)} + \|\mathbf{W}_{\mathcal{S}}\|_{\max}^{(M^2)} \right)$$
$$+ \lambda_n C_{\mathbf{\Gamma}^2} \left(\max_{(j,l)\in\mathcal{S}} \tau_{jl} \right).$$

Using the condition of the lemma, we have

$$\|\tilde{\mathbf{Z}}_{\mathcal{S}^c}\|_{\max}^{(M^2)} < \frac{1 + C_{\mathbf{\Gamma}^2}}{a_1 \lambda_n} \left(\|\mathbf{W}\|_{\max}^{(M)} + \|\mathbf{R}(\mathbf{\Delta})\|_{\max}^{(M)} \right) + \frac{a_2}{a_1} C_{\mathbf{\Gamma}^2},$$

which is no greater than 1. □

Our next step is to relate the behavior of the remainder term (A.9) to the deviation $\mathbf{\Delta} = \tilde{\mathbf{\Omega}} - \mathbf{\Omega}_0$.

Lemma A.7 (Control of the remainder)**.** *Suppose that* $\|\mathbf{\Delta}\|_{\max}^{(M)} \leq \frac{1}{3C_{\mathbf{\Sigma}} d}$ *holds. Then the matrix*

$$\mathbf{J} := \sum_{s=0}^{\infty} (-1)^s (\mathbf{\Omega}_0^{-1} \mathbf{\Delta})^s$$

satisfies $\|\mathbf{J}\|_{\infty}^{(M)} \leq \frac{3}{2}$*. Moreover, the remainder* $\mathbf{R}(\mathbf{\Delta})$ *is equal to*

$$\mathbf{\Omega}_0^{-1} \mathbf{\Delta} \mathbf{\Omega}_0^{-1} \mathbf{\Delta} \mathbf{J} \mathbf{\Omega}_0^{-1}$$

and has its M*-block* ℓ_∞ *vector norm satisfying*

$$\|\mathbf{R}(\mathbf{\Delta})\|_{\max}^{(M)} \leq \frac{3}{2} C_{\mathbf{\Sigma}}^3 d \left(\|\mathbf{\Delta}\|_{\max}^{(M)} \right)^2. \tag{A.12}$$

Proof. We write the remainder in the equivalent form

$$\mathbf{R}(\mathbf{\Delta}) = (\mathbf{\Omega}_0 + \mathbf{\Delta})^{-1} - \mathbf{\Omega}_0^{-1} + \mathbf{\Omega}_0^{-1} \mathbf{\Delta} \mathbf{\Omega}_0^{-1}. \tag{A.13}$$

By the submultiplicativity of the $\|\cdot\|_\infty^{(M)}$, we have

$$\|\mathbf{\Omega}_0^{-1}\mathbf{\Delta}\|_\infty^{(M)} \leq \|\mathbf{\Omega}_0^{-1}\|_\infty^{(M)}\|\mathbf{\Delta}\|_\infty^{(M)} \leq dC_{\mathbf{\Sigma}}\|\mathbf{\Delta}\|_{\max}^{(M)} < \frac{1}{3}, \tag{A.14}$$

where we have used $\|\mathbf{\Delta}\|_{\max}^{(M)} \leq \frac{1}{3C_{\mathbf{\Sigma}}d}$, and the fact that for each j, $\mathbf{\Delta}$ has at most d nonzero blocks $\mathbf{\Delta}_{jl}$. Consequently, we have the convergent matrix expansion (Schechter, 1996, p. 627)

$$\begin{aligned}
(\mathbf{\Omega}_0 + \mathbf{\Delta})^{-1} &= \left[\mathbf{\Omega}_0(\mathbf{I} + \mathbf{\Omega}_0^{-1}\mathbf{\Delta})\right]^{-1} \\
&= (\mathbf{I} + \mathbf{\Omega}_0^{-1}\mathbf{\Delta})^{-1}\mathbf{\Omega}_0^{-1} \\
&= \sum_{s=0}^{\infty}(-1)^s(\mathbf{\Omega}_0^{-1}\mathbf{\Delta})^s\mathbf{\Omega}_0^{-1} \\
&= \mathbf{\Omega}_0^{-1} - \mathbf{\Omega}_0^{-1}\mathbf{\Delta}\mathbf{\Omega}_0^{-1} + \mathbf{\Omega}_0^{-1}\mathbf{\Delta}\mathbf{\Omega}_0^{-1}\mathbf{\Delta}\mathbf{J}\mathbf{\Omega}_0^{-1}
\end{aligned} \tag{A.15}$$

Substituting (A.15) into (A.13) yields

$$\mathbf{R}(\mathbf{\Delta}) = \mathbf{\Omega}_0^{-1}\mathbf{\Delta}\mathbf{\Omega}_0^{-1}\mathbf{\Delta}\mathbf{J}\mathbf{\Omega}_0^{-1}.$$

We now prove the bound on $\mathbf{R}(\mathbf{\Delta})$ as follows. Let the block matrix $\mathbf{e}_j \in \mathbb{R}^{pM\times M}$ with identity matrix in the j-th block and zero matrix elsewhere. Then,

$$\begin{aligned}
\|\mathbf{R}(\mathbf{\Delta})\|_{\max}^{(M)} &= \max_{j,l}\|\mathbf{e}_j^\intercal\mathbf{\Omega}_0^{-1}\mathbf{\Delta}\mathbf{\Omega}_0^{-1}\mathbf{\Delta}\mathbf{J}\mathbf{\Omega}_0^{-1}\mathbf{e}_l\|_F \\
&\leq \max_j\|\mathbf{\Delta}\mathbf{\Omega}_0^{-1}\mathbf{e}_j\|_{\max}^{(M)}\max_l\|\mathbf{\Omega}_0^{-1}\mathbf{\Delta}\mathbf{J}\mathbf{\Omega}_0^{-1}\mathbf{e}_l\|_1^{(M)}, \quad (\text{by } A.2c) \\
&\leq \max_j\|\mathbf{\Omega}_0^{-1}\mathbf{e}_j\|_1^{(M)}\|\mathbf{\Delta}\|_{\max}^{(M)}\max_l\|\mathbf{\Omega}_0^{-1}\mathbf{\Delta}\mathbf{J}\mathbf{\Omega}_0^{-1}\mathbf{e}_l\|_1^{(M)}, \quad (\text{by } A.2d) \\
&= \|\mathbf{\Omega}_0^{-1}\|_\infty^{(M)}\|\mathbf{\Delta}\|_{\max}^{(M)}\|\mathbf{\Omega}_0^{-1}\mathbf{J}^\intercal\mathbf{\Delta}\mathbf{\Omega}_0^{-1}\|_\infty^{(M)}, \quad (\text{by } A.2e) \\
&\leq C_{\mathbf{\Sigma}}^3\|\mathbf{\Delta}\|_{\max}^{(M)}\|\mathbf{J}^\intercal\|_\infty^{(M)}\|\mathbf{\Delta}\|_\infty^{(M)}, \quad (\text{by } A.2f) \\
&\leq C_{\mathbf{\Sigma}}^3 d\left(\|\mathbf{\Delta}\|_{\max}^{(M)}\right)^2\|\mathbf{J}^\intercal\|_\infty^{(M)}.
\end{aligned}$$

Note that by (A.14), we have

$$\|\mathbf{J}^{\intercal}\|_{\infty}^{(M)} \leq \sum_{s=0}^{\infty} \left(\|\boldsymbol{\Delta}\boldsymbol{\Omega}_0^{-1}\|_{\infty}^{(M)}\right)^s \leq \frac{1}{1 - \|\boldsymbol{\Delta}\|_{\infty}^{(M)}\|\boldsymbol{\Omega}_0^{-1}\|_{\infty}^{(M)}} < \frac{3}{2}.$$

Thus, (A.12) holds. □

Our next lemma provides control on the deviation $\boldsymbol{\Delta} = \tilde{\boldsymbol{\Omega}} - \boldsymbol{\Omega}_0$, measured in the M-block elementwise norm.

Lemma A.8 (Control of $\boldsymbol{\Delta}$). *Suppose that*

$$\max_{(j,l)\in\mathcal{S}} \tau_{jl} < a_2 \quad \textit{and} \quad r := 2C_{\boldsymbol{\Gamma}}\left(\|\mathbf{W}\|_{\max}^{(M)} + \lambda_n a_2\right) \leq \min\left\{\frac{1}{3C_{\boldsymbol{\Sigma}}d}, \frac{1}{3C_{\boldsymbol{\Sigma}}^3 C_{\boldsymbol{\Gamma}}d}\right\}.$$

Then, there exists $\boldsymbol{\Delta} \in \mathbb{R}^{pM\times pM}$ *such that* $\tilde{\boldsymbol{\Omega}} = \boldsymbol{\Delta} + \boldsymbol{\Omega}_0$ *and* $\|\boldsymbol{\Delta}\|_{\max}^{(M)} \leq r$.

Proof. By arguing the same way as in Lemma A.5, there exists a unique solution to the restricted adaptive fglasso problem (A.7) and it is equal to $\tilde{\boldsymbol{\Omega}}$ if and only if $\tilde{\boldsymbol{\Omega}}_{\mathcal{S}^c} = \mathbf{0}$ and it is the root of the first-order derivative equation

$$\hat{\boldsymbol{\Sigma}}_{\mathcal{S}} - (\tilde{\boldsymbol{\Omega}}^{-1})_{\mathcal{S}} + \lambda_n(\mathbf{T} \otimes \mathbf{1}_M\mathbf{1}_M^{\intercal})_{\mathcal{S}} \circ \tilde{\mathbf{Z}}_{\mathcal{S}} = \mathbf{0},$$

where $\tilde{\mathbf{Z}}_{\mathcal{S}}$ is a member of the subdifferential $\partial(\sum_{j=1}\|\tilde{\boldsymbol{\Omega}}_{jl}\|_F)/\partial\boldsymbol{\Omega}_{\mathcal{S}}$.

Let $\mathcal{A} = \{\boldsymbol{\Omega} \succ \mathbf{0} : \boldsymbol{\Omega}_{\mathcal{S}^c} = \mathbf{0}\}$ and $\mathcal{B} = \{\boldsymbol{\Delta} \in \mathbb{R}^{pM\times pM} : \boldsymbol{\Delta}_{\mathcal{S}^c} = \mathbf{0}\}$. We define the functions $\mathbf{G} : \mathcal{A} \to \mathbb{R}^{|\mathcal{S}|^2M^2}$and $\mathbf{F} : \mathcal{B} \to \mathbb{R}^{|\mathcal{S}|^2M^2}$ as

$$\begin{aligned}
\mathbf{G}(\boldsymbol{\Omega}) &= -[\boldsymbol{\Omega}^{-1}]_{\mathcal{S}} + \hat{\boldsymbol{\Sigma}}_{\mathcal{S}} + \lambda_n(\mathbf{T} \otimes \mathbf{1}_M\mathbf{1}_M^{\intercal})_{\mathcal{S}} \circ \tilde{\mathbf{Z}}_{\mathcal{S}}, \\
\mathbf{F}(\boldsymbol{\Delta}) &= -{\boldsymbol{\Gamma}_{\mathcal{S}\mathcal{S}}}^{-1}\mathbf{G}(\boldsymbol{\Omega}_0 + \boldsymbol{\Delta}) + \boldsymbol{\Delta}_{\mathcal{S}}.
\end{aligned} \tag{A.16}$$

Finally, we define the closed ball

$$\mathcal{C}(\mathcal{B}, r) = \left\{\boldsymbol{\Delta} \in \mathcal{B} : \|\boldsymbol{\Delta}_{\mathcal{S}}\|_{\max}^{(M^2)} \leq r\right\}.$$

Observe that

$$\begin{aligned}\mathbf{G}(\boldsymbol{\Omega}_0+\boldsymbol{\Delta}) &= -[(\boldsymbol{\Omega}_0+\boldsymbol{\Delta})^{-1}]_{\mathcal{S}} + \hat{\boldsymbol{\Sigma}}_{\mathcal{S}} + \lambda_n(\mathbf{T}\otimes \mathbf{1}_M\mathbf{1}_M^{\intercal})_{\mathcal{S}} \circ \tilde{\mathbf{Z}}_{\mathcal{S}} \\ &= -[\mathbf{R}(\boldsymbol{\Delta}) - \boldsymbol{\Omega}_0^{-1}\boldsymbol{\Delta}\boldsymbol{\Omega}_0^{-1}]_{\mathcal{S}} + \mathbf{W}_{\mathcal{S}} + \lambda_n(\mathbf{T}\otimes \mathbf{1}_M\mathbf{1}_M^{\intercal})_{\mathcal{S}} \circ \tilde{\mathbf{Z}}_{\mathcal{S}} \\ &= -\mathbf{R}_{\mathcal{S}} + \boldsymbol{\Gamma}_{\mathcal{S}\mathcal{S}}\boldsymbol{\Delta}_{\mathcal{S}} + \mathbf{W}_{\mathcal{S}} + \lambda_n(\mathbf{T}\otimes \mathbf{1}_M\mathbf{1}_M^{\intercal})_{\mathcal{S}} \circ \tilde{\mathbf{Z}}_{\mathcal{S}}. \end{aligned} \tag{A.17}$$

Substituting (A.17) into (A.16), we obtain

$$\mathbf{F}(\boldsymbol{\Delta}) = \underbrace{\boldsymbol{\Gamma}_{\mathcal{S}\mathcal{S}}^{-1}\mathbf{R}_{\mathcal{S}}}_{\mathbf{T}_1} - \underbrace{\boldsymbol{\Gamma}_{\mathcal{S}\mathcal{S}}^{-1}[\mathbf{W}_{\mathcal{S}} + \lambda_n(\mathbf{T}\otimes \mathbf{1}_M\mathbf{1}_M^{\intercal})_{\mathcal{S}} \circ \tilde{\mathbf{Z}}_{\mathcal{S}}]}_{\mathbf{T}_2}.$$

Let $\boldsymbol{\Delta} \in \mathcal{C}(\mathcal{B}, r)$. Then, by Lemma A.7 and the conditions of this lemma about r, we get

$$\|\mathbf{T}_1\|_{\max}^{(M^2)} \leq C_{\boldsymbol{\Gamma}}\|\mathbf{R}_{\mathcal{S}}\|_{\max}^{(M^2)} \leq C_{\boldsymbol{\Gamma}}\frac{3}{2}C_{\boldsymbol{\Sigma}}^3 d r^2 \leq \frac{r}{2}.$$

Concerning $\mathbf{T}_2$, we have

$$\|\mathbf{T}_2\|_{\max}^{(M^2)} \leq C_{\boldsymbol{\Gamma}}\left(\|\mathbf{W}\|_{\max}^{(M)} + \lambda_n \max_{(j,l)\in\mathcal{S}} \tau_{jl}\right) \leq \frac{r}{2}.$$

Combining the two bounds, we get $\|\mathbf{F}(\boldsymbol{\Delta})\| \leq r$ for every $\boldsymbol{\Delta} \in \mathcal{C}(\mathcal{B}, r)$, which means that $\mathbf{F}(\mathcal{C}(\mathcal{B}, r)) \subset \mathcal{C}(\mathcal{B}, r)$. By Brouwer's fixed point theorem (Ortega and Rheinboldt, 2000, p. 161), there exists $\boldsymbol{\Delta} \in \mathcal{C}(\mathcal{B}, r)$ such that

$$\mathbf{F}(\boldsymbol{\Delta}) = \boldsymbol{\Delta} \Leftrightarrow \mathbf{G}(\boldsymbol{\Omega}_0+\boldsymbol{\Delta}) = \mathbf{0} \Leftrightarrow \tilde{\boldsymbol{\Omega}} = \boldsymbol{\Omega}_0 + \boldsymbol{\Delta},$$

which completes the proof. □

Let $\mathbb{N}$ denote the set of positive integers and let the true covariance matrix be $\boldsymbol{\Sigma}_0 = (\sigma_{0,jl})$. To prove the tail bounds of $\|\mathbf{W}\|_{\max}^{(M)}$ and $\|\mathbf{R}(\boldsymbol{\Delta})\|_{\max}^{(M)}$, we make use of

the following useful definition.

Definition A.1 (Tail condition)**.** *The random vector* $\mathbf{X} \in \mathbb{R}^{p_n M_n}$ *satisfies the tail condition* $\mathcal{T}(f, \nu_\star)$ *if there exists a constant* $\nu_\star \in (0, \infty]$ *and a function* $f : \mathbb{N} \times (0, \infty) \to (0, \infty)$ *such that for any* $i, j = 1, \ldots, p_n M_n$*, we have*

$$P(|\hat{\sigma}_{ij} - \sigma_{0,ij}| \geq \delta) \leq \frac{1}{f(n, \delta)}, \quad \textit{for all } \delta \in (0, 1/\nu_\star].$$

Given a larger sample size n, we expect the tail probability bound $1/f(n, \delta)$ to be smaller, or equivalently, for the tail function $f(n, \delta)$ to be larger. Accordingly, we require that f is monotonically increasing in n, so that for each fixed $\delta > 0$, we can define the inverse function

$$\bar{n}_f(\delta, r) = \text{argmax}\{n : f(n, \delta) \leq r\}. \tag{A.18}$$

Similarly, we expect that f is monotonically increasing in δ, so that for each fixed n, we can define the inverse in the second argument

$$\bar{\delta}_f(n, r) = \text{argmax}\{\delta : f(n, \delta) \leq r\}, \tag{A.19}$$

where $r \in [1, \infty)$. For future reference, we note a simple consequence of the monotonicity of the tail function f, that is

$$n > \bar{n}_f(\delta, r) \quad \text{for some } \delta > 0 \quad \Longrightarrow \quad \bar{\delta}_f(n, r) < \delta. \tag{A.20}$$

It can be proven (Theorem 1; Qiao et al., 2019) that under condition 1, one such tail function is

$$f(n, \delta) = \frac{1}{C_2} \exp\left\{C_1 n^{1-2\alpha(1+\beta)} \delta^2\right\},$$

for some constants $C_1, C_2 > 0$ and any $0 < \delta \leq C_1$. Define

$$\delta_1 = \frac{\min\left\{\frac{1}{3C_{\mathbf{\Sigma}}d_n}, \frac{1}{3C_{\mathbf{\Sigma}}^3 C_{\mathbf{\Gamma}} d_n}\right\}}{2M_n C_{\mathbf{\Gamma}}\left[1 + \frac{2a_2(1+C_{\mathbf{\Gamma}^2})}{a_1 - a_2 C_{\mathbf{\Gamma}^2}}\right]}, \quad \delta_2 = \frac{1}{6M_n d_n C_{\mathbf{\Sigma}}^3 C_{\mathbf{\Gamma}}^2\left[1 + \frac{2a_2(1+C_{\mathbf{\Gamma}^2})}{a_1 - a_2 C_{\mathbf{\Gamma}^2}}\right]^2}$$

Lemma A.9. *Suppose that*

$$\|\mathbf{W}\|_{\max}^{(M)} \leq \frac{\lambda_n\left(a_1 - a_2 C_{\mathbf{\Gamma}^2}\right)}{2\left(1 + C_{\mathbf{\Gamma}^2}\right)} \quad \textit{and} \quad \max_{(j,l)\in\mathcal{S}} \tau_{jl} < a_2,$$

for a_1, a_2 specified in condition 2. Then, for any $\gamma > 2$,

$$n > \bar{n}_f\left(\min\{C_1, \delta_1, \delta_2\}, (M_n p_n)^\gamma\right), \quad \lambda_n = \frac{2(1 + C_{\mathbf{\Gamma}^2}) M_n \bar{\delta}_f\left(n, (M_n p_n)^\gamma\right)}{a_1 - a_2 C_{\mathbf{\Gamma}^2}},$$

we have,

$$\|\mathbf{R}(\mathbf{\Delta})\|_{\max}^{(M)} \leq \frac{\lambda_n\left(a_1 - a_2 C_{\mathbf{\Gamma}^2}\right)}{2\left(1 + C_{\mathbf{\Gamma}^2}\right)}.$$

Proof. Observe that

$$2C_{\mathbf{\Gamma}}(\|\mathbf{W}\|_{\max}^{(M)} + \lambda_n a_2) \leq 2C_{\mathbf{\Gamma}}\left[1 + \frac{2a_2(1 + C_{\mathbf{\Gamma}^2})}{a_1 - a_2 C_{\mathbf{\Gamma}^2}}\right] M_n \bar{\delta}_f\left(n, (M_n p_n)^\gamma\right).$$

From the lower bound on n and the monotonicity of f, we have

$$2C_{\mathbf{\Gamma}}\left[1 + \frac{2a_2(1 + C_{\mathbf{\Gamma}^2})}{a_1 - a_2 C_{\mathbf{\Gamma}^2}}\right] M_n \bar{\delta}_f\left(n, (M_n p_n)^\gamma\right) \leq \min\left\{\frac{1}{3C_{\mathbf{\Sigma}}d_n}, \frac{1}{3C_{\mathbf{\Sigma}}^3 C_{\mathbf{\Gamma}} d_n}\right\}.$$

By Lemmas A.7 and A.8, we have

$$\begin{aligned}\|\mathbf{R}(\mathbf{\Delta})\|_{\max}^{(M)} &\leq 6d_n C_{\mathbf{\Sigma}}^3 C_{\mathbf{\Gamma}}^2\left[1 + \frac{2a_2(1 + C_{\mathbf{\Gamma}^2})}{a_1 - a_2 C_{\mathbf{\Gamma}^2}}\right]^2 M_n^2 \bar{\delta}_f\left(n, (M_n p_n)^\gamma\right)^2 \\ &\leq M_n \bar{\delta}_f\left(n, (M_n p_n)^\gamma\right) = \frac{\lambda_n\left(a_1 - a_2 C_{\mathbf{\Gamma}^2}\right)}{2\left(1 + C_{\mathbf{\Gamma}^2}\right)},\end{aligned}$$

where the last inequality follows from the lower bound on the sample size n and the monotonicity of f. □

Lemma A.10. *Suppose condition 1 holds. For any* $\gamma > 2$, $n > \bar{n}_f\left(C_1, (M_n p_n)^\gamma\right)$ *and*

$$\lambda_n = \frac{2(1 + C_{\mathbf{\Gamma}^2}) M_n \bar{\delta}_f\left(n, (M_n p_n)^\gamma\right)}{a_1 - a_2 C_{\mathbf{\Gamma}^2}},$$

we have,

$$\|\mathbf{W}\|_{\max}^{(M)} \leq \frac{\lambda_n\left(a_1 - a_2 C_{\mathbf{\Gamma}^2}\right)}{2\left(1 + C_{\mathbf{\Gamma}^2}\right)},$$

with probability greater than $1 - (M_n p_n)^{2-\gamma}$.

Proof. By Lemma 14 of Qiao et al. (2019), we have

$$P\left(\|\mathbf{W}\|_{\max}^{(M)} \geq M_n \bar{\delta}_f\left(n, (M_n p_n)^\gamma\right)\right) \leq (Mp)^{2-\gamma}.$$

Thus, for

$$\lambda_n = \frac{2(1 + C_{\mathbf{\Gamma}^2}) M_n \bar{\delta}_f\left(n, (M_n p_n)^\gamma\right)}{a_1 - a_2 C_{\mathbf{\Gamma}^2}},$$

we have

$$\|\mathbf{W}\|_{\max}^{(M)} \leq \frac{\lambda_n\left(a_1 - a_2 C_{\mathbf{\Gamma}^2}\right)}{2\left(1 + C_{\mathbf{\Gamma}^2}\right)},$$

with probability greater than $1 - (M_n p_n)^{2-\gamma}$. □

We are now ready to prove graph selection consistency.

Theorem A.1. *Suppose conditions 1 and 2 hold, $\gamma > 2$,*

$$\lambda_n = \frac{2(1+C_{\mathbf{\Gamma}^2})M_n}{a_1 - a_2 C_{\mathbf{\Gamma}^2}} \sqrt{\frac{\log C_2 + \gamma \log(M_n p_n)}{C_1 n^{1-2\alpha(1+\beta)}}},$$

and

$$\min_{(j,l)\in\mathcal{S}} \|\mathbf{\Omega}_{0,jl}\|_F > \min\left\{ \frac{1}{3C_{\mathbf{\Sigma}} d_n}, \frac{1}{3C_{\mathbf{\Sigma}}^3 C_{\mathbf{\Gamma}} d_n} \right\}.$$

Then, for all n satisfying the lower bound

$$n^{1-2\alpha(1+\beta)} > \frac{\log[C_2(M_n p_n)^\gamma]}{C_1} \max\left\{ \frac{1}{C_1}, \frac{2M_n C_{\mathbf{\Gamma}}\left[1+\frac{2a_2(1+C_{\mathbf{\Gamma}^2})}{a_1-a_2C_{\mathbf{\Gamma}^2}}\right]}{\min\left\{\frac{1}{3C_{\mathbf{\Sigma}}d_n}, \frac{1}{3C_{\mathbf{\Sigma}}^3 C_{\mathbf{\Gamma}} d_n}\right\}}, 6M_n d_n C_{\mathbf{\Gamma}}^2 C_{\mathbf{\Gamma}^2}\left[1+\frac{2a_2(1+C_{\mathbf{\Gamma}^2})}{a_1-a_2C_{\mathbf{\Gamma}^2}}\right]^2 \right\}^2,$$

we have $\hat{\mathcal{S}} = \mathcal{S}$ with probability greater than $1 - 3(M_n p_n)^{2-\gamma}$.

Proof. By Lemmas A.6 and A.9 we can see that

$$\left\{\hat{\mathbf{\Omega}} = \tilde{\mathbf{\Omega}}\right\} \supset \left\{ \max_{(j,l)\in\mathcal{S}} \tau_{jl} < a_2 \right\} \cap \left\{ \min_{(j,l)\in\mathcal{S}^c} \tau_{jl} > a_1 \right\} \cap \left\{ \|\mathbf{W}\|_{\max}^{(M)} \leq \frac{\lambda_n (a_1 - a_2 C_{\mathbf{\Gamma}^2})}{2(1+C_{\mathbf{\Gamma}^2})} \right\}.$$

Hence, by Lemmas A.6, A.9 and A.10, we have

$$P\left(\hat{\mathbf{\Omega}} = \tilde{\mathbf{\Omega}}\right) \geq 1 - 3(M_n p_n)^{2-\gamma}. \tag{A.21}$$

To find the lower bound of n so that (A.21) is satisfied we use Definition A.18, with $\delta = \min\{C_1, \delta_1, \delta_2\}$ and $r = (M_n p_n)^\gamma$. Then we fix an $n > \bar{n}_f(\min\{C_1, \delta_1, \delta_2\}, (M_n p_n)^\gamma)$ and find $\bar{\delta}_f(n, (M_n p_n)^\gamma)$ using Definition A.19, relationship (A.20), and substituting in λ_n.

Conditioning on the event $\hat{\mathbf{\Omega}} = \tilde{\mathbf{\Omega}}$, we have $\mathcal{S}^c \subset \hat{\mathcal{S}}^c$. By Lemma A.8, for any

$(j,l) \in \hat{\mathcal{S}}^c \cap \mathcal{S}$, we have

$$\|\mathbf{\Omega}_{0,jl}\|_F = \|\mathbf{\Omega}_{0,jl} - \hat{\mathbf{\Omega}}_{jl}\|_F = \|\mathbf{\Omega}_{0,jl} - \tilde{\mathbf{\Omega}}_{jl}\|_F = \|\mathbf{\Delta}_{jl}\| \leq \min\left\{\frac{1}{3C_{\mathbf{\Sigma}}d_n}, \frac{1}{3C_{\mathbf{\Sigma}}^3 C_{\mathbf{\Gamma}} d_n}\right\},$$

which is a contradiction. Thus, $\hat{\mathcal{S}} = \mathcal{S}$. □

Appendix B

Proofs of Chapter 2

B.1 Proof of Theorem 4

Define the intermediate objective function:

$$\frac{1}{2n}\sum_{k=1}^{K}\|\mathbb{X}^{(k)}(\mathbf{I}-\mathbf{\Theta}^{(k)})\|_F^2+\|\mathbf{H}\|_1+\lambda_1\lambda_2\sum_{k=1}^{K}\|\mathbf{\Gamma}^{(k)}\|_1, \tag{B.1}$$

for $\mathbf{H}$ and $\mathbf{\Gamma}$ specified in subsection 2.2.2. We first show that the objective functions (2.4) and (B.1) are equivalent.

Lemma B.1. *If* $(\hat{\mathbf{H}},\hat{\mathbf{\Gamma}})$ *is a local minimizer of* (2.4)*, then there exists a local minimizer* $(\tilde{\mathbf{H}},\tilde{\mathbf{\Gamma}})$ *of* (B.1) *such that* $\tilde{\mathbf{H}}\circ\tilde{\mathbf{\Gamma}}^{(k)}=\hat{\mathbf{H}}\circ\hat{\mathbf{\Gamma}}^{(k)}$ *for all* k*. Conversely, if* $(\tilde{\mathbf{H}},\tilde{\mathbf{\Gamma}})$ *is a local minimizer of* (B.1)*, then there exists a local minimizer* $(\hat{\mathbf{H}},\hat{\mathbf{\Gamma}})$ *of* (2.4) *such that* $\hat{\mathbf{H}}\circ\hat{\mathbf{\Gamma}}^{(k)}=\tilde{\mathbf{H}}\circ\tilde{\mathbf{\Gamma}}^{(k)}$ *for all* k*.*

Proof. Let $Q_1(\lambda_1,\lambda_2,\mathbf{H},\mathbf{\Gamma})$ and $Q_2(\lambda_1\lambda_2,\mathbf{H},\mathbf{\Gamma})$ denote the objective functions (2.4) and (B.1), respectively. Observe that

$$\begin{aligned}Q_1(\lambda_1,\lambda_2,\mathbf{H},\mathbf{\Gamma})&=Q_2(\lambda_1\lambda_2,\lambda_1\mathbf{H},\lambda_1^{-1}\mathbf{\Gamma}),\\ Q_2(\lambda_1\lambda_2,\mathbf{H},\mathbf{\Gamma})&=Q_1(\lambda_1,\lambda_2,\lambda_1^{-1}\mathbf{H},\lambda_1\mathbf{\Gamma}).\end{aligned}$$

Since $(\hat{\mathbf{H}}, \hat{\mathbf{\Gamma}})$ is a local minimizer of $Q_1(\lambda_1, \lambda_2, \cdot, \cdot)$, there exists $\delta > 0$ such that for every $(\mathbf{H}, \mathbf{\Gamma})$ with

$$\|\mathbf{H} - \hat{\mathbf{H}}\|_1 + \sum_{k=1}^{K} \|\mathbf{\Gamma}^{(k)} - \hat{\mathbf{\Gamma}}^{(k)}\|_1 < \delta,$$

we have

$$Q_1(\lambda_1, \lambda_2, \hat{\mathbf{H}}, \hat{\mathbf{\Gamma}}) \leq Q_1(\lambda_1, \lambda_2, \mathbf{H}, \mathbf{\Gamma}).$$

Let $0 < \delta^* \leq \delta \min(\lambda_1, \lambda_1^{-1})$, and define $(\tilde{\mathbf{H}}, \tilde{\mathbf{\Gamma}}) = (\lambda_1 \hat{\mathbf{H}}, \lambda_1^{-1} \hat{\mathbf{\Gamma}})$. Then, for any $(\mathbf{H}, \mathbf{\Gamma})$ satisfying

$$\|\mathbf{H} - \tilde{\mathbf{H}}\|_1 + \sum_{k=1}^{K} \|\mathbf{\Gamma}^{(k)} - \tilde{\mathbf{\Gamma}}^{(k)}\|_1 < \delta^*,$$

we have

$$\|\lambda_1^{-1}\mathbf{H} - \hat{\mathbf{H}}\|_1 + \sum_{k=1}^{K} \|\lambda_1 \mathbf{\Gamma}^{(k)} - \hat{\mathbf{\Gamma}}^{(k)}\|_1 \leq \frac{\|\mathbf{H} - \tilde{\mathbf{H}}\|_1 + \sum_{k=1}^{K} \|\mathbf{\Gamma}^{(k)} - \tilde{\mathbf{\Gamma}}^{(k)}\|_1}{\min(\lambda_1, \lambda_1^{-1})} \leq \delta.$$

Thus

$$Q_1(\lambda_1, \lambda_2, \hat{\mathbf{H}}, \hat{\mathbf{\Gamma}}) \leq Q_1(\lambda_1, \lambda_2, \lambda_1^{-1}\mathbf{H}, \lambda_1 \mathbf{\Gamma}) \Rightarrow Q_2(\lambda_1\lambda_2, \tilde{\mathbf{H}}, \tilde{\mathbf{\Gamma}}) \leq Q_2(\lambda_1\lambda_2, \mathbf{H}, \mathbf{\Gamma}),$$

which means that $(\tilde{\mathbf{H}}, \tilde{\mathbf{\Gamma}})$ is a local minimizer of (B.1). The other direction is proven similarly. □

Lemma B.2. *Suppose* $(\hat{\mathbf{H}}, \hat{\mathbf{\Gamma}})$ *is a local minimizer of* (B.1) *and* $\hat{\mathbf{\Theta}}^{(k)} = \hat{\mathbf{H}} \circ \hat{\mathbf{\Gamma}}^{(k)}$ *for all* k*. Then for any* l, j *the following statements hold true:*

1. $\hat{\eta}_{lj} = 0$ *if and only if* $\hat{\gamma}_{lj}^{(k)} = 0$ *for all* $k = 1, \ldots, K$*;*

2. If $\hat{\eta}_{lj} \neq 0$, *then* $\hat{\eta}_{lj} = \left(\lambda_1\lambda_2 \sum_{k=1}^{K} |\hat{\theta}_{lj}(k)|\right)^{1/2}$.

Proof. 1. If $l = j$, by definition $\gamma_{jj}^{(k)} = \eta_{jj} = 0$ for all k. Suppose then $l \neq j$ and $\eta_{lj} = 0$, then $\gamma_{lj}^{(1)}, \ldots, \gamma_{lj}^{(K)}$ only appear in the third term in (B.1). Thus, in order to minimize $Q_2(\lambda_1\lambda_2, \cdot, \cdot)$, we need $\gamma_{lj}^{(k)} = 0$ for all $k = 1, \ldots, K$. The reverse implication can be proved similarly.

2. Suppose $\hat{\eta}_{lj} \neq 0$ and let

$$c = \frac{\left(\lambda_1\lambda_2 \sum_{k=1}^{K} |\hat{\theta}_{lj}(k)|\right)^{1/2}}{\hat{\eta}_{lj}}.$$

We will show $c = 1$. By definition

$$\hat{\gamma}_{lj}^{(k)} = c\frac{\hat{\theta}_{lj}(k)}{\left(\lambda_1\lambda_2 \sum_{k=1}^{K} |\hat{\theta}_{lj}(k)|\right)^{1/2}}.$$

Suppose $c > 1$. Since $(\hat{\mathbf{H}}, \hat{\mathbf{\Gamma}})$ is a local minimizer of $Q_2(\lambda_1\lambda_2, \cdot, \cdot)$, there exists $\delta > 0$ such that for all $(\mathbf{H}, \mathbf{\Gamma})$ with

$$\|\mathbf{H} - \hat{\mathbf{H}}\|_1 + \sum_{k=1}^{K} \|\mathbf{\Gamma}^{(k)} - \hat{\mathbf{\Gamma}}^{(k)}\|_1 < \delta,$$

we have $Q_2(\lambda_1\lambda_2, \hat{\mathbf{H}}, \hat{\mathbf{\Gamma}}) \leq Q_2(\lambda_1\lambda_2, \mathbf{H}, \mathbf{\Gamma})$. Furthermore, there exists $\delta^* \in (1, c)$ slightly greater that 1, such that for $(\tilde{\mathbf{H}}, \tilde{\mathbf{\Gamma}})$ defined by

$$\begin{cases} \tilde{\eta}_{l'j'} = \hat{\eta}_{l'j'} \text{ and } \tilde{\gamma}_{l'j'}^{(k)} = \hat{\gamma}_{l'j'}^{(k)}, & (l', j') \neq (l, j) \\ \tilde{\eta}_{lj} = \delta^*\hat{\eta}_{lj} \text{ and } \tilde{\gamma}_{lj}^{(k)} = \frac{1}{\delta^*}\hat{\gamma}_{lj}^{(k)} \end{cases}$$

we have

$$\|\tilde{\mathbf{H}} - \hat{\mathbf{H}}\|_1 + \sum_{k=1}^{K} \|\tilde{\mathbf{\Gamma}}^{(k)} - \hat{\mathbf{\Gamma}}^{(k)}\|_1 < \delta.$$

But this implies

$$Q_2(\lambda_1\lambda_2, \hat{\mathbf{H}}, \hat{\mathbf{\Gamma}}) - Q_2(\lambda_1\lambda_2, \tilde{\mathbf{H}}, \tilde{\mathbf{\Gamma}}) = (1-\delta^*)\hat{\eta}_{lj} + \left(1 - \frac{1}{\delta^*}\right)\lambda_1\lambda_2 \sum_{k=1}^{K} |\hat{\gamma}_{lj}^{(k)}|$$
$$= \frac{1}{c}(\delta^* - 1)\left(\frac{c^2}{\delta^*} - 1\right)\left(\lambda_1\lambda_2 \sum_{k=1}^{K} |\hat{\theta}_{lj}(k)|\right)^{1/2} > 0,$$

which is impossible because $(\hat{\mathbf{H}}, \hat{\mathbf{\Gamma}})$ is a local minimizer. Hence $c \leq 1$. Following the same argument we can show $c \geq 1$. Thus $c = 1$. □

Lemma B.3. *If* $\{a_l^{(k)} : l \in \mathcal{V}\}$ *and* $\{b_l^{(k)} : l \in \mathcal{V}\}$ *satisfy*

$$0 < |b_l^{(k)}| < \frac{1}{K} \min\left\{|a_l^{(k)}| : a_l^{(k)} \neq 0, l \in \mathcal{V}, k = 1, \ldots, K\right\}$$

for all l, k*, then, for each* $l \in \mathcal{V}$*,*

$$\left|\left(\sum_{k=1}^{K} |a_l^{(k)} + b_l^{(k)}|\right)^{1/2} - \left(\sum_{k=1}^{K} |b_l^{(k)}|\right)^{1/2}\right| \leq \frac{1}{2} \frac{\sum_{k=1}^{K} |b_l^{(k)}|}{\left(\sum_{k=1}^{K} |a_l^{(k)}| - \sum_{k=1}^{K} |b_l^{(k)}|\right)^{1/2}}.$$

Proof. First note that for any $0 < x < y$, we have

$$|\sqrt{y+x} - \sqrt{y}| \leq |\sqrt{y-x} - \sqrt{y}| \leq \frac{1}{2}\frac{x}{\sqrt{y-x}}. \tag{B.2}$$

By the triangular inequality,

$$\sum_{k=1}^{K} |a_l^{(k)}| - \sum_{k=1}^{K} |b_l^{(k)}| \leq \sum_{k=1}^{K} |a_l^{(k)} + b_l^{(k)}| \leq \sum_{k=1}^{K} |a_l^{(k)}| + \sum_{k=1}^{K} |b_l^{(k)}|.$$

Therefore, for each $l \in \mathcal{V}$,

$$\left|\left(\sum_{k=1}^{K} |a_l^{(k)} + b_l^{(k)}|\right)^{1/2} - \left(\sum_{k=1}^{K} |a_l^{(k)}|\right)^{1/2}\right|$$

is less than or equal to the maximum of

$$\left| \left(\sum_{k=1}^{K} |a_l^{(k)}| - \sum_{k=1}^{K} |b_l^{(k)}| \right)^{1/2} - \left(\sum_{k=1}^{K} |a_l^{(k)}| \right)^{1/2} \right|,$$

and

$$\left| \left(\sum_{k=1}^{K} |a_l^{(k)}| + \sum_{k=1}^{K} |b_l^{(k)}| \right)^{1/2} - \left(\sum_{k=1}^{K} |a_l^{(k)}| \right)^{1/2} \right|.$$

According to (B.2), the maximum is less than or equal to

$$\frac{1}{2} \frac{\sum_{k=1}^{K} |b_l^{(k)}|}{\left(\sum_{k=1}^{K} |a_l^{(k)}| - \sum_{k=1}^{K} |b_l^{(k)}| \right)^{1/2}},$$

as desired. □

We are now ready to establish the equivalence between the objective functions (2.4) and (2.5), for which it suffices to show the equivalence of (2.5) and (B.1), as we do next.

Lemma B.4. *If* $(\hat{\mathbf{H}}, \hat{\mathbf{\Gamma}})$ *is a local minimizer of* (B.1)*, then there exists a local minimizer* $\hat{\mathbf{\Theta}}$ *of* (2.5) *such that* $\hat{\mathbf{\Theta}}^{(k)} = \hat{\mathbf{H}} \circ \hat{\mathbf{\Gamma}}^{(k)}$ *for all* k*. Conversely, if* $\hat{\mathbf{\Theta}}$ *is a local minimizer of* (2.5)*, then there exists a local minimizer* $(\hat{\mathbf{H}}, \hat{\mathbf{\Gamma}})$ *of* (B.1) *such that* $\hat{\mathbf{H}} \circ \hat{\mathbf{\Gamma}}^{(k)} = \hat{\mathbf{\Theta}}^{(k)}$ *for all* k*.*

Proof. Let $Q_3(\lambda_1\lambda_2, \mathbf{\Theta})$ denote the objective function (2.5), and let also $\lambda = \lambda_1\lambda_2$. Suppose $(\hat{\mathbf{H}}, \hat{\mathbf{\Gamma}})$ is a local minimizer of (B.1). Then, there exists $\delta > 0$ such that for

all $(\mathbf{H}, \mathbf{\Gamma})$ with

$$\|\mathbf{H} - \hat{\mathbf{H}}\|_1 + \sum_{k=1}^{K} \|\mathbf{\Gamma}^{(k)} - \hat{\mathbf{\Gamma}}^{(k)}\|_1 < \delta,$$

we have

$$Q_2(\lambda, \hat{\mathbf{H}}, \hat{\mathbf{\Gamma}}) \leq Q_2(\lambda, \mathbf{H}, \mathbf{\Gamma}).$$

Let $\hat{\mathbf{\Theta}}$ be the estimator associated with $(\hat{\mathbf{H}}, \hat{\mathbf{\Gamma}})$, that is, $\hat{\mathbf{\Theta}}^{(k)} = \hat{\mathbf{H}} \circ \hat{\mathbf{\Gamma}}^{(k)}$ for all k. Before proceeding further, we need to define some constants. Let

$$\begin{aligned}
a &= \frac{1}{K} \min \left\{ |\hat{\theta}_{lj}(k)| : \hat{\theta}_{lj}(k) \neq 0, l, j \in \mathcal{V}, k = 1, \ldots, K \right\}, \\
b &= \max \left\{ |\hat{\theta}_{lj}(k)| : l, j \in \mathcal{V}, k = 1, \ldots, K \right\}, \\
c &= 2p^2 \max \left\{ \lambda^{1/2}, \lambda^{-1/2}, \left(\frac{\lambda}{2Ka} \right)^{1/2}, \left(\frac{2}{\lambda Ka} \right)^{1/2} + \left(\frac{b}{\lambda K^2 a^2} \right)^{1/2} \right\}.
\end{aligned}$$

Let also $0 < \delta^* \leq \min\left(\frac{Ka}{2}, a, 1, c^{-2}\delta^2\right)$ and $\mathbf{D} = \left(\mathbf{D}^{(1)}, \ldots, \mathbf{D}^{(K)}\right)$, where $\mathbf{D}^{(k)} = (d_{lj}^{(k)})$ and satisfies $0 < |d_l^{(k)}|$ for all l, k and $\sum_{k=1}^{K} \|\mathbf{D}^{(k)}\|_1 < \delta^*$. Let $\tilde{\mathbf{\Theta}} = \hat{\mathbf{\Theta}} + \mathbf{D}$. Then

$$\sum_{k=1}^{K} \|\tilde{\mathbf{\Theta}}^{(k)} - \hat{\mathbf{\Theta}}^{(k)}\|_1 < \delta^*.$$

This means that $\tilde{\mathbf{\Theta}}$ is a generic element of the ball with radius δ^* and center $\hat{\mathbf{\Theta}}$.

Define $(\tilde{\mathbf{H}}, \tilde{\mathbf{\Gamma}})$ by

$$\tilde{\eta}_{lj} = \left(\lambda \sum_{k=1}^{K} |\hat{\theta}_{lj} + d_{lj}^{(k)}| \right)^{1/2} \quad \text{and} \quad \tilde{\gamma}_{lj}^{(k)} = \frac{\hat{\theta}_{lj}(k) + d_{lj}^{(k)}}{\left(\lambda \sum_{k=1}^{K} |\hat{\theta}_{lj}(k) + d_{lj}^{(k)}| \right)^{1/2}}$$

for all l, j, k. By Lemma B.2,

$$Q_2(\lambda, \hat{\mathbf{H}}, \hat{\mathbf{\Gamma}}) = Q_3(\lambda, \hat{\mathbf{\Theta}}),$$

and by the definition of $(\tilde{\mathbf{H}}, \tilde{\mathbf{\Gamma}})$,

$$Q_2(\lambda, \tilde{\mathbf{H}}, \tilde{\mathbf{\Gamma}}) = Q_3(\lambda, \tilde{\mathbf{\Theta}}).$$

If $\hat{\eta}_{lj} = 0$, then $\hat{\gamma}_{lj}^{(k)} = \hat{\theta}_{lj}(k) = 0$ for all k. Hence

$$|\tilde{\eta}_{lj} - \hat{\eta}_{lj}| = \left(\lambda \sum_{k=1}^{K} |d_{lj}^{(k)}|\right)^{1/2} \leq \left(\lambda \sum_{k=1}^{K} \|\mathbf{D}^{(k)}\|_1\right)^{1/2} < \lambda^{1/2}\delta^{*1/2}$$

and

$$\sum_{k=1}^{K} |\tilde{\gamma}_{lj}^{(k)} - \hat{\gamma}_{lj}^{(k)}| = \frac{\sum_{k=1}^{K} |d_{lj}^{(k)}|}{\left(\lambda \sum_{k=1}^{K} |d_{lj}^{(k)}|\right)^{1/2}} \leq \lambda^{-1/2} \left(\sum_{k=1}^{K} \|\mathbf{D}^{(k)}\|_1\right)^{1/2} < \lambda^{-1/2}\delta^{*1/2}$$

If $\hat{\eta}_{lj} \neq 0$, then by Lemma B.3,

$$\begin{aligned} |\tilde{\eta}_{lj} - \hat{\eta}_{lj}| &\leq \frac{\lambda^{1/2}}{2} \frac{\sum_{k=1}^{K} |d_{lj}^{(k)}|}{\left(\sum_{k=1}^{K} |\hat{\theta}_{lj}(k)| - \sum_{k=1}^{K} |d_{lj}^{(k)}|\right)^{1/2}} \\ &\leq \frac{\lambda^{1/2}}{2K^{1/2}} \frac{\delta^*}{\sqrt{a - \frac{a}{2}}} \leq \left(\frac{\lambda}{2Ka}\right)^{1/2} \delta^{*1/2} \end{aligned}$$

and

$$\sum_{k=1}^{K} |\tilde{\gamma}_{lj}^{(k)} - \hat{\gamma}_{lj}^{(k)}| \leq I_1 + I_2,$$

where

$$I_1 = \lambda^{-1/2}\frac{\sum_{k=1}^{K}|d_{lj}^{(k)}|}{\left(\sum_{k=1}^{K}|\hat{\theta}_{lj}(k)+d_{lj}^{(k)}|\right)^{1/2}} \leq \lambda^{-1/2}\frac{\sum_{k=1}^{K}|d_{lj}^{(k)}|}{\left(\sum_{k=1}^{K}|\hat{\theta}_{lj}(k)|-\sum_{k=1}^{K}|d_{lj}^{(k)}|\right)^{1/2}}$$
$$\leq \lambda^{-1/2}\frac{\delta^*}{K^{1/2}\sqrt{a-\frac{a}{2}}} \leq \left(\frac{2}{\lambda K a}\right)^{1/2}\delta^{*1/2}$$

and

$$I_2 = \lambda^{-1/2}\sum_{k=1}^{K}\left|\frac{\hat{\theta}_{lj}(k)}{\left(\sum_{k=1}^{K}|\hat{\theta}_{lj}(k)+d_{lj}^{(k)}|\right)^{1/2}} - \frac{\hat{\theta}_{lj}(k)}{\left(\sum_{k=1}^{K}|\hat{\theta}_{lj}(k)|\right)^{1/2}}\right|$$
$$= \lambda^{-1/2}\left(\sum_{k=1}^{K}|\hat{\theta}_{lj}(k)|\right)^{1/2}\frac{\left|\left(\sum_{k=1}^{K}|\hat{\theta}_{lj}(k)+d_{lj}^{(k)}|\right)^{1/2}-\left(\sum_{k=1}^{K}|\hat{\theta}_{lj}(k)|\right)^{1/2}\right|}{\left(\sum_{k=1}^{K}|\hat{\theta}_{lj}(k)+d_{lj}^{(k)}|\right)^{1/2}}$$
$$\leq \lambda^{-1/2}\frac{\left(\sum_{k=1}^{K}|\hat{\theta}_{lj}(k)|\right)^{1/2}}{\left(\sum_{k=1}^{K}|\hat{\theta}_{lj}(k)|-\sum_{k=1}^{K}|d_{lj}^{(k)}|\right)^{1/2}}\lambda^{-1/2}|\tilde{\eta}_{lj}-\hat{\eta}_{lj}|$$
$$\leq \lambda^{-1/2}\frac{b^{1/2}}{\sqrt{Ka-\frac{Ka}{2}}}\frac{1}{\sqrt{2Ka}}\delta^{*1/2} \leq \left(\frac{b}{\lambda K^2 a^2}\right)^{1/2}\delta^{*1/2}.$$

Therefore,

$$\|\tilde{\mathbf{H}}-\hat{\mathbf{H}}\|_1 + \sum_{k=1}^{K}\|\tilde{\mathbf{\Gamma}}^{(k)}-\hat{\mathbf{\Gamma}}^{(k)}\|_1 < \delta.$$

Thus,

$$Q_2(\lambda,\hat{\mathbf{H}},\hat{\mathbf{\Gamma}}) \leq Q_2(\lambda,\tilde{\mathbf{H}},\tilde{\mathbf{\Gamma}}) \Rightarrow Q_3(\lambda,\hat{\mathbf{\Theta}}) \leq Q_3(\lambda,\tilde{\mathbf{\Theta}}),$$

which means that $\hat{\mathbf{\Theta}}$ is a local minimizer of (2.5). The other direction can be proved similarly. □

Combining Lemma B.1 and B.4, yields Theorem 4.

B.2 Proof of Proposition 1

Proof. By the Woodbury identity,

$$(\mathbb{X}_{-j}^\intercal \mathbb{X}_{-j} + nb\mathbf{I}_{p-1})^{-1} = \frac{1}{nb}\mathbf{I}_{p-1} - \frac{1}{nb}\mathbb{X}_{-j}^\intercal(\mathbb{X}_{-j}\mathbb{X}_{-j}^\intercal + nb\mathbf{I}_n)^{-1}\mathbb{X}_{-j}.$$

By the Sherman-Morrison identity, we get

$$\begin{aligned}(\mathbb{X}_{-j}\mathbb{X}_{-j}^\intercal + nb\mathbf{I}_n)^{-1} &= (\mathbb{X}\mathbb{X}^\intercal + nb\mathbf{I}_n - \mathbb{X}_j\mathbb{X}_j^\intercal)^{-1} \\ &= (\mathbb{X}\mathbb{X}^\intercal + nb\mathbf{I}_n)^{-1} + \frac{(\mathbb{X}\mathbb{X}^\intercal + nb\mathbf{I}_n)^{-1}\mathbb{X}_j\mathbb{X}_j^\intercal(\mathbb{X}\mathbb{X}^\intercal + nb\mathbf{I}_n)^{-1}}{1 - \mathbb{X}_j^\intercal(\mathbb{X}\mathbb{X}^\intercal + nb\mathbf{I}_n)^{-1}\mathbb{X}_j}.\end{aligned}$$

Combining the two,we have the desired result. □

B.3 Proof of Proposition 2

Proof. The most computationally expensive part of (2.11) is that of computing $\mathbf{v}^{t+1}$, which consists of two components. The first component is

$$(\mathbb{Z}^\intercal\mathbb{Z} + nb\mathbf{I}_{p(p-1)})^{-1}\mathbb{Z}^\intercal\mathbb{Y} = \begin{bmatrix} (\mathbb{X}_{-1}^\intercal\mathbb{X}_{-1} + nb\mathbf{I}_{p-1})^{-1}\mathbb{X}_{-1}^\intercal\mathbb{X}_1 \\ \vdots \\ (\mathbb{X}_{-p}^\intercal\mathbb{X}_{-p} + nb\mathbf{I}_{p-1})^{-1}\mathbb{X}_{-p}^\intercal\mathbb{X}_p \end{bmatrix}.$$

Using proposition 1, it is easy to see that the computational complexity of

$$(\mathbb{X}_{-j}^\intercal\mathbb{X}_{-j} + nb\mathbf{I}_{p-1})^{-1}\mathbb{X}_{-j}^\intercal\mathbb{X}_j$$

is $\mathcal{O}(np)$. Therefore, the computational complexity of

$$(\mathbb{Z}^\intercal\mathbb{Z} + nb\mathbf{I}_{p(p-1)})^{-1}\mathbb{Z}^\intercal\mathbb{Y}$$

is $\mathcal{O}(np^2)$. Similarly, the computational complexity of the second component

$$(\mathbb{Z}^\intercal\mathbb{Z} + nb\mathbf{I}_{p(p-1)})^{-1}nb(\mathbf{r}^t - \mathbf{u}^t)$$

is $\mathcal{O}(np^2)$, which concludes the proof. □

B.4 Proof of Theorem 5

Lemma B.5. *Suppose* $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)^\intercal$ *follows* $\mathcal{N}_n(\mathbf{0}, \sigma^2 \mathbf{I}_n)$. *Then, there exists a constant* $M > 0$ *such that*

$$f(t) = \sup_{\|\boldsymbol{a}\|_2 \leq 1} P\left(|\boldsymbol{a}^\intercal \boldsymbol{\epsilon}| > t\right) \leq 2 \exp\left(-Mt^2\right)$$

Proof. The proof of this result can be found in the supplementary material of Huang et al. (2008). However, for completeness we include it here.

For $d \in [1, +\infty)$, define $\psi_d(x) = \exp(x^d) - 1$. For any random variable $\mathbf{X}$, its ψ_d-Orlicz norm is defined as

$$\|X\|_{\psi_d} = \inf\left\{C \in (0, +\infty) : \mathrm{E}\left[\psi_d\left(\frac{|X|}{C}\right)\right] \leq 1\right\}.$$

Since ϵ_i is Gaussian with mean zero and variance σ^2, by exercise 2.7 of Boucheron et al. (2013) we have

$$P(|\epsilon_i| > t) \leq \exp\left(-\frac{t^2}{2\sigma^2}\right).$$

By Lemma 2.2.1 of Van Der Vaart and Wellner (1996) we have $\|\epsilon_i\|_{\psi_2} \leq 2\sigma$. Let $\boldsymbol{a} \in \mathbb{R}^n$, satisfying $\|\boldsymbol{a}\|_2 \leq 1$. By proposition A.1.6 of Van Der Vaart and Wellner (1996), there exists $K_2 > 0$ which only depends on d such that

$$\|\boldsymbol{a}^\intercal \boldsymbol{\epsilon}\|_{\psi_2} \leq K_2 \left[\mathrm{E}|\boldsymbol{a}^\intercal \boldsymbol{\epsilon}| + \left(\sum_{i=1}^n \|a_i \epsilon_i\|_{\psi_2}^2\right)^{1/2}\right].$$

Since

$$\mathrm{E}(\boldsymbol{a}^\intercal \boldsymbol{\epsilon}) = \mathrm{E}\left(\sum_{i=1}^n a_i^2 \epsilon_i^2 + \sum_{i \neq j} a_i a_j \epsilon_i \epsilon_j\right) = \sum_{i=1}^n a_i^2 \sigma^2 \leq \sigma^2,$$

by Jensen's inequality

$$\mathrm{E}|\boldsymbol{a}^\intercal\boldsymbol{\epsilon}| \leq \left[\mathrm{E}\left(\sum_{i=1}^{n} a_i\epsilon_i\right)^2\right]^{1/2} \leq \sigma.$$

Concerning the second term,

$$\left(\sum_{i=1}^{n}\|a_i\epsilon_i\|_{\psi_2}^2\right)^{1/2} = \left(\sum_{i=1}^{n}|a_i|^2\|\epsilon_i\|_{\psi_2}^2\right)^{1/2} \leq 2\sigma\|\boldsymbol{a}\|_2 \leq 2\sigma.$$

Therefore, $\|\boldsymbol{a}^\intercal\boldsymbol{\epsilon}^{(k)}\|_{\psi_2} \leq K_2(\sigma^2+2\sigma)$. By the definition of $\|X\|_{\psi_2}$ it is true that

$$\mathrm{E}\left[\exp\left(\frac{|X|^2}{\|X\|_{\psi_2}^2}\right)\right] \leq 2.$$

Then, for $t>0$, with the help of Markov's inequality

$$P(|X|>t) = P\left(\exp\left\{\frac{|X|^2}{\|X\|_{\psi_2}^2}\right\} > \exp\left\{\frac{t^2}{\|X\|_{\psi_2}^2}\right\}\right) \leq 2\exp\left(\frac{t^2}{\|X\|_{\psi_2}^2}\right).$$

This means that

$$P(|\boldsymbol{a}^\intercal\boldsymbol{\epsilon}|>t) \leq 2\exp\left(-\frac{t^2}{\|\boldsymbol{a}^\intercal\boldsymbol{\epsilon}\|_{\psi_2}^2}\right) \leq 2\exp\left(-\frac{t^2}{K_2^2(\sigma^2+2\sigma)^2}\right).$$

Define $M = K_2^{-2}(\sigma^2+2\sigma)^{-2}$. Then

$$f(t) = \sup_{\|\boldsymbol{a}\|_2\leq 1} P(|\boldsymbol{a}^\intercal\boldsymbol{\epsilon}|>t) \leq 2\exp\left(-Mt^2\right).$$

□

Lemma B.6. *If Assumption 4 is satisfied, then*

$$P(\|\mathbf{s}_j^{(k)}\|_2 \geq t) \leq q_n \exp\left\{-C\left(\frac{4t^2}{M_1 q_n} - 1\right)\right\}, \quad \textit{for} \quad t \geq \frac{(M_1 q_n)^{1/2}}{2},$$
$$\textit{and} \quad P(\tau_{lj}^{-1} \geq t) \leq \exp\left\{-C\left(\frac{t^2}{4} - M_2\right)\right\}, \quad \textit{for} \quad t \geq 2M_2^{1/2}.$$

Proof. By the definition of $\mathbf{s}_j^{(k)}$,

$$\begin{aligned}
P(\|\mathbf{s}_j^{(k)}\|_2 \geq t) &= P\left(\sum_{l=1}^{q_{nj}^{(k)}} \frac{1}{\sum_{k=1}^K |\tilde{\theta}_{lj}^{(k)}|} \geq 4t^2\right) \\
&\leq \sum_{l=1}^{q_{nj}^{(k)}} P\left(\frac{1}{\sum_{k=1}^K |h_{lj}^{(k)}|} \frac{\sum_{k=1}^K |h_{lj}^{(k)}|}{\sum_{k=1}^K |\tilde{\theta}_{lj}^{(k)}|} \geq \frac{4t^2}{q_{nj}^{(k)}}\right) \\
&\leq \sum_{l=1}^{q_{nj}^{(k)}} P\left(M_1 \frac{\sum_{k=1}^K |h_{lj}^{(k)}|}{\sum_{k=1}^K |\tilde{\theta}_{lj}^{(k)}|} \geq \frac{4t^2}{q_n}\right) \\
&= \sum_{l=1}^{q_{nj}^{(k)}} P\left(\frac{\sum_{k=1}^K |h_{lj}^{(k)}|}{\sum_{k=1}^K |\tilde{\theta}_{lj}^{(k)}|} - 1 \geq \frac{4t^2}{M_1 q_n} - 1\right) \\
&\leq q_n P\left(\max_{\substack{1\leq j\leq p_n \\ 1\leq l\leq p_n - 1}} \left|\frac{\sum_{k=1}^K |h_{lj}^{(k)}|}{\sum_{k=1}^K |\tilde{\theta}_{lj}^{(k)}|} - 1\right| \geq \frac{4t^2}{M_1 q_n} - 1\right) \\
&\leq q_n \exp\left\{-C\left(\frac{4t^2}{M_1 q_n} - 1\right)\right\},
\end{aligned}$$

for $t > (M_1 q_n)^{1/2}/2$.

Also, by the definition of τ_{lj},

$$\begin{aligned}
P(\tau_{lj}^{-1} \geq t) = P(\tau_{lj}^{-2} \geq t^2) &= P\left(4\sum_{k=1}^K |\tilde{\theta}_{lj}^{(k)}| \geq t^2\right) \\
&= P\left(\sum_{k=1}^K |\tilde{\theta}_{lj}^{(k)}| - \sum_{k=1}^K |h_{lj}^{(k)}| \geq \frac{t^2}{4} - \sum_{k=1}^K |h_{lj}^{(k)}|\right)
\end{aligned}$$

$$\leq P\left(\max_{\substack{1\leq j\leq p_n\\1\leq l\leq p_n-1}}\left|\sum_{k=1}^{K}|\tilde{\theta}_{lj}^{(k)}|-\sum_{k=1}^{K}|h_{lj}^{(k)}|\right|\geq\frac{t^2}{4}-M_2\right)$$
$$\leq \exp\left\{-C\left(\frac{t^2}{4}-M_2\right)\right\},$$

for $t\geq 2M_2^{1/2}$. □

We are now ready to prove Theorem 5.

Proof. Since k does not depend on n and the proof is the same for all k, we omit k from this proof. Let $\hat{\boldsymbol{\theta}}_j$ be the unique solution of

$$\underset{\boldsymbol{\theta}_j\in\mathbb{R}^{p_n-1}}{\operatorname{argmin}}\left(\frac{1}{2n}\|\mathbb{X}_j-\mathbb{X}_{-j}\boldsymbol{\theta}_j\|_2^2+\lambda_n\sum_{l=1}^{p_n-1}\tau_{lj}|\theta_{lj}|\right). \tag{B.3}$$

Note that

$$\left\{\hat{\mathcal{E}}=\mathcal{E}\right\}\supseteq\bigcap_{j=1}^{p_n}\left\{\hat{\boldsymbol{\theta}}_j=_s\boldsymbol{\theta}_j\right\}.$$

Therefore,

$$P\left(\hat{\mathcal{E}}\neq\mathcal{E}\right)\leq P\left(\bigcup_{j=1}^{p_n}\left\{\hat{\boldsymbol{\theta}}_j\neq_s\boldsymbol{\theta}_j\right\}\right)\leq\sum_{j=1}^{p_n}P\left(\hat{\boldsymbol{\theta}}_j\neq_s\boldsymbol{\theta}_j\right).$$

Let $\boldsymbol{\theta}_j=\left(\boldsymbol{\theta}_{j;1}^\intercal,\boldsymbol{\theta}_{j;2}^\intercal\right)^\intercal$ and $\hat{\boldsymbol{\theta}}_j=\left(\hat{\boldsymbol{\theta}}_{j;1}^\intercal,\hat{\boldsymbol{\theta}}_{j;2}^\intercal\right)^\intercal$, where $\boldsymbol{\theta}_{j;1},\hat{\boldsymbol{\theta}}_{j;1}\in\mathbb{R}^{q_{nj}}$ and $\boldsymbol{\theta}_{j;2},\hat{\boldsymbol{\theta}}_{j;2}\in\mathbb{R}^{s_{nj}}$. Denote the objective function in (B.3) by $L(\boldsymbol{\theta}_j)$, and $\partial L(\boldsymbol{\theta}_j)/\partial\theta_{lj}$ its subdifferential with respect to the l-th component θ_{lj} (Clarke, 1990). For $l=1,\ldots,p_n-1$, the KKT conditions (Kuhn and Tucker, 2014) are given by

$$0\in\frac{\partial L(\hat{\boldsymbol{\theta}}_j)}{\partial\theta_{lj}}\Leftrightarrow\begin{cases}(\mathbb{X}_{-j})_l^\intercal(\mathbb{X}_j-\mathbb{X}_{-j}\hat{\boldsymbol{\theta}}_j)=n\lambda_n\tau_{lj}\operatorname{sgn}(\hat{\theta}_{lj}), & \hat{\theta}_{lj}\neq 0\\ \left|(\mathbb{X}_{-j})_l^\intercal(\mathbb{X}_j-\mathbb{X}_{-j}\hat{\boldsymbol{\theta}}_j)\right|\leq n\lambda_n\tau_{lj}, & \hat{\theta}_{lj}=0\end{cases}.$$

Thus

$$\hat{\boldsymbol{\theta}}_j =_s \boldsymbol{\theta}_j \Rightarrow \begin{cases} \mathbb{X}_{-j;1}^\intercal[\mathbb{X}_j - \mathbb{X}_{-j;1}\hat{\boldsymbol{\theta}}_{j;1}] = n\lambda_n \mathbf{s}_j \\ \left|(\mathbb{X}_{-j})_l^\intercal[\mathbb{X}_j - \mathbb{X}_{-j;1}\hat{\boldsymbol{\theta}}_{j;1}]\right| \leq n\lambda_n\tau_{lj}, \quad l = q_{nj}+1,\ldots,p_n-1 \end{cases}.$$

Define the matrix

$$\hat{\mathbf{H}}_{jj} = \mathbf{I}_n - n^{-1}\mathbb{X}_{-j;1}\hat{\boldsymbol{\Sigma}}_{jj}^{-1}\mathbb{X}_{-j;1}^\intercal,$$

and let $\boldsymbol{\epsilon} = (\epsilon_1,\ldots,\epsilon_n)^\intercal$ be the vector of errors mentioned in section 2. Then,

$$\begin{aligned} \mathbb{X}_j &= \mathbb{X}_{-j;1}\boldsymbol{\theta}_{j;1} + \boldsymbol{\epsilon} \\ \Rightarrow \mathbb{X}_{-j;1}^\intercal\mathbb{X}_j &= \mathbb{X}_{-j;1}^\intercal\mathbb{X}_{-j;1}\boldsymbol{\theta}_{j;1} + \mathbb{X}_{-j;1}^\intercal\boldsymbol{\epsilon}, \end{aligned}$$

and

$$\hat{\boldsymbol{\theta}}_j =_s \boldsymbol{\theta}_j \Rightarrow \begin{cases} \hat{\boldsymbol{\theta}}_{j;1} = \boldsymbol{\theta}_{j;1} + \frac{1}{n}\hat{\boldsymbol{\Sigma}}_{jj}^{-1}\left(\mathbb{X}_{-j;1}^\intercal\boldsymbol{\epsilon} - n\lambda_n\mathbf{s}_j\right) \\ \left|(\mathbb{X}_{-j})_l^\intercal\left[\hat{\mathbf{H}}_{jj}\boldsymbol{\epsilon} + \lambda_n\mathbb{X}_{-j;1}\hat{\boldsymbol{\Sigma}}_{jj}^{-1}\mathbf{s}_j\right]\right| \leq n\lambda_n\tau_{lj}, \quad l = q_{nj}+1,\ldots,p_n-1 \end{cases}.$$

Let $\mathbf{e}_l$ be the unit vector in the direction of the l-th coordinate. Then,

$$\hat{\theta}_{lj} = \theta_{lj} + \frac{1}{n}\mathbf{e}_l^\intercal\hat{\boldsymbol{\Sigma}}_{jj}^{-1}\left(\mathbb{X}_{-j;1}^\intercal\boldsymbol{\epsilon} - n\lambda_n\mathbf{s}_j\right).$$

For $l = 1,\ldots,q_{nj}$, a necessary condition for $\operatorname{sgn}(\hat{\theta}_{lj}) \neq \operatorname{sgn}(\theta_{lj})$ is

$$|\theta_{lj}| < \frac{1}{n}\left|\mathbf{e}_l^\intercal\hat{\boldsymbol{\Sigma}}_{jj}^{-1}\left(\mathbb{X}_{-j;1}^\intercal\boldsymbol{\epsilon} - n\lambda_n\mathbf{s}_j\right)\right|.$$

For $l = q_{nj}+1,\ldots,p_n-1$, a necessary condition for $\hat{\theta}_{lj} \neq 0$ is

$$\left|(\mathbb{X}_{-j})_l^\intercal\left(\hat{\mathbf{H}}_{jj}\boldsymbol{\epsilon} + \lambda_n\mathbb{X}_{-j;1}\hat{\boldsymbol{\Sigma}}_{jj}^{-1}\mathbf{s}_j\right)\right| > n\lambda_n\tau_{lj}.$$

Therefore,

$$
\begin{aligned}
P(\hat{\boldsymbol{\theta}}_j \neq_s \boldsymbol{\theta}_j) &\leq P\left(\bigcup_{l=1}^{q_{nj}} \left\{ \frac{1}{n} \left| \mathbf{e}_l^\intercal \hat{\boldsymbol{\Sigma}}_{jj}^{-1} \left(\mathbb{X}_{-j;1}^\intercal \boldsymbol{\epsilon} - n\lambda_n \mathbf{s}_j \right) \right| > |\theta_{lj}| \right\} \right) \\
&\quad + P\left(\bigcup_{l=q_{nj}+1}^{p_n-1} \left\{ \left| (\mathbb{X}_{-j})_l^\intercal \left(\hat{\mathbf{H}}_{jj} \boldsymbol{\epsilon} + \lambda_n \mathbb{X}_{-j;1} \hat{\boldsymbol{\Sigma}}_{jj}^{-1} \mathbf{s}_j \right) \right| > n\lambda_n \tau_{lj} \right\} \right) \\
&\leq P(B_1) + P(B_2) + P(B_3) + P(B_4),
\end{aligned}
$$

where

$$
\begin{aligned}
P(B_1) &= P\left(\bigcup_{l=1}^{q_{nj}} \left\{ \frac{1}{n} \left| \mathbf{e}_l^\intercal \hat{\boldsymbol{\Sigma}}_{jj}^{-1} \mathbb{X}_{-j;1}^\intercal \boldsymbol{\epsilon} \right| > \frac{|\theta_{lj}|}{2} \right\} \right) \\
P(B_2) &= P\left(\bigcup_{l=1}^{q_{nj}} \left\{ \lambda_n \left| \mathbf{e}_l^\intercal \hat{\boldsymbol{\Sigma}}_{jj}^{-1} \mathbf{s}_j \right| > \frac{|\theta_{lj}|}{2} \right\} \right) \\
P(B_3) &= P\left(\bigcup_{l=q_{nj}+1}^{p_n-1} \left\{ \left| (\mathbb{X}_{-j})_l^\intercal \hat{\mathbf{H}}_{jj} \boldsymbol{\epsilon} \right| > \frac{n\lambda_n \tau_{lj}}{2} \right\} \right) \\
P(B_4) &= P\left(\bigcup_{l=q_{nj}+1}^{p_n-1} \left\{ \left| (\mathbb{X}_{-j})_l^\intercal \mathbb{X}_{-j;1} \hat{\boldsymbol{\Sigma}}_{jj}^{-1} \mathbf{s}_j \right| > \frac{n\tau_{lj}}{2} \right\} \right)
\end{aligned}
$$

Concerning $P(B_1)$, using Assumption 3 we get

$$
\frac{1}{n^2} \left|\left| \mathbf{e}_l^\intercal \hat{\boldsymbol{\Sigma}}_{jj}^{-1} \mathbb{X}_{-j;1}^\intercal \right|\right|_2^2 = \frac{1}{n} \mathbf{e}_l^\intercal \hat{\boldsymbol{\Sigma}}_{jj}^{-1} \mathbf{e}_l \leq \frac{1}{n} \upsilon_{nj}^{-1} \leq \frac{1}{n} \xi^{-1}.
$$

Therefore,

$$
\left|\left| \left(\frac{\xi}{n} \right)^{1/2} \mathbf{e}_l^\intercal \hat{\boldsymbol{\Sigma}}_{jj}^{-1} \mathbb{X}_{-j;1}^\intercal \right|\right|_2 \leq 1,
$$

and with the help of Lemma B.5,

$$\begin{aligned}
P(B_1) &\leq \sum_{l=1}^{q_{nj}} P\left(\left|\left(\frac{\xi}{n}\right)^{1/2} \mathbf{e}_l^\intercal \hat{\mathbf{\Sigma}}_{jj}^{-1} \mathbb{X}_{-j;1}^\intercal \boldsymbol{\epsilon}\right| \geq \frac{(n\xi)^{1/2}}{2}|\theta_{lj}|\right) \\
&\leq \sum_{l=1}^{q_{nj}} P\left(\left|\left(\frac{\xi}{n}\right)^{1/2} \mathbf{e}_l^\intercal \hat{\mathbf{\Sigma}}_{jj}^{-1} \mathbb{X}_{-j;1}^\intercal \boldsymbol{\epsilon}\right| \geq \frac{(n\xi)^{1/2}}{2} b_n\right) \\
&\leq \sum_{l=1}^{q_{nj}} f(\sqrt{n\xi} b_n/2) \leq q_n f(\sqrt{n\xi} b_n/2) \\
&\leq q_n \exp\left\{-\frac{M\xi}{4} n b_n^2\right\}.
\end{aligned}$$

Concerning $P(B_2)$, by the Cauchy-Schwarz inequality,

$$\lambda_n \left|\mathbf{e}_l^\intercal \hat{\mathbf{\Sigma}}_{jj}^{-1} \mathbf{s}_j\right| \leq \lambda_n \left|\left|\mathbf{e}_l^\intercal \hat{\mathbf{\Sigma}}_{jj}^{-1}\right|\right|_2 \|\mathbf{s}_j\|_2 \leq \frac{\lambda_n \|\mathbf{s}_j\|_2}{\xi}.$$

Thus, by Lemma B.6, with $t = \xi b_n / 2\lambda_n$,

$$P(B_2) \leq P\left(\|\mathbf{s}_j\|_2 \geq \frac{\xi}{2}\frac{b_n}{\lambda_n}\right) \leq q_n \exp\left\{-C\left(\frac{\xi^2}{M_1}\frac{b_n^2}{\lambda_n^2 q_n} - 1\right)\right\}.$$

Note that by Assumption 6, the condition

$$\frac{\xi b_n}{2\lambda_n} > \frac{(M_1 q_n)^{1/2}}{2},$$

is satisfied for sufficiently large n.

Concerning $P(B_3)$, since $\hat{\mathbf{H}}_{jj}$ is a projection, its operator norm is bounded by 1. Thus, by (2.13),

$$\|(\mathbb{X}_{-j})_l^\intercal \hat{\mathbf{H}}_{jj}\|_2 \leq \|(\mathbb{X}_{-j})_l\|_2 = n^{1/2} \Rightarrow \|n^{-1/2}(\mathbb{X}_{-j})_l^\intercal \hat{\mathbf{H}}_{jj}\|_2 \leq 1.$$

Therefore, with the help of Lemma B.5 and Lemma B.6, for sufficiently large n, we

have

$$
\begin{aligned}
P(B_3) &\leq \sum_{l=q_{nj}+1}^{p_n-1} P\left(\left|(\mathbb{X}_{-j})_l^\intercal \hat{\mathbf{H}}_{jj}\boldsymbol{\epsilon}\right| > \frac{n\lambda_n\tau_{lj}}{2}\right) \\
&\leq \sum_{l=q_{nj}+1}^{p_n-1} P(\tau_{lj}^{-1} \geq n^{1/2}) + \sum_{l=q_{nj}+1}^{p_n-1} P\left(\left|(\mathbb{X}_{-j})_l^\intercal \hat{\mathbf{H}}_{jj}\boldsymbol{\epsilon}\right| \geq \frac{n^{1/2}\lambda_n}{2}\right) \\
&= \sum_{l=q_{nj}+1}^{p_n-1} P(\tau_{lj}^{-1} \geq n^{1/2}) + \sum_{l=q_{nj}+1}^{p_n-1} P\left(\left|\frac{1}{n^{1/2}}(\mathbb{X}_{-j})_l^\intercal \hat{\mathbf{H}}_{jj}\boldsymbol{\epsilon}\right| \geq \frac{\lambda_n}{2}\right) \\
&\leq s_n \exp\left\{-C\left(\frac{n}{4} - M_2\right)\right\} + s_n \exp\left\{-\frac{M}{4}\lambda_n^2\right\}.
\end{aligned}
$$

Concerning $P(B_4)$, with the help of Lemma B.6 and Assumption 5, for sufficiently large n, we have,

$$
\begin{aligned}
P(B_4) &\leq \sum_{l=q_{nj}+1}^{p_n-1} P(\tau_{lj}^{-1} \geq n^{1/2}) + \sum_{l=q_{nj}+1}^{p_n-1} P\left(\left|(\mathbb{X}_{-j})_l^\intercal \mathbb{X}_{-j;1}\hat{\boldsymbol{\Sigma}}_{jj}^{-1}\mathbf{s}_j\right| \geq \frac{n^{1/2}}{2}\right) \\
&\leq s_n \exp\left\{-C\left(\frac{n}{4} - M_2\right)\right\} + s_n \exp\left\{-\frac{C}{4}n\right\}.
\end{aligned}
$$

Combining the upper bounds for $P(B_1), P(B_2), P(B_3), P(B_4)$, we have, for sufficiently large n,

$$
\begin{aligned}
\sum_{j=1}^{p_n} P\left(\hat{\boldsymbol{\theta}}_j \neq_s \boldsymbol{\theta}_j\right) \leq\ & \exp\left\{\log(p_n)\left[2 - \frac{M\xi}{4}\frac{nb_n^2}{\log p_n}\right]\right\} \\
&+ \exp\left\{\log(p_n)\left[2 - \frac{C\xi^2}{M_1}\frac{b_n^2}{\lambda_n^2 q_n \log p_n} + \frac{C}{\log p_n}\right]\right\} \\
&+ \exp\left\{\log(p_n)\left[2 - \frac{M}{4}\frac{\lambda_n^2}{\log p_n}\right]\right\} \\
&+ \exp\left\{\log(p_n)\left[2 - \frac{C}{4}\frac{n}{\log p_n}\right]\right\} \\
&+ 2\exp\left\{\log(p_n)\left[2 - \frac{C}{4}\frac{n}{\log p_n} + \frac{CM_2}{\log p_n}\right]\right\},
\end{aligned}
$$

where the right-hand side converges to 0 by Assumption 6. □

Bibliography

Bell, H. E. (1965). Gershgorin's theorem and the zeros of polynomials. *The American Mathematical Monthly*, 72(3):292–295.

Boucheron, S., Lugosi, G., and Massart, P. (2013). *Concentration inequalities: A nonasymptotic theory of independence*. Oxford university press.

Boyd, S., Parikh, N., Chu, E., Peleato, B., Eckstein, J., et al. (2011). Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine learning*, 3(1):1–122.

Boyd, S. and Vandenberghe, L. (2004). *Convex optimization*. Cambridge university press.

Cai, T. T., Li, H., Liu, W., and Xie, J. (2016). Joint estimation of multiple high-dimensional precision matrices. *Statistica Sinica*, 26(2):445.

Candes, E., Tao, T., et al. (2007). The dantzig selector: Statistical estimation when p is much larger than n. *The annals of Statistics*, 35(6):2313–2351.

Clarke, F. H. (1990). *Optimization and nonsmooth analysis*. SIAM.

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

Dempster, A. P. (1972). Covariance selection. *Biometrics*, pages 157–175.

Drton, M. and Perlman, M. D. (2004). Model selection for gaussian concentration graphs. *Biometrika*, 91(3):591–602.

Guo, J., Levina, E., Michailidis, G., and Zhu, J. (2011). Joint estimation of multiple graphical models. *Biometrika*, 98(1):1–15.

Hsing, T. and Eubank, R. (2015). *Theoretical foundations of functional data analysis, with an introduction to linear operators*. John Wiley & Sons.

Huang, J., Ma, S., and Zhang, C.-H. (2008). Adaptive lasso for sparse high-dimensional regression models. *Statistica Sinica*, pages 1603–1618.

Kuhn, H. W. and Tucker, A. W. (2014). Nonlinear programming. In *Traces and emergence of nonlinear programming*, pages 247–258. Springer.

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

Li, B., Kim, M. K., Altman, N., et al. (2010). On dimension folding of matrix-or array-valued statistical objects. *The Annals of Statistics*, 38(2):1094–1121.

Li, B. and Solea, E. (2018). A nonparametric graphical model for functional data with application to brain networks based on fmri. *Journal of the American Statistical Association*, 113(524):1637–1655.

Liu, H., Han, F., Yuan, M., Lafferty, J., Wasserman, L., et al. (2012). High-dimensional semiparametric gaussian copula graphical models. *The Annals of Statistics*, 40(4):2293–2326.

Liu, H., Lafferty, J., and Wasserman, L. (2009). The nonparanormal: Semiparametric estimation of high dimensional undirected graphs. *Journal of Machine Learning Research*, 10(Oct):2295–2328.

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

Ortega, J. M. and Rheinboldt, W. C. (2000). *Iterative solution of nonlinear equations in several variables*. SIAM.

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

Ramsay, J. and Silverman, B. W. (2001). Functional data analysis.

Rocha, G. V., Zhao, P., and Yu, B. (2008). A path following algorithm for sparse pseudo-likelihood inverse covariance estimation (splice). *arXiv preprint arXiv:0807.3734*.

Saegusa, T. and Shojaie, A. (2016). Joint estimation of precision matrices in heterogeneous populations. *Electronic journal of statistics*, 10(1):1341.

Schechter, E. (1996). *Handbook of Analysis and its Foundations*. Academic Press.

Scheinberg, K., Ma, S., and Goldfarb, D. (2010). Sparse inverse covariance selection via alternating linearization methods. *arXiv preprint arXiv:1011.0097*.

Sinoquet, C. (2014). *Probabilistic graphical models for genetics, genomics, and postgenomics*. OUP Oxford.

Solea, E. and Li, B. (2019). Copula gaussian graphical models for functional data.

Spivak, M. (1980). Calculus. houston, tx: Publish or perish.

Van Der Vaart, A. W. and Wellner, J. A. (1996). Weak convergence. In *Weak convergence and empirical processes*, pages 16–28. Springer.

Xue, L., Zou, H., et al. (2012). Regularized rank-based estimation of high-dimensional nonparanormal graphical models. *The Annals of Statistics*, 40(5):2541–2571.

Yuan, M. and Lin, Y. (2007). Model selection and estimation in the gaussian graphical model. *Biometrika*, 94(1):19–35.

Zhang, F. (2006). *The Schur complement and its applications*, volume 4. Springer Science & Business Media.

Zhu, H., Strawn, N., and Dunson, D. B. (2016). Bayesian graphical models for multivariate functional data. *The Journal of Machine Learning Research*, 17(1):7157–7183.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American statistical association*, 101(476):1418–1429.

Zou, H. and Li, R. (2008). One-step sparse estimates in nonconcave penalized likelihood models. *Annals of statistics*, 36(4):1509.

Curriculum Vitae

Ilias Moysidis

Email: iliamous92@gmail.com **Phone:** (718) 564-0756
LinkedIn: https://www.linkedin.com/in/iliasmoysidis
Languages: Greek (Native), English (Fluent)

Education

Pennsylvania State University — **State College, PA**
PhD in Statistics, GPA: 3.7/4.0 — August 2016 – August 2021 (Expected)

Aristotle University of Thessaloniki — **Thessaloniki, Greece**
Bachelor's Diploma in Mathematics, GPA: 8.7/10 — September 2010 – November 2014

Professional Experience

Pennsylvania State University — **State College, PA**
Instructor — January 2019 – May 2019
Teaching Assistant — August 2016 – December 2020
Research Assistant — January 2021 – May 2021

Hellenic Army — **Athens, Thessaloniki, Greece, Larnaka, Cyprus**
Corporal — November 2015 – August 2016

Private Tutor — **Thessaloniki, Greece**
High School and Undergraduate Mathematics — September 2014 – June 2015

Honors and Awards

- Nikolaos Danikas award for outstanding grades in the compulsory courses of the Aristotle University of Thessaloniki Mathematics Department
- Compensatory grant for outstanding grades in the first two years of study in the Aristotle University of Thessaloniki Mathematics Department

Publications

- Joint Functional Graphical Models (submitted, April 23rd, Journal of Machine Learning Research)
- Simultaneous Estimation of Graphical Models by Neighborhood Selection (submitted, September 27th, Journal of the Royal Statistical Society: Series B)

Research Interests

- Graphical Models, Penalized Methods, Variable Selection, Nonsmooth optimization