

Estimation and Inference for High-Dimensional Gaussian Graphical Models with Structural Constraints

by

Jing Ma

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Statistics)
in the University of Michigan
2015

Doctoral Committee:

Professor George Michailidis, Co-Chair
Professor Kerby Shedden, Co-Chair
Professor Bin Nan
Professor Ji Zhu

© Jing Ma 2015
All Rights Reserved

For all the people

ACKNOWLEDGEMENTS

First and foremost, I would like to thank my mentor Professor George Michailidis for guiding me through the years. George is a serious thinker. His attitudes towards statistical methodologies and scientific applications have a great impact on me which has finally directed me towards an academic career. I am very grateful to him for his kindness and patience in helping me with job applications, correcting my English writing, and introducing me to a field that I feel excited about.

I would like to thank Professor Ji Zhu for his insightful comments about my research and presentations, as well as his encouragement through the years. I would also like to thank Professor Kerby Shedden for his extremely helpful instructions on applied statistics and teaching me effective communication of statistics to scientists. I am very grateful to my collaborator Professor Ali Shojaie for his guidance on statistical computation, and Professor Bin Nan for serving on my thesis committee.

Finally, I would also like to express my gratitude to my parents for their love and support; in particular, I would like to thank my best friend David Jones for always being on my side. He gives me confidence and encouragement in both research and life.

TABLE OF CONTENTS

DEDICATION	ii
ACKNOWLEDGEMENTS	iii
LIST OF FIGURES	vi
LIST OF TABLES	viii
LIST OF ABBREVIATIONS	x
ABSTRACT	xii
CHAPTER	
I. Introduction	1
1.1 Gaussian Graphical Models	1
1.1.1 Nodewise Regression	2
1.1.2 Penalized Maximum Likelihood Estimation	3
1.1.3 Covariance Estimation based on Undirected Graph	3
1.1.4 Sparse Partial Correlation Estimation	4
1.1.5 Applications of GGM	5
1.2 Outline	5
II. Network-Based Pathway Enrichment Analysis with Incomplete Network Information	8
2.1 Background	8
2.2 Network Estimation Under External Information Constraints	11
2.3 NetGSA with Estimated Network Information	18
2.3.1 Efficient Estimation of Model Parameters	18
2.3.2 Joint Pathway Enrichment and Differential Network Analysis Using NetGSA	21
2.4 Simulation Results	23
2.5 Applications to Genomics and Metabolomics	28

2.6	Discussion	30
2.7	Software	32
2.8	Proof of Theorem II.4	32
2.9	Proof of Theorem II.8	41
2.10	Derivation for Newton's Method	44
2.11	Additional Simulation Results	47
III. Joint Structural Estimation of Multiple Graphical Models		52
3.1	Background	52
3.2	The Joint Structural Estimation Method	55
3.2.1	An Illustrative Example	56
3.2.2	The General Case	57
3.2.3	Choice of Tuning Parameters	58
3.3	Theoretical Results	59
3.3.1	Estimation Consistency	59
3.3.2	Graph Selection Consistency	62
3.4	Performance Evaluation	65
3.4.1	Simulation Study 1	65
3.4.2	Simulation Study 2	68
3.4.3	Simulation Study 3	72
3.5	Application to Climate Modeling	73
3.6	Discussion	80
3.7	Proof of Theorem III.1	80
3.7.1	Regression	81
3.7.2	Selecting Edge Set	88
3.7.3	Refitting	90
3.8	Proof of Theorem III.2	95
BIBLIOGRAPHY		98

LIST OF FIGURES

Figure

2.1	A graph showing the varying structure of pathways 5–8 from null (left) to alternative (right) in Experiment 2. Dashed lines represent edges that are present in only one condition.	24
3.1	Image plots of the adjacency matrices for all four graphical models. The black color represents presence of an edge. The structured sparsity pattern is encoded in $\mathcal{G} = \{(1, 2), (3, 4), (1, 3), (2, 4)\}$, i.e. each pair of graphical models in \mathcal{G} share a subset of edges.	55
3.2	Simulation study 1: left panel shows the image plot of the adjacency matrix corresponding to the shared structure across all graphs. Each black cell indicates presence of an edge. The right panel shows the receiver operating characteristic (ROC) curves for sample size $n_k = 50$: Graphical Lasso (Glasso) (dotted), Joint Estimation Method - <i>Guo et al.</i> (2011) (JEM-G) (dotdash), Group Graphical Lasso (GGL) (solid), Fused Graphical Lasso (FGL) (dashed), Joint Structural Estimation Method (JSEM) (longdash).	66
3.3	Simulation study 2: image plots of the adjacency matrices from all graphical models. Graphs in the same row share the same connectivity pattern at the bottom right block, whereas graphs in the same column share the same pattern at remaining locations.	68
3.4	Simulation study 2: ROC curves for sample size $n_k = 100$: Glasso (dotted), JEM-G (dotdash), GGL (solid), FGL (dashed), JSEM (longdash). The misspecification ratio ρ varies from (left to right): 0, 0.2, 0.4 (top row) and 0.6, 0.8, 1 (bottom row).	70
3.5	The selected 27 locations based on climate classification. The solid line separates the south and north of North America and corresponds to latitude 39 N.	76

3.6	Estimated climate networks at the six distinct climate zones using JSEM, with edges shared across all locations solid and differential edges dashed.	77
3.7	Estimated climate networks at the six distinct climate zones using JEM-G, with edges shared across all locations solid and differential edges dashed.	78
3.8	Estimated climate networks at the six distinct climate zones using GGL, with edges shared across all locations solid and differential edges dashed.	79

LIST OF TABLES

Table

2.1	Deviance measures for network estimation in experiment 1 and 2. FPR(%), false positive rate in percentage; FNR(%), false negative rate in percentage; MCC, Matthews correlation coefficient; Fnorm, Frobenius norm loss. The best cases are highlighted in bold.	25
2.2	Powers based on false discovery rate with $q^* = 0.05$ in experiment 1. 0.2/0.8 refer to NetGSA with 20%/80% external information; E refers to NetGSA with the exact networks; T refers to the true power; GSA-s/GSA-c refer to Gene Set Analysis with self-contained/competitive null hypothesis in 1000 permutations, respectively. True powers are highlighted in bold.	26
2.3	Powers based on false discovery rate with $q^* = 0.05$ in experiment 2. 0.2/0.8 refer to NetGSA with 20%/80% external information; E refers to NetGSA with the exact networks; T refers to the true power; GSA-s/GSA-c refer to Gene Set Analysis with self-contained/competitive null hypothesis in 1000 permutations, respectively. True powers are highlighted in bold.	27
2.4	Timings (in seconds) for NetGSA. Density (%) refers to density of studied networks in percentage; Newton’s method refers to NetGSA implemented with Newton’s method; L-BFGS-B refers to NetGSA implemented with the method of <i>Byrd et al.</i> (1995).	28
2.5	p -values for the pathways in the metabolomics data, with false discovery rate correction at $q^* = 0.01$. NetGSA refers to Network-based Gene Set Analysis; GSA-s/GSA-c refer to Gene Set Analysis with self-contained/competitive null hypothesis in 3000 permutations, respectively.	29

2.6	<i>p</i> -values for the pathways in the microarray data, with false discovery rate correction at $q^* = 0.001$. NetGSA refers to Network-based Gene Set Analysis; GSA-s/GSA-c refer to Gene Set Analysis with self-contained/competitive null hypothesis in 3000 permutations, respectively.	30
2.7	Deviance measures for network estimation in experiment 3 and 4. FPR(%), false positive rate in percentage; FNR(%), false negative rate in percentage; MCC, Matthews correlation coefficient; Fnorm, Frobenius norm loss. The best cases are highlighted in bold.	49
2.8	Powers based on false discovery rate with $q^* = 0.05$ in experiment 3. 0.2/0.8 refer to NetGSA with 20%/80% external information; E refers to NetGSA with the exact networks; T refers to the true power; GSA-s/GSA-c refer to Gene Set Analysis with self-contained/competitive null hypothesis in 1000 permutations, respectively. True powers are highlighted in bold.	50
2.9	Powers based on false discovery rate with $q^* = 0.05$ in experiment 4. 0.2/0.8 refer to NetGSA with 20%/80% external information; E refers to NetGSA with the exact networks; T refers to the true power; GSA-s/GSA-c refer to Gene Set Analysis with self-contained/competitive null hypothesis in 1000 permutations, respectively. True powers are highlighted in bold.	51
3.1	Performance of different regularization methods for estimating graphical models in Simulation Study 1: average FP, FN, SHD, F1 and FL (SE) for sample size $n_k = 50$. The best cases are highlighted in bold.	68
3.2	Performance of different regularization methods for estimating graphical models in Simulation Study 2: average FP, FN, SHD, F1 and FL (SE) for sample size $n_k = 100$. The best cases are highlighted in bold.	71
3.3	Performance of JSEM and thresholded JSEM with misspecified groups ($\rho = 0.3$): average FP, FN, SHD, F1 and FL (SE) for sample size $n_k = 200$. The better cases are highlighted in bold.	72

LIST OF ABBREVIATIONS

AER	average aerosol optical depth
AUC	area under the curve
CLD	cloud cover
CH₄	methane
CO	carbon monoxide
CO₂	carbon dioxide
DTR	diurnal temperature range
FGL	Fused Graphical Lasso
FRS	frost days
GGL	Group Graphical Lasso
GGM	Gaussian graphical models
Glasso	Graphical Lasso
H₂	hydrogen
JEM-G	Joint Estimation Method - <i>Guo et al.</i> (2011)
JGL	Joint Graphical Lasso
JSEM	Joint Structural Estimation Method
NetGSA	Network-based Gene Set Analysis
PET	potential evapotranspiration
PRE	precipitation
ROC	receiver operating characteristic

SOL solar radiation

TMN minimum temperature

TMP mean temperature

TMX maximum temperature

VAP vapor pressure

WET rainday counts

ABSTRACT

Estimation and Inference for High-Dimensional Gaussian Graphical Models with Structural Constraints

by

Jing Ma

Co-Chairs: Professor George Michailidis and Professor Kerby Shedden

This work discusses several aspects of estimation and inference for high-dimensional Gaussian graphical models and consists of two main parts.

The first part considers network-based pathway enrichment analysis based on incomplete network information. Pathway enrichment analysis has become a key tool for biomedical researchers to gain insight into the underlying biology of differentially expressed genes, proteins and metabolites. We propose a constrained network estimation framework that combines network estimation based on cell- and condition-specific high-dimensional Omics data with interaction information from existing databases. The resulting pathway topology information is subsequently used to provide a framework for simultaneous testing of differences in expression levels of pathway members, as well as their interactions. We study the asymptotic properties of the proposed network estimator and the test for pathway enrichment, and investigate its small sample performance in simulated experiments and illustrate it on two cancer data sets.

The second part of the thesis is devoted to reconstructing multiple graphical mod-

els simultaneously from high-dimensional data. We develop methodology that *jointly* estimates multiple Gaussian graphical models, assuming that there exists prior information on how they are structurally related. The proposed method consists of two steps: in the first one, we employ neighborhood selection to obtain estimated edge sets of the graphs using a group lasso penalty. In the second step, we estimate the nonzero entries in the inverse covariance matrices by maximizing the corresponding Gaussian likelihood. We establish the consistency of the proposed method for sparse high-dimensional Gaussian graphical models and illustrate its performance using simulation experiments. An application to a climate data set is also discussed.

CHAPTER I

Introduction

1.1 Gaussian Graphical Models

Graphical models are probabilistic ones that capture conditional dependence relationships between a set of random variables. Specifically, the random variables are represented by the nodes of a graph, while its edges reflect the relationships amongst themselves. An important class of such models is the Gaussian one, where the random variables are assumed to be jointly normally distributed. For this model, conditional independence relationships between variables are captured through the zero entries of the inverse covariance matrix (or precision matrix). Specifically, let X be a p -dimensional multivariate normal random vector where

$$X = (X_1, \dots, X_p) \sim \mathcal{N}(\mu, \Sigma).$$

For $1 \leq i \neq j \leq p$, X_i and X_j is said to be conditionally independent given all the remaining variables if the corresponding entry in the precision matrix $\Omega = \Sigma^{-1}$ is zero. Denote by $\mathcal{G} = (V, E)$ the underlying graph. An edge between the nodes X_i and X_j in the graph implies that they are conditionally dependent, and corresponds to a non-zero entry in the precision matrix. To identify the graph, one only needs to select the corresponding inverse covariance matrix.

Earlier work on the problem includes *Dempster* (1972a), where thresholding to zero of the elements of the precision matrix is employed in a low-dimensional setting, thus achieving a balance between the fit and the cost. When the number of variables p is relatively small, *Drton and Perlman* (2004) suggest pairwise hypothesis testing of the partial correlation to select a model with conservative overall confidence level. Their approach requires the sample covariance matrix to be positive definite and is not appropriate in high-dimensional settings where the number of variables p is much larger than the number of observations n . More recently, there has been a large amount of work on estimating Gaussian graphical models (GGM) from high-dimensional data subject to sparsity constraints, an attractive feature that reduces the number of parameters to be estimated and produce more interpretable results.

1.1.1 Nodewise Regression

Meinshausen and Bühlmann (2006) introduced a penalized regression model to estimate the skeleton (edge set) of the underlying graph. Specifically, for each node $i = 1, \dots, p$ in the graphical model, consider the optimal prediction of the random variable X_i as a linear combination of the remaining variables:

$$\boldsymbol{\theta}_i = \arg \min_{\boldsymbol{\theta}_i \in \mathbb{R}^p: \theta_{ii}=0} \mathbb{E} \left(X_i - \sum_{j \neq i} \theta_{ij} X_j \right)^2,$$

where θ_{ij} ($j \neq i$) are the regression coefficients. The matrix (θ_{ij}) is determined by the inverse covariance matrix $\Omega = (\omega_{ij})$. Specifically, it holds that $\theta_{ij} = -\omega_{ij}/\omega_{ii}$, for all $j \neq i$. The set of nonzero coefficients of $\boldsymbol{\theta}_i$ is thus the same as the set of nonzero entries in the row vector of ω_{ij} ($j \neq i$), which defines the set of neighbors of node i . Using an l_1 -penalized regression, the authors estimated the neighborhood for each node and combined the estimates to obtain the underlying graph. They further established that the nodewise regression approach yields consistent estimation of the

skeleton (edge set) of sparse high-dimensional graphs, under the regime $p = O(n^\alpha)$ and the neighborhood stability condition.

1.1.2 Penalized Maximum Likelihood Estimation

Work on penalized log-likelihood approaches includes *Yuan and Lin (2007)*, *Banerjee et al. (2008)* that employed the following objective function:

$$\min_{\Omega \succ 0} \{ \text{tr}(\hat{\Sigma}\Omega) - \log \det(\Omega) + \lambda \sum_{i \neq j} |\omega_{ij}| \}, \quad (1.1)$$

where $\hat{\Sigma}$ is the empirical covariance matrix and λ the regularization parameter. The l_1 penalty leads to desired sparsity, provided that an appropriate penalty parameter is chosen. *Friedman et al. (2008)* developed a simple and fast algorithm *Graphical lasso*, which uses a block coordinate descent approach to solve (1.1).

Parallel to algorithmic work there has been a large body of theoretical work establishing norm consistency and model selection consistency properties of the proposed estimator. For the solution $\hat{\Omega}$ to the problem (1.1), *Rothman et al. (2008)* established that its convergence rate in the Frobenius norm is $O(\sqrt{\|\Omega^-\|_0 \log p/n})$ for appropriately chosen λ , where $\|\Omega^-\|_0$ represents the number of non-zero off-diagonal entries in Ω . *Raskutti et al. (2009)* studied sufficient conditions for model selection consistency, i.e. the l_1 -regularized Gaussian maximum likelihood estimator of (1.1) recovers the edge set of the underlying graph with high probability, under the incoherence condition on the Fisher information of the model.

1.1.3 Covariance Estimation based on Undirected Graph

The work by *Zhou et al. (2011)* combines the nodewise regression approach with the idea of thresholding and maximum likelihood refitting to estimate the covariance matrix and its inverse. The proposed method consists of the following two steps:

- Infer the edge set \hat{E} through the regression coefficients $\hat{\theta}_{ij}$, where $\hat{\theta}_{ij}$ are estimated using the threshold lasso algorithm (Zhou, 2010).
- Refit the model via maximum likelihood

$$\min_{\Omega > 0} \{ \text{tr}(\hat{\Sigma}\Omega) - \log \det(\Omega),$$

subject to the constraints in \hat{E} .

It is argued that the first step requires a much weaker restricted eigenvalue condition (Bickel *et al.*, 2009) for consistent recovery of the edge set, i.e. the conditional dependency relationships among variables, compared to the neighborhood stability condition (Meinshausen and Bühlmann, 2006). The proposed method is further shown to yield fast convergence rates with respect to the operator and Frobenius norm for the covariance matrix and its inverse.

1.1.4 Sparse Partial Correlation Estimation

In the case of Gaussian graphical models, the partial correlation ρ_{ij} between node i and j is $\rho_{ij} = -\omega_{ij}/\sqrt{\omega_{ii}\omega_{jj}}$. Thus, ρ_{ij} is nonzero if and only if ω_{ij} is nonzero, or equivalently, node i and j are conditionally dependent given all the remaining ones. Moreover, the partial correlation coefficient quantifies the correlation/interaction between two variables while conditioning on others. Peng *et al.* (2009) introduced SPACE that directly estimates the partial correlations by taking into account the symmetric nature of the problem. Their approach aims at solving the following optimization

$$\min_{\Omega} \left\{ \frac{1}{2} \sum_{i=1}^p \|\mathbf{X}_i - \sum_{j \neq i} \rho_{ij} \sqrt{\frac{\omega_{jj}}{\omega_{ii}}} \mathbf{X}_j\|_2^2 + \lambda \sum_{1 \leq i < j \leq p} |\rho_{ij}| \right\},$$

which, after proper rearrangement of variables, becomes the ℓ_1 regularized lasso problem.

1.1.5 Applications of GGM

Gaussian graphical models have found applications in diverse fields including analysis of Omics data (*Perroud et al.*, 2006; *Pujana et al.*, 2007; *Putluri et al.*, 2011), as well as reconstruction of gene regulatory networks (*Wille et al.*, 2004; *Dehmer and Emmert-Streib*, 2008, chapter 6).

An interesting and important application that requires the knowledge of the underlying network is Network-based Gene Set Analysis (NetGSA) (*Shojaie and Michailidis*, 2009, 2010). In biomedical research, a pathway is defined as a set of functionally related genes, proteins or metabolites. Pathway enrichment analysis has become a key tool for biomedical researchers to gain insight into the underlying biology of differentially expressed genes, proteins and metabolites. It reduces complexity and provides a system-level view of changes in cellular activity in response to treatments and/or progression of disease states. Methods that use pathway network information have been shown to outperform simpler methods that only take into account pathway membership. However, despite significant progress in understanding the association amongst members of biological pathways, and expansion of data bases containing information about interactions of biomolecules, the existing network information may be incomplete or inaccurate, and is not cell-type or disease condition-specific. The work in Chapter II allows researchers to perform pathway enrichment analysis based on the incomplete biomolecular interactions in databases.

1.2 Outline

Chapter II discusses network-based pathway enrichment analysis with incomplete network information. We propose a method that explicitly incorporates external structural information, available in carefully curated biological databases, in the widely used Gaussian graphical model. The resulting estimates are then incorpo-

rated into Network-based Gene Set Analysis (NetGSA), which provides a rigorous statistical framework for simultaneous testing of differences in expression levels of pathway members as well as their interactions, sometimes referred to as differential network biology (*Ideker and Krogan, 2012*).

In the second part, we consider the computational aspect of NetGSA. The main bottleneck in applying the NetGSA methodology arises from the estimation of mixed effects linear parameters – specifically the variance components – for thousands of variables. We develop efficient computational methods for estimation of these parameters based on a profile likelihood approach and thus allow researchers to tackle much larger scale problems, involving thousands of genes as opposed to a few hundred that was the case with the previously available algorithm.

Chapter III studies joint structural estimation of multiple graphical models, motivated from an important application in biomedical research. For example, gene networks for different subtypes of a certain disease share common patterns; i.e. there are *shared common links*, as well as shared *absence of links* between the models (subtypes’ networks). While separate estimation of individual models without taking the known pattern into consideration ignores the common structure, estimating one single model would mask the differences that could prove critical in understanding subtypes. The available approaches usually assume that all graphical models are *globally* related. However, in many settings different relationships between subsets of the node sets exist between different graphical models; such an application is discussed in Section 3.5. We introduce a method that allows one to specify complex substructures from external knowledge. Using the framework of Gaussian graphical models, we formulate the problem as jointly estimating the dependence relationships between the nodes, encoded in the inverse covariance matrices, subject to the substructure constraints. Theoretical analysis indicates the proposed approach recovers consistently the shared and individual structures with faster convergence rate compared to existing methods,

under some technical conditions. Moreover, a thresholded variation of the proposed estimator outperforms existing methods even when the prior substructures are slightly misspecified.

CHAPTER II

Network-Based Pathway Enrichment Analysis with Incomplete Network Information

2.1 Background

Recent advances in high throughput technologies have transformed biomedical research by enabling comprehensive monitoring of complex biological systems. By profiling the activity of different molecular compartments (genomic, proteomic, metabolomic), one can delineate complex mechanisms that play a key role in biological processes or the development of distinct phenotypes. These technological advances have been accompanied by methodological ones, the most notable being adopting a systems perspective in analyzing such systems. Pathway analysis represents a key component in the analysis process, and has been used successfully in generating new biological hypotheses, as well as in determining whether specific pathways are associated with particular phenotypes. Examples include analysis of pathways involved in initiation and progression of cancer and other complex diseases (*Cui et al.*, 2006; *Wilson et al.*, 2010), discovering novel transcriptional effects and co-regulated genes (*Palomero et al.*, 2006; *Huarte et al.*, 2010; *Green et al.*, 2011), and understanding the basic biological processes in model organisms (*Gottwein et al.*, 2007; *Baur et al.*, 2006; *Houstis et al.*, 2006). See *Huang et al.* (2008) for additional examples of applications.

Pathway analysis methods have evolved since the seminal work by *Subramanian et al.* (2005) that vastly popularized the approach. As pointed out in the review paper by *Khatri et al.* (2012), earlier techniques such as over-representation analysis (*Al-Shahrour et al.*, 2005; *Beißbarth and Speed*, 2004), and gene set analysis (*Subramanian et al.*, 2005; *Efron and Tibshirani*, 2007) treat each pathway as a set of biomolecules. These methods assess whether members of a given pathway have higher than expected levels of activity, either by counting the number of differentially active members, or by also accounting for the relative rankings of pathway members and/or the magnitude of their associations with the phenotype. On the other hand, more recent and statistically powerful methods take into consideration the interactions between the biomolecules. These interactions are increasingly available from carefully curated biological databases, including the Kyoto Encyclopedia of Genes and Genomes (*Kanehisa and Goto*, 2000), Reactome (*Joshi-Tope et al.*, 2003), RegulonDB (*Huerta et al.*, 1998) and BioCarta (*Nishimura*, 2001).

A network topology based method that exhibits superior statistical power in identifying differential activity of pathways was proposed in *Shojaie and Michailidis* (2009, 2010). The Network-based Gene Set Analysis (NetGSA) method also allows testing for potential changes in the network structure under different experimental or disease conditions. However, it requires *a priori* knowledge of interactions of the members of pathways, which despite rapid progress remains highly incomplete and occasionally unreliable (see e.g. *Zaki et al.* (2013) and references therein). Moreover, existing network information often determines molecular interactions in the normal state of the cell, and does not provide any insight into condition/disease-specific alterations in interactions amongst components of biological systems.

On the other hand, increased availability of large sample collections of high-dimensional Omics data (e.g. from The Cancer Genome Atlas, <http://cancergenome.nih.gov/>), coupled with the development of network estimation techniques based on graphical

models (*Lauritzen, 1996*) offers the possibility to validate and complement existing network information, and to obtain estimates of condition-specific molecular interactions in the cell. Such an approach for leveraging existing knowledge to enhance the analysis of low signal-to-noise biological datasets was advocated in *Ideker et al. (2011)*.

The first contribution of this paper is the development of an efficient algorithm for constrained network estimation, together with establishing the consistency of the obtained estimates, as a function of existing network information. Estimation of high dimensional networks subject to hard (or soft) constraints on conditional dependence relationships among random variables represents a canonical problem in the context of graphical models, and the proposed method for addressing this problem is of independent interest. By incorporating the condition specific network estimates from the proposed method into the NetGSA framework we also provide a rigorous statistical framework for assessing alterations in biological pathways, sometimes referred to as differential network biology (*Ideker and Krogan, 2012*).

A second objective of this study is to scale up the NetGSA estimation algorithm to very large size networks. The main bottleneck in applying the NetGSA methodology arises from the estimation of mixed effects linear parameters – specifically the variance components – for thousands of variables. We develop efficient computational methods for estimation of these parameters based on a profile likelihood approach. In particular, we employ a Cholesky factorization of the covariance matrices to speed up matrix inversions, and use it to develop an efficient algorithm based on Newton’s method with backtracking line search (*Boyd and Vandenberghe, 2004, page 487*) for step size selection. To supply reliable starting points for this algorithm, we further develop an approximate method-of-moment-type estimator.

This study is strongly motivated by our work on metabolic profiling of cancer and the identification of enriched pathways. Unlike gene expression data, identification

and measurement of metabolites by mass spectrometry techniques is challenging, resulting in reliable measurements for a few hundred metabolites, and hence incomplete coverage of the underlying biochemical pathways. The small number of metabolites in each pathway, and the incomplete coverage of the metabolites particularly hinders the application of over-representation and gene set analysis methods in this setting. In our experience, only topology-based pathway enrichment techniques, such as NetGSA, are capable of reliably delineating pathway activity, as illustrated in Section 2.5.

The remainder of the paper is organized as follows. Section 2.2 presents network estimation based on a Gaussian graphical model under external information constraints and establishes the consistency of the method, while Section 2.3 discusses scaling up the algorithm for the NetGSA mixed effects linear model to large scale networks. The performance of the developed methodology is evaluated in Section 2.4 and is illustrated on two real data sets in Section 2.5. Section 2.6 concludes the chapter with some discussions. Technical details and additional simulation results are provided in Section 2.8, 2.9, 2.10 and 2.11, respectively.

2.2 Network Estimation Under External Information Constraints

Gaussian graphical models (*Lauritzen, 1996, Chapter 5*) are widely used in biological applications to model the interactions among components of biological systems (*Dehmer and Emmert-Streib, 2008, chapter 6*). Specifically, the partial correlation structure corresponding to a molecular network can be represented by an undirected graph $G = (V, E)$ with V and E being the set of nodes (biomolecules) and edges (interactions), respectively. The edge set E corresponds to the $p \times p$ precision, or inverse covariance, matrix Ω , whose nonzero elements $\omega_{ii'}$ refer to edges between nodes i and i' , and indicate that i and i' are conditionally dependent given all other nodes in

the network. Further, the magnitude of the partial correlation $\mathbf{A}_{ii'} = -\omega_{ii'}/\sqrt{\omega_{ii}\omega_{i'i'}}$ determines the strength (positive or negative) of the conditional association between the respective nodes.

As discussed in Section 2.1, the availability of large collections of samples for different disease states and biological processes together with carefully curated information of biomolecular interactions enables the estimation of network structures within the setting of Gaussian graphical models. However, the presence of this externally given network information provides a novel and unexplored modification of the corresponding network estimation problem. Denote by E^c the set of node pairs not connected in the network, i.e. $\omega_{ii'} = 0$. Then, the external information can be represented by the following two subsets

$$E_1 = \{(i, i') \in E : i \neq i', \omega_{ii'} \neq 0\}, \quad E_0 = \{(i, i') \in E^c : i \neq i', \omega_{ii'} = 0\}.$$

In words, E_1 contains known edges, while E_0 contains node pairs where it is known that no interaction exists between them. The external information available in E_1 does not imply exact knowledge of the magnitude of $\omega_{ii'}$ nor $\mathbf{A}_{ii'}$.

Suppose we observe an $m \times p$ data matrix $\mathbf{Z} = (\mathbf{Z}_1, \dots, \mathbf{Z}_p)$, where each row represents one sample from a p -variate Gaussian distribution $\mathcal{N}(0, \Omega^{-1})$. Our goal is then to estimate the network structure, or equivalently the precision matrix Ω , subject to external information encoded in E_1 and E_0 . It follows immediately that the partial correlation $\mathbf{A} = \mathbf{I}_p - \mathbf{D}^{-1/2}\Omega\mathbf{D}^{-1/2}$, where $\mathbf{D} = \text{diag}(\Omega)$ and \mathbf{I}_p is the p -identity matrix. When $E_1 = E$ and $E_0 = E^c$, the problem becomes that of covariance selection (*Dempster, 1972b*), which has been studied extensively in the literature. However, to the best of our knowledge, the problem of estimating Ω (and the partial correlation matrix \mathbf{A}) when E_1 and E_0 only contain partial information has not been investigated before.

In this section, we assume that the m observations used for estimating condition-specific networks are separate from those used for pathway enrichment analysis (highlighted by the use of \mathbf{Z}_i 's and m to denote the random variables and sample size, respectively). However, our theoretical analysis in the next section indicates that when sample sizes are large enough, network estimation can be performed using the same set of samples used for pathway enrichment. The framework proposed in this section reduces the potential bias in small sample settings, and takes advantage of the additional publicly available samples, in lieu of reliable network information. On the other hand, while this problem is seemingly similar to matrix completion (*Candes and Recht, 2009; Cai et al., 2010*), the two problems are fundamentally different in nature. In particular, the goal of matrix completion is to complete the remaining entries from the partially observed $m \times p$ matrix \mathbf{Z} , under some structural assumptions on \mathbf{Z} , such as low-rankness (*Candes and Recht, 2009*). On the other hand, in the setting of graphical models, the entries of the adjacency matrix are estimated based on observations on the nodes of the graph.

In biological settings, both the structure of the network, as well as strengths of associations may be condition-specific. Therefore, we need to accurately estimate the nonzero entries in Ω to recover both the structure of the network and the strength of associations between nodes. In the absence of any external information, the ℓ_1 -penalized negative log-likelihood estimate of Ω is obtained by solving

$$\arg \min_{\Omega \succ 0} \left\{ \text{tr}(\Omega \hat{\Sigma}) - \log \det \Omega + \lambda \|\Omega\|_1 \right\}, \quad (2.1)$$

wherein $\hat{\Sigma}$ is the empirical covariance matrix of the data, $\|\Omega\|_1 = \sum_{i \neq i'} |\omega_{ii'}|$ denotes the ℓ_1 norm of the parameters, and λ is the regularization parameter. In the presence of external information, the problem can be cast as the following constrained

optimization one

$$\min_{\Omega \succ 0} \left\{ \text{tr}(\Omega \hat{\Sigma}) - \log \det \Omega \right\}, \quad (2.2)$$

subject to

$$\sum_{i \neq i', (i, i') \notin E_0 \cup E_1} |\omega_{ii'}| \leq t, \quad \omega_{ii'} = 0, (i, i') \in E_0, \quad \omega_{ii'} \neq 0, (i, i') \in E_1.$$

In the following, we present a two-step procedure to solve the constrained optimization problem (2.2). The proposed approach combines the neighborhood selection (*Meinshausen and Bühlmann, 2006*) with constrained maximum likelihood estimation. It exploits the fact that the estimated neighbors of each node using neighborhood selection coincide with the nonzero entries of the inverse covariance matrix (*Friedman et al., 2008*). Specifically, in neighborhood selection (*Meinshausen and Bühlmann, 2006*), the structure of the network is estimated by finding the optimal set of predictors when regressing the random variable \mathbf{Z}_i corresponding to node $i \in V$ on all other variables, using an l_1 -penalized linear regression. The coefficients for this optimal prediction $\boldsymbol{\theta}_i$ are closely related to the entries of the inverse covariance matrix: for all $i' \neq i$, $\theta_{ii'} = -\omega_{ii'}/\omega_{ii}$. The set of nonzero coefficients of $\boldsymbol{\theta}_i$ is thus the same as the set of nonzero entries in the row vector of $\omega_{ii'}$ ($i' \neq i$), which defines the set of neighbors of node i .

Let J_1^i and J_0^i denote the set of (potential) neighbors of node i for which external information is available: J_1^i is the set of nodes which are known to be in the neighborhood of i , and J_0^i is the set of nodes which are known to be not connected to i . Let \mathbf{Z}_{-i} denote the submatrix obtained by removing the i th column of \mathbf{Z} . Assume all columns of \mathbf{Z} are centered and scaled to have norm 1. Denote by \mathcal{S}_+^p the set of all $p \times p$ positive definite matrices and $\mathcal{S}_E^p = \{\Omega \in \mathbb{R}^{p \times p} : \omega_{ii'} = 0, \text{ for all } (i, i') \notin E \text{ where } i \neq i'\}$. The proposed algorithm proceeds in two steps.

(i) Estimate the network structure \hat{E} . For every node i , find

$$\hat{\boldsymbol{\theta}}_i = \arg \min_{\boldsymbol{\theta}_i \in \mathbb{R}^p: \theta_{ii}=0} \frac{1}{m} \|\mathbf{Z}_i - \mathbf{Z}_{-i} \boldsymbol{\theta}_i\|_2^2 + 2\lambda \sum_{i' \neq i} t_{i'} |\theta_{ii'}|, \quad (2.3)$$

where the penalty weights $t_{i'} = 0$, $i' \in J_1^i$; $t_{i'} = \infty$, $i' \in J_0^i$ and $t_{i'} = 1$ elsewhere. An edge (i, i') is estimated if $\hat{\theta}_{ii'} \neq 0$ or $\hat{\theta}_{i'i} \neq 0$.

(ii) Given the structure \hat{E} , estimate the inverse covariance matrix $\hat{\Omega}$ by

$$\hat{\Omega} = \arg \min_{\Omega \in \mathcal{S}_+^p \cap \mathcal{S}_E^p} \left\{ \text{tr}(\hat{\Sigma} \Omega) - \log \det \Omega \right\}. \quad (2.4)$$

Remark II.1. In this algorithm, the first step estimates the coefficients $\boldsymbol{\theta}_i$ for optimal prediction, such that penalization respects the external information constraints. In practice, one can adjust the weights $t_{i'}$ ($i' \neq i$) to allow for uncertainty in the amount of information available regarding the network of interest. The second step focuses on estimation of the magnitude of nonzero entries in the precision matrix Ω , conditional on the estimated network topology. The optimization problems in both steps are convex and can be solved efficiently using existing software.

The proposed estimator enjoys nice theoretical properties under certain regulatory conditions. Before presenting the main result, we introduce some additional notations. Let Σ_0 be the covariance matrix in the true model and $\Omega_0 = \Sigma_0^{-1}$. For $i = 1, \dots, p$, let $s^i = \|\boldsymbol{\theta}_i\|_0 - |J_1^i|$, where $\|\boldsymbol{\theta}_i\|_0 = \#\{i' : \theta_{ii'} \neq 0\}$ is the l_0 norm. Hence, s^i represents the number of nonzero coordinates after excluding the known ones in each regression. Write $s = \max_{i=1, \dots, p} s^i$ and $S = \sum_{i=1}^p \|\boldsymbol{\theta}_i\|_0$. For a subset $J \subset \{1, \dots, p\}$, let \mathbf{Z}_J be the submatrix by removing the columns whose indices are not in J . We make the following assumptions.

Assumption II.2. *There exist $\phi_1, \phi_2 > 0$ such that*

$$0 < \phi_2 \leq \phi_{\min}(\Sigma_0) \leq \phi_{\max}(\Sigma_0) \leq 1/\phi_1 < \infty.$$

And there exists $\varsigma^2 > 0$ such that for all i , $\text{var}(Z_i | Z_{-i}) = 1/\omega_{0,ii} \geq \varsigma^2$.

Assumption II.3. *Let J and \tilde{J} be disjoint subsets of $\{1, \dots, p\}$. Denote by \mathbf{P}_J the projection matrix onto the column space of \mathbf{Z}_J . There exists $\kappa(s) > 0$ such that*

$$\min_{|\tilde{J}| \leq s} \min_{\substack{\boldsymbol{\delta} \in \mathbb{R}^p \\ \|\boldsymbol{\delta}_{\tilde{J}^c}\|_1 \leq 3\|\boldsymbol{\delta}_{\tilde{J}}\|_1}} \frac{\|(\mathbf{I}_p - \mathbf{P}_J)\mathbf{Z}\boldsymbol{\delta}\|_2}{\sqrt{m}\|\boldsymbol{\delta}_{\tilde{J}}\|_2} \geq \kappa(s) > 0. \quad (2.5)$$

Assumption II.2 is a regulatory condition that explicitly excludes singular or near-singular covariance matrices. Assumption II.3 is adapted from the restricted eigenvalue assumptions in *Bickel et al.* (2009) to allow for presence of external information on relevant indices in the subset J . For example, if $J = J_1^i$ and $\tilde{J} = \{1, \dots, p\} \setminus \{\{i\} \cup J\}$, then (2.5) says that the eigenvalues of the projected matrix $(\mathbf{I}_p - \mathbf{P}_J)\mathbf{Z}$ on the restricted set $\{\boldsymbol{\delta} \in \mathbb{R}^p : |\tilde{J}| \leq s, \|\boldsymbol{\delta}_{\tilde{J}^c}\|_1 \leq c_0\|\boldsymbol{\delta}_{\tilde{J}}\|_1\}$ are bounded away from 0.

Let $0 \leq r < 1$ represent the percentage of available external information, which is defined as $(|E_0| + |E_1|)/\{p(p-1)/2\}$. Next, we state our main result.

Theorem II.4. *Suppose Assumption II.2 and Assumption II.3 with $\kappa(2s)$ are satisfied. For constants $c_1 > 4$ and $0 < k_1 < 1$, assume also that*

$$16c_1 \sqrt{\frac{(1-r)S \log(p-rp)}{m}} \leq k_1 \phi_1 \kappa^2(2s), \quad (2.6)$$

where S is the total number of nonzero parameters excluding the diagonal. Consider $\hat{\Omega}$ defined in (2.4). Then, with probability at least $1 - p^{2-c_1^2/8}$, under appropriately

chosen λ , we have

$$\|\hat{\Omega} - \Omega_0\|_2 \leq \|\hat{\Omega} - \Omega_0\|_F = O\left(\sqrt{\frac{S \log(p - rp)}{m}}\right). \quad (2.7)$$

Remark II.5. The convergence rate in (2.7) indicates an improvement of the order of $\{S \log(1 - r)^{-1}/m\}^{1/2}$ in the presence of external information. The assumption in (2.6) is a regulatory condition that ensures the positive definiteness of Ω_0 when restricted to the estimated edge set under the chosen λ . The proof utilizes techniques from *Bickel et al.* (2009) and *Zhou et al.* (2011) and is given in Section 2.8.

Let \mathbf{A}_0 be the partial correlation matrix in the true model, i.e. $\mathbf{A}_0 = \mathbf{I}_p - \mathbf{D}_0^{-1/2} \Omega_0 \mathbf{D}_0^{-1/2}$, where $\mathbf{D}_0 = \text{diag}(\Omega_0)$. The following corollary is an immediate result of Theorem II.4.

Corollary II.6. *Let assumptions in Theorem II.4 be satisfied. Assume further that $S = o(m/\log(p - rp))$. For $\hat{\Omega}$ defined in (2.4), let $\hat{\mathbf{A}}$ be the corresponding partial correlation matrix. Then, with probability at least $1 - p^{2-c_1^2/8}$, under appropriately chosen λ , we have*

$$\|\hat{\mathbf{A}} - \mathbf{A}_0\|_2 = o(1).$$

Remark II.7. The result in Corollary II.6 implies that under certain regulatory conditions, the error in the condition-specific network estimate \hat{A} is negligible. This proves essential for establishing power properties of NetGSA with estimated network information, as shown in the next section. The proof of Corollary II.6 is available in Section 2.8.

The tuning parameter λ in the first step of the proposed algorithm is important in selecting the correct structure of the network, which will further influence the magnitude of the network interactions in the second step. Accurate estimation of these magnitudes are crucial for topology-based pathway enrichment methods. We

propose to select λ via cross validation to minimize the squared prediction error from all p regressions. Specifically, the cross validation score for the i th regression (2.3) is defined as

$$\text{CV}_i(\lambda) = \sum_{j=1}^m \{\mathbf{Z}_{ji} - \mathbf{Z}_{j,-i} \hat{\boldsymbol{\theta}}_i(j)\}^2,$$

where $\hat{\boldsymbol{\theta}}_i(j)$ is the estimated regression coefficient vector after removing the j th sample of $(\mathbf{Z}_i, \mathbf{Z}_{-i})$. We minimize $\text{CV}(\lambda) = \sum_{i=1}^p \text{CV}_i(\lambda)$ to select the optimal λ .

2.3 NetGSA with Estimated Network Information

In this section, we discuss how (condition-specific) estimates of bimolecular interactions from Section 2.2 can be incorporated into the NetGSA framework to obtain a rigorous inference procedure for both pathway enrichment and differential network analysis. To this end, we formally define the NetGSA methodology based on undirected Gaussian graphical models and address estimation of variance parameters in the corresponding mixed linear model framework in Section 2.3.1 and present an updated algorithm that significantly improves computational speed and stability of the method. In Section 2.3.2, we discuss how the constrained-network estimation procedure of Section 2.2 can be combined with the updated estimation procedure of Section 2.3.1 to rigorously infer differential activities of biological pathways, as well as changes in their network structures.

2.3.1 Efficient Estimation of Model Parameters

Consider p genes (proteins/metabolites) whose activity levels across n samples are organized in a $p \times n$ matrix \mathcal{D} . In the framework of NetGSA, the effect of genes (proteins/metabolites) in the network are captured using a latent variable model (*Shojaie and Michailidis, 2010*). Denote by \mathbf{Y} an arbitrary column of the data matrix, and decompose the observed data into signal, \mathbf{X} , plus noise, $\boldsymbol{\varepsilon}$, i.e. $\mathbf{Y} = \mathbf{X} + \boldsymbol{\varepsilon}$.

The latent variable model assumes that the signal \mathbf{X} follows a multivariate normal distribution with partial correlation matrix \mathbf{A} . Decompose the signal as $\mathbf{X} = \Lambda\boldsymbol{\gamma}$ such that $\boldsymbol{\gamma} \sim \mathcal{N}_p(\boldsymbol{\mu}, \sigma_\gamma^2 \mathbf{I}_p)$ and Λ is the lower triangular matrix that satisfies $\Lambda\Lambda^T = (\mathbf{I}_p - \mathbf{A})^{-1}$.

Assume that $\boldsymbol{\gamma}$ and $\boldsymbol{\varepsilon}$ are independent and $\boldsymbol{\varepsilon}$ is also normally distributed; specifically, $\boldsymbol{\varepsilon} \sim \mathcal{N}_p(\mathbf{0}, \sigma_\varepsilon^2 \mathbf{I}_p)$. The NetGSA model can be summarized in vector notation as

$$\mathbf{Y} = \Lambda\boldsymbol{\gamma} + \boldsymbol{\varepsilon}. \quad (2.8)$$

The NetGSA methodology allows for more complex models, including time course observations. For expositional clarity, we present the methodology in the setting of two experimental conditions and consider the general case where $\mathbf{A}^{(k)} \neq \mathbf{A}^{(k')}$. Let $\mathbf{Y}_j^{(k)}$ ($j = 1, \dots, n; k = 1, 2$) be the j th sample in the expression data under condition k (j th column of data matrix \mathcal{D}), with the first n_1 columns of \mathcal{D} corresponding to condition 1 (control) and the remaining $n_2 = n - n_1$ columns to condition 2 (treatment). Denote by $\Lambda^{(k)}$ the influence matrix and $\boldsymbol{\mu}^{(k)}$ the mean vector under condition k . The NetGSA framework considers a latent variable model of the form

$$\begin{aligned} \mathbf{Y}_j^{(1)} &= \Lambda^{(1)}\boldsymbol{\mu}^{(1)} + \Lambda^{(1)}\boldsymbol{\gamma}_j + \boldsymbol{\varepsilon}_j, & (j = 1, \dots, n_1), \\ \mathbf{Y}_j^{(2)} &= \Lambda^{(2)}\boldsymbol{\mu}^{(2)} + \Lambda^{(2)}\boldsymbol{\gamma}_j + \boldsymbol{\varepsilon}_j, & (j = n_1 + 1, \dots, n). \end{aligned}$$

Here, $\boldsymbol{\gamma}_j$ is the vector of (unknown) random effects, and $\boldsymbol{\varepsilon}_j$ is the vector of random errors. They are independent and normally distributed with mean $\mathbf{0}$ and variances $\sigma_\gamma^2 \mathbf{I}_p$ and $\sigma_\varepsilon^2 \mathbf{I}_p$, respectively.

Inference in NetGSA requires estimation of the mean parameters $\boldsymbol{\mu}^{(1)}$ and $\boldsymbol{\mu}^{(2)}$, which depend on estimates of the variance components σ_γ^2 and σ_ε^2 . In practice, the variance components can be estimated via maximum likelihood or restricted maximum likelihood, which can be computationally demanding for large networks. To

ensure stability, the earlier version of the NetGSA considered profiling out one of the variance components and implemented an algorithm from *Byrd et al. (1995)*, which uses a limited-memory modification of the Broyden–Fletcher–Goldfarb–Shanno quasi-Newton method to optimize the profile log-likelihood. However, the above implementation has a few issues. The first issue is its high computational cost due to the inefficient evaluation of matrix inverses and determinants. Moreover, the algorithm from *Byrd et al. (1995)* requires finite values of the objective function within the supplied box constraints, which is often not satisfied, even after the constraints are adjusted to be within a small range of the optimal estimate. This is particularly the case when the underlying networks are large. To extend the applicability of the NetGSA, we consider using Newton’s method for estimating the variance parameters based on the profile log-likelihood (see Section 2.10 for more details) to improve both the computational efficiency and stability. In particular, we make the following two key improvements for implementation of Newton’s method.

First, it is clear that $\text{Var}(\mathbf{Y}_j^{(k)}) = \sigma_\epsilon^2 \{\mathbf{I}_p + \tau \Lambda^{(k)} (\Lambda^{(k)})^T\} = \sigma_\epsilon^2 \Sigma^{(k)}$, where $\tau = \sigma_\gamma^2 / \sigma_\epsilon^2$. Since the profile log-likelihood as well as its gradient and Hessian matrix with respect to τ all depend on $\Sigma^{(k)}$ ($k = 1, 2$) and their inverses, we choose to invert from their Cholesky decompositions $\Sigma^{(k)} = \mathbf{U}^T \mathbf{U}$, where \mathbf{U} is an upper triangular matrix. The inversion of the triangular matrices results in significant speedup and the inverses of the original matrices can then be computed as $(\Sigma^{(k)})^{-1} = (\mathbf{U}^{-1})(\mathbf{U}^{-1})^T$. In the meantime, we also simplify the calculation of determinant of $\Sigma^{(k)}$ since $\det(\Sigma^{(k)}) = \det(\mathbf{U})^2$, which is necessary for evaluating the profile log-likelihood.

Second, quality of the starting point as well as step sizes will both affect convergence of Newton’s method. To select a good starting point, we use a method-of-moment-type estimate of the variance components. Specifically, denote the residuals $\mathbf{R}_j = \mathbf{Y}_j^{(k)} - \Lambda^{(k)} \hat{\boldsymbol{\mu}}^{(k)}$ for $j = 1, \dots, n$, where $\hat{\boldsymbol{\mu}}^{(k)}$ is the estimate of $\boldsymbol{\mu}^{(k)}$. Assume that there is a single variance σ_ϵ^2 that applies to all ϵ_j ($j = 1, \dots, n$) and variances of γ_j

are different. The variance of \mathbf{R}_j can be decomposed as $(\sigma_\gamma^2)_j + \sigma_\epsilon^2$. We then take the minimum of $\text{Var}(\mathbf{R}_j)$ as the estimate of σ_ϵ^2 and average of the remaining variances as the estimate of σ_γ^2 . Their ratio is used as the initial value for τ . The approximation runs very fast and does not add much computational cost to the method. To find the appropriate step sizes, we use backtracking line search as described in *Boyd and Vandenberghe* (2004, page 464).

With the above two modifications, Newton’s method is then implemented to optimize the profile log-likelihood and returns an estimate of τ . Estimates of $\hat{\sigma}_\gamma^2$ and $\hat{\sigma}_\epsilon^2$ follow immediately (see Section 2.10). Once estimates of the variance components are available, one can derive estimates of the mean parameters $\hat{\boldsymbol{\mu}}^{(1)}$ and $\hat{\boldsymbol{\mu}}^{(2)}$ similarly as in *Shojaie and Michailidis* (2009, 2010).

2.3.2 Joint Pathway Enrichment and Differential Network Analysis Using NetGSA

To test for pathway enrichment with NetGSA, let \mathbf{b} be a row binary vector determining the membership of genes in a pre-specified pathway P . *Shojaie and Michailidis* (2009) show that the contrast vector (*Searle, 1971*) $\boldsymbol{\ell} = (-\mathbf{b}\boldsymbol{\Lambda}^{(1)} \cdot \mathbf{b}, \mathbf{b}\boldsymbol{\Lambda}^{(2)} \cdot \mathbf{b})$ – with \cdot denoting the Hadamard product – satisfies the constraint $\mathbf{1}^T \boldsymbol{\ell} = 0$ and tests the enrichment of pathway P . The advantage of this contrast vector is that it isolates influences from nodes outside the pathways of interest. Let $\boldsymbol{\beta}$ be the concatenated vector of means $\boldsymbol{\mu}^{(1)}$ and $\boldsymbol{\mu}^{(2)}$. The null hypothesis of no pathway activity vs the alternative of pathway activation then becomes

$$H_0 : \boldsymbol{\ell}\boldsymbol{\beta} = 0, \quad H_1 : \boldsymbol{\ell}\boldsymbol{\beta} \neq 0. \tag{2.9}$$

This general framework allows for test of pathway enrichment in arbitrary subnetworks, while automatically adjusting for overlap among pathways. In addition, the

above choice of contrast vector $\boldsymbol{\ell}$ accommodates changes in the network structure. Such changes have been found to play a significant role in development and initiation of complex diseases (*Chuang et al.*, 2012), and NetGSA is currently the only method that systematically combines the changes in expression levels and network structures, when testing for pathway enrichment. However, the applicability of the existing NetGSA framework (*Shojaie and Michailidis*, 2009, 2010) is limited by the assumption of known network structure. Here we show that NetGSA with estimated network information provides a valid inference framework for pathway enrichment and differential network analysis.

The significance of individual contrast vectors in (2.9) can be tested using the following Wald test statistic

$$TS = \frac{\boldsymbol{\ell}\hat{\boldsymbol{\beta}}}{\text{SE}(\boldsymbol{\ell}\hat{\boldsymbol{\beta}})}, \quad (2.10)$$

where $\text{SE}(\boldsymbol{\ell}\hat{\boldsymbol{\beta}})$ represents the standard error of $\boldsymbol{\ell}\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\beta}}$ is the estimate of $\boldsymbol{\beta}$. Both $\boldsymbol{\ell}$ and $\text{SE}(\boldsymbol{\ell}\hat{\boldsymbol{\beta}})$ depend on the underlying networks, which are estimated using data from the two experimental conditions. Under the null hypothesis, TS follows approximately a t -distribution whose degrees of freedom can be estimated using the Satterthwaite approximation method (*Shojaie and Michailidis*, 2010).

For $k = 1, 2$, let $\mathbf{Z}^{(k)}$ of dimension $m_k \times p$ be the data matrix under condition k . Denote S_k the number of nonzero off-diagonal entries in the partial correlation matrix $\mathbf{A}_0^{(k)}$ from the true model, and r_k the percentage of external information. We obtain the following result.

Theorem II.8. *Let assumptions in Theorem II.4 be satisfied and $S_k = o(m_k/\log(p - r_k p))$ under each condition k ($k = 1, 2$). Consider the inverse covariance matrices $\hat{\Omega}^{(k)}$ estimated from (2.3) and (2.4) of Section 2.2. Then the test statistic in (2.10) based on the corresponding networks $\hat{\mathbf{A}}^{(k)}$ is an asymptotically most powerful unbiased test for (2.9).*

Remark II.9. Theorem 2.1 of *Shojaie and Michailidis (2010)* says that NetGSA is robust to uncertainty in network information. Specifically, *Shojaie and Michailidis (2010)* show that if the error in network information $\Delta_{\mathbf{A}_0^{(k)}} = \hat{\mathbf{A}}^{(k)} - \mathbf{A}_0^{(k)}$ satisfies $\|\Delta_{\mathbf{A}_0^{(k)}}\|_2 = o_{\mathbb{P}}(1)$, then NetGSA is an asymptotically most powerful unbiased test for (2.9). The result in Theorem II.8 establishes this property for (partially) estimated networks using the consistency of our proposed network estimation procedure in Theorem II.4 and Corollary II.6. A detailed proof can be found in Section 2.8.

2.4 Simulation Results

We present two experiments to demonstrate the performance of the proposed network estimation procedure, as well as its impact on NetGSA. We refer readers to Section 2.11 for additional simulation scenarios – in particular settings with large number of variable p – and discussions.

Our first experiment is based on a undirected network of size $p = 64$. There are 8 subnetworks, each corresponding to a subgraph/pathway of 8 members. Under the null, all subnetworks have the same topology, which was generated from a scale-free random graph, and all nodes have mean expression values 1. To allow for interactions between subnetworks, there is 20% probability for subnetworks to connect to each other. Under the alternative, the proportion of nodes that have mean changes of magnitude 1 is 0%, 40%, 40% and 50% for subnetwork 1–4. The same applies to subnetworks 5–8.

Our second experiment considers a network of size $p = 160$ with a similar design, except that there are 20 members in each subnetwork. Mean expression values for all nodes are the same under the null. Under the alternative, we allow 0%, 40%, 60% and 80% of the nodes to have mean changes of magnitude 0.3 for subnetworks 1–4. Subnetworks 5–8 follow the same pattern. Here an important comparison is to see whether NetGSA is able to detect small but coordinated changes in mean expression

levels.

In both experiments, we also allowed the structures in subnetworks 5–8 under the alternative to differ from their null equivalent by 10% to simultaneously test pathway enrichment and differential network structure. Fig. 2.1 shows the slight modification in the topology for subnetworks 5–8, from the null to the alternative hypothesis in the second experiment.

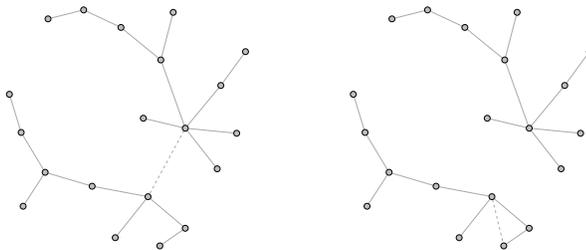


Figure 2.1: A graph showing the varying structure of pathways 5–8 from null (left) to alternative (right) in Experiment 2. Dashed lines represent edges that are present in only one condition.

To illustrate how external information about the network structure facilitates the estimation, we let the percentage of information r vary from 0 to 1. When r is less than 1, we estimated the adjacency matrices using the proposed two-step procedure and filled in the nonzero edges with the estimated weights. When full knowledge of the network topology is given ($r = 1$), one only needs to apply the second step to estimate the edge weights. Table 2.1 compares the estimated networks with the true model under several deviance measures based on 200 replications, with a sample size $m = 40$ for experiment 1 and $m = 100$ for experiment 2. The Matthews correlation coefficient exhibits a clear increasing trend, while the Frobenius norm loss a clear decreasing trend, both indicating the improvement in estimation when the percentage of external information r increases.

Table 2.1: Deviance measures for network estimation in experiment 1 and 2. FPR(%), false positive rate in percentage; FNR(%), false negative rate in percentage; MCC, Matthews correlation coefficient; Fnorm, Frobenius norm loss. The best cases are highlighted in bold.

	r	$p = 64$				$p = 160$			
		FPR(%)	FNR(%)	MCC	Fnorm	FPR(%)	FNR(%)	MCC	Fnorm
Null	0.0	10.80	9.83	0.42	0.76	5.74	1.49	0.41	0.41
	0.2	9.69	10.54	0.44	0.73	5.08	1.50	0.53	0.38
	0.8	3.52	4.72	0.67	0.46	1.77	0.98	0.64	0.22
Alternative	0.0	10.04	8.62	0.44	0.68	4.72	1.01	0.45	0.37
	0.2	9.02	8.37	0.46	0.65	4.15	1.28	0.47	0.35
	0.8	3.13	2.78	0.70	0.40	1.46	0.62	0.68	0.20

Next, we evaluated the performance of NetGSA in detecting pathway enrichment by comparing it with Gene Set Analysis (*Efron and Tibshirani, 2007*), which tests either a competitive or self-contained null hypothesis. While a self-contained null hypothesis permutes the samples and compares the gene set in the pathway with itself, a competitive null hypothesis permutes the genes and compares the set of genes in the pathway with a set of genes not in the pathway. Gene Set Analysis recommends using the competitive null approach to take into consideration the distribution of individual gene set scores, which are used to determine the test statistics.

Table 2.2 and 2.3 present, respectively, the estimated powers for each pathway in the two experiments from 200 replicates, given the differences in mean expression levels and/or subnetwork structures described above. Here we used 16 samples for each condition in experiment 1 and 40 in experiment 2, which are different from the datasets used for network estimation. The powers were calculated as the proportion of replicates that show differential changes, based on the false discovery rate controlling procedure in *Benjamini and Hochberg (1995)* with a q -value of 0.05. For NetGSA, we looked at scenarios when there is 20% and 80% external structural information, and used the estimated networks to detect enrichment for each pathway. We also included the scenario when the exact networks with correct edge weights are provided, in which

Table 2.2: Powers based on false discovery rate with $q^* = 0.05$ in experiment 1. 0.2/0.8 refer to NetGSA with 20%/80% external information; E refers to NetGSA with the exact networks; T refers to the true power; GSA-s/GSA-c refer to Gene Set Analysis with self-contained/competitive null hypothesis in 1000 permutations, respectively. True powers are highlighted in bold.

Pathway	$p = 64$					
	0.2	0.8	E	T	GSA-s	GSA-c
1	0.03	0.03	0.01	0.05	0.07	0.00
2	0.20	0.14	0.04	0.14	0.10	0.00
3	0.50	0.56	0.60	0.78	0.59	0.01
4	0.83	0.75	0.93	0.98	0.70	0.22
5	0.35	0.37	0.08	0.13	0.09	0.00
6	0.49	0.37	0.34	0.47	0.28	0.00
7	0.71	0.71	0.87	0.92	0.67	0.07
8	0.73	0.68	0.85	0.91	0.60	0.07

case only the variance components and mean expression values are estimated from the mixed linear model. True powers for each pathway were calculated when all unknown parameters were substituted with their corresponding known values. As shown in Table 2.2 and 2.3, results from NetGSA with the exact networks agree with the true powers in both experiments, reflecting low powers for pathways 1 and 2, slightly higher powers for 5 and 6 due to change of pathway topology, high powers for 3 and 4 due to change of mean expression levels and highest powers for pathways 7 and 8 for both changes in mean and structure. When the exact networks are unknown, we can still see improvement in estimating powers for pathways 2, and 3 in experiment 1, and for pathways 2, 3, 4, and 5 in experiment 2 as the percentage of external information increases from 20% to 80%. Comparing the estimated powers for the same pathway in the two experiments also confirms our hypothesis that NetGSA is able to identify large changes in only a few genes of the pathway, as well as weak but coordinated changes in the pathway. In contrast, Gene Set Analysis with competitive null approach fails to identify most of the differentially expressed pathways. Gene Set Analysis with self-contained null hypothesis recognizes mostly correctly the pathways

Table 2.3: Powers based on false discovery rate with $q^* = 0.05$ in experiment 2. 0.2/0.8 refer to NetGSA with 20%/80% external information; E refers to NetGSA with the exact networks; T refers to the true power; GSA-s/GSA-c refer to Gene Set Analysis with self-contained/competitive null hypothesis in 1000 permutations, respectively. True powers are highlighted in bold.

Pathway	$p = 160$					
	0.2	0.8	E	T	GSA-s	GSA-c
1	0.01	0.02	0.02	0.05	0.06	0.07
2	0.45	0.35	0.20	0.22	0.34	0.03
3	0.58	0.60	0.72	0.73	0.81	0.00
4	0.64	0.67	0.88	0.94	0.94	0.05
5	0.68	0.55	0.26	0.32	0.12	0.03
6	0.73	0.71	0.62	0.66	0.34	0.01
7	0.79	0.77	0.98	0.99	0.88	0.01
8	0.83	0.81	0.99	1.00	0.95	0.10

that are significantly differentially expressed, although with lower powers for pathways 3, 4, 7 and 8 in experiment 1 and 6 in experiment 2.

Finally, to evaluate the computational efficiency of NetGSA with the updated algorithm based on Newton’s method, we compared it with the earlier version of NetGSA implemented with an algorithm from *Byrd et al. (1995)*. Four different scenarios were considered, including the two experiments described above and another two from Section 2.11. The comparison was based on the average elapsed time of NetGSA in 100 replicates. All timings were carried out under R version 3.0.2 on a Intel Xeon 2.00 GHz processor. Table 2.4 presents the results. In general, we see NetGSA with the updated algorithm runs significantly faster (two times or more) than the previous implementation. The updated implementation is also more stable in terms of evaluating the profile log-likelihood and its gradient, which is especially important when the underlying network is large. In contrast, the earlier version with the method from *Byrd et al. (1995)* failed to run successfully for large p because the gradient of the profile log-likelihood was evaluated to be infinite within the supplied box constraints.

Table 2.4: Timings (in seconds) for NetGSA. Density (%) refers to density of studied networks in percentage; Newton’s method refers to NetGSA implemented with Newton’s method; L-BFGS-B refers to NetGSA implemented with the method of *Byrd et al.* (1995).

p	Density (%)	(1) Newton’s method	(2) L-BFGS-B	Ratio of (2) to (1)
64	3.42	0.23	0.40	1.74
160	1.29	1.87	7.08	3.79
160	5.16	1.53	6.34	4.14
400	1.13	22.61	NA	NA

2.5 Applications to Genomics and Metabolomics

In this section, we discuss applications of the proposed NetGSA to genomic and metabolomic data to demonstrate its potential in revealing biological insights. The metabolomics data set (*Putluri et al.*, 2011) examines changes in the metabolic profile between 58 cancer and adjacent benign tissue specimens through an untargeted mass spectrometry data acquisition strategy. There are two groups of tissue specimens, with 31 samples from the cancer class and 28 from the benign class. The total number of metabolites detected is 63. Here we focused on estimating the network of metabolic interactions, enhanced by information gleaned from the Kyoto Encyclopedia of Genes and Genomes (*Kanehisa and Goto*, 2000). To select the estimated networks for both conditions, we performed 5-fold cross validation. We also tested for differential activity of biochemical pathways extracted from the Kyoto Encyclopedia of Genes and Genomes using the same set of data. Shown in Table 2.5 are estimated p -values after false discovery rate correction with a q -value of 0.01 for the significant pathways selected from NetGSA. These identified pathways include those that describe altered utilization of amino acids and their aromatic counterparts, as well as metabolism of fatty acids and intermediates of tricarboxylic acid cycle (TCA) which were followed up for biological insights in the original study *Putluri et al.* (2011). Among all the selected pathways, fatty acid biosynthesis and phenylalanine,

tyrosine and tryptophan biosynthesis were not identified by Gene Set Analysis with the self-contained null hypothesis. On the other hand, Gene Set Analysis with the competitive null (the recommended setting) failed to report any pathway as being significantly enriched. This again confirms our hypothesis that incorporating pathway topology information allows sophisticated enrichment methods in detecting important regulatory pathways.

Table 2.5: p -values for the pathways in the metabolomics data, with false discovery rate correction at $q^* = 0.01$. NetGSA refers to Network-based Gene Set Analysis; GSA-s/GSA-c refer to Gene Set Analysis with self-contained/competitive null hypothesis in 3000 permutations, respectively.

Pathway	NetGSA	GSA-s	GSA-c
Fatty acid biosynthesis	< 0.001	1.000	1.000
Purine metabolism	0.009	0.009	0.612
Pyrimidine metabolism	0.001	< 0.001	0.395
Glycine, serine and threonine metabolism	< 0.001	0.001	0.672
Tryptophan metabolism	< 0.001	< 0.001	0.338
Phenylalanine, tyrosine and tryptophan biosynthesis	0.002	1.000	1.000
beta-Alanine metabolism	< 0.001	< 0.001	0.338
Aminoacyl-tRNA biosynthesis	0.004	< 0.001	0.458
ABC transporters	0.004	< 0.001	0.624

For the second application, we consider data from *Subramanian et al.* (2005), which consists of gene expression profiles of 5217 genes for 62 normal and 24 lung cancer patients. We excluded genes that are not present in the 186 pathways from the Kyoto Encyclopedia of Genes and Genomes data base as well as those which do not have recorded network information, which leaves us with 1416 genes. We then performed 5-fold cross validation to estimate the underlying interaction networks for both normal and lung cancer conditions based on the external topology information from the BioGRID Database.

To test for pathway enrichment, we considered a subset of pathways from the Kyoto Encyclopedia data base that describe signaling and biochemical mechanisms and restricted their membership to be at least 5, so that Gene Set Analysis could be

applicable. This reduces the number of pathways tested to 61. Table 2.6 presents the p -values for the significant pathways identified from all three methods based on false discovery rate correction at 0.001, sorted with respect to results from using NetGSA. It turns out that Gene Set Analysis does not consider any of the pathways as differentially active, whichever null hypothesis is used. In comparison, the small p -values from NetGSA suggest these 15 pathways could be of interest for further investigation. Of particular biological interest is the identification of the TGF-beta signaling pathway, that has been linked to biological mechanisms for onset and progression of lung cancer (see e.g. *Ischenko et al.* (2014) and references therein).

Table 2.6: p -values for the pathways in the microarray data, with false discovery rate correction at $q^* = 0.001$. NetGSA refers to Network-based Gene Set Analysis; GSA-s/GSA-c refer to Gene Set Analysis with self-contained/competitive null hypothesis in 3000 permutations, respectively.

Pathway	NetGSA	GSA-s	GSA-c
Glycolysis / Gluconeogenesis	< 0.001	0.360	0.520
Citrate cycle (TCA cycle)	< 0.001	0.457	0.357
Fructose and mannose metabolism	< 0.001	0.308	0.408
Galactose metabolism	< 0.001	0.404	0.466
Alanine, aspartate and glutamate metabolism	< 0.001	0.404	0.565
Tyrosine metabolism	< 0.001	0.483	0.441
beta-Alanine metabolism	< 0.001	0.404	1.000
Glutathione metabolism	< 0.001	0.283	0.357
Ether lipid metabolism	< 0.001	0.338	0.379
ErbB signaling pathway	< 0.001	0.035	0.249
TGF-beta signaling pathway	< 0.001	0.404	0.518
VEGF signaling pathway	< 0.001	0.360	0.441
NOD-like receptor signaling pathway	< 0.001	0.308	0.427
RIG-I-like receptor signaling pathway	0.001	0.308	0.357
B cell receptor signaling pathway	< 0.001	0.338	0.441

2.6 Discussion

This chapter introduces a constrained network estimation method for incorporating externally available interaction information based on high-dimensional Omics

data. Under mild assumptions on sample sizes, the proposed approach yields reliable condition-specific estimates for the underlying networks, and can also conveniently accommodate uncertainty in external information. In our simulations, we notice that the proposed two-step procedure is more robust than the one-step constrained maximum likelihood estimation (a functionality offered in the R-package `glasso`) in recovering the partial correlations, because the latter requires sophisticated specification of tuning parameters to satisfy the positive definiteness property of the estimate while taking into consideration the structural constraints.

Another alternative for recovering the underlying network is to use `space` (Peng *et al.*, 2009) that utilizes the symmetric nature of the partial correlation matrix. By incorporating the external structural information, the original lasso problem becomes a generalized lasso and can be solved by existing software.

In the framework of NetGSA, it is recommended that the expression data \mathcal{D} for testing pathway activity and the data \mathbf{Z} for estimating the partial correlation networks are two separate data sets in order to reduce potential bias. It is important to have sufficient samples in \mathbf{Z} for reliable estimation of the underlying networks. The expression data \mathcal{D} can be of a much smaller size compared to \mathbf{Z} . The choice of Λ as the Cholesky factor of the covariance matrix is mainly for interpretation purpose. One can also permute the nodes in the network and obtain similar results on testing for gene set enrichment.

As detailed in *Shojaie and Michailidis* (2010), the NetGSA methodology is a general framework that can be extended to situations where more than two experimental conditions are considered. The underlying networks can be the partial correlations among variables of interest, as discussed in the current context, and can also be the physical interactions among different components of the system. The updated NetGSA algorithm enables pathway enrichment analysis at a much larger scale, significantly enhancing the applicability of the method.

2.7 Software

The proposed method has been implemented in the R-package `netgsa` available on CRAN.

2.8 Proof of Theorem II.4

To prove our main results, we need some additional notations. Define $\tilde{\Omega}_0 = \text{diag}(\Omega_0) + \Omega_{0, E \cap \hat{E}}$, where E and \hat{E} are the true and the estimated edge set, respectively. By definition, $\tilde{\Omega}_0$ and \mathbf{A}_0 will be different at position (i, i') only when the edge (i, i') is falsely rejected. We first derive an upper bound for the size of \hat{E} and $\|\tilde{\Omega}_0 - \Omega_0\|_F$. To do this, we show that the regression problem (2.3) is essentially a lasso problem, and then invoke the oracle inequalities from Theorem 7.2 of *Bickel et al.* (2009). To simplify the notation, we drop the superscript i for sets J_0, J_1 in the i th regression, but they should be understood as J_0^i, J_1^i , respectively.

Let $\tilde{J} = V \setminus \{i\} \cup J_0 \cup J_1$ represent the set of indices for which there is no information available. Denote by $\mathbf{P}_{J_1} = \mathbf{Z}_{J_1}(\mathbf{Z}_{J_1}^T \mathbf{Z}_{J_1})^{-1} \mathbf{Z}_{J_1}^T$ the projection onto the column space of \mathbf{Z}_{J_1} . The following lemma is needed in the proof of Theorem II.11 below.

Lemma II.10. *For $i = 1, \dots, p$, denote $\boldsymbol{\xi}^i = \mathbf{Z}_i - \sum_{i' \neq i} \theta_{ii'} \mathbf{Z}_{i'}$, where $\boldsymbol{\theta}_i$ is the optimal prediction coefficient vector in the i th regression. Consider the event*

$$\mathcal{F}_i := \left\{ \mathbf{Z} : \|\mathbf{Z}_{\tilde{J}}^T (\mathbf{I}_p - \mathbf{P}_{J_1}) \boldsymbol{\xi}^i / m\|_\infty \leq \frac{c_1}{2} \sqrt{\frac{\log(p - rp)}{m \mathbf{A}_{0,ii}}} \right\}$$

with a constant $c_1 > 4$, where $\omega_{0,ii}$ is the i th diagonal element of the true inverse covariance matrix Ω_0 . Define the event $\mathcal{F} = \bigcap_{i=1}^p \mathcal{F}_i$. Then $\mathbb{P}(\mathcal{F}) > 1 - p^{2-c_1^2/8}$.

The proof of Lemma II.10 will be provided shortly. Denote by Λ_{\max} the maximal eigenvalue of $\mathbf{Z}^T \mathbf{Z} / m$. Conditional on event \mathcal{F} , we have the following results on

controlling the size of \hat{E} and the Frobenius norm of the deviance, $\|\tilde{\Omega}_0 - \Omega_0\|_F$.

Theorem II.11. *Suppose all conditions in Theorem II.4 are satisfied. Then on event \mathcal{F} , for appropriately chosen λ , we have*

$$|\hat{E}| \leq \frac{64\Lambda_{\max}}{\kappa^2(s)}(1-r)S + rS, \quad (2.11)$$

and

$$\|\tilde{\Omega}_0 - \Omega_0\|_F \leq c_3 \sqrt{\frac{S \log(p-rp)}{m}} \leq k_1 \phi_1, \quad (2.12)$$

where $c_3 = 16c_1\sqrt{1-r}/\kappa^2(2s)$.

Remark II.12. The result indicates that the cardinality of the estimated edge set is upper bounded by a function of r , the percentage of the external information. The bound for $|\hat{E}|$ also depends on the restricted eigenvalue $\kappa(s)$, which is necessarily positive by the assumption that $\kappa(2s) > 0$. Two extreme cases occur when (i) $r = 0$, i.e. we do not observe any information, thus reducing problem (2.3) to the original neighborhood selection in *Meinshausen and Bühlmann (2006)*; (ii) $r = 1$, i.e. the exact network topology is known and hence $\hat{E} = E$. On the other hand, the upper bound for $\|\tilde{\Omega}_0 - \Omega_0\|_F$ decreases as r increases, i.e. when more external information becomes available. However, since the coefficients also need to be estimated, this deviance always stays positive, even when $r = 1$.

Proof of Theorem II.11. Recall that \mathbf{P}_{J_1} is the projection matrix onto the column space of \mathbf{Z}_{J_1} . Let $\tilde{\mathbf{Y}} = (\mathbf{I}_p - \mathbf{P}_{J_1})\mathbf{Z}_i$ be the projection of \mathbf{Z}_i onto the orthogonal space of \mathbf{Z}_{J_1} and $\tilde{\mathbf{Z}} = (\mathbf{I}_p - \mathbf{P}_{J_1})\mathbf{Z}_{\tilde{J}}$. With some algebra, the problem (2.3) is equivalent to solving

$$\min_{\boldsymbol{\theta}_{\tilde{J}}} \frac{1}{m} \|\tilde{\mathbf{Y}} - \tilde{\mathbf{Z}}\boldsymbol{\theta}_{\tilde{J}}\|_2^2 + 2\lambda \|\boldsymbol{\theta}_{\tilde{J}}\|_1, \quad (2.13)$$

which is a lasso problem. It suffices to focus mainly on the set \tilde{J} , as false positive and negative errors will only occur on this set.

To apply Theorem 7.2 of *Bickel et al. (2009)*, we also need to bound the maximum eigenvalue of the matrix $\tilde{\mathbf{Z}}^T \tilde{\mathbf{Z}}/m$. Consider the eigendecomposition of the projection $\mathbf{I}_p - \mathbf{P}_{J_1} = \mathbf{U}\mathbf{D}\mathbf{U}^T$, where \mathbf{D} is the diagonal matrix composed of eigenvalues and \mathbf{U} is orthogonal. As $\mathbf{I}_p - \mathbf{P}_{J_1}$ is also a projection matrix, the diagonals of \mathbf{D} are either 0 or 1. It then follows that

$$\begin{aligned} \phi_{\max}(\tilde{\mathbf{Z}}^T \tilde{\mathbf{Z}}/m) &= \phi_{\max}(\mathbf{Z}_j^T \mathbf{U}\mathbf{D}\mathbf{U}^T \mathbf{Z}_j/m) \leq \phi_{\max}(\mathbf{Z}_j^T \mathbf{U}\mathbf{U}^T \mathbf{Z}_j/m) \\ &\leq \phi_{\max}(\mathbf{Z}_j^T \mathbf{Z}_j/m) \leq \Lambda_{\max}. \end{aligned}$$

Recall s^i is the number of nonzero coordinates after excluding the known ones in each regression and $s = \max_i s^i$. Under the assumption that $\kappa(2s) > 0$, we also have that $\kappa(s^i) \geq \kappa(s) > 0$ for $s^i \leq s$. Let $\hat{\boldsymbol{\theta}}_{i,\bar{j}}^i$ be the lasso estimator in (2.13) with

$$\lambda = c_1 \left\{ \frac{\log(p - rp)}{m\omega_{0,ii}} \right\}^{1/2} \quad (2.14)$$

for $c_1 > 4$. Conditioned on event \mathcal{F} , we can invoke Theorem 7.2 of *Bickel et al. (2009)* and obtain simultaneously for all i ,

$$\|\hat{\boldsymbol{\theta}}_{i,\bar{j}}\|_0 \leq \frac{64\Lambda_{\max}}{\kappa^2(s)} s^i, \quad (2.15)$$

and

$$\|\hat{\boldsymbol{\theta}}_{i,\bar{j}} - \boldsymbol{\theta}_{i,\bar{j}}\|_2 \leq \frac{16c_1}{\omega_{0,ii}\kappa^2(2s)} \sqrt{\frac{s^i \log(p - rp)}{m}}. \quad (2.16)$$

Combining (2.15) with the number of known edges s_1^i as given in J_1^i , we get

$$|\hat{E}| \leq \sum_{i=1}^p \{ \|\hat{\boldsymbol{\theta}}_{i,\bar{j}}\|_0 + |J_1^i| \} \leq \frac{64\Lambda_{\max}}{\kappa^2(s)} \sum_{i=1}^p s^i + \sum_{i=1}^p s_1^i.$$

The upper bound in (2.11) follows immediately, since by definition the number of known and unknown edges are $\sum_{i=1}^p s_1^i = rS$ and $\sum_{i=1}^p s^i = (1 - r)S$, respectively.

To control $\|\tilde{\Omega}_0 - \Omega_0\|_F$, recall that for every i , $\omega_{0,ii'} = -\theta_{ii'}\omega_{0,ii}$. Using the bound in (2.16), we have

$$\begin{aligned} \|\tilde{\Omega}_0 - \Omega_0\|_F^2 &= \sum_{i=1}^p \sum_{i' \in J(\boldsymbol{\theta}_i) \cap J(\hat{\boldsymbol{\theta}}_i)^c} (\theta_{ii'}\omega_{0,ii})^2 = \sum_{i=1}^p \omega_{0,ii}^2 \sum_{i' \in J(\boldsymbol{\theta}_i) \cap J(\hat{\boldsymbol{\theta}}_i)^c} |\theta_{ii'} - \hat{\theta}_{ii'}|^2 \\ &\leq \sum_{i=1}^p \omega_{0,ii}^2 \|\boldsymbol{\theta}_{i,\tilde{J}} - \hat{\boldsymbol{\theta}}_{i,\tilde{J}}\|_2^2 \leq \left\{ \frac{16c_1}{\kappa^2(2s)} \right\}^2 \frac{(1-r)S \log(p-rp)}{m}. \end{aligned}$$

The last inequality in (2.12) follows from condition (2.6) in Theorem II.4. \square

Proof of Lemma II.10. For every i , it is easy to verify that $\boldsymbol{\xi}^i$ is normally distributed with mean $\mathbf{0}$ and variance $1/\omega_{0,ii}\mathbf{I}_m$. Define random variables $\Upsilon_{ii'} = (\omega_{0,ii}/m)^{1/2}\mathbf{Z}_{i'}^T\boldsymbol{\xi}^i$ for $i \neq i'$. Then, $\mathbf{Z}_{i'}^T\mathbf{Z}_{i'}/m = 1$ implies that $\Upsilon_{ii'} \sim \mathcal{N}(0, 1)$. Let λ be defined as in (2.14). Using the fact that $|\mathbf{Z}_{i'}^T(\mathbf{I}_p - \mathbf{P}_{J_1})\boldsymbol{\xi}^i/m|$ is stochastically smaller than $|\mathbf{Z}_{i'}^T\boldsymbol{\xi}^i/m|$ for all $i' \in \tilde{J}$ and an elementary bound on the tails of Gaussian distributions

$$\begin{aligned} \mathbb{P}(\mathcal{F}^c) &\leq \sum_{i=1}^p \sum_{i' \in \tilde{J}} \mathbb{P}(\{|\mathbf{Z}_{i'}^T(\mathbf{I}_p - \mathbf{P}_{J_1})\boldsymbol{\xi}^i/m| > \lambda/2\}) \\ &\leq \sum_{i=1}^p \sum_{i' \in \tilde{J}} \mathbb{P}(|\Upsilon_{ii'}| > (m\omega_{0,ii})^{1/2}\lambda/2) \leq \sum_{i=1}^p \sum_{i' \in \tilde{J}} \exp\{-m\omega_{0,ii}\lambda^2/8\} \\ &\leq p(p-rp) \exp\{-c_1^2 \log(p-rp)/8\} \leq p^{2-c_1^2/8}. \end{aligned}$$

Therefore, $\mathbb{P}(\mathcal{F}) > 1 - p^{2-c_1^2/8}$. \square

With Lemma II.10 and Theorem II.11, we are ready to prove our main results in Theorem II.4. The following proof is adapted from *Zhou et al.* (2011).

Proof of Theorem II.4. Consider $\hat{\mathbf{A}}$ defined in (2.4). It suffices to show that on the event \mathcal{F}

$$\|\hat{\Omega} - \tilde{\Omega}_0\|_F = O(\{S \log(p-rp)/m\}^{1/2}),$$

since by triangle inequality and Theorem II.11, we can conclude

$$\|\hat{\Omega} - \Omega_0\|_F \leq \|\hat{\Omega} - \tilde{\Omega}_0\|_F + \|\tilde{\Omega}_0 - \Omega_0\|_F \leq O(\{S \log(p - rp)/m\}^{1/2}).$$

Denote $\tilde{\Sigma}_0 = \tilde{\Omega}_0^{-1}$, which is positive definite since by Theorem II.11,

$$\phi_{\min}(\tilde{\Omega}_0) \geq \phi_{\min}(\Omega_0) - \|\tilde{\Omega}_0 - \Omega_0\|_2 \geq \phi_{\min}(\Omega_0) - \|\tilde{\Omega}_0 - \Omega_0\|_F \geq \phi_1 - k_1 \phi_1 > 0. \quad (2.17)$$

Given $\tilde{\Omega}_0 \in \mathcal{S}_+^p \cap \mathcal{S}_{\hat{E}}^p$, define a new convex set:

$$\mathcal{U}_m(\tilde{\Omega}_0) = \{\mathbf{B} - \tilde{\Omega}_0 \mid \mathbf{B} \in \mathcal{S}_+^p \cap \mathcal{S}_{\hat{E}}^p\} \subset \mathcal{S}_{\hat{E}}^p.$$

Let

$$Q(\Omega) = \text{tr}(\Omega \hat{\Sigma}) - \text{tr}(\tilde{\Omega}_0 \hat{\Sigma}) - \log \det \Omega + \log \det \tilde{\Omega}_0.$$

Since the estimate $\hat{\Omega}$ minimizes $Q(\Omega)$, $\hat{\Delta} = \hat{\Omega} - \tilde{\Omega}_0$ minimizes $G(\Delta) = Q(\Delta + \tilde{\Omega}_0)$.

The main idea of this proof is as follows. For a sufficiently large $M > 0$, consider sets

$$\mathcal{T}_1 = \{\Delta \in \mathcal{U}_m(\tilde{\Omega}_0), \|\Delta\|_F = Mr_m\}, \quad \mathcal{T}_2 = \{\Delta \in \mathcal{U}_m(\tilde{\Omega}_0), \|\Delta\|_F \leq Mr_m\},$$

where

$$r_m = \{S \log(p - rp)/m\}^{1/2}.$$

Note that \mathcal{T}_1 is non-empty. Indeed, consider $\mathbf{B}_\epsilon = \epsilon \tilde{\Omega}_0$ for $\epsilon = Mr_m / \|\tilde{\Omega}_0\|_F$. Then $\mathbf{B}_\epsilon = (1 + \epsilon)\tilde{\Omega}_0 - \tilde{\Omega}_0 \in \mathcal{U}_m(\tilde{\Omega}_0)$, hence $\mathbf{B}_\epsilon \in \mathcal{T}_1$. Denote by $\bar{0}$ the matrix of all zero entries. It is clear that $G(\Delta)$ is convex, and $G(\hat{\Delta}) \leq G(\bar{0}) = Q(\tilde{\Omega}_0) = 0$. Thus if we can show that $G(\Delta) > 0$ for all $\Delta \in \mathcal{T}_1$, the minimizer $\hat{\Delta}$ must be inside \mathcal{T}_2 and hence

$\|\hat{\Delta}\|_F \leq Mr_m$. To see this, note that the convexity of $Q(\Omega)$ implies that

$$\inf_{\|\Delta\|_F=Mr_m} Q(\tilde{\Omega}_0 + \Delta) > Q(\tilde{\Omega}_0) = 0.$$

There exists therefore a local minimizer in the ball $\{\tilde{\Omega}_0 + \Delta : \|\Delta\|_F \leq Mr_m\}$, or equivalently, for $\hat{\Delta} \in \mathcal{T}_2$, i.e. $\|\hat{\Delta}\|_F \leq Mr_m$.

In the remainder of the proof, we focus on

$$G(\Delta) = Q(\Delta + \tilde{\Omega}_0) = \text{tr}(\Delta \hat{\Sigma}) - \log \det(\Delta + \tilde{\Omega}_0) + \log \det \tilde{\Omega}_0. \quad (2.18)$$

Applying a Taylor expansion to $\log \det(\tilde{\Omega}_0 + \Delta)$ in (2.18) gives

$$\begin{aligned} & \log \det(\tilde{\Omega}_0 + \Delta) - \log \det \tilde{\Omega}_0 \\ &= \frac{d}{dt} \log \det(\tilde{\Omega}_0 + t\Delta) \Big|_{t=0} \Delta + \int_0^1 (1-t) \frac{d^2}{dt^2} \log \det(\tilde{\Omega}_0 + t\Delta) dt \\ &= \text{tr}(\Delta \tilde{\Sigma}_0) - \text{vec}(\Delta)^T \left\{ \int_0^1 (1-t) (\tilde{\Omega}_0 + t\Delta)^{-1} \otimes (\tilde{\Omega}_0 + t\Delta)^{-1} dt \right\} \text{vec}(\Delta), \end{aligned} \quad (2.19)$$

where $\text{vec}(\Delta)$ denotes the vectorized Δ , and \otimes is the Kronecker product. For $\Delta \in \mathcal{T}_1$, let K_1 be the integral term in (2.19), and define

$$K_2 = \text{tr} \left\{ \Delta (\hat{\Sigma} - \Sigma_0) \right\}, \quad K_3 = \text{tr} \left\{ \Delta (\tilde{\Sigma}_0 - \Sigma_0) \right\}.$$

We can then write

$$G(\Delta) = K_1 + \text{tr}(\Delta \hat{\Sigma}) - \text{tr}(\Delta \tilde{\Sigma}_0) = K_1 + K_2 - K_3.$$

Next, we bound each of the terms K_1, K_2 and K_3 to find a lower bound for $G(\Delta)$.

First consider K_2 . Since the diagonal elements of $\hat{\Sigma}$ and Σ_0 are the same after

scaling,

$$|K_2| \leq \left| \sum_{i \neq i'} (\hat{\Sigma}_{ii'} - \Sigma_{0,ii'}) \Delta_{ii'} \right|.$$

By Lemma A.3 of *Bickel and Levina* (2008), there exists a positive constant c_2 depending on $\phi_{\max}(\Sigma_0)$ such that

$$\max_{i \neq i'} |\hat{\Sigma}_{ii'} - \Sigma_{0,ii'}| \leq c_2 \{\log(p - rp)/m\}^{1/2},$$

with probability tending to 1. Let $\Delta^+ = \text{diag}(\Delta)$ be the matrix of diagonal elements of Δ , and write $\Delta^- = \Delta - \Delta^+$. Then, K_2 is bounded by

$$|K_2| \leq c_2 \{\log(p - rp)/m\}^{1/2} \|\Delta^-\|_1. \quad (2.20)$$

For K_3 , we can use the upper bound for $\|\tilde{\Omega}_0 - \Omega_0\|_F$ in (2.12), and the lower bound for $\phi_{\min}(\tilde{\Omega}_0)$ in (2.17), to write,

$$|K_3| \leq \|\Delta\|_F \|\tilde{\Sigma}_0 - \Sigma_0\|_F \leq \|\Delta\|_F \frac{\|\tilde{\Omega}_0 - \Omega_0\|_F}{\phi_{\min}(\tilde{\Omega}_0) \phi_{\min}(\Omega_0)} \quad (2.21)$$

$$\leq \|\Delta\|_F \frac{c_3 \{S \log(p - rp)/m\}^{1/2}}{(1 - k_1) \phi_1^2}. \quad (2.22)$$

The second inequality in (2.21) comes from the rotation invariant property of Frobenius norm, i.e.

$$\|\tilde{\Sigma}_0 - \Sigma_0\|_F = \|\Sigma_0(\Omega_0 - \tilde{\Omega}_0)\tilde{\Sigma}_0\|_F \leq \phi_{\max}(\Sigma_0) \|\Omega_0 - \tilde{\Omega}_0\|_F \phi_{\max}(\tilde{\Sigma}_0).$$

Using (2.12), we can also obtain an upper bound for the maximum eigenvalue of $\tilde{\Omega}_0$:

$$\phi_{\max}(\tilde{\Omega}_0) \leq \phi_{\max}(\Omega_0) + \|\tilde{\Omega}_0 - \Omega_0\|_2 \leq \phi_{\max}(\Omega_0) + \|\tilde{\Omega}_0 - \Omega_0\|_F \leq \frac{1}{\phi_2} + k_1 \phi_1.$$

Since $r_m \rightarrow 0$, there exists a sufficiently large $k_2 > 0$ such that for $\Delta \in \mathcal{T}_1$,

$$\|\Delta\|_2 \leq \|\Delta\|_F = Mr_m < \frac{1}{\phi_2} k_2.$$

Following *Rothman et al.* (2008, Page 502, proof of Theorem 1), a lower bound for K_1 can be found as

$$\begin{aligned} K_1 &\geq \|\Delta\|_F^2 / \{2(\phi_{\max}(\tilde{\Omega}_0) + \|\Delta\|_2)^2\} \\ &\geq \|\Delta\|_F^2 / \{2(1/\phi_2 + k_1\phi_1 + k_2/\phi_2)^2\} = \frac{\phi_2^2}{2(1 + k_1\phi_1\phi_2 + k_2)^2} \|\Delta\|_F^2. \end{aligned} \quad (2.23)$$

Combining (2.20), (2.22) and (2.23),

$$\begin{aligned} G(\Delta) &\geq \frac{\phi_2^2}{2(1 + k_1\phi_1\phi_2 + k_2)^2} \|\Delta\|_F^2 - c_2 \{\log(p - rp)/m\}^{1/2} \|\Delta^-\|_1 \\ &\quad - \frac{c_3 \{S \log(p - rp)/m\}^{1/2}}{(1 - k_1)\phi_1^2} \|\Delta\|_F. \end{aligned}$$

For $\Delta \in \mathcal{T}_1$, applying Cauchy-Schwarz inequality yields

$$\|\Delta^-\|_1 \leq (|\hat{E}|)^{1/2} \|\Delta^-\|_F.$$

We thus have

$$\begin{aligned} G(\Delta) &\geq \frac{\phi_2^2}{2(1 + k_1\phi_1\phi_2 + k_2)^2} \|\Delta\|_F^2 - c_2 \{|\hat{E}| \log(p - rp)/m\}^{1/2} \|\Delta^-\|_F \\ &\quad - \frac{c_3}{(1 - k_1)\phi_1^2} \{S \log(p - rp)/m\}^{1/2} \|\Delta\|_F \\ &\geq \|\Delta\|_F^2 \left\{ \frac{\phi_2^2}{2(1 + k_1\phi_1\phi_2 + k_2)^2} - \frac{c_2}{M} \{|\hat{E}|/S\}^{1/2} - \frac{c_3}{M(1 - k_1)\phi_1^2} \right\} > 0, \end{aligned}$$

for M sufficiently large. □

Proof of Corollary II.6. Under the assumptions in Theorem II.4, we have

$$\|\Delta_{\Omega_0}\| = \|\hat{\Omega} - \Omega_0\|_2 = O_{\mathbb{P}}(\{S \log(p - rp)/m\}^{1/2}) = o_{\mathbb{P}}(1).$$

Therefore the partial correlation matrix corresponding to $\hat{\Omega}$ can be written as

$$\hat{\mathbf{A}} = \mathbf{I}_p - \hat{\mathbf{D}}^{-1/2} \hat{\Omega} \hat{\mathbf{D}}^{-1/2} = \mathbf{A}_0 + \mathbf{D}_0^{-1/2} \Omega_0 \mathbf{D}_0^{-1/2} - (\hat{\mathbf{D}})^{-1/2} \hat{\Omega} \hat{\mathbf{D}}^{-1/2} = \mathbf{A}_0 + \Delta_{\mathbf{A}_0},$$

where

$$\begin{aligned} \Delta_{\mathbf{A}_0} &= \mathbf{D}_0^{-1/2} \Omega_0 \mathbf{D}_0^{-1/2} - (\hat{\mathbf{D}})^{-1/2} \hat{\Omega} \hat{\mathbf{D}}^{-1/2} \\ &= \mathbf{D}_0^{-1/2} (\Omega_0 - \hat{\Omega}) \mathbf{D}_0^{-1/2} + \mathbf{D}_0^{-1/2} \hat{\Omega} (\mathbf{D}_0^{-1/2} - \hat{\mathbf{D}}^{-1/2}) + (\mathbf{D}_0^{-1/2} - \hat{\mathbf{D}}^{-1/2}) \hat{\Omega} \hat{\mathbf{D}}^{-1/2}. \end{aligned} \tag{2.24}$$

Next we show that each of the summands on the right hand side of (2.24) has ℓ_2 norm $o_{\mathbb{P}}(1)$ and conclude thus $\|\Delta_{\mathbf{A}_0}\|_2 = o_{\mathbb{P}}(1)$.

By Assumption 2, the diagonal entries of Ω_0 satisfy $\omega_{0,ii} \geq \phi_{\min}(\Omega_0) \geq \phi_1$ for all $i = 1, \dots, p$. Thus, $\|\mathbf{D}_0^{-1/2}\|_2 = \max_i \omega_{0,ii}^{-1/2} \leq \phi_1^{-1/2}$. It follows that

$$\|\mathbf{D}_0^{-1/2} (\Omega_0 - \hat{\Omega}) \mathbf{D}_0^{-1/2}\|_2 \leq \|\mathbf{D}_0^{-1/2}\|_2^2 \|\Omega_0 - \hat{\Omega}\|_2 = o_{\mathbb{P}}(1).$$

For the remaining two terms, first notice that $\|\mathbf{D}_0 - \hat{\mathbf{D}}\|_2 \leq \|\mathbf{D}_0 - \hat{\mathbf{D}}\|_F \leq \|\Omega_0 - \hat{\Omega}\|_F = o_{\mathbb{P}}(1)$. Therefore,

$$\begin{aligned} \|\mathbf{D}_0^{-1/2} - \hat{\mathbf{D}}^{-1/2}\|_2 &= \max_{i=1, \dots, p} |\omega_{0,ii}^{-1/2} - \hat{\omega}_{ii}^{-1/2}| = \max_{i=1, \dots, p} \left| \frac{\omega_{0,ii}^{1/2} - \hat{\omega}_{ii}^{1/2}}{\omega_{0,ii}^{1/2} \hat{\omega}_{ii}^{1/2}} \right| \\ &= \max_{i=1, \dots, p} \left| \frac{\omega_{0,ii} - \hat{\omega}_{ii}}{\omega_{0,ii}^{1/2} \hat{\omega}_{ii}^{1/2} (\omega_{0,ii}^{1/2} + \hat{\omega}_{ii}^{1/2})} \right| \leq \phi_1^{-1} (\phi_1 - o_{\mathbb{P}}(1))^{-1/2} \|\mathbf{D}_0 - \hat{\mathbf{D}}\|_2, \end{aligned}$$

where the last inequality comes from that fact that

$$\min_i |\hat{\omega}_{ii}| = \min_i |\hat{\omega}_{ii} - \omega_{0,ii} + \omega_{0,ii}| \geq \min_i |\omega_{0,ii}| - \max_i |\hat{\omega}_{ii} - \omega_{0,ii}| \geq \phi_1 - o_{\mathbb{P}}(1).$$

Hence, $\|\mathbf{D}_0^{-1/2} - \hat{\mathbf{D}}^{-1/2}\|_2 = o_{\mathbb{P}}(1)$. Note further,

$$\|\hat{\Omega}\|_2 = \|\hat{\Omega} - \Omega_0 + \Omega_0\|_2 \leq \|\Omega_0\|_2 + \|\hat{\Omega} - \Omega_0\|_2 = \|\Omega_0\|_2 + o_{\mathbb{P}}(1)$$

is bounded above. It follows thus,

$$\begin{aligned} \|\mathbf{D}_0^{-1/2} \hat{\Omega} (\mathbf{D}_0^{-1/2} - \hat{\mathbf{D}}^{-1/2})\|_2 &\leq \|\mathbf{D}_0^{-1/2}\|_2 \|\hat{\Omega}\|_2 \|\mathbf{D}_0^{-1/2} - \hat{\mathbf{D}}^{-1/2}\|_2 = o_{\mathbb{P}}(1), \\ \|(\mathbf{D}_0^{-1/2} - \hat{\mathbf{D}}^{-1/2}) \hat{\Omega} \hat{\mathbf{D}}^{-1/2}\|_2 &\leq \|\mathbf{D}_0^{-1/2} - \hat{\mathbf{D}}^{-1/2}\|_2 \|\hat{\Omega}\|_2 \|\hat{\mathbf{D}}^{-1/2}\|_2 = o_{\mathbb{P}}(1). \end{aligned}$$

This completes the proof. □

2.9 Proof of Theorem II.8

The following proof of Theorem II.8 adapts from that of Theorem 2.1 in *Shojaie and Michailidis (2010)*.

Proof of Theorem II.8. Consider the special case where the row vector $\mathbf{b} = \mathbf{1}^T$, i.e. the whole network is tested as one pathway. The general case when $\mathbf{b} \neq \mathbf{1}^T$ follows from a similar argument.

For the partial correlation $\mathbf{A}_0^{(k)}$ ($k = 1, 2$) defined in Section 2.3.2, it holds that $\Lambda^{(k)}(\Lambda^{(k)})^T = (\mathbf{I}_p - \mathbf{A}_0^{(k)})^{-1} = \sum_{t=0}^{\infty} (\mathbf{A}_0^{(k)})^t$. Hence

$$\begin{aligned} \hat{\Lambda}^{(k)}(\hat{\Lambda}^{(k)})^T &= \sum_{t=0}^{\infty} (\hat{\mathbf{A}}^{(k)})^t = \sum_{t=0}^{\infty} (\mathbf{A}_0^{(k)})^t + \sum_{t=1}^{\infty} \sum_{u=1}^t \binom{t}{u} (\mathbf{A}_0^{(k)})^{t-u} (\Delta_{\mathbf{A}_0^{(k)}})^u \\ &= \Lambda^{(k)}(\Lambda^{(k)})^T + \Delta_{\Lambda^{(k)}}. \end{aligned}$$

For $\hat{\mathbf{A}}^{(k)}$ defined under the assumptions in Theorem II.4 and II.8, we have $\|\Delta_{\mathbf{A}_0^{(k)}}\|_2 = o_{\mathbb{P}}(1)$ by Corollary II.6. Thus, $\|\Delta_{\Lambda^{(k)}}\|_2 = o_{\mathbb{P}}(1)$.

Using results from *Shojaie and Michailidis* (2010), the test statistic in (3.12) can be written as

$$TS = \frac{\mathbf{b}(\bar{\mathbf{Y}}^{(2)} - \bar{\mathbf{Y}}^{(1)})}{\sqrt{\hat{\sigma}_{\gamma}^2 \left[\mathbf{b} \left\{ \frac{1}{n_1} \hat{\Lambda}^{(1)} (\hat{\Lambda}^{(1)})^T + \frac{1}{n_2} \hat{\Lambda}^{(2)} (\hat{\Lambda}^{(2)})^T \right\} \mathbf{b}^T \right] + \hat{\sigma}_{\varepsilon}^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right) \mathbf{b} \mathbf{b}^T}},$$

where $\bar{\mathbf{Y}}^{(k)}$ is the mean expression of genes in the experimental condition k . *Shojaie and Michailidis* (2010) show that TS is an asymptotically most powerful unbiased test for (3.11) when the correct network information is provided. Therefore, to establish the result in Theorem II.8, it suffices to show that the denominator of TS is a consistent estimator.

In the following, we first consider the log-likelihood $l_F(\vartheta; \hat{\Lambda})$ based on the estimated networks $\hat{\Lambda} = (\hat{\Lambda}^{(1)}, \hat{\Lambda}^{(2)})$ and correct variance components $\vartheta = (\sigma_{\gamma}^2, \sigma_{\varepsilon}^2)$. We then establish that the maximum likelihood estimator $\hat{\vartheta}_{\hat{\Lambda}} \rightarrow_P \vartheta$ as $\hat{\Lambda}^{(k)} (\hat{\Lambda}^{(k)})^T \rightarrow_P \Lambda^{(k)} (\Lambda^{(k)})^T$ for both k . Hence the denominator of TS is consistent and TS is an asymptotically most powerful unbiased test for (3.11).

Let $\hat{\mathbf{W}}^{(k)} = \sigma_{\gamma}^2 \hat{\Lambda}^{(k)} (\hat{\Lambda}^{(k)})^T + \sigma_{\varepsilon}^2 \mathbf{I}_p$ for $k = 1, 2$. Up to a constant, the negative log-likelihood

$$l_F(\vartheta; \hat{\Lambda}) = \frac{n_1}{2n} l(\vartheta; \hat{\Lambda}^{(1)}) + \frac{n_2}{2n} l(\vartheta; \hat{\Lambda}^{(2)})$$

with

$$l(\vartheta; \hat{\Lambda}^{(1)}) = \log \det(\hat{\mathbf{W}}^{(1)}) + \frac{1}{n_1} \sum_{j=1}^{n_1} \mathbf{R}_j^T (\hat{\mathbf{W}}^{(1)})^{-1} \mathbf{R}_j,$$

$$l(\vartheta; \hat{\Lambda}^{(2)}) = \log \det(\hat{\mathbf{W}}^{(2)}) + \frac{1}{n_2} \sum_{j=1+n_1}^n \mathbf{R}_j^T (\hat{\mathbf{W}}^{(2)})^{-1} \mathbf{R}_j,$$

where $\mathbf{R}_j = \mathbf{Y}_j^{(1)} - \bar{\mathbf{Y}}^{(1)}$ ($j = 1, \dots, n_1$) and $\mathbf{R}_j = \mathbf{Y}_j^{(2)} - \bar{\mathbf{Y}}^{(2)}$ ($j = 1 + n_1, \dots, n$). We

treat $l(\vartheta; \hat{\Lambda}^{(1)})$ first. In particular, we can approximate $l(\vartheta; \hat{\Lambda}^{(1)})$ using its one-term Taylor expansion around $\mathbf{W}^{(1)}$

$$l(\vartheta; \hat{\Lambda}^{(1)}) = l(\vartheta; \Lambda^{(1)}) + \text{tr} [\nabla_{\mathbf{W}^{(1)}} l(\vartheta; \Lambda^{(1)})^T \Delta_{\mathbf{W}^{(1)}}] + o(\|\Delta_{\mathbf{W}^{(1)}}\|_2^2),$$

where $\nabla_{\mathbf{W}^{(1)}} l(\vartheta; \Lambda^{(1)})$ is the gradient of $l(\vartheta; \Lambda^{(1)})$ with respect to $\mathbf{W}^{(1)}$ and

$$\nabla_{\mathbf{W}^{(1)}} l(\vartheta; \Lambda^{(1)}) = (\mathbf{W}^{(1)})^{-1} - n_1^{-1} \sum_{j=1}^{n_1} (\mathbf{W}^{(1)})^{-1} \mathbf{R}_j \mathbf{R}_j^T (\mathbf{W}^{(1)})^{-1}.$$

Let $\Gamma = \Delta_{\mathbf{W}^{(1)}} / \|\Delta_{\mathbf{W}^{(1)}}\|_2$ and denote

$$g(\vartheta) = \text{tr} [\nabla_{\mathbf{W}^{(1)}} l(\vartheta; \Lambda^{(1)})^T \Gamma] = \text{tr} [(\mathbf{W}^{(1)})^{-1} \Gamma] - n_1^{-1} \sum_{j=1}^{n_1} \mathbf{R}_j^T (\mathbf{W}^{(1)})^{-1} \Gamma (\mathbf{W}^{(1)})^{-1} \mathbf{R}_j.$$

then

$$l(\vartheta; \hat{\Lambda}^{(1)}) = l(\vartheta; \Lambda^{(1)}) + g(\vartheta) \|\Delta_{\mathbf{W}^{(1)}}\|_2 + o(\|\Delta_{\mathbf{W}^{(1)}}\|_2^2).$$

Using von Neumann's trace inequality (*Mirsky, 1975*), we can bound the first term in $g(\vartheta)$ by

$$\begin{aligned} |\text{tr} [(\mathbf{W}^{(1)})^{-1} \Gamma]| &\leq \sum_{i=1}^p \varsigma_{[i]}((\mathbf{W}^{(1)})^{-1}) \varsigma_{[i]}(\Gamma) \\ &\leq p \varsigma_{[1]}((\sigma_\gamma^2 \Lambda^{(1)} (\Lambda^{(1)})^T + \sigma_\varepsilon^2 \mathbf{I}_p)^{-1}) \varsigma_{[1]}(\Gamma) \\ &= p \frac{1}{\phi_{\min}(\sigma_\gamma^2 \Lambda^{(1)} (\Lambda^{(1)})^T + \sigma_\varepsilon^2 \mathbf{I}_p)} \varsigma_{[1]}(\Gamma), \end{aligned}$$

where $\varsigma_{[i]}(\mathbf{A})$ denotes the i th largest singular value of \mathbf{A} . By construction, $\varsigma_{[1]}(\Gamma) = 1$ and $\phi_{\min}(\sigma_\gamma^2 \Lambda^{(1)} (\Lambda^{(1)})^T + \sigma_\varepsilon^2 \mathbf{I}_p) \geq \sigma_\varepsilon^2$. Hence $|\text{tr}[(\mathbf{W}^{(1)})^{-1} \Gamma]| \leq p/\sigma_\varepsilon^2$. On the other

hand, with probability tending to 1,

$$\begin{aligned} n_1^{-1} \sum_{j=1}^{n_1} \mathbf{R}_j^T (\mathbf{W}^{(1)})^{-1} \Gamma (\mathbf{W}^{(1)})^{-1} \mathbf{R}_j &\leq \|(\mathbf{W}^{(1)})^{-1} \Gamma (\mathbf{W}^{(1)})^{-1}\|_2 n_1^{-1} \sum_{j=1}^{n_1} \mathbf{R}_j^T \mathbf{R}_j \\ &\leq \|(\mathbf{W}^{(1)})^{-1}\|_2^2 \|\Gamma\|_2 n_1^{-1} \sum_{j=1}^{n_1} \mathbf{R}_j^T \mathbf{R}_j = \sigma_\epsilon^{-4} \mathbb{E}(\|\mathbf{R}_j\|_2^2), \end{aligned}$$

where the last step follows from the strong law of large numbers. This implies that $g(\vartheta)$ is bounded for nontrivial σ_ϵ^2 . Note also $\Delta_{\mathbf{W}^{(1)}} = \hat{\mathbf{W}}^{(1)} - \mathbf{W}^{(1)} = \sigma_\gamma^2 \{\hat{\Lambda}^{(1)} (\hat{\Lambda}^{(1)})^T - \Lambda^{(1)} (\Lambda^{(1)})^T\} = \sigma_\gamma^2 \Delta_{\Lambda^{(k)}}$. Hence $g(\vartheta) \|\Delta_{\mathbf{W}^{(1)}}\|_2 = g(\vartheta) \sigma_\gamma^2 \|\Delta_{\Lambda^{(k)}}\|_2 = o_{\mathbb{P}}(1)$. Therefore $l(\vartheta; \hat{\Lambda}^{(1)}) = l(\vartheta; \Lambda^{(1)}) + o_{\mathbb{P}}(1)$, and similarly one can show that $l(\vartheta; \hat{\Lambda}^{(2)}) = l(\vartheta; \Lambda^{(2)}) + o_{\mathbb{P}}(1)$. They together imply that

$$l_F(\vartheta; \hat{\Lambda}) = l_F(\vartheta; \Lambda) + o_{\mathbb{P}}(1).$$

Now conditioning on the event $\{l_F(\vartheta; \hat{\Lambda}) = l_F(\vartheta; \Lambda)\}$, the estimate of the variance components is $\hat{\vartheta} = \arg \min_{\vartheta} l_F(\vartheta; \Lambda)$. Since $l_F(\vartheta; \Lambda)$ is convex with respect to ϑ , M-estimation results in *Haberman* (1989) imply that $\mathbb{P}(\hat{\vartheta} = \vartheta) = 1$ and hence $\hat{\vartheta} \rightarrow_P \vartheta$ as $\hat{\Lambda}^{(k)} (\hat{\Lambda}^{(k)})^T \rightarrow_P \Lambda^{(k)} (\Lambda^{(k)})^T$ for both k . It follows immediately that the denominator of the test statistic TS is a consistent estimator as $\hat{\Lambda}^{(k)} (\hat{\Lambda}^{(k)})^T \rightarrow_P \Lambda^{(k)} (\Lambda^{(k)})^T$ for both k . This concludes the proof. \square

2.10 Derivation for Newton's Method

The implementation of Newton's method requires the gradient and the Hessian of the objective function, i.e. the profile log-likelihood. Here we provide details about how to calculate the gradient and Hessian based on the profile log-likelihood when profiling out σ_ϵ . The derivation follows similarly when profiling out σ_γ .

Let $N = np$ be the total number of observations for all genes. Recall that for $k = 1, 2$, $\Sigma^{(k)} = \mathbf{I}_p + \tau \Lambda^{(k)} (\Lambda^{(k)})^T$ with $\tau = \sigma_\gamma^2 / \sigma_\epsilon^2$. The residuals $\mathbf{R}_j = \mathbf{Y}_j^{(k)} - \Lambda^{(k)} \hat{\boldsymbol{\mu}}^{(k)}$

for $j = 1, \dots, n$, where $\hat{\boldsymbol{\mu}}^{(k)}$ is the estimate of $\boldsymbol{\mu}^{(k)}$. Given the observations $\mathbf{Y}_1, \dots, \mathbf{Y}_n$ (with the first n_1 samples from condition 1 and the remaining $n_2 = n - n_1$ samples from condition 2), the nonconstant part of the “full” log-likelihood l_F is

$$l_F(\sigma_\varepsilon, \tau \mid \mathbf{Y}_1, \dots, \mathbf{Y}_n) = -\frac{1}{2} \left\{ n_1 \log \det(\sigma_\varepsilon^2 \Sigma^{(1)}) + n_2 \log \det(\sigma_\varepsilon^2 \Sigma^{(2)}) \right\} \\ - \frac{1}{2} \sigma_\varepsilon^{-2} \left\{ \sum_{j=1}^{n_1} \mathbf{R}_j^T (\Sigma^{(1)})^{-1} \mathbf{R}_j + \sum_{j=n_1+1}^n \mathbf{R}_j^T (\Sigma^{(2)})^{-1} \mathbf{R}_j \right\},$$

Similarly, the nonconstant part of the log-likelihood using restricted maximum likelihood is

$$l_R(\sigma_\varepsilon, \tau \mid \mathbf{Y}_1, \dots, \mathbf{Y}_n) = l_F(\sigma_\varepsilon, \tau \mid \mathbf{Y}_1, \dots, \mathbf{Y}_n) \\ - \frac{1}{2} \log \det \{ n_1 \sigma_\varepsilon^{-2} (\Lambda^{(1)})^T (\Sigma^{(1)})^{-1} \Lambda^{(1)} \} - \frac{1}{2} \log \det \{ n_2 \sigma_\varepsilon^{-2} (\Lambda^{(2)})^T (\Sigma^{(2)})^{-1} \Lambda^{(2)} \}.$$

To simplify the computation, we solve for σ_ε^2 as a function of τ . The maximum likelihood estimate of σ_ε^2 is

$$\hat{\sigma}_\varepsilon^2 = \frac{1}{N} \left\{ \sum_{j=1}^{n_1} \mathbf{R}_j^T (\Sigma^{(1)})^{-1} \mathbf{R}_j + \sum_{j=n_1+1}^n \mathbf{R}_j^T (\Sigma^{(2)})^{-1} \mathbf{R}_j \right\}, \quad (2.25)$$

whereas its restricted maximum likelihood estimate is given by

$$\hat{\sigma}_\varepsilon^2 = \frac{1}{N - 2p} \left\{ \sum_{j=1}^{n_1} \mathbf{R}_j^T (\Sigma^{(1)})^{-1} \mathbf{R}_j + \sum_{j=n_1+1}^n \mathbf{R}_j^T (\Sigma^{(2)})^{-1} \mathbf{R}_j \right\}. \quad (2.26)$$

Substituting σ_ε^2 with the corresponding estimate, we obtain the profile log-likelihood

$$p_F(\tau \mid \mathbf{Y}_1, \dots, \mathbf{Y}_n) = -\frac{1}{2} (n_1 \log \det \Sigma^{(1)} + n_2 \log \det \Sigma^{(2)}) \\ - \frac{1}{2} N \log \left\{ \sum_{j=1}^{n_1} \mathbf{R}_j^T (\Sigma^{(1)})^{-1} \mathbf{R}_j + \sum_{j=n_1+1}^n \mathbf{R}_j^T (\Sigma^{(2)})^{-1} \mathbf{R}_j \right\}, \quad (2.27)$$

for maximum likelihood, and

$$\begin{aligned}
p_R(\tau \mid \mathbf{Y}_1, \dots, \mathbf{Y}_n) &= -\frac{1}{2}(n_1 \log \det \Sigma^{(1)} + n_2 \log \det \Sigma^{(2)}) \\
&\quad - \frac{1}{2}(N - 2p) \log \left\{ \sum_{j=1}^{n_1} \mathbf{R}_j^T (\Sigma^{(1)})^{-1} \mathbf{R}_j + \sum_{j=n_1+1}^n \mathbf{R}_j^T (\Sigma^{(2)})^{-1} \mathbf{R}_j \right\} \\
&\quad - \frac{1}{2} \log \det \{n_1 (\Lambda^{(1)})^T (\Sigma^{(1)})^{-1} \Lambda^{(1)}\} - \frac{1}{2} \log \det \{n_2 (\Lambda^{(2)})^T (\Sigma^{(2)})^{-1} \Lambda^{(2)}\}, \quad (2.28)
\end{aligned}$$

for restricted maximum likelihood.

As $\Sigma^{(k)}$ ($k = 1, 2$) are the only terms that depend on τ , we first look at the derivatives of

$$\log \det \Sigma^{(k)}, \quad \mathbf{R}_j^T (\Sigma^{(k)})^{-1} \mathbf{R}_j, \quad \log \det \{(\Lambda^{(k)})^T (\Sigma^{(k)})^{-1} \Lambda^{(k)}\},$$

with respect to τ . Denote

$$\mathbf{B}^{(k)} = (\Sigma^{(k)})^{-1} \frac{d\Sigma^{(k)}}{d\tau} (\Sigma^{(k)})^{-1}, \quad \mathbf{H}^{(k)} = (\Lambda^{(k)})^T (\Sigma^{(k)})^{-1} \Lambda^{(k)}.$$

Then

$$\begin{aligned}
\frac{d \log \det(\Sigma^{(k)})}{d\tau} &= \text{tr} \left\{ (\Sigma^{(k)})^{-1} \frac{d\Sigma^{(k)}}{d\tau} \right\}, \\
\frac{d^2 \log \det(\Sigma^{(k)})}{d\tau^2} &= \text{tr} \left\{ -(\mathbf{B}^{(k)})^T \frac{d\Sigma^{(k)}}{d\tau} + (\Sigma^{(k)})^{-1} \frac{d^2 \Sigma^{(k)}}{d\tau^2} \right\}, \\
\frac{d \mathbf{R}_j^T (\Sigma^{(k)})^{-1} \mathbf{R}_j}{d\tau} &= -\mathbf{R}_j^T \mathbf{B}^{(k)} \mathbf{R}_j, \quad \frac{d^2 \mathbf{R}_j^T (\Sigma^{(k)})^{-1} \mathbf{R}_j}{d\tau^2} = -\mathbf{R}_j^T \frac{d\mathbf{B}^{(k)}}{d\tau} \mathbf{R}_j, \\
\frac{d \log \det \mathbf{H}^{(k)}}{d\tau} &= -\text{tr} \left\{ (\mathbf{H}^{(k)})^{-1} (\Lambda^{(k)})^T \mathbf{B}^{(k)} \Lambda^{(k)} \right\},
\end{aligned}$$

$$\begin{aligned}
\frac{d^2 \log \det \mathbf{H}^{(k)}}{d\tau^2} &= -\text{tr} \left\{ (\mathbf{H}^{(k)})^{-1} (\Lambda^{(k)})^T \mathbf{B}^{(k)} \Lambda^{(k)} (\mathbf{H}^{(k)})^{-1} (\Lambda^{(k)})^T \mathbf{B}^{(k)} \Lambda^{(k)} \right\} \\
&\quad - \text{tr} \left\{ (\mathbf{H}^{(k)})^{-1} (\Lambda^{(k)})^T \frac{d\mathbf{B}^{(k)}}{d\tau} \Lambda^{(k)} \right\},
\end{aligned}$$

where

$$\frac{d\mathbf{B}^{(k)}}{d\tau} = -(\Sigma^{(k)})^{-1} \left\{ 2 \frac{d\Sigma^{(k)}}{d\tau} (\Sigma^{(k)})^{-1} \frac{d\Sigma^{(k)}}{d\tau} - \frac{d^2 \Sigma^{(k)}}{d\tau^2} \right\} (\Sigma^{(k)})^{-1}.$$

Given the covariance $\Sigma^{(k)}$ ($k = 1, 2$) defined in Section 3, we can further simplify the above derivatives and obtain

$$\begin{aligned} \frac{d \log \det \Sigma^{(k)}}{d\tau} &= \text{tr} \{ \mathbf{H}^{(k)} \}, & \frac{d^2 \log \det \Sigma^{(k)}}{d\tau^2} &= -\text{tr} \{ \mathbf{H}^{(k)} \mathbf{H}^{(k)} \}, \\ \frac{d \mathbf{R}_j^T (\Sigma^{(k)})^{-1} \mathbf{R}_j}{d\tau} &= -\mathbf{R}_j^T (\Sigma^{(k)})^{-1} \Lambda^{(k)} (\Lambda^{(k)})^T (\Sigma^{(k)})^{-1} \mathbf{R}_j \\ \frac{d^2 \mathbf{R}_j^T (\Sigma^{(k)})^{-1} \mathbf{R}_j}{d\tau^2} &= 2\mathbf{R}_j^T (\Sigma^{(k)})^{-1} \Lambda^{(k)} \mathbf{H}^{(k)} (\Lambda^{(k)})^T (\Sigma^{(k)})^{-1} \mathbf{R}_j, \\ \frac{d \log \det \mathbf{H}^{(k)}}{d\tau} &= -\text{tr} \{ \mathbf{H}^{(k)} \}, & \frac{d^2 \log \det \mathbf{H}^{(k)}}{d\tau^2} &= \text{tr} \{ \mathbf{H}^{(k)} \mathbf{H}^{(k)} \}. \end{aligned}$$

With the above quantities, one can then calculate the gradient and Hessian of the profile log-likelihood p_R for restricted maximum likelihood and use Newton's method to obtain an estimate of τ . Estimate of $\hat{\sigma}_\epsilon^2$ is calculated from (2.26), and $\hat{\sigma}_\gamma^2 = \hat{\tau} \hat{\sigma}_\epsilon^2$. Estimation with maximum likelihood follows similarly by applying Newton's method to p_F and utilizing (2.25).

2.11 Additional Simulation Results

To benchmark the performance of the proposed network estimation procedure as well as NetGSA, we carried out another two experiments, which we describe as the third and fourth experiment following the earlier two in Section 2.4. These are also the two experiments we mentioned when comparing the running time of NetGSA with different variance estimation algorithms.

Our third experiment considers a undirected network with $p = 160$. The simula-

tion design is similar to that in experiment 2, except that each of the 8 subnetworks has a denser structure. Specifically, there are 80 edges connecting the 20 genes in each subnetwork under the null. The probability of an interaction between subnetworks is 0.3. Under the alternative, there is an increase of 0.6 in mean expression values for varying proportions of genes (0%, 30%, 50% and 90%) for pathways 1–4 and 5–8. Moreover, half of the interactions in the latter four subnetworks disappear.

The fourth experiment is about networks of size $p = 400$, which also illustrates that the proposed method scales well with the size of the networks based on implementation of the updated optimization algorithm. Again the topology is similar to previous scenarios so that it consists of 20 subnetworks, each corresponding to a pathway with 20 genes. The probability of an interaction between pathways is also 0.3. All subnetworks have the same topology and were generated as scale-free random graphs such that there are 40 edges linking the 20 genes. We then divided the 20 subnetworks into two groups, with the first 10 in the first group, and the last 10 in the second group. Under the null, mean expression values for all subnetworks were set to be 1. Under the alternative, the first 6 subnetworks in each group remained to have the same mean expression values, but 20%, 30%, 30% and 40% of genes in the last four subnetworks had 0.5 unit higher expression values, respectively. In addition, subnetwork structure for the second group under the alternative differed from their null equivalent by 22.5%. This experiment is also of interest because we created a setting where there are enough pathways in order for the permutation based Gene Set Analysis to calibrate the number of permutations required.

Table 2.7 presents the deviance measures for estimating the networks with 200 replicates and sample sizes of 300 for both $p = 160$ and $p = 400$, when varying levels of external information are available. In both experiments, we see performance improvement in Matthews correlation coefficient and Frobenius norm loss as the structural information of the networks r increases. Under the alternative of $p = 160$, the

slightly better performance in terms of Frobenius norm loss is due to the sparser network structure relative to the null.

Table 2.7: Deviance measures for network estimation in experiment 3 and 4. FPR(%), false positive rate in percentage; FNR(%), false negative rate in percentage; MCC, Matthews correlation coefficient; Fnorm, Frobenius norm loss. The best cases are highlighted in bold.

		$p = 160$				$p = 400$			
	r	FPR(%)	FNR(%)	MCC	Fnorm	FPR(%)	FNR(%)	MCC	Fnorm
Null	0.0	7.78	4.75	0.59	0.65	2.87	5.85	0.51	0.38
	0.2	6.81	5.03	0.61	0.63	2.44	8.15	0.53	0.37
	0.8	2.60	4.41	0.78	0.49	0.81	5.63	0.74	0.25
Alternative	0.0	5.60	2.95	0.61	0.46	2.58	5.74	0.53	0.36
	0.2	4.72	3.67	0.64	0.45	2.18	7.95	0.56	0.35
	0.8	1.47	3.68	0.83	0.34	0.70	5.91	0.76	0.24

Table 2.8 shows the estimated powers after correcting for false discovery rate in the third experiment with $p = 160$. While Gene Set Analysis with the competitive null hypothesis tends to suggest that none of the pathways is significantly differentially expressed under the alternative, its equivalent with the self-contained null overestimates powers for most pathways. In comparison, NetGSA with exact network information slightly underestimates, but mostly correctly the significance of each subnetwork. Moreover, the differences in powers between each pair of pathways (1 and 5, 2 and 6, 3 and 7, as well as 4 and 8) indicate that the topologies for each pair are different, since both had the same amount of changes in mean expression values. When the exact networks are unknown, we see improvement in detected powers for pathway 3, 4 and 8 as the structural information increases from 20% to 80%, which suggests that a small amount of external knowledge is beneficial for making reliable inference using the network-based method.

The estimated powers after correcting for false discovery rate in experiment 4 are shown separately in Table 2.9, as there are 20 pathways with varying parameters. When the exact networks with the correct partial correlation coefficients are known,

Table 2.8: Powers based on false discovery rate with $q^* = 0.05$ in experiment 3. 0.2/0.8 refer to NetGSA with 20%/80% external information; E refers to NetGSA with the exact networks; T refers to the true power; GSA-s/GSA-c refer to Gene Set Analysis with self-contained/competitive null hypothesis in 1000 permutations, respectively. True powers are highlighted in bold.

Pathway	$p = 160$					
	0.2	0.8	E	T	GSA-s	GSA-c
1	0.04	0.04	0.04	0.05	0.12	0.00
2	0.52	0.50	0.44	0.54	0.69	0.00
3	0.64	0.70	0.80	0.87	0.95	0.00
4	0.72	0.76	0.97	0.99	0.99	0.00
5	0.55	0.52	0.12	0.17	0.48	0.00
6	0.48	0.50	0.20	0.25	0.43	0.00
7	0.60	0.54	0.52	0.64	0.78	0.00
8	0.70	0.72	0.80	0.88	0.91	0.00

NetGSA returns estimated powers that match the true powers very well, with high powers for pathways 8, 9 and 10 which have significant changes in mean expression values, moderately high powers for pathways 11–16 that have significant changes in structures and high powers for pathways 17–20 with both changes. When there is 20% external information on the underlying pathway topology, NetGSA is able to identify mostly correctly the powers for pathway 7, 9, 10, and 17–20, with slight overestimation for other pathways. However, the overall trend suggests that pathways 11–16 have higher powers than 1–6, which is consistent with the true power. There is also improvement when 80% structural information is known, although the improvement is minor compared to the amount of structural information required. In comparison, Gene Set Analysis with the self-contained null hypothesis also performs well in recognizing correctly the differentially expressed pathways. On the other hand, Gene Set Analysis with the competitive null is still not able to identify any differential expression among all 20 pathways. The conflicting results from Gene Set Analysis with different null hypotheses also raise concerns as to which version to choose in practice.

Table 2.9: Powers based on false discovery rate with $q^* = 0.05$ in experiment 4. 0.2/0.8 refer to NetGSA with 20%/80% external information; E refers to NetGSA with the exact networks; T refers to the true power; GSA-s/GSA-c refer to Gene Set Analysis with self-contained/competitive null hypothesis in 1000 permutations, respectively. True powers are highlighted in bold.

Pathway	$p = 400$					
	0.2	0.8	E	T	GSA-s	GSA-c
1	0.05	0.06	0.05	0.05	0.08	0.00
2	0.38	0.27	0.00	0.05	0.08	0.00
3	0.56	0.41	0.01	0.05	0.04	0.00
4	0.53	0.46	0.02	0.05	0.06	0.00
5	0.54	0.40	0.02	0.05	0.02	0.00
6	0.62	0.56	0.03	0.05	0.10	0.01
7	0.82	0.74	0.99	0.99	0.97	0.00
8	0.62	0.70	0.42	0.48	0.51	0.00
9	0.77	0.85	1.00	1.00	1.00	0.00
10	0.94	0.94	1.00	1.00	1.00	0.00
11	0.81	0.70	0.29	0.47	0.67	0.00
12	0.71	0.75	0.36	0.47	0.68	0.00
13	0.71	0.78	0.31	0.47	0.65	0.00
14	0.73	0.80	0.32	0.48	0.62	0.00
15	0.79	0.67	0.30	0.47	0.62	0.00
16	0.79	0.64	0.33	0.48	0.68	0.00
17	0.86	0.90	1.00	1.00	1.00	0.00
18	0.84	0.73	0.93	0.97	1.00	0.00
19	0.88	0.86	1.00	1.00	1.00	0.00
20	0.93	0.94	1.00	1.00	1.00	0.06

CHAPTER III

Joint Structural Estimation of Multiple Graphical Models

3.1 Background

As discussed in Chapter I, there has been a lot of work on estimating a *single* graphical model. More recently, the focus has shifted to joint estimation of multiple graphs due to the availability of heterogeneous data (see discussion in *Guo et al. (2011)*). *Guo et al. (2011)* introduced a joint estimation method by adding a hierarchical penalty to the log-likelihood and is thus able to recover both the common and the individual zeros in the precision matrices. *Danaher et al. (2014)* proposed a joint graphical lasso to estimate multiple related graphical models by maximizing the log-likelihood with generalized fused lasso or group lasso penalties, which can be solved efficiently by a standard alternating directions method of multipliers algorithm (*Boyd et al., 2011*). Both joint estimation methods rely on the assumption that there exists only a single common structure across all graphs. *Peterson et al. (2014)* introduced a Bayesian approach that links the estimation of the graphs via a Markov random field prior for common structures. Further, a spike-and-slab prior is placed on the parameters that measures the similarity between graphs, thus relaxing the assumption on sharing of structures across all graphs. Recent work by *Zhu et al. (2014)* investigates

the joint estimation problem by pursuing the element-wise clustering of the network structure over multiple graphs using a truncated ℓ_1 penalty on the pairwise differences between the precision matrices.

Despite recent advances in joint estimation algorithms, theoretical properties of the resulting estimators have not been fully investigated. For example, *Guo et al.* (2011) discussed asymptotic properties of the resulting estimator by establishing recovery results of the common zeros across multiple precision matrices, which is the focus of their method. *Zhu et al.* (2014) focused mainly on consistent estimation of the entry-wise clustering structures with a brief mention of consistency of precision matrices in a special temporal setting; however, no theoretical guarantees are provided for more general settings. Finally, many papers only present algorithms for joint estimation of the Gaussian graphical models under consideration, but no theoretical properties of the estimates (e.g. *Chiquet et al.* (2011); *Danaher et al.* (2014); *Mohan et al.* (2014)).

In this chapter, we investigate estimation of multiple graphical models under complex structural relationships, assuming that there exists prior information on their specification. In many applications, such information is available and may come from prior knowledge in the literature of relationships among different node subsets of the graphical models under consideration, or from clustering of all graphs. The approach allows sharing common sub-graph components between different models and does not require sharing of values for the same element across multiple inverse covariance matrices. The proposed method, called JSEM (**J**oint **S**tructural **E**stimation **M**ethod), leverages structured sparsity patterns as illustrated in Section 3.2 and is a two-step procedure. In the first step, we infer the sparse graphical models by incorporating the available structure through a group lasso penalty. In the second step, we maximize the Gaussian log-likelihood subject to the edge set constraints obtained from the previous step. We establish that the proposed estimator is consistent and establish a

fast rate of convergence with respect to the Frobenius norm for the estimated inverse covariance matrices. We also establish the graph selection consistency property of JSEM under appropriately specified structured sparsity. When the structured sparsity pattern is slightly misspecified, we provide a modified estimator that reduces the number of false positive edges identified due to prior information misspecification. The numerical work shows that JSEM exhibits superior performance in controlling both the number of false positive and false negative edges compared to available methods. Moreover, JSEM is computationally appealing as the number of graphs increases. Finally, we illustrate the method on a real data set dealing with climate modeling, where the structural relationships between the various graphical models reflects geographical information. Our results highlight the different roles forcing factors on climate play at different regions of the United States.

In summary, we develop a very general method for the problem of joint estimation of multiple Gaussian graphical models. The method can incorporate detailed structural information regarding relationships between subsets of the graphical models, while in the absence of such information reduces to the group graphical lasso procedure of *Danaher et al. (2014)*. Further, we rigorously establish the consistent recovery of the edge sets for JSEM, under suitable regularity conditions. Finally, a modified estimator allows consistent recovery even in the presence of misspecification of the structural relationships, thus further enhancing the applicability of JSEM.

This chapter is organized as follows. Section 3.2 discusses the structural relationships model used in this work and present the estimation procedure. Section 3.3 presents the theoretical properties of the proposed method, followed by simulation studies in Section 3.4 and a real data analysis on climate modeling in Section 3.5. We conclude with a discussion in Section 3.6. Most details of the theoretical analysis and proofs are relegated to Section 3.7 and 3.8.

3.2 The Joint Structural Estimation Method

Suppose we are interested in estimating K Gaussian graphical models from their corresponding K data sets, assuming that the models exhibit complex relationships between their edge sets. The data in the k -th model are organized in a $n_k \times p$ matrix $\mathbf{X}^k = (\mathbf{X}_1^k, \dots, \mathbf{X}_p^k)$, where each row represents one observation from $\mathcal{N}(0, \Sigma^k)$, $k = 1, \dots, K$. Without loss of generality, we assume the observations from each model are centered and standardized. For ease of presentation, it is assumed in the following that the sample size $n_k = n$ for all $k = 1, \dots, K$, but the modeling framework can easily accommodate unequal sample sizes. Our goal is to estimate jointly $\Omega^k = (\Sigma^k)^{-1}$ for all k , under the assumption that the K corresponding graphs are related via a structured sparsity pattern \mathcal{G} . For example, consider climate models capturing relationships between climate forcing variables defined over a pre-specified spatial domain. Models that belong to the same climate zone may exhibit greater similarity in their graph structures than those from different zones. Thus, one can define \mathcal{G} based on their spatial locations. Figure 3.1 gives an illustration of the structured sparsity among four graphical models in terms of their adjacency matrices. This pattern indicates that sharing of structures may occur at different subsets of the edge set, which motivates us to develop a joint estimation method that can incorporate this rich and complex structural information.

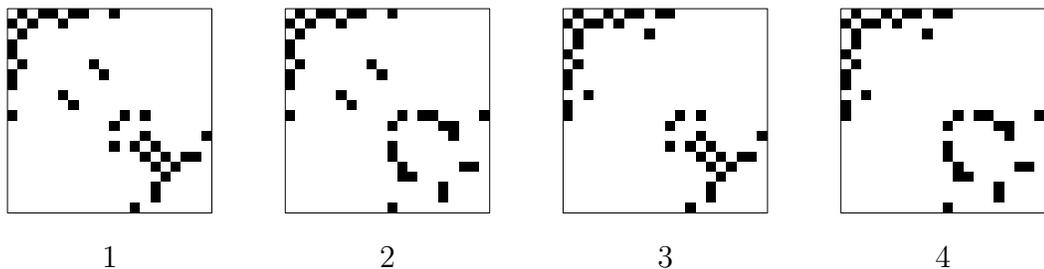


Figure 3.1: Image plots of the adjacency matrices for all four graphical models. The black color represents presence of an edge. The structured sparsity pattern is encoded in $\mathcal{G} = \{(1, 2), (3, 4), (1, 3), (2, 4)\}$, i.e. each pair of graphical models in \mathcal{G} share a subset of edges.

3.2.1 An Illustrative Example

We first illustrate how to extend the idea of neighborhood selection (*Meinshausen and Bühlmann, 2006*) to multiple graphical models using the example in Figure 3.1. For $k = 1, \dots, K$, let $(\theta_{ij}^k)_{p \times p}$ be the matrix of regression coefficients in graph k and $\boldsymbol{\theta}_i^k$ the vector of all θ_{ij}^k ($j \neq i$) for node $i = 1, \dots, p$. Unless otherwise stated, all vectors are assumed to be column vectors. For node i in a single graph k , neighborhood selection suggests estimating the coefficients $\boldsymbol{\theta}_i^k$ by

$$\min_{\boldsymbol{\theta}_i^k} \frac{1}{n} \|\mathbf{X}_i^k - \mathbf{X}_{-i}^k \boldsymbol{\theta}_i^k\|^2 + 2\lambda \sum_{j \neq i} |\theta_{ij}^k|,$$

where \mathbf{X}_{-i}^k is \mathbf{X}^k with the i th column removed, $\|\cdot\|$ represents the standard Euclidean norm and λ is the regularization parameter. To achieve joint estimation, consider the following regularized regression problem

$$\min_{\Theta_i} \frac{1}{n} \sum_{k=1}^K \|\mathbf{X}_i^k - \mathbf{X}_{-i}^k \boldsymbol{\theta}_i^k\|^2 + 2P_\lambda(\Theta_i), \quad (3.1)$$

where $K = 4$, $\Theta_i = (\boldsymbol{\theta}_i^1, \dots, \boldsymbol{\theta}_i^K)$ and $P_\lambda(\Theta_i)$ is a regularization term to be determined next. Note that each column of Θ_i represents the regression coefficients from one graphical model and each row of Θ_i corresponds to the four coefficients at the same (i, j) pair.

The penalty $P_\lambda(\Theta_i)$ is chosen based on information from the structured sparsity pattern \mathcal{G} in Figure 3.1. Specifically, depending on j relative to i , we can group the coefficients in the j th row of Θ_i as

$$\underbrace{(\theta_{ij}^1, \theta_{ij}^2, \theta_{ij}^3, \theta_{ij}^4)}_{\boldsymbol{\theta}_{ij}^{[1,2]}} \quad \text{or} \quad \underbrace{(\theta_{ij}^1, \theta_{ij}^3, \theta_{ij}^2, \theta_{ij}^4)}_{\boldsymbol{\theta}_{ij}^{[1,3]}} \quad \underbrace{\phantom{(\theta_{ij}^1, \theta_{ij}^3, \theta_{ij}^2, \theta_{ij}^4)}}_{\boldsymbol{\theta}_{ij}^{[2,4]}}$$

and set $P_\lambda(\Theta_i)$ to be the group lasso penalty

$$\sum_{j \neq i} \sum_{g=[1,2],[3,4]} \lambda_{ij}^g \|\boldsymbol{\theta}_{ij}^{[g]}\| \quad \text{or} \quad \sum_{j \neq i} \sum_{g=[1,3],[2,4]} \lambda_{ij}^g \|\boldsymbol{\theta}_{ij}^{[g]}\|.$$

The group lasso penalty forces the two coefficients in each group to be zero or nonzero at the same time, leading to the same structure for graphical models belonging to the same group.

The solution $\hat{\Theta}_i$ to (3.1) for $i = 1, \dots, p$ can then be used for graph selection.

3.2.2 The General Case

Denote the structured sparsity pattern by $\mathcal{G} = \bigcup_{1 \leq i < j \leq p} \mathcal{G}^{ij}$, where the union is over all $p(p-1)/2$ pairs of potential edges. Each \mathcal{G}^{ij} is a partition of the set $\{1, 2, \dots, K\}$ and consists of prior knowledge on the structural similarity for the (i, j) th pair across models. For example, $\mathcal{G}^{ij} = \{(1, 2), (3, \dots, K)\}$ means that the graphs 1 and 2 exhibit the same structure at (i, j) , whereas the remaining ones behave the same at (i, j) . It is possible for all graphs to have the edge (i, j) or not have the edge (i, j) at the same time, but we do not impose this restriction. Therefore the pattern \mathcal{G} allows a more flexible structural relationships among multiple graphical models.

For $1 \leq i < j \leq p$ and a group $g \in \mathcal{G}^{ij}$, denote by $\boldsymbol{\theta}_{ij}^{[g]}$ the vector $(\theta_{ij}^k)_{k \in g}$, a concatenation of all regression coefficients from graphs in g . The grouping for the regression coefficients $(\theta_{ij}^1, \dots, \theta_{ij}^K)$ is determined by \mathcal{G}^{ij} . Under correctly specified \mathcal{G} , all coefficients in the same group should be zero or nonzero simultaneously. For $k = 1, \dots, K$, let $E^k = \{(i, j) : 1 \leq i < j \leq p, \theta_{ij}^k \neq 0\}$ be the set of undirected edges in graph k . Denote by \mathcal{S}_+^p the set of all positive definite matrices of size $p \times p$ and $\mathcal{S}_E^p = \{\Omega \in \mathcal{R}^{p \times p} : \omega_{ij} = 0, \text{ for all } (i, j) \notin E \text{ where } i \neq j\}$.

The *Joint Structural Estimation Method* (JSEM) proceeds in the following two steps.

- (1) For $k = 1, \dots, K$, we infer the sparse graphs \hat{E}^k through the following group lasso estimator, i.e. for $i = 1, \dots, p$,

$$\min_{\Theta_i} \frac{1}{n} \sum_{k=1}^K \|\mathbf{X}_i^k - \mathbf{X}_{-i}^k \boldsymbol{\theta}_i^k\|^2 + 2 \sum_{j \neq i} \sum_{g \in \mathcal{G}^{ij}} \lambda_{ij}^g \|\boldsymbol{\theta}_{ij}^{[g]}\|. \quad (3.2)$$

\hat{E}^k is estimated to be the set $\{(i, j) : \max(\hat{\theta}_{ij}^k, \hat{\theta}_{ji}^k) \neq 0\}$.

- (2) We refit the model by

$$\min_{\Omega^k \in \mathcal{S}_+^p \cap \mathcal{S}_{\hat{E}^k}^p} \sum_{k=1}^K \{\text{tr}(\hat{\Sigma}^k \Omega^k) - \log \det(\Omega^k)\}. \quad (3.3)$$

Note that problems (3.2) and (3.3) are both convex and can thus be solved by available convex optimization algorithms. In this work, we use the R-package `grpreg` (Breheny and Huang, 2009) for implementation of the group lasso penalized optimization (3.2) and `glasso` (Friedman et al., 2008) for solving (3.3).

3.2.3 Choice of Tuning Parameters

Like any other penalty-based method, JSEM requires selecting the tuning parameters λ_{ij}^g for all p regressions in (3.2). For our purpose, it suffices to use the same λ for all 3-tuples (i, j, g) , which significantly simplifies the computation. In simulations, we generate a validation dataset and select λ by maximizing the log-likelihood on the validation data using $\hat{\Omega}_\lambda^k$ ($k = 1, \dots, K$) estimated from the training data. In practice, we recommend using the Bayesian information criterion (BIC) coupled with stability selection (Meinshausen and Bühlmann, 2010; Shah and Samworth, 2012) to select graphical models that are both stable and interpretable. Specifically, for a given λ , we define BIC for the proposed method as

$$\text{BIC}(\lambda) = \sum_{k=1}^K \left\{ \text{tr}(\hat{\Sigma}^k \hat{\Omega}_\lambda^k) - \log \det(\hat{\Omega}_\lambda^k) + \frac{\log(n_k)}{n_k} df_k \right\},$$

where $\hat{\Omega}_\lambda^k$ ($k = 1, \dots, K$) are the estimated precision matrices from the data and the degrees of freedom $df_k = \#\{(i, j) : i < j, \hat{\omega}_{\lambda, ij}^k \neq 0\}$. The optimal tuning parameter is thus $\lambda^* = \arg \min_\lambda \text{BIC}(\lambda)$.

3.3 Theoretical Results

In this section, we establish the theoretical properties of JSEM; specifically, the norm consistency of the estimated inverse covariance matrices, as well as the consistent recovery of the edge sets of the various graphical models under consideration based on the structured sparsity pattern \mathcal{G} .

3.3.1 Estimation Consistency

Under the pattern \mathcal{G} , the set $\{(j, g) : j \neq i, g \in \mathcal{G}^{ij}\}$ defines a partition of the index set $\mathbb{N}_{(p-1)K}^i$ in G_i groups, where $\mathbb{N}_{(p-1)K}^i = \{(j, k) : j \neq i, k = 1, \dots, K\}$ and $1 \leq G_i \leq (p-1)K$. Let $J(\Theta_i) = \{(j, g) : j \neq i, g \in \mathcal{G}^{ij}, \boldsymbol{\theta}_{ij}^{[g]} \neq 0\}$ be the set of nonzero groups in the i th regression. We assume an overall sparsity at the group level, i.e. the size of $J(\Theta_i)$ is $s_i \ll G_i$. Let

$$G_0 = \max_{i=1, \dots, p} G_i, \quad s_0 = \max_{i=1, \dots, p} s_i, \quad S_0 = \sum_{i=1}^p s_i,$$

and $|g|$ be the size of the group g with $|g_{\max}| = \max_{g \in \mathcal{G}} |g|$.

Let $\mathbb{M}(p, K)$ represent the set of all $p \times K$ matrices. For $\Delta = (\boldsymbol{\delta}^1, \dots, \boldsymbol{\delta}^K) \in \mathbb{M}(p, K)$ and a group $g \subset \{1, \dots, K\}$, denote by $\boldsymbol{\delta}_j^{[g]}$ the vector composed of all δ_j^k for which $k \in g$. Write $\mathcal{J} = \{J(\Theta_1), \dots, J(\Theta_p)\}$, the collection of sets of nonzero groups in all p regressions. For any $J \in \mathcal{J}$, denote Δ_J the nonzero matrix in $\mathbb{M}(p, K)$, which has the same coordinates as Δ on J and zero elsewhere. Let J^c denote the complement of the index set J . Write $\bar{0}$ the zero matrix in $\mathbb{M}(p, K)$. We make the following assumptions.

A1: For $0 < s < G_0$, there exists $\kappa = \kappa(s) > 0$, such that

$$\min_{J \in \mathcal{J}, |J| \leq s} \min_{\Delta \in \mathcal{F}_J} \frac{\sum_{k=1}^K \|\mathbf{X}^k \boldsymbol{\delta}^k\|^2 / n}{\|\Delta_J\|_F^2} \geq \kappa^2(s), \quad (3.4)$$

where for i satisfying $J(\Theta_i) = J$, \mathcal{F}_J is defined as

$$\mathcal{F}_J = \{\Delta : \Delta \in \mathbb{M}(p, K) \setminus \{\bar{0}\}, \sum_{(j,g) \in J^c} \lambda_{ij}^g \|\boldsymbol{\delta}_j^{[g]}\| \leq 3 \sum_{(j,g) \in J} \lambda_{ij}^g \|\boldsymbol{\delta}_j^{[g]}\|\}.$$

A2: For every $k = 1, \dots, K$ and $i = 1, \dots, p$, $\text{Var}(X_i^k) = 1$. Further, there exist constants c_0, d_0 such that for every k ,

$$0 < 1/c_0 \leq \phi_{\min}(\Sigma_0^k) \leq \phi_{\max}(\Sigma_0^k) \leq 1/d_0 < \infty.$$

Assumption A1 is a generalization of the Restricted Eigenvalue assumption for the Lasso in *Bickel et al. (2009)* to the group lasso setting in our problem and requires the super design matrix $\text{diag}(\mathbf{X}^1, \dots, \mathbf{X}^K)$ to be well conditioned over the restricted set of vectors.

The equal variance requirement in assumption A2 can be easily achieved by appropriate scaling of the data. The second part of the assumption explicitly excludes singular or nearly singular covariance matrices and guarantees that Ω_0^k exists for every model $k = 1, \dots, K$.

Now we are ready to state our main result.

Theorem III.1. *Consider $\hat{\Omega}^k$ ($k = 1, \dots, K$) defined in (3.3). Let Assumption A1 with $s = 2s_0$ and Assumption A2 be satisfied. For every regression defined in (3.2), choose*

$$\lambda_{ij}^g = \frac{2}{\sqrt{nd_0}} \left(\sqrt{|g_{\max}|} + \frac{\pi}{\sqrt{2}} \sqrt{q \log G_0} \right), \quad (3.5)$$

with $q > 1$. Then with probability at least $1 - 2pG_0^{1-q}$, we have

$$\frac{1}{K} \sum_{k=1}^K \|\hat{\Omega}^k - \Omega_0^k\|_F \leq \mathcal{O} \left(\sqrt{\frac{S_0}{nK}} \left\{ \sqrt{|g_{\max}|} + \frac{\pi}{\sqrt{2}} \sqrt{q \log G_0} \right\} \right), \quad (3.6)$$

where G_0 is the maximum number of groups in all regressions, S_0 is the total number of relevant groups and $|g_{\max}|$ is the maximum group size.

Note the rate in (3.6) improves over estimating each precision matrix separately as long as the sparsity pattern \mathcal{G} is appropriately specified and nontrivial, i.e. there exists structural similarity among the considered graphical models. For example, if all K graphs share the same structure, then $|g_{\max}| = K$ and $G_0 = p - 1$. Thus JSEM achieves a convergence rate of the order of

$$\mathcal{O} \left(\sqrt{\frac{S_0}{n}} \left\{ 1 + \frac{\pi}{\sqrt{2}} \sqrt{\frac{q \log(p-1)}{K}} \right\} \right).$$

In contrast, separate estimation of Ω^k is known to be of the order of

$$\mathcal{O} \left(\sqrt{\sum_k \|\Omega^{k,-}\|_0 \log p / (nK)} \right),$$

where $\|\Omega^{k,-}\|_0$ denotes the number of nonzero off-diagonal entries in Ω^k and \sum_k is a short notation for $\sum_{k=1}^K$. Thus JSEM has a lower estimation error rate than separate estimation if $S_0 \asymp \|\Omega^{k,-}\|_0$, where \asymp means that the expressions on both sides are of the same order. On the other hand, the rate in (3.6) could be worse if the sparsity pattern \mathcal{G} is highly misspecified such that the number of nonzero parameters $S_0 > \sum_k \|\Omega^{k,-}\|_0$.

3.3.2 Graph Selection Consistency

To understand how JSEM performs in selecting the edge sets of the graphical models, it suffices to focus on each of the group lasso estimation problems (3.2) as consistent graph selection relies on consistent variable selection in all p regressions. Unlike the sign consistency in the lasso setting (*Zhao and Yu, 2006*), variable selection properties with a group lasso penalty are much more complicated because the latter selects whole groups rather than individual variables (see *Basu et al. (2012)* and the discussion therein). The *Basu et al. (2012)* paper offers a generalization and introduces the notion of direction consistency for the group lasso. Specifically, for a nonzero vector $\boldsymbol{\xi}$, its direction vector is defined as $D(\boldsymbol{\xi}) = \boldsymbol{\xi}/\|\boldsymbol{\xi}\|$ and $D(\mathbf{0}) = \mathbf{0}$. An estimator $\hat{\Theta}_i$ of (3.2) is *direction consistent* at rate α_n if for a sequence of positive real numbers $\alpha_n \rightarrow 0$,

$$\mathbb{P}(\|D(\hat{\boldsymbol{\theta}}_{ij}^{[g]}) - D(\boldsymbol{\theta}_{0,ij}^{[g]})\| < \alpha_n, \forall (j, g) \in J(\Theta_{0,i}); \hat{\boldsymbol{\theta}}_{ij}^{[g]} = \mathbf{0}, \forall (j, g) \notin J(\Theta_{0,i})) \rightarrow 1,$$

as $n, p \rightarrow \infty$. In general, direction consistency does not guarantee sign consistency, especially when there are multiple members within one group. However, if the group is selected, all the members within the group are selected, which is sufficient for joint neighborhood selection for each node and subsequent selection of graphs. Motivated by the above idea, we establish the graph selection consistency property of JSEM in Theorem III.2, which can be conveniently modified to adjust for the misspecification in the prior information \mathcal{G} . Before we present the main result, we need more notations.

Consider the group lasso estimation problem (3.2) for node i . For simplicity, we discuss the estimation consistency properties with a common tuning parameter λ for all (j, g) . For $k = 1, \dots, K$, denote $\mathbf{X}_{I_k}^k$ the $n \times |I_k|$ sub-matrix consisting of all relevant variables from the k th model. In other words, for all $j \in I_k$, there exists a group $g \ni k$ such that $(j, g) \in J(\Theta_{0,i})$. Note the dependency of each index set I_k

on i is made implicit here for notational convenience. Further, let $\boldsymbol{\xi}^k \in \mathbb{R}^{|I_k|}$ be a vector indexed by I_k . The following assumption adapts the *Uniform Irrepresentability Condition (IC)* in *Basu et al. (2012)* to our setting:

A3: There exists a positive constant η such that for all $\boldsymbol{\xi} = ((\boldsymbol{\xi}^1)^T, \dots, (\boldsymbol{\xi}^K)^T)^T \in \mathbb{R}^{\sum_k |I_k|}$ with $\max_{(j,g) \in J(\Theta_{0,i})} \|\boldsymbol{\xi}_j^{[g]}\| \leq 1$ and all $(j, g) \notin J(\Theta_{0,i})$

$$\left(\sum_{k \in g} \left[(\mathbf{X}_j^k)^T \mathbf{X}_{I_k}^k \{ (\mathbf{X}_{I_k}^k)^T \mathbf{X}_{I_k}^k \}^{-1} \boldsymbol{\xi}^k \right]^2 \right)^{1/2} \leq 1 - \eta. \quad (3.7)$$

Note the group level constraint (3.7) is required to hold for all p regressions and is less stringent than the IC for the selection consistency of lasso.

Theorem III.2. *Let Assumption A1 with $s = s_0$, A2 and A3 be satisfied. Assume further that the sparsity pattern \mathcal{G} is correctly specified. For every regression defined in (3.2), choose*

$$\lambda \geq \max_{i, (j,g) \notin J(\Theta_{0,i})} \frac{1}{\eta} \frac{1}{\sqrt{nd_0}} \left(\sqrt{|g|} + \frac{\pi}{\sqrt{2}} \sqrt{q \log G_0} \right), \quad (3.8)$$

$$\alpha_n \geq \max_{i, (j,g) \in J(\Theta_{0,i})} \frac{1}{\kappa(s_0)} \frac{1}{\|\boldsymbol{\theta}_{0,ij}^{[g]}\|} \left\{ \lambda \frac{\sqrt{s_0}}{\kappa(s_0)} + \frac{1}{\sqrt{nd_0}} \left(\sqrt{|g|} + \frac{\pi}{\sqrt{2}} \sqrt{q \log G_0} \right) \right\}, \quad (3.9)$$

with $q > 1$. Then with probability at least $1 - 4pG_0^{1-q}$, we have simultaneously for all i

1. $\hat{\boldsymbol{\theta}}_{ij}^{[g]} = \mathbf{0}$, for all $(j, g) \notin J(\Theta_{0,i})$,
2. $\|\hat{\boldsymbol{\theta}}_{ij}^{[g]} - \boldsymbol{\theta}_{0,ij}^{[g]}\| < \alpha_n \|\boldsymbol{\theta}_{0,ij}^{[g]}\|$, and hence $\|D(\hat{\boldsymbol{\theta}}_{ij}^{[g]}) - D(\boldsymbol{\theta}_{0,ij}^{[g]})\| < 2\alpha_n$ for all $(j, g) \in J(\Theta_{0,i})$.

Further, if $\alpha_n < 1$, then with the same probability,

$$\hat{E}^k = \{(i, j) : 1 \leq i < j \leq p, \max(\hat{\theta}_{ij}^k, \hat{\theta}_{ji}^k) \neq 0\} \quad (3.10)$$

estimates correctly the true edge set E_0^k for all $k = 1, \dots, K$.

Note the choice of λ in (3.8) is of the same order as the tuning parameter required for estimation consistency in Theorem III.1. With the above choice of λ , α_n can be chosen to be of the order of $\mathcal{O}(\sqrt{s_0}(\sqrt{|g_{\max}|} + \sqrt{\log G_0})/\sqrt{n})$. A proof of Theorem III.2 can be found in the Section 3.8.

The results in Theorem III.2 are stated under appropriately specified \mathcal{G} . When \mathcal{G} is misspecified, it is possible that not all the members within a group have nonzero effects. However, the group lasso penalty may fail to exclude members with actual zero effect within the misspecified group, leading to the recovery of spurious edges. The following result implies that the property of direction consistency helps identify influential members within a group, i.e. those with noticeable nonzero effects.

Corollary III.3. *Let Assumption A1 with $s = s_0$, A2 and A3 be satisfied. For every regression defined in (3.2), choose λ and α_n as in Theorem III.2. Define*

$$\hat{\theta}_{ij}^{k,thr} = \hat{\theta}_{ij}^k \mathbf{1}\{\hat{\theta}_{ij}^k / \|\hat{\boldsymbol{\theta}}_{ij}^{[g]}\| > 2\alpha_n\}, \quad \forall k \in g, \quad \forall (j, g) \in J(\Theta_{0,i}).$$

If for all $g \in \mathcal{G}$, $\min_{k \in g} \theta_{0,ij}^k / \|\boldsymbol{\theta}_{0,ij}^{[g]}\| > 2\alpha_n$, then with probability at least $1 - 4pG_0^{1-q}$,

$$\hat{E}^{k,thr} = \{(i, j) : 1 \leq i < j \leq p, \max(\hat{\theta}_{ij}^{k,thr}, \hat{\theta}_{ji}^{k,thr}) \neq 0\}$$

estimates correctly the true edge set E_0^k for all $k = 1, \dots, K$.

The result in Corollary III.3 implies immediately that JSEM with an additional thresholding step on the estimated direction vectors $D(\|\hat{\boldsymbol{\theta}}_{ij}^{[g]}\|)$ can be applied to reduce false discoveries and thus improve selection of the edge sets under moderate level of misspecification of the structured pattern \mathcal{G} .

3.4 Performance Evaluation

We present three simulation studies to evaluate the performance of JSEM. The first study considers a single common structure across all graphical models, while the second one features a more complex structured sparsity pattern. Other methods compared include the separate estimation method Glasso, where the *Graphical lasso* by *Friedman et al.* (2008) is applied to each graphical model separately, joint estimation by *Guo et al.* (2011), denoted by JEM-G, and the Joint Graphical Lasso denote by Joint Graphical Lasso (JGL) by *Danaher et al.* (2014). Our results show that JSEM outperforms competing methods in both settings, even when the structured pattern is moderately misspecified. The third study compares JSEM with its thresholded version under misspecified \mathcal{G} using the experimental settings of the first two studies.

3.4.1 Simulation Study 1

In our first simulation, we consider $K = 5$, with each graphical model comprising of $p = 100$ variables. The covariance matrices Σ^k ($k = 1, \dots, K$) are constructed as follows: we first generate a scale-free network with edge set E_0 as the common structure shared across all graphs, shown in the left panel of Figure 3.2. To generate the edge set E^k , we randomly pick a pair of $(i, j), i < j$ such that $(i, j) \notin E_0$ and add it to E^k . This procedure is repeated $\rho|E_0|$ times for each k , where ρ is a positive number corresponding to the ratio of individual edges to common ones. In this example, we take $\rho = 0.1$ to allow a high level of structural similarity across graphs. Thus, all graphical models have the same degree of sparsity, with 108 or 2.2% of all possible edges present. Note that due to the sparse structure of each graph, the proportion of shared non-edges (i.e. common zeros in the adjacency matrices) among all models is 98%. Given the edge set E^k , we then construct the inverse covariance matrix with the nonzero off-diagonal entries in Ω^k being uniformly generated from the $[-1, -0.5] \cup [0.5, 1]$ interval. The positive definiteness of Ω^k is guaranteed by

setting the diagonal elements to be $|\phi_{\min}(\Omega^k)| + 0.1$. The covariance matrix Σ^k is then determined by

$$\Sigma_{ij}^k = (\Omega^k)_{ij}^{-1} / \sqrt{(\Omega^k)_{ii}^{-1}(\Omega^k)_{jj}^{-1}}.$$

By construction, each Σ^k corresponds to the correlation matrix for the k th graphical model. The sparsity pattern supplied for JSEM is $\mathcal{G} = \{1, \dots, K\}$, i.e. assuming all graphical models share the same structure. Thus, the parameter ρ indicates the amount of pattern misspecification as compared to the true edge set structure.

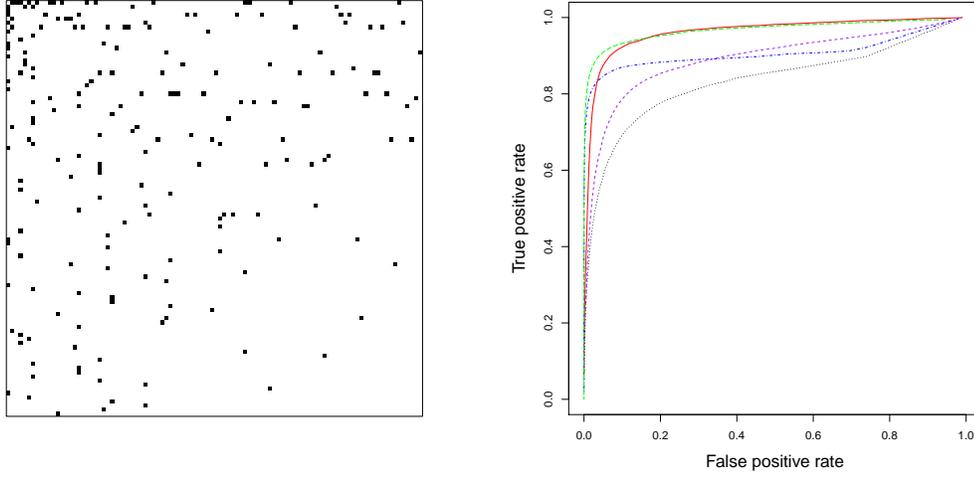


Figure 3.2: Simulation study 1: left panel shows the image plot of the adjacency matrix corresponding to the shared structure across all graphs. Each black cell indicates presence of an edge. The right panel shows the ROC curves for sample size $n_k = 50$: Glasso (dotted), JEM-G (dotdash), GGL (solid), FGL (dashed), JSEM (longdash).

To compare the overall performance of all methods, we generate $n_k = 50$ samples from each $k = 1, \dots, K$ and compute the average false positive and true positive rates of the estimated inverse covariance matrices over a fine grid of tuning parameters from 20 replications. This gives the ROC curves as shown in the right panel of Figure 3.2. JGL provides two options for constraining the similarity among multiple graphical models, i.e. GGL and FGL, corresponding to the group graphical lasso and fused

lasso regularization, respectively. Since each of the two methods in JGL requires two tuning parameters, one for controlling the *sparsity* of individual graph and the other for controlling the *similarity* across all graphs, we compute the ROC curves over a fine grid of the sparsity regularization parameter while fixing the similarity regularization at four different levels (from low to high similarity), and plot the one that has the largest value of area under the curve (AUC). In this simulation, it turns out that GGL performs the best when there is only regularization on the similarity, i.e. a *group lasso* penalty on the same entry across all K inverse covariance matrices, which we expect to behave close to JSEM we propose. In the right panel of Figure 3.2, the ROC curve for GGL falls slightly below that of JSEM. In comparison, FGL does not perform as well. The best curve we get from FGL shows some advantage over the separate estimation Glasso, but mostly falls below curves from other joint estimation methods. JEM-G performs well and is very competitive compared to GGL and JSEM for very low false positive and high true positive rates, but starts falling behind when the false positive rate is greater than 5%. In this example, JSEM performs the best with the highest ROC curve throughout the domain.

Next, we computed the estimators from different methods on a training dataset with $n_k = 50$ samples for each $k = 1, \dots, K$, using the tuning parameters selected by maximizing the log-likelihood of a separate validation dataset generated from the same distribution and of the same size. Results are summarized in table 3.1, which compares the estimated inverse covariance matrices with the population version in the true model based on 50 replications under falsely discovered edges (FP), falsely deleted edges (FN), structural hamming distance (SHD), F_1 score (F1) and Frobenius norm loss (FL). F_1 score measures the accuracy of a test by summarizing information from both FP and FN, where it reaches its best value at 1 and worst at 0. The results indicate that although GGL and FGL are good at identifying true edges (low FN), they tend to produce a high number of false positives. In comparison, the proposed

method JSEM achieves a balance and obtains the highest F_1 score, as well as the lowest Frobenius norm loss. The JEM-G performs slightly worse, but still well above the other three methods.

Table 3.1: Performance of different regularization methods for estimating graphical models in Simulation Study 1: average FP, FN, SHD, F1 and FL (SE) for sample size $n_k = 50$. The best cases are highlighted in bold.

Method	FP	FN	SHD	F1	FL
Glasso	411(9)	36(2)	447(9)	0.24(0.01)	0.63(0.01)
JEM-G	24(3)	29(3)	53(5)	0.75(0.02)	0.32(0.03)
GGL	1482(24)	6(1)	1488(24)	0.12(0.002)	0.55(0.01)
FGL	653(16)	20(2)	674(17)	0.21(0.01)	0.60(0.01)
JSEM	21(5)	24(3)	45(6)	0.79(0.03)	0.27(0.02)

3.4.2 Simulation Study 2

In our second study, we consider a more structured pattern with $K = 10$ graphs. Each graphical model contains $p = 50$ variables. Figure 3.3 shows heatmaps of the adjacency matrices of the 10 models.

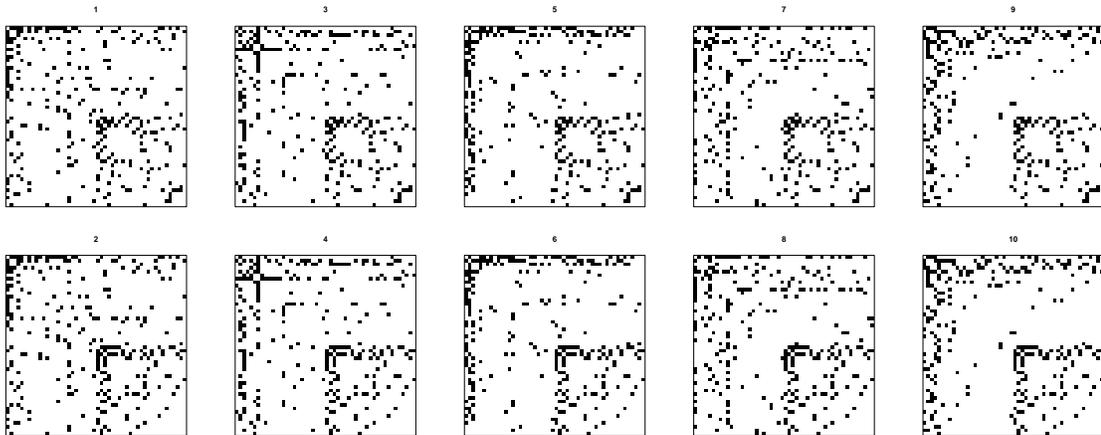


Figure 3.3: Simulation study 2: image plots of the adjacency matrices from all graphical models. Graphs in the same row share the same connectivity pattern at the bottom right block, whereas graphs in the same column share the same pattern at remaining locations.

This pattern is constructed as follows: we first generate the adjacency matrices corresponding to five distinct p -dimensional scale-free networks, so that the adjacency matrices in each column of the plot are the same. Next, we replace the connectivity structure of the bottom right diagonal block of size $p/2$ by $p/2$ within each adjacency matrix with that of another two distinct $p/2$ -dimensional scale-free networks, so that graphical models in each row exhibit the same connectivity pattern, but across rows behave differently in the bottom right block of their adjacency matrices. Note that by replacing the connectivity structure among the second half of nodes, the relationships between the first half and the second half of nodes are also altered. In summary, this structured pattern illustrates how different subsets of the edge set across multiple graphical models can be similar, as well as exhibit differences in their topologies; to the best of our knowledge, such complex relationships have not been studied in the literature. In this setting, the proportion of shared non-edges (common zeros in the precision matrices) among all graphical models is about 60%.

Once the adjacency matrix or equivalently the edge set E^k is constructed, we generate the covariance and inverse covariance matrices similarly to our first simulation study. We also study the effect of misspecification in the input sparsity pattern by varying $\rho = 0, 0.2, 0.4, 0.6, 0.8, 1$.

At each level of pattern misspecification, we generate $n_k = 100$ independent samples for each $k = 1, \dots, K$ and compare the ROC curves from different methods based on 20 replications in Figure 3.4. Again, the ROC curves for GGL and FGL are optimized first with respect to the similarity regularization in terms of AUC. When $\rho < 0.6$, the results show a superior performance of JSEM, since it effectively incorporates available prior information across the various graphical models. JEM-G also yields a reasonably high ROC curve by taking advantage of the shared non-edges among all models. When $\rho \geq 0.6$, JSEM starts suffering from the large amount of pattern misspecification and behaving not much better than even the separate esti-

mation method Glasso, which is the case for other joint estimation methods as well. In all cases, GGL and FGL behave about the same or worse than Glasso, due to the complex edge set structures shared only within subsets of all graphical models.

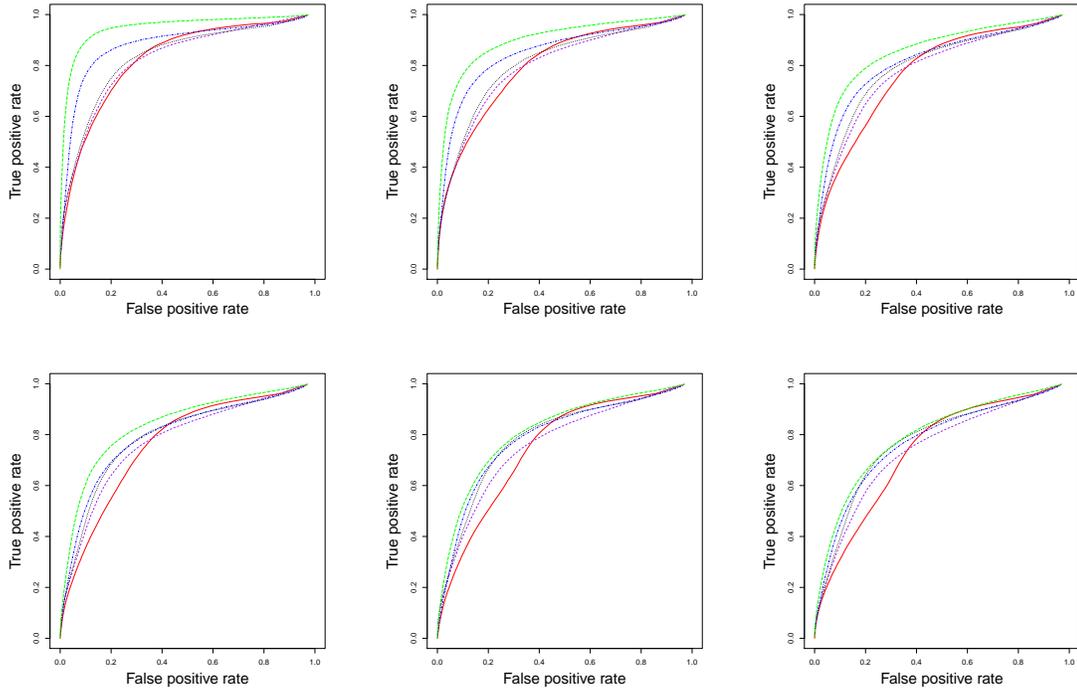


Figure 3.4: Simulation study 2: ROC curves for sample size $n_k = 100$: Glasso (dotted), JEM-G (dotdash), GGL (solid), FGL (dashed), JSEM (long-dash). The misspecification ratio ρ varies from (left to right): 0, 0.2, 0.4 (top row) and 0.6, 0.8, 1 (bottom row).

We then compare the performance of different methods in identifying the true graphs and estimating the inverse covariance matrices at the optimal choice of tuning parameters. Table 3.2 shows the deviance measures between the estimated and the true inverse covariance matrices based on 50 replications for varying levels of pattern misspecification. In all cases, GGL and FGL have low FN, but very high FP, thus resulting in low F_1 scores. For $\rho < 0.6$, JSEM gives a much better control over false positives and yields the highest F_1 score and lowest Frobenius norm loss. JEM-G is also very competitive in controlling false positive edges and comes next in overall

performance. When $\rho \geq 0.6$, the advantage of using a joint estimation method begins to diminish due to the large amount of misspecification and separate estimation is recommended.

Table 3.2: Performance of different regularization methods for estimating graphical models in Simulation Study 2: average FP, FN, SHD, F1 and FL (SE) for sample size $n_k = 100$. The best cases are highlighted in bold.

ρ	Method	FP	FN	SHD	F1	FL
0	Glasso	300(4)	26(1)	326(5)	0.41(0.005)	0.54(0.01)
	JEM-G	120(4)	31(1)	152(4)	0.59 (0.01)	0.38(0.02)
	GGL	640(10)	8 (1)	648(9)	0.29 (0.003)	0.56(0.01)
	FGL	519(9)	16(1)	535(9)	0.31(0.004)	0.61(0.01)
	JSEM	55 (3)	28(2)	83 (4)	0.73 (0.01)	0.30 (0.01)
0.2	Glasso	319(5)	34(1)	352(5)	0.43(0.004)	0.52(0.01)
	JEM-G	168(5)	42(2)	210(5)	0.54(0.01)	0.36(0.01)
	GGL	523(6)	18 (1)	541(6)	0.35(0.003)	0.54(0.01)
	FGL	503(9)	23(1)	526(10)	0.35(0.01)	0.61(0.01)
	JSEM	107 (4)	40(2)	147 (5)	0.63 (0.01)	0.33 (0.01)
0.4	Glasso	302(4)	44(2)	346(5)	0.46(0.01)	0.49(0.01)
	JEM-G	213(6)	53(2)	266(6)	0.51(0.01)	0.37(0.01)
	GGL	536(5)	21 (1)	558(6)	0.38(0.003)	0.50(0.01)
	FGL	490(9)	31(1)	521(9)	0.38(0.005)	0.58(0.01)
	JSEM	165 (5)	49(2)	214 (5)	0.57 (0.01)	0.35 (0.01)
0.6	Glasso	316(4)	49(2)	364(4)	0.49(0.004)	0.47(0.01)
	JEM-G	262(6)	58(2)	319(6)	0.51(0.01)	0.36 (0.01)
	GGL	542(6)	25 (2)	566(6)	0.41(0.003)	0.48(0.01)
	FGL	476(8)	38(2)	514(8)	0.42(0.004)	0.59(0.004)
	JSEM	206 (5)	56(2)	261 (6)	0.56 (0.01)	0.37(0.01)
0.8	Glasso	338(4)	49(2)	387(4)	0.51(0.003)	0.46(0.01)
	JEM-G	263(4)	65(2)	327(5)	0.53(0.01)	0.38 (0.01)
	GGL	589(5)	22 (1)	611(5)	0.43(0.002)	0.49(0.01)
	FGL	466(7)	44(2)	510(7)	0.45(0.004)	0.60(0.004)
	JSEM	240 (4)	64(2)	304 (5)	0.55 (0.01)	0.39(0.01)
1.0	Glasso	331(5)	61(2)	392(5)	0.52(0.005)	0.46(0.01)
	JEM-G	257 (6)	83(2)	340(5)	0.53(0.01)	0.38 (0.01)
	GGL	576(6)	29 (1)	605(6)	0.45(0.003)	0.49(0.01)
	FGL	454(9)	55(2)	509(9)	0.46(0.01)	0.60(0.005)
	JSEM	259(5)	75(2)	334 (6)	0.54 (0.01)	0.40(0.01)

While evaluating the performance in estimating multiple graphical models, we notice that GGL and FGL are very computationally demanding compared to JEM-G

and JSEM; especially FGL due to the fused penalty when the number of models K is large. This might limit their applicability in practice.

3.4.3 Simulation Study 3

Finally, we illustrate how direction consistency helps improve the estimation of graphical models using the previous two experimental settings. Table 3.3 presents the performance of thresholded JSEM when \mathcal{G} is moderately misspecified with individual to common ratio $\rho = 0.3$, based on 50 replications. The tuning parameter λ is chosen via maximum likelihood over a separate validation dataset. At the optimal λ , the within group thresholding parameter $\alpha_n = n^{-0.25}/2$ is again selected via maximum likelihood. Note that we use a larger sample size $n_k = 200$ in both settings to ensure that the Uniform IC required for direction consistency holds. The advantage of thresholding within groups is obvious in both settings, where the thresholded JSEM significantly reduces the number of false positive edges with only a small loss in the presence of false negative edges. One may notice the slight increase in Frobenius norm loss for thresholded JSEM, which is likely due to the increased false negative edges. Nevertheless, the thresholded version of JSEM obtains higher F_1 scores, indicating an overall improvement in the structural estimates of all graphs.

Table 3.3: Performance of JSEM and thresholded JSEM with misspecified groups ($\rho = 0.3$): average FP, FN, SHD, F1 and FL (SE) for sample size $n_k = 200$. The better cases are highlighted in bold.

Method	$K = 5, p = 100$ $\mathcal{G} = \{1, 2, 3, 4, 5\}$					$K = 10, p = 40$ \mathcal{G} as in Figure 3.3				
	FP	FN	SHD	F1	FL	FP	FN	SHD	F1	FL
JSEM	76(6)	15(1)	91(6)	0.71	0.15	51(3)	4(0.4)	55(3)	0.71	0.19
ThJSEM	32(4)	21(1)	54(4)	0.80	0.16	36(2)	6(0.5)	41(2)	0.76	0.20

3.5 Application to Climate Modeling

To illustrate the performance of our joint estimation method in inferring real-world networks, we apply JSEM on a climate dataset to study climate forcing at multiple locations in North America. Recent assessments from the Intergovernmental Panel on Climate Change (IPCC, *Stocker et al.*, 2013) indicate multiple lines of evidence for climate change in the past century and these changes have caused significant impacts on natural and human systems. One common approach towards understanding the climate system has been attribution studies of detected changes to internal and external forcing mechanisms (such as solar radiation, greenhouse gases, etc.) using simulated climate models. *Lozano et al.* (2009) used spatial-temporal modeling to study the attribution of climate forcing mechanisms from observed data. In this work, we provide an alternative to learn the complex interactions among climate forcing factors exhibited across different climate zones based on observed data.

The data we use in this study data come from multiple sources and are collected under different resolutions for varying lengths of time periods. Specifically, the sources we consider include:

- (1) CRU: Climate Research Unit provides monthly climatology data (<http://www.cru.uea.ac.uk/cru/data>) for 10 surface variables including mean temperature (TMP), diurnal temperature range (DTR), maximum temperature (TMX), minimum temperature (TMN), precipitation (PRE), vapor pressure (VAP), cloud cover (CLD), rainday counts (WET), potential evapotranspiration (PET) and frost days (FRS) from 1901 to 2013 at the 0.5 degree latitude and longitude resolution. Note these high-resolution gridded datasets are constructed using not only directly observed data, but also derived and estimated values with well-known formulae wherever the observed data are not available (see details in *Harris et al.* (2014)).

- (2) NASA: The Goddard Earth Sciences Data and Information Services Center (GES DISC) from the National Aeronautics and Space Administration (NASA) has collected aerosol measurements using Moderate Resolution Imaging Spectroradiometer (MODIS) on satellites. The dataset obtained from Terra satellite consists of monthly average aerosol optical depth (AER) at the 1 degree latitude by 1 degree longitude resolution from March 2000 to August 2014.
- (3) NCDC: The National Solar Radiation Database (NSRDB) 1991-2010 (a collaborative project between The National Renewable Energy Laboratory (NREL) and the National Climatic Data Center (NCDC)) provides statistical summaries for solar data (<ftp://ftp.ncdc.noaa.gov/pub/data/nsrdb-solar/>) from 860 different locations across the United States. The locations are recorded using their latitude, longitude and altitude. We used measurements for global horizontal solar radiation (SOL) at 242 class I stations that have high-quality data.
- (4) NOAA: The climate data center of National Oceanic and Atmospheric Administration (NOAA) has archived the trace gases data, including carbon dioxide (CO₂), carbon monoxide (CO), methane (CH₄) and hydrogen (H₂), from 170 worldwide stations (<http://www.esrl.noaa.gov/gmd/dv/ftpdata.html>). These datasets consist of measurements spanning different time periods, with CO₂ ranging from 1968 to 2013 (the longest) and H₂ from 1992 to 2005 (the shortest). In addition, they come with relatively low resolution compared to other variables due to the limited number of stations.

To ensure compatibility and consistency among multiple data sources, we performed the following pre-processing:

- (1) Normalization: We first transformed each dataset into monthly observations in a standard format including longitude, latitude, altitude (when available),

date, variable, value, unit, and source. We focus on a 54-month time period from January 2001 to June 2005 where data for all variables are available.

- (2) Interpolation and smoothing: We interpolated the monthly data from NCDC and NOAA onto a common 2.5 by 2.5 degree grid for North America using thin plate splines. Since the data from CRU and NASA were provided for a finer resolution grid, thin plate splines were used to first interpolate the data onto a grid of the same resolution as the source data. Then we performed spatial averaging to get data on the common 2.5 by 2.5 degree grid.
- (3) Seasonality and autocorrelation: We reduced the short-term autocorrelation by aggregating the time series for each variable at each location into bins of 3-month intervals and taking first differences on the quarterly data. The resulting data, which consists of 17 measurements, are assumed to be independent samples for the corresponding variable at the specified location.

Next, we randomly select $K = 27$ locations spanning all types of climate from the 2.5 by 2.5 degree grid of North America (see Figure 3.5). This gives us an $n \times p$ matrix at each of the 27 locations, corresponding to $n = 17$ observations for the $p = 16$ variables on climate forcing. At each location, the conditional dependency network is of dimension $p \times p$, which has $16 \times 15/2 = 120$ edges to be inferred.

Our goal is to infer the conditional dependency networks for all locations simultaneously based on available spatial information, obtained from the classification of climate zones in *Kottek et al.* (2006). Specifically, it is assumed that AER and SOL have one common connectivity pattern with other variables in the geographical south of North America and another common pattern in the north. The definition of the south and north is given in Figure 3.5. Variables on greenhouse gases (CO₂, CO, CH₄ and H₂) are assumed to interact with other variables (except AER and SOL) in the same fashion within each of the four climate groups, i.e. midlatitude desert, semiarid

steppe, humid subtropical and humid continental. The connectivity patterns among all remaining variables are assumed to be the same within each of the six distinct climate zones in Figure 3.5.

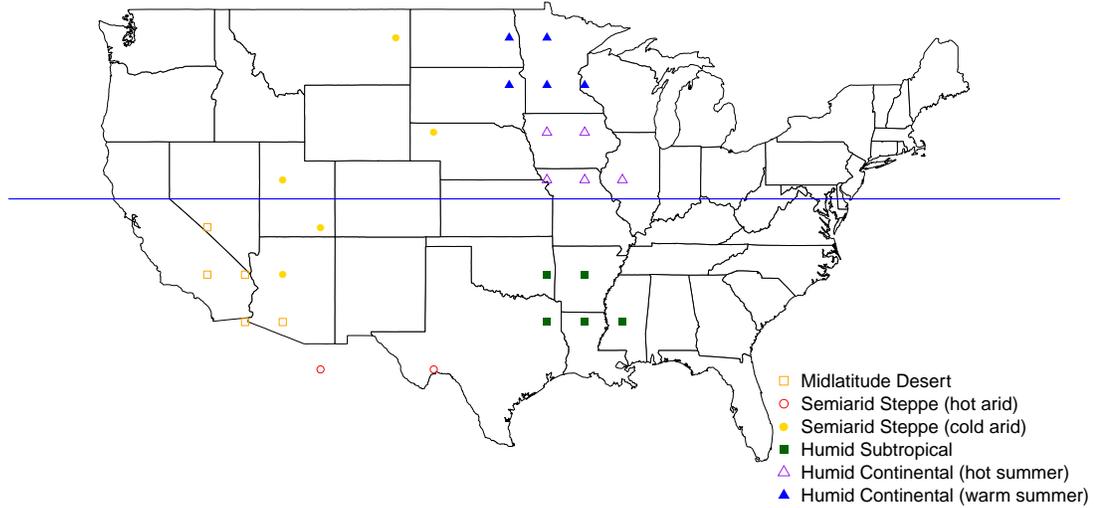


Figure 3.5: The selected 27 locations based on climate classification. The solid line separates the south and north of North America and corresponds to latitude 39 N.

Since there is no separate validation data available, we used BIC on the normalized data to select the tuning parameter λ for the proposed JSEM. At the optimal λ , we applied our method coupled with complementary pairs stability selection (*Shah and Samworth, 2012*) to identify the interaction networks at the 27 locations. To perform stability selection, we run our method 50 times on two randomly drawn complementary pairs of size 8 and 9, and kept only edges that are selected over 70% of the time.

Figure 3.6 shows the estimated networks at the six distinct climate zones. Although we do not impose the assumption on sharing of a single common structure across all locations, there are common edges (solid) identified for all climate zones, reflecting key features of climate forcing regardless of the location. Such relationships

are consistent with how the corresponding climate forcing variables are defined as well as how the data are collected (*Harris et al.*, 2014). The Midlatitude and Semi-arid Steppe climate zones have an edge between DTR and CLD, indicating that they are correlated conditional on all other variables. Similar relationships have also been found over drier regions in *Zhou et al.* (2009). In addition, we notice that the inferred networks at neighboring climate zones are more similar, such as Semi-arid Steppe (hot arid and cold arid), or Humid Continental (hot summer and warm summer), whereas those with dramatically different climate show significantly different connectivity patterns. These common and individual interactions can prove critical in understanding the mechanisms of climate forcing, and facilitate decision making in maintaining the best environmental results.

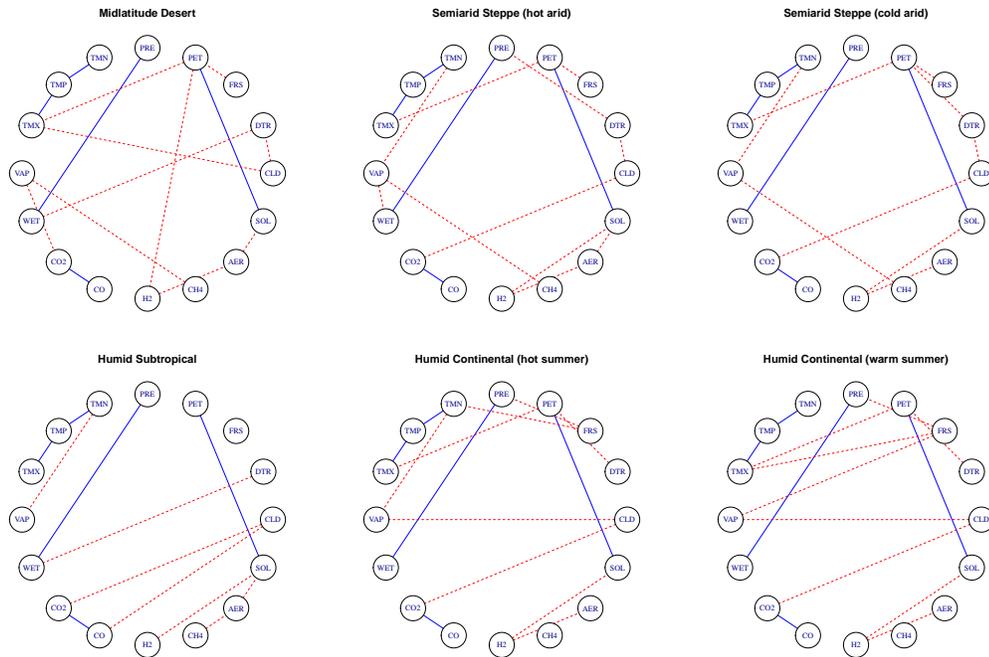


Figure 3.6: Estimated climate networks at the six distinct climate zones using JSEM, with edges shared across all locations solid and differential edges dashed.

As a comparison, we also applied other joint estimation methods JEM-G and GGL on the same data. Here we do not present the result from FGL due to the

extremely high computational cost caused by the fused penalty with a large K .

For each of JEM-G and GGL, we used BIC on the normalized data to select the optimal tuning parameters and coupled each method with complementary pairs stability selection (*Shah and Samworth, 2012*) to infer the related climate networks. As in the case of JSEM, we run each method 50 times on two randomly drawn complementary pairs of size 8 and 9 and kept only edges that are selected above a certain threshold. The selection probability used for JSEM is 70%. However, as the two simulation studies both indicate JEM-G and GGL tend to produce higher false positives, especially GGL, we increased the probability threshold for JEM-G and GGL to 90% and 100%, respectively. The results are shown in Figure 3.7 and 3.8.

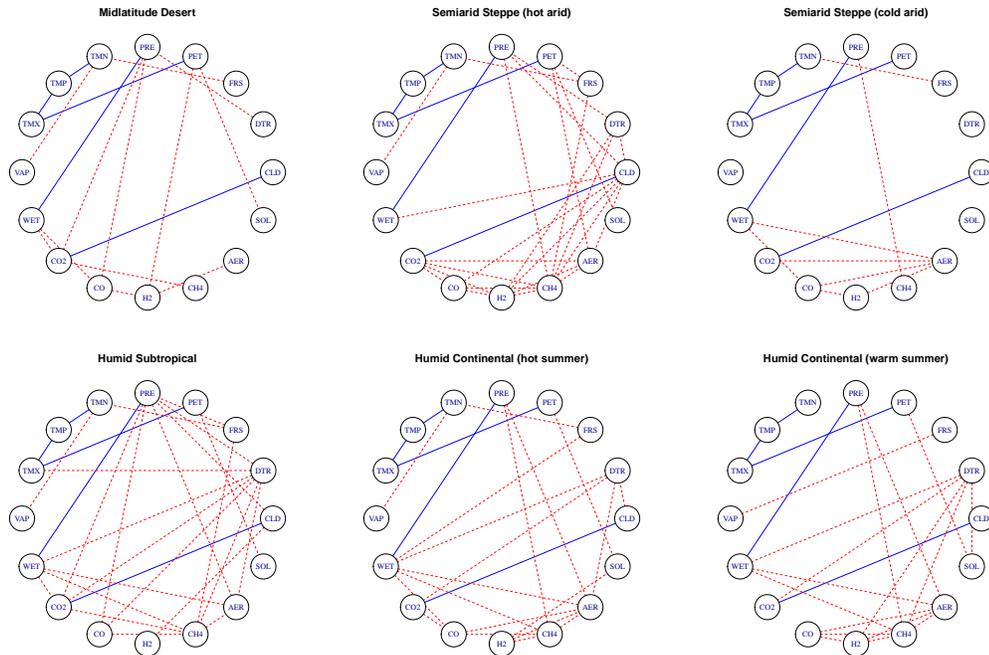


Figure 3.7: Estimated climate networks at the six distinct climate zones using JEM-G, with edges shared across all locations solid and differential edges dashed.

One can see clearly that the estimated networks between JEM-G and GGL exhibit quite different connectivity patterns from those inferred from JSEM. In particular, the results from GGL seem to suggest strong conditional dependence structure among

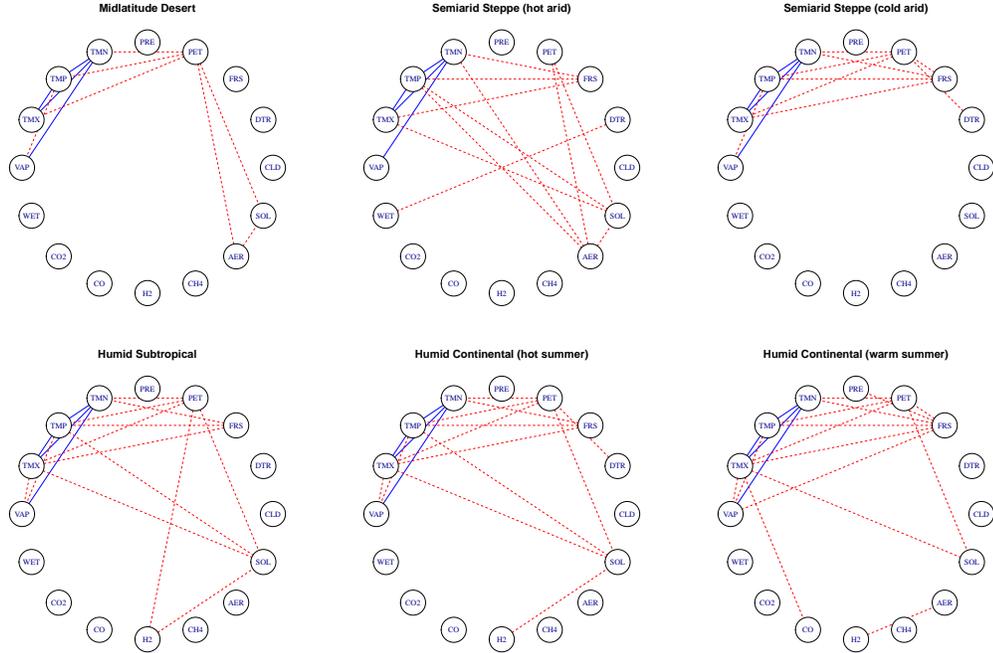


Figure 3.8: Estimated climate networks at the six distinct climate zones using GGL, with edges shared across all locations solid and differential edges dashed.

a subset of variables, which distinguishes itself from JEM-G and JSEM. On the other hand, the results from JEM-G and JSEM are more similar. For example, common edges identified using JEM-G, such as TMN–TMP, TMP–TMX, PRE–WET, also show up under JSEM. The common edge between CLD and CO₂ is found at all locations except Midlatitude Desert under JSEM, whereas the edge between PET and SOL identified using JSEM exists everywhere except at Semiarid Steppe (cold arid) under JEM-G. Note although JEM-G does not require external information on the structural relationships across graphs, the inferred networks respect roughly the spatial pattern of all climate zones. For instance, Humid Continental (hot summer and cool summer) are more similar.

3.6 Discussion

This work introduced a joint structural estimation method that incorporates *a priori* known structural relationships between multiple graphical models. Under appropriately specified sparsity patterns, the proposed method borrows information across models wherever there is sharing of structures or substructures, leading to improved performance in network estimation. Further, when the structured sparsity pattern is moderately misspecified, we establish that an additional step of hard thresholding on the estimated groups of coefficients obtained from the penalized regression modeling employed helps control the type-I error introduced by the misspecification. In practice, if not all entry-wise structural relationships across multiple graphical models are available, it is recommended to add restrictions at mainly edge pairs that are likely to share the same structures instead of providing a highly misspecified structured sparsity pattern. Therefore, the proposed method works well in situations where there is a large number of graphical models, but external similarity information is available only for sub-components of the models.

3.7 Proof of Theorem III.1

For convenience, we shall use \sum_k as a short notation for $\sum_{k=1}^K$ throughout the proof when it is clear.

The first lemma is borrowed from (*Basu et al.*, 2012, Lemma A.2). We state the result here for completeness. Please refer to their paper for proof of the lemma.

Lemma III.4. *Let $Z_{k \times 1} \sim \mathcal{N}(0, \Sigma)$. Then for any $t > 0$, the following inequalities hold:*

$$\mathbb{P}(\|Z\| - \mathbb{E}\|Z\| > t) \leq 2 \exp\left(-\frac{2t^2}{\pi^2 \|\Sigma\|}\right), \quad \mathbb{E}\|Z\| \leq \sqrt{k} \sqrt{\|\Sigma\|}.$$

To prove the rate of convergence in Theorem III.1, we look at three key steps: nodewise regression in subsection 3.7.1, selecting the edge set in 3.7.2 and maximum

likelihood refitting in 3.7.3.

3.7.1 Regression

For $j \neq i, g \in \mathcal{G}^{ij}, k \in g$, let $\varepsilon_i^k = \mathbf{X}_i^k - \sum_{j \neq i} \theta_{0,ij}^k \mathbf{X}_j^k$. Denote $\zeta_{ij}^k = \langle \varepsilon_i^k, \mathbf{X}_j^k \rangle / n$ and $\boldsymbol{\zeta}_{ij}^{[g]} = (\zeta_{ij}^k)_{k \in g} \in \mathbb{R}^{|g|}$. Consider the random event $\mathcal{A} = \bigcap_{i,j \neq i,g} \mathcal{A}_{ij}^g$, where $\mathcal{A}_{ij}^g = \{2\|\boldsymbol{\zeta}_{ij}^{[g]}\| \leq \lambda_{ij}^g\}$. The next lemma provides a concentration bound for the random event \mathcal{A} used in the proof of Theorem III.1.

Lemma III.5. *Consider the random event $\mathcal{A} = \bigcap_{i,j \neq i,g} \mathcal{A}_{ij}^g$, where $\mathcal{A}_{ij}^g = \{2\|\boldsymbol{\zeta}_{ij}^{[g]}\| \leq \lambda_{ij}^g\}$. For each combination of $(i, j \neq i, g)$, choose*

$$\lambda_{ij}^g \geq \max_{k \in g} \frac{2}{\sqrt{n\omega_{0,ii}^k}} \left(\sqrt{|g|} + \frac{\pi}{\sqrt{2}} \sqrt{q \log G_0} \right). \quad (3.11)$$

where $q > 1$ and G_0 is the maximum number of groups in all regressions. Then

$$\mathbb{P}(\mathcal{A}) \geq 1 - 2pG_0^{1-q}.$$

Proof of Lemma III.5. By Bonferroni inequality, $\mathbb{P}(\mathcal{A}^c) \leq \sum_{i,j \neq i,g} \mathbb{P}(\{\mathcal{A}_{ij}^g\}^c)$. For any 3-tuple of $(i, j \neq i, g)$, it suffices to find an upper bound for $\mathbb{P}(\{\mathcal{A}_{ij}^g\}^c)$. Denote $\boldsymbol{\Psi}_j^k = (\mathbf{X}_j^k)^T \mathbf{X}_j^k / n$ and $\boldsymbol{\Phi}_j^k = \mathbf{X}_j^k (\mathbf{X}_j^k)^T / n$, both of rank 1. The eigendecomposition of $\boldsymbol{\Phi}_j^k$ is $\boldsymbol{\Phi}_j^k = \mathbf{Q}^k \mathbf{V}^k (\mathbf{Q}^k)^T$, where \mathbf{Q}^k is the orthogonal matrix whose columns are the eigenvectors of $\boldsymbol{\Phi}_j^k$ and \mathbf{V}^k is the diagonal matrix whose diagonal elements are the corresponding eigenvalues. It is clear that the only non-zero eigenvalue of $\boldsymbol{\Phi}_j^k$ is given by $\gamma_j^k = \|\mathbf{X}_j^k\|^2 / n = 1$. Let Q_1^k be the eigenvector corresponding to γ_j^k . Therefore

$$\begin{aligned} \|\boldsymbol{\zeta}_{ij}^{[g]}\|^2 &= \sum_{k \in g} (\zeta_{ij}^k)^2 = \sum_{k \in g} \frac{1}{n^2} (\varepsilon_i^k)^T \mathbf{X}_j^k (\mathbf{X}_j^k)^T \varepsilon_i^k = \frac{1}{n} \sum_{k \in g} (\varepsilon_i^k)^T \mathbf{Q}^k \mathbf{V}^k (\mathbf{Q}^k)^T \varepsilon_i^k, \\ &= \frac{1}{n} \sum_{k \in g} (\varepsilon_i^k)^T Q_1^k \gamma_j^k (Q_1^k)^T \varepsilon_i^k = \frac{1}{n} \|Z^{[g]}\|^2, \end{aligned}$$

where $Z^{[g]} = (Z^k)_{k \in g}$ with $Z^k = (Q_1^k)^T \varepsilon_i^k$. By definition of ε_i^k , $\text{Var}(Z^k) = 1/\omega_{0,ii}^k$ and $\text{Var}(Z^{[g]})$ is a diagonal matrix with the diagonal $(1/\omega_{0,ii}^k)_{k \in g}$. Note that the independence of Z^k and $Z^{k'}$ ($k \neq k'$) comes from the fact that ε_i^k and $\varepsilon_i^{k'}$ are independent. Therefore

$$\mathbb{P}(\{\mathcal{A}_{ij}^g\}^c) = \mathbb{P}(\|Z^{[g]}\|/\sqrt{n} > \lambda_{ij}^g/2) = \mathbb{P}(\|Z^{[g]}\| - \mathbb{E}\|Z^{[g]}\| > \sqrt{n}\lambda_{ij}^g/2 - \mathbb{E}\|Z^{[g]}\|).$$

Applying Lemma III.4,

$$\begin{aligned} \mathbb{P}(\{\mathcal{A}_{ij}^g\}^c) &\leq \mathbb{P}(\|\|Z^{[g]}\| - \mathbb{E}\|Z^{[g]}\|\| > \sqrt{n}\lambda_{ij}^g/2 - \mathbb{E}\|Z^{[g]}\|) \\ &\leq 2 \exp \left\{ -\frac{2}{\pi^2 \|\text{Var}(Z^{[g]})\|} \left(\frac{\sqrt{n}\lambda_{ij}^g}{2} - \mathbb{E}\|Z^{[g]}\| \right)^2 \right\}. \end{aligned}$$

Choose λ_{ij}^g such that the right-hand side of above inequality is less than $2G_0^{-q}$ for some positive parameter q . Then

$$\lambda_{ij}^g \geq \frac{2}{\sqrt{n}} \left(\mathbb{E}\|Z^{[g]}\| + \frac{\pi}{\sqrt{2}} \sqrt{q \log G_0} \sqrt{\|\text{Var}(Z^{[g]})\|} \right),$$

and is satisfied if

$$\lambda_{ij}^g \geq \max_{k \in g} \frac{2}{\sqrt{n\omega_{0,ii}^k}} \left(\sqrt{|g|} + \frac{\pi}{\sqrt{2}} \sqrt{q \log G_0} \right).$$

With the above choice of λ_{ij}^g ,

$$\mathbb{P}(\mathcal{A}^c) \leq \sum_{i=1}^p \sum_{j \neq i} \sum_{g \in \mathcal{G}^{ij}} \mathbb{P}(\{\mathcal{A}_{ij}^g\}^c) \leq 2pG_0^{1-q},$$

or equivalently, $\mathbb{P}(\mathcal{A}) \geq 1 - 2pG_0^{1-q}$. □

By Lemma III.5, if we choose λ_{ij}^g as

$$\lambda_{ij}^g \geq \max_{k \in g} \frac{2}{\sqrt{n\omega_{0,ii}^k}} \left(\sqrt{|g|} + \frac{\pi}{\sqrt{2}} \sqrt{q \log G_0} \right) \quad (3.12)$$

with $q > 1$, then $\mathbb{P}(\mathcal{A}) \geq 1 - 2pG_0^{1-q}$. The next theorem establishes oracle bounds for $\hat{\Theta}_i - \Theta_{0,i}$ under the chosen λ_{ij}^g .

Theorem III.6. *For $i = 1, \dots, p$, consider the problem (3.2) and choose λ_{ij}^g as in (3.12). Let $\hat{\Theta}_i$ be the solution to problem (3.2). If Assumption A1 holds with $\kappa^2 = \kappa^2(s_0)$, then for any solution $\hat{\Theta}_i$ of problem (3.2), we have on the event \mathcal{A}*

$$\sum_{j \neq i, g \in \mathcal{G}^{ij}} \|\hat{\theta}_{ij}^{[g]} - \theta_{0,ij}^{[g]}\| \leq \frac{16}{\kappa^2 \lambda_{\min}} \sum_{(j,g) \in J(\Theta_{0,i})} (\lambda_{ij}^g)^2, \quad (3.13)$$

$$\mathcal{M}(\hat{\Theta}_i) \leq \frac{64\phi_{\max}}{\kappa^2 \lambda_{\min}^2} \sum_{(j,g) \in J(\Theta_{0,i})} (\lambda_{ij}^g)^2, \quad (3.14)$$

where $\lambda_{\min} = \min_{i,j \neq i, g \in \mathcal{G}^{ij}} \lambda_{ij}^g$, $\mathcal{M}(\hat{\Theta}_i) = |J(\hat{\Theta}_i)|$ and ϕ_{\max} is the maximal eigenvalue of $(\mathbf{X}^k)^T \mathbf{X}^k / n$ for all $k = 1, \dots, K$. If, in addition, Assumption A1 holds with $\kappa^2(2s_0)$, then for any solution $\hat{\Theta}_i$ of problem (3.2) we have that

$$\|\hat{\Theta}_i - \Theta_{0,i}\|_F \leq \frac{4\sqrt{10}}{\kappa^2(2s_0)} \frac{\sum_{(j,g) \in J(\Theta_{0,i})} (\lambda_{ij}^g)^2}{\lambda_{\min} \sqrt{s_i}}. \quad (3.15)$$

Remark III.7. By Assumption A2, $\omega_{0,ii}^k \geq \phi_{\min}(\Omega_0^k) = \phi_{\max}^{-1}(\Sigma_0^k) \geq d_0$ for all i, k . Thus, (3.12) implies that we can choose $\lambda_{ij}^g = \lambda_{\max}$ as

$$\lambda_{\max} = \frac{2}{\sqrt{nd_0}} \left(\sqrt{|g_{\max}|} + \frac{\pi}{\sqrt{2}} \sqrt{q \log G_0} \right), \quad (3.16)$$

with $q > 1$ for all 3-tuples (i, j, g) . Then we can rewrite the oracle inequalities in

(3.14) and (3.15) as

$$\mathcal{M}(\hat{\Theta}_i) \leq \frac{64\phi_{\max}}{\kappa^2} s_i, \quad (3.17)$$

$$\|\hat{\Theta}_i - \Theta_{0,i}\|_F \leq \frac{8\sqrt{10}}{\kappa^2(2s_0)\sqrt{d_0}} \left(\sqrt{|g_{\max}|} + \frac{\pi}{\sqrt{2}} \sqrt{q \log G_0} \right) \sqrt{\frac{s_i}{n}}. \quad (3.18)$$

Proof of Theorem III.6. For all $\Theta_i \in \mathbb{M}(p-1, K)$, by an adaptive argument of Lemma 3.1 of *Lounici et al.* (2011), it is straightforward to verify the following:

$$\begin{aligned} & \sum_{k=1}^K \frac{1}{n} \|\mathbf{X}_{-i}^k(\hat{\boldsymbol{\theta}}_i^k - \boldsymbol{\theta}_{0,i}^k)\|^2 + \sum_{j \neq i} \sum_{g \in \mathcal{G}^{ij}} \lambda_{ij}^g \|\hat{\boldsymbol{\theta}}_{ij}^{[g]} - \boldsymbol{\theta}_{ij}^{[g]}\| \\ & \leq \sum_{k=1}^K \frac{1}{n} \|\mathbf{X}_{-i}^k(\boldsymbol{\theta}_i^k - \boldsymbol{\theta}_{0,i}^k)\|^2 + 4 \sum_{(j,g) \in J(\theta_i)} \lambda_{ij}^g \min \left(\|\boldsymbol{\theta}_{ij}^{[g]}\|, \|\hat{\boldsymbol{\theta}}_{ij}^{[g]} - \boldsymbol{\theta}_{ij}^{[g]}\| \right), \end{aligned} \quad (3.19)$$

$$\left\{ \sum_{k \in g} \langle n^{-1} \mathbf{X}_j^k, \mathbf{X}_{-i}^k(\hat{\boldsymbol{\theta}}_i^k - \boldsymbol{\theta}_{0,i}^k) \rangle^2 \right\}^{1/2} \leq \frac{3\lambda_{ij}^g}{2}, \quad (3.20)$$

$$\mathcal{M}(\hat{\Theta}_i) \leq \frac{4\phi_{\max}}{\lambda_{\min}^2} \sum_{k=1}^K \frac{1}{n_k} \|\mathbf{X}_{-i}^k(\hat{\boldsymbol{\theta}}_i^k - \boldsymbol{\theta}_{0,i}^k)\|^2, \quad (3.21)$$

where λ_{\min} and ϕ_{\max} are defined in Theorem III.6.

Let Δ be a matrix in $\mathbb{M}(p, K)$ such that $\delta_j^k = \hat{\boldsymbol{\theta}}_{ij}^k - \boldsymbol{\theta}_{0,ij}^k$ for $j \neq i$ and $\delta_i^k = 0$ for all k . We would like to first find an upper bound for B^2 , where

$$B^2 := \sum_k \frac{1}{n} \|\mathbf{X}_{-i}^k(\hat{\boldsymbol{\theta}}_i^k - \boldsymbol{\theta}_{0,i}^k)\|^2 = \sum_k \frac{1}{n} \|\mathbf{X}^k \boldsymbol{\delta}^k\|^2.$$

By the inequality (3.19) with $\Theta_i = \Theta_{0,i}$, we have, on the event \mathcal{A} , that

$$\sum_{j \neq i} \sum_{g \in \mathcal{G}^{ij}} \lambda_{ij}^g \|\boldsymbol{\delta}_j^{[g]}\| \leq B^2 + \sum_{j \neq i} \sum_{g \in \mathcal{G}^{ij}} \lambda_{ij}^g \|\boldsymbol{\delta}_j^{[g]}\| \leq 4 \sum_{(j,g) \in J(\Theta_{0,i})} \lambda_{ij}^g \|\boldsymbol{\delta}_j^{[g]}\|. \quad (3.22)$$

Therefore

$$\sum_{(j,g) \in J(\Theta_{0,i})^c} \lambda_{ij}^g \|\boldsymbol{\delta}_j^{[g]}\| \leq 3 \sum_{(j,g) \in J(\Theta_{0,i})} \lambda_{ij}^g \|\boldsymbol{\delta}_j^{[g]}\|,$$

and $\Delta \in \mathcal{F}$, the restricted set defined in Assumption A1. Under Assumption A1 with $\kappa = \kappa(s_0)$,

$$B^2 \geq \kappa^2 \|\Delta_J\|_F^2 = \kappa^2 \sum_{(j,g) \in J(\Theta_{0,i})} \|\boldsymbol{\delta}_j^{[g]}\|^2. \quad (3.23)$$

Combing (3.22), (3.23) and Cauchy-Schwarz inequality, we have

$$\begin{aligned} B^2 &\leq 4 \sum_{(j,g) \in J(\Theta_{0,i})} \lambda_{ij}^g \|\boldsymbol{\delta}_j^{[g]}\| \leq 4 \left\{ \sum_{(j,g) \in J(\Theta_{0,i})} (\lambda_{ij}^g)^2 \right\}^{1/2} \left(\sum_{(j,g) \in J(\Theta_{0,i})} \|\boldsymbol{\delta}_j^{[g]}\|^2 \right)^{1/2} \\ &\leq 4 \left\{ \sum_{(j,g) \in J(\Theta_{0,i})} (\lambda_{ij}^g)^2 \right\}^{1/2} \frac{B}{\kappa}, \end{aligned} \quad (3.24)$$

or equivalently

$$B^2 = \sum_k \frac{1}{n} \|\mathbf{X}_{-i}^k (\hat{\boldsymbol{\theta}}_i^k - \boldsymbol{\theta}_{0,i}^k)\|^2 \leq \frac{16}{\kappa^2} \sum_{(j,g) \in J(\Theta_{0,i})} (\lambda_{ij}^g)^2. \quad (3.25)$$

For the inequality (3.13), by (3.22), Cauchy-Schwarz inequality and (3.25),

$$\begin{aligned} \sum_{j \neq i} \sum_{g \in \mathcal{G}^{ij}} \|\boldsymbol{\delta}_j^{[g]}\| &\leq \frac{1}{\lambda_{\min}} \sum_{j \neq i} \sum_{g \in \mathcal{G}^{ij}} \lambda_{ij}^g \|\boldsymbol{\delta}_j^{[g]}\| \leq \frac{4}{\lambda_{\min}} \sum_{(j,g) \in J(\Theta_{0,i})} \lambda_{ij}^g \|\boldsymbol{\delta}_j^{[g]}\| \\ &\leq \frac{4}{\lambda_{\min}} \left\{ \sum_{(j,g) \in J(\Theta_{0,i})} \|\boldsymbol{\delta}_j^{[g]}\|^2 \right\}^{1/2} \left\{ \sum_{(j,g) \in J(\Theta_{0,i})} (\lambda_{ij}^g)^2 \right\}^{1/2} \\ &\leq \frac{4}{\lambda_{\min}} \frac{B}{\kappa} \left\{ \sum_{(j,g) \in J(\Theta_{0,i})} (\lambda_{ij}^g)^2 \right\}^{1/2} \\ &\leq \frac{16}{\kappa^2 \lambda_{\min}} \sum_{(j,g) \in J(\Theta_{0,i})} (\lambda_{ij}^g)^2. \end{aligned}$$

(3.14) follows readily from (3.21) and (3.25)

$$\mathcal{M}(\hat{\Theta}_i) \leq \frac{4\phi_{\max}}{\lambda_{\min}^2} B^2 \leq \frac{64\phi_{\max}}{\kappa^2 \lambda_{\min}^2} \sum_{(j,g) \in J(\Theta_{0,i})} (\lambda_{ij}^g)^2.$$

Finally, we prove (3.15). Let $J_0 = J(\Theta_{0,i})$ and J_1 denote the set of indices in J_0^c corresponding to the s_i largest values of $\lambda_{ij}^g \|\delta_j^{[g]}\|$. The dependence of J_0 and J_1 on i is made implicit here for clarity. Let $J_{01} = J_0 \cup J_1$. So $|J_{01}| \leq 2s_i$. Let (j_ℓ, g_ℓ) be the index of the ℓ th largest element of the set $\{\lambda_{ij}^g \|\delta_j^{[g]}\| : (j, g) \in J_0^c\}$. Then

$$\lambda_{ij_\ell}^{g_\ell} \|\Delta_{ij_\ell}^{[g_\ell]}\| \leq \sum_{(j,g) \in J_0^c} \frac{\lambda_{ij}^g \|\delta_j^{[g]}\|}{\ell}.$$

Combining with the fact that $\Delta \in \mathcal{F}$, we have on the event \mathcal{A} ,

$$\begin{aligned} \sum_{(j,g) \in J_{01}^c} \left(\lambda_{ij}^g \|\delta_j^{[g]}\| \right)^2 &\leq \sum_{(j,g) \in J_0^c} \left(\lambda_{ij}^g \|\delta_j^{[g]}\| \right)^2 \leq \sum_{\ell=s_i+1}^{\infty} \frac{\left(\sum_{(j,g) \in J_0^c} \lambda_{ij}^g \|\delta_j^{[g]}\| \right)^2}{\ell^2} \\ &\leq \frac{1}{s_i} \left(\sum_{(j,g) \in J_0^c} \lambda_{ij}^g \|\delta_j^{[g]}\| \right)^2 \leq \frac{9}{s_i} \left(\sum_{(j,g) \in J_0} \lambda_{ij}^g \|\delta_j^{[g]}\| \right)^2 \\ &\leq \frac{9}{s_i} \sum_{(j,g) \in J_0} (\lambda_{ij}^g)^2 \|\Delta_{J_0}\|_F^2 \leq \frac{9}{s_i} \sum_{(j,g) \in J_0} (\lambda_{ij}^g)^2 \|\Delta_{J_{01}}\|_F^2. \end{aligned}$$

It follows immediately that

$$\lambda_{\min}^2 \sum_{(j,g) \in J_{01}^c} \|\delta_j^{[g]}\|^2 \leq \frac{9}{s_i} \sum_{(j,g) \in J_0} (\lambda_{ij}^g)^2 \|\Delta_{J_{01}}\|_F^2.$$

Hence

$$\begin{aligned}
\|\hat{\Theta}_i - \Theta_{0,i}\|_F^2 &= \sum_{j \neq i} \sum_{g \in \mathcal{G}^{ij}} \|\delta_j^{[g]}\|^2 = \|\Delta_{J_{01}}\|_F^2 + \|\Delta_{J_{01}^c}\|_F^2 \\
&\leq \|\Delta_{J_{01}}\|_F^2 + \frac{9}{s_i \lambda_{\min}^2} \sum_{(j,g) \in J_0} (\lambda_{ij}^g)^2 \|\Delta_{J_{01}}\|_F^2 \\
&\leq \frac{10}{s_i \lambda_{\min}^2} \sum_{(j,g) \in J_0} (\lambda_{ij}^g)^2 \|\Delta_{J_{01}}\|_F^2.
\end{aligned} \tag{3.26}$$

On the other hand, (3.24) implies that

$$B^2 \leq 4 \left\{ \sum_{(j,g) \in J_0} (\lambda_{ij}^g)^2 \right\}^{1/2} \|\Delta_{J_0}\|_F \leq 4 \left\{ \sum_{(j,g) \in J_0} (\lambda_{ij}^g)^2 \right\}^{1/2} \|\Delta_{J_{01}}\|_F.$$

Under Assumption A1 with $s = 2s_0$, we have

$$B^2 \geq \kappa^2(2s_0) \|\Delta_{J_{01}}\|_F^2.$$

So

$$\|\Delta_{J_{01}}\|_F^2 \leq \frac{B^2}{\kappa^2(2s_0)} \leq \frac{4}{\kappa^2(2s_0)} \left\{ \sum_{(j,g) \in J_0} (\lambda_{ij}^g)^2 \right\}^{1/2} \|\Delta_{J_{01}}\|_F,$$

which implies

$$\|\Delta_{J_{01}}\|_F \leq \frac{4}{\kappa^2(2s_0)} \left\{ \sum_{(j,g) \in J_0} (\lambda_{ij}^g)^2 \right\}^{1/2}.$$

Plugging the above in (3.26), we obtain

$$\|\hat{\Theta}_i - \Theta_{0,i}\|_F^2 \leq \left\{ \frac{4\sqrt{10}}{\kappa^2(2s_0)} \right\}^2 \left\{ \frac{\sum_{(j,g) \in J_0} (\lambda_{ij}^g)^2}{\lambda_{\min} \sqrt{s_i}} \right\}^2,$$

or equivalently

$$\|\hat{\Theta}_i - \Theta_{0,i}\|_F \leq \frac{4\sqrt{10}}{\kappa^2(2s_0)} \frac{\sum_{(j,g) \in J(\Theta_{0,i})} (\lambda_{ij}^g)^2}{\lambda_{\min} \sqrt{s_i}}.$$

□

3.7.2 Selecting Edge Set

Given the estimates $\hat{\Theta}_i$ ($i = 1, \dots, p$), define \hat{E}^k as in (3.10) the estimated set of edges in graph $k = 1, \dots, K$. For every k , let $\tilde{\Omega}^k = \text{diag}(\Omega_0^k) + \Omega_{0, E_0^k \cap \hat{E}^k}^k$ and $\tilde{\Sigma}^k = (\tilde{\Omega}^k)^{-1}$. Let

$$C_{\text{bias}} = \frac{8\sqrt{10}c_0}{\kappa^2(2s_0)\sqrt{d_0}}.$$

The following corollary is an immediate result of (3.17) and (3.18).

Corollary III.8. *Consider \hat{E}^k ($k = 1, \dots, K$) selected in (3.10). Suppose all conditions in Theorem III.1 are satisfied. Choose $\lambda_{ij}^g = \lambda_{\max}$ as defined in (3.16) with $q > 1$. Then we have on the event \mathcal{A}*

$$|\hat{E}^k| \leq \frac{64\phi_{\max}}{\kappa^2(s_0)} S_0, \quad k = 1, \dots, K, \quad (3.27)$$

and

$$\frac{1}{K} \sum_k \|\tilde{\Omega}^k - \Omega_0^k\|_F \leq \frac{1}{\sqrt{K}} \left\{ \sum_k \|\tilde{\Omega}^k - \Omega_0^k\|_F^2 \right\}^{1/2} \leq C_{\text{bias}} \sqrt{\frac{S_0}{nK}} \left(\sqrt{|g_{\max}|} + \frac{\pi}{\sqrt{2}} \sqrt{q \log G_0} \right), \quad (3.28)$$

where G_0 is the maximum number of groups in all p regressions, S_0 is the total number of relevant groups, and $|g_{\max}|$ is the maximum group size.

Remark III.9. The bound in (3.27) says that the cardinality of the estimated set of edges is at most of the order of S_0 and proves essential in controlling the error rate of the maximum likelihood estimate $\hat{\Omega}^k$ in the refitting step. Further, the second inequality in (3.28) implies

$$\left\{ \sum_k \|\tilde{\Omega}^k - \Omega_0^k\|_F^2 \right\}^{1/2} \leq \tau_1 d_0,$$

provided the sample size n satisfies for $0 < \tau_1 < 1$,

$$n \geq S_0 \left(\sqrt{|g_{\max}|} + \frac{\pi}{\sqrt{2}} \sqrt{q \log G_0} \right)^2 \left(\frac{C_{\text{bias}}}{\tau_1 d_0} \right)^2. \quad (3.29)$$

It follows immediately that on the event \mathcal{A} , $\tilde{\Omega}^k$ is positive definite for all $k = 1, \dots, K$.

Indeed, by Assumption A2,

$$\begin{aligned} \phi_{\min}(\tilde{\Omega}^k) &\geq \phi_{\min}(\Omega_0^k) - \|\tilde{\Omega}^k - \Omega_0^k\| \geq \phi_{\min}(\Omega_0^k) - \|\tilde{\Omega}^k - \Omega_0^k\|_F \\ &\geq \phi_{\min}(\Omega_0^k) - \left\{ \sum_k \|\tilde{\Omega}^k - \Omega_0^k\|_F^2 \right\}^{1/2} \geq (1 - \tau_1) d_0 > 0. \end{aligned} \quad (3.30)$$

In addition, we have an upper bound for the maximum eigenvalue of $\tilde{\Omega}^k$,

$$\begin{aligned} \phi_{\max}(\tilde{\Omega}^k) &\leq \phi_{\max}(\Omega_0^k) + \|\tilde{\Omega}^k - \Omega_0^k\| \leq \phi_{\max}(\Omega_0^k) + \|\tilde{\Omega}^k - \Omega_0^k\|_F \\ &\leq \phi_{\max}(\Omega_0^k) + \left\{ \sum_k \|\tilde{\Omega}^k - \Omega_0^k\|_F^2 \right\}^{1/2} \leq c_0 + \tau_1 d_0 < \infty. \end{aligned} \quad (3.31)$$

Proof of Corollary III.8. By definition, $\omega_{0,ij}^k = -\theta_{0,ij}^k \omega_{0,ii}^k$ for all $j \neq i$ and $k = 1, \dots, K$. Further, under Assumption A2, $\omega_{0,ii}^k \leq \phi_{\max}(\Omega_0^k) = \phi_{\min}^{-1}(\Sigma_0^k) \leq c_0$ for all i, k . Therefore

$$\begin{aligned} \sum_k \|\tilde{\Omega}^k - \Omega_0^k\|_F^2 &= \sum_k \sum_{i=1}^p \sum_{j \in J(\theta_{0,i}) \cap J(\hat{\theta}_i)^c} (\theta_{0,ij}^k \omega_{0,ii}^k)^2 \\ &= \sum_{i=1}^p \sum_{j \in J(\theta_{0,i}) \cap J(\hat{\theta}_i)^c} \sum_{g \in \mathcal{G}^{ij}} \sum_{k \in g} (\theta_{0,ij}^k \omega_{0,ii}^k)^2 \\ &\leq c_0^2 \sum_{i=1}^p \sum_{j \in J(\theta_{0,i}) \cap J(\hat{\theta}_i)^c} \sum_{g \in \mathcal{G}^{ij}} \|\theta_{0,ij}^{[g]}\|^2 \\ &\leq c_0^2 \sum_{i=1}^p \sum_{j \neq i} \sum_{g \in \mathcal{G}^{ij}} \|\theta_{0,ij}^{[g]} - \hat{\theta}_{ij}^{[g]}\|^2. \end{aligned}$$

Under Assumption A1 with $s = 2s_0$, applying Theorem III.6 with $\lambda_{ij}^g = \lambda_{\max}$ in (3.16),

$$\sum_{j \neq i} \sum_{g \in \mathcal{G}^{ij}} \|\boldsymbol{\theta}_{0,ij}^{[g]} - \hat{\boldsymbol{\theta}}_{ij}^{[g]}\|^2 \leq \left\{ \frac{4\sqrt{10}}{\kappa^2(2s_0)} \lambda_{\max} \right\}^2 s_i.$$

Therefore,

$$\sum_k \|\tilde{\Omega}^k - \Omega_0^k\|_F^2 \leq \left\{ \frac{4\sqrt{10}c_0}{\kappa^2(2s_0)} \lambda_{\max} \right\}^2 \sum_{i=1}^p s_i = \left\{ \frac{4\sqrt{10}c_0}{\kappa^2(2s_0)} \lambda_{\max} \right\}^2 S_0.$$

It follows immediately that

$$\begin{aligned} \frac{1}{K} \sum_k \|\tilde{\Omega}^k - \Omega_0^k\|_F &\leq \frac{1}{\sqrt{K}} \left\{ \sum_k \|\tilde{\Omega}^k - \Omega_0^k\|_F^2 \right\}^{1/2} \leq \frac{4\sqrt{10}c_0}{\kappa^2(2s_0)} \lambda_{\max} \sqrt{\frac{S_0}{K}} \\ &\leq C_{\text{bias}} \sqrt{\frac{S_0}{nK}} \left(\sqrt{|g_{\max}|} + \frac{\pi}{\sqrt{2}} \sqrt{q \log G_0} \right). \end{aligned}$$

To bound the estimated edge set \hat{E}^k , notice if there exists (i, j, k) such that $\hat{\boldsymbol{\theta}}_{ij}^k \neq 0$, then $\hat{\boldsymbol{\theta}}_{ij}^{[g]} \neq \mathbf{0}$, where $g \ni k$. Hence $\mathcal{M}(\hat{\boldsymbol{\theta}}_i^k) \leq \mathcal{M}(\hat{\Theta}_i)$ for all k . By (3.14), the upper bound for \hat{E}^k is thus

$$|\hat{E}^k| \leq \sum_{i=1}^p \mathcal{M}(\hat{\boldsymbol{\theta}}_i^k) \leq \sum_{i=1}^p \frac{64\phi_{\max}}{\kappa^2(s_0)\lambda_{\min}^2} \sum_{(j,g) \in J(\Theta_0,i)} (\lambda_{ij}^g)^2 = \frac{64\phi_{\max}}{\kappa^2(s_0)} \sum_{i=1}^p s_i \leq \frac{64\phi_{\max}}{\kappa^2(s_0)} S_0.$$

□

3.7.3 Refitting

Proof of Theorem III.1. Let

$$r_n = C_{\text{bias}} \sqrt{\frac{S_0}{n}} \left(\sqrt{|g_{\max}|} + \frac{\pi}{\sqrt{2}} \sqrt{q \log G_0} \right).$$

In view of Corollary III.8, it suffices to show that

$$\sum_k \|\hat{\Omega}^k - \tilde{\Omega}^k\|_F^2 \leq \mathcal{O}(r_n^2),$$

since by Cauchy-Schwarz inequality,

$$\frac{1}{K} \sum_k \|\hat{\Omega}^k - \tilde{\Omega}^k\|_F \leq \frac{1}{\sqrt{K}} \left\{ \sum_k \|\hat{\Omega}^k - \tilde{\Omega}^k\|_F^2 \right\}^{1/2},$$

and by triangle inequality,

$$\frac{1}{K} \sum_k \|\hat{\Omega}^k - \Omega_0^k\|_F \leq \frac{1}{K} \sum_k \|\hat{\Omega}^k - \tilde{\Omega}^k\|_F + \frac{1}{K} \sum_k \|\tilde{\Omega}^k - \Omega_0^k\|_F.$$

For $k = 1, \dots, K$, let $\Delta^k = \Omega^k - \tilde{\Omega}^k \in \mathbb{M}(p, p)$ and $\hat{\Delta}^k = \hat{\Omega}^k - \tilde{\Omega}^k$. Let

$$Q(\Omega) = \sum_k \left\{ \text{tr}(\hat{\Sigma}^k \Omega^k) - \log \det(\Omega^k) - \text{tr}(\hat{\Sigma}^k \tilde{\Omega}^k) + \log \det(\tilde{\Omega}^k) \right\}.$$

Since $(\hat{\Omega}^k)_{k=1}^K$ minimizes $Q(\Omega)$, $(\hat{\Delta}^k)_{k=1}^K$ minimizes $G(\Delta) = Q(\tilde{\Omega} + \Delta)$.

For $k = 1, \dots, K$, define a sequence of convex sets

$$\mathcal{U}_n(\tilde{\Omega}^k) = \{\Gamma - \tilde{\Omega}^k \mid \Gamma \in \mathcal{S}_+^p \cap \mathcal{S}_{\hat{E}^k}^p\}.$$

The main idea of the proof is as follows. For a sufficiently large $M > 0$, consider the set

$$\mathcal{T}_n = \{(\Delta^1, \dots, \Delta^K) : \Delta^k \in \mathcal{U}_n(\tilde{\Omega}^k), \sum_k \|\Delta^k\|_F^2 = Mr_n^2\}.$$

It is clear that $G(\Delta)$ is a convex function and $G(\hat{\Delta}) \leq G(\mathbf{0}) = 0$. Thus if we can show $\inf_{\Delta \in \mathcal{T}_n} G(\Delta) > 0$, the minimizer $\hat{\Delta}$ must be inside the ball defined by \mathcal{T}_n .

That is $\sum_k \|\hat{\Delta}^k\|_F^2 \leq Mr_n^2$. To see this, note that the convexity of $Q(\Omega)$ implies that $\inf_{\Delta \in \mathcal{T}_n} Q(\tilde{\Omega} + \Delta) > Q(\tilde{\Omega}) = 0$. There exists therefore a local minimizer in the ball

$\{\tilde{\Omega}^k + \Delta^k : \sum_k \|\Delta^k\|_F^2 \leq Mr_n^2\}$, or equivalently, $\sum_k \|\hat{\Delta}^k\|_F^2 \leq Mr_n^2$.

In the remainder of the proof, we focus on

$$G(\Delta) = \sum_k \left\{ \text{tr}(\hat{\Sigma}^k \Delta^k) - \log \det(\tilde{\Omega}^k + \Delta^k) + \log \det(\tilde{\Omega}^k) \right\}.$$

Applying Taylor expansion to the logarithm terms in the above equation, we have

$$\begin{aligned} & \log \det(\tilde{\Omega}^k + \Delta^k) - \log \det(\tilde{\Omega}^k) \\ &= \text{tr}(\tilde{\Sigma}^k \Delta^k) + \text{vec}(\Delta^k)^T \left\{ \int_0^1 (1-t)(\tilde{\Omega}^k + t\Delta^k)^{-1} \otimes (\tilde{\Omega}^k + t\Delta^k)^{-1} dt \right\} \text{vec}(\Delta^k), \end{aligned}$$

where \otimes is the Kronecker product, and $\text{vec}(\Delta^k)$ is Δ^k vectorized to match the dimensions of the Kronecker product. Therefore, we can rewrite $G(\Delta) = L_1 - L_2 + L_3$, with

$$\begin{aligned} L_1 &= \sum_k \text{tr} \{ (\hat{\Sigma}^k - \Sigma_0^k) \Delta^k \}, \\ L_2 &= \sum_k \text{tr} \{ (\tilde{\Sigma}^k - \Sigma_0^k) \Delta^k \}, \\ L_3 &= \sum_k \text{vec}(\Delta^k)^T \left\{ \int_0^1 (1-t)(\tilde{\Omega}^k + t\Delta^k)^{-1} \otimes (\tilde{\Omega}^k + t\Delta^k)^{-1} dt \right\} \text{vec}(\Delta^k). \end{aligned}$$

Next we bound each term separately.

Recall for every k , Σ_0^k and $\hat{\Sigma}^k$ represent the correlation and the sample correlation matrix, respectively. Since $\phi_{\max}(\Sigma_0^k) \leq 1/d_0$ for all k , by Lemma 14 of *Zhou et al.* (2011) [see details on page 3003],

$$\mathbb{P} \left\{ |\hat{\sigma}_{ij}^k - \sigma_{0,ij}^k| \geq t \right\} \leq \exp \left(- \frac{3nt^2}{10\{1 + (\sigma_{0,ij}^k)^2\}} \right) \leq \exp \left(- \frac{3nt^2}{20} \right), \quad (3.32)$$

for $0 \leq t \leq \{1 + (\sigma_{0,ij}^k)^2\}/2$. Thus if we choose for some $c_1 > 0$

$$t = c_1 \sqrt{\frac{1}{K}} \sqrt{\frac{1}{n}} \left(\sqrt{|g_{\max}|} + \frac{\pi}{\sqrt{2}} \sqrt{q \log G_0} \right),$$

then $\max_{k,i \neq j} |\hat{\sigma}_{ij}^k - \sigma_{0,ij}^k| \leq t$ with probability tending to 1, provided that the sample size satisfies

$$n \geq \frac{4c_1^2}{K} \left(\sqrt{|g_{\max}|} + \frac{\pi}{\sqrt{2}} \sqrt{q \log G_0} \right)^2. \quad (3.33)$$

Write $\Delta^k = \Delta^{k,+} + \Delta^{k,-}$ such that $\Delta^{k,+}$ is the diagonal matrix which has the same diagonal elements as Δ^k and $\Delta^{k,-}$ consists of the off-diagonal elements. Then

$$\begin{aligned} |L_1| &\leq \sum_k \sum_{i \neq j} |\hat{\sigma}_{ij}^k - \sigma_{0,ij}^k| \|\Delta_{ij}^k\| \leq c_1 \sqrt{\frac{1}{nK}} \left(\sqrt{|g_{\max}|} + \frac{\pi}{\sqrt{2}} \sqrt{q \log G_0} \right) \sum_k \|\Delta^{k,-}\|_1 \\ &\leq c_1 \sqrt{\frac{1}{n}} \left(\sqrt{|g_{\max}|} + \frac{\pi}{\sqrt{2}} \sqrt{q \log G_0} \right) \max_k |2\hat{E}^k|^{1/2} \left(\sum_k \|\Delta^k\|_F^2 \right)^{1/2} \\ &\leq \frac{8\sqrt{2}c_1 \sqrt{\phi_{\max}}}{C_{\text{bias}} \kappa(s_0)} r_n \left(\sum_k \|\Delta^k\|_F^2 \right)^{1/2}. \end{aligned}$$

To bound the second term, since $\phi_{\min}(\tilde{\Omega}^k)$ ($k = 1, \dots, K$) is bounded below by (3.30),

$$\begin{aligned} |L_2| &\leq \sum_k |\langle \tilde{\Sigma}^k - \Sigma_0^k, \Delta^k \rangle| \leq \sum_k \|\tilde{\Sigma}^k - \Sigma_0^k\|_F \|\Delta^k\|_F \leq \sum_k \|\Delta^k\|_F \frac{\|\tilde{\Omega}^k - \Omega_0^k\|_F}{\phi_{\min}(\tilde{\Omega}^k) \phi_{\min}(\Omega_0^k)} \\ &\leq \frac{1}{(1 - \tau_1) d_0^2} \left(\sum_k \|\Delta^k\|_F^2 \right)^{1/2} \left(\sum_k \|\tilde{\Omega}^k - \Omega_0^k\|_F^2 \right)^{1/2} \leq \frac{r_n}{(1 - \tau_1) d_0^2} \left(\sum_k \|\Delta^k\|_F^2 \right)^{1/2}, \end{aligned} \quad (3.34)$$

where the last inequality in (3.34) comes from the rotation invariant property of the Frobenius norm.

Finally we bound L_3 . Suppose for a small constant $0 < \tau_2 < 1$ such that $\tau_1 + \tau_2 < 1$,

the sample size n satisfies

$$n \geq MS_0 \left(\sqrt{|g_{\max}|} + \frac{\pi}{\sqrt{2}} \sqrt{q \log G_0} \right)^2 \left(\frac{C_{\text{bias}}}{\tau_2 d_0} \right)^2, \quad (3.35)$$

then $\sqrt{M}r_n \leq \tau_2 d_0$. By (3.31), $\phi_{\max}(\tilde{\Omega}^k)$ is bounded above by $c_0 + \tau_1 d_0$. Therefore for $\Delta \in \mathcal{T}_n$,

$$\begin{aligned} \phi_{\max}(\tilde{\Omega}^k + \Delta^k) &\leq c_0 + \tau_1 d_0 + \|\Delta^k\| \leq c_0 + \tau_1 d_0 + \|\Delta^k\|_F \\ &\leq c_0 + \tau_1 d_0 + \left(\sum_k \|\Delta^k\|_F^2 \right)^{1/2} \leq c_0 + (\tau_1 + \tau_2) d_0, \\ \phi_{\min}(\tilde{\Omega}^k + \Delta^k) &\geq (1 - \tau_1) d_0 - \|\Delta^k\| \geq (1 - \tau_1) d_0 - \|\Delta^k\|_F \\ &\geq (1 - \tau_1) d_0 - \left(\sum_k \|\Delta^k\|_F^2 \right)^{1/2} \geq (1 - \tau_1 - \tau_2) d_0 > 0. \end{aligned}$$

For $\tilde{\Omega}^k$ and Δ^k defined above, *Zhou et al.* (2011) showed that $\tilde{\Omega}^k + t\Delta^k \succ 0, t \in [0, 1]$, for all $k = 1, \dots, K$ on the event \mathcal{A} . Thus, following similar arguments as in *Rothman et al.* (2008, page 502), we have

$$\begin{aligned} |L_3| &\geq \frac{1}{2} \sum_k \phi_{\min}^2(\tilde{\Omega}^k + \Delta^k)^{-1} \|\Delta^k\|_F^2 = \frac{1}{2} \sum_k \phi_{\max}^{-2}(\tilde{\Omega}^k + \Delta^k) \|\Delta^k\|_F^2 \\ &\geq \frac{1}{2(c_0 + \tau_1 d_0 + \tau_2 d_0)^2} \sum_k \|\Delta^k\|_F^2. \end{aligned}$$

Combining the above three bounds, we thus have

$$\begin{aligned} G(\Delta) &\geq |L_3| - |L_1| - |L_2| \\ &\geq \frac{1}{2(c_0 + \tau_1 d_0 + \tau_2 d_0)^2} \sum_k \|\Delta^k\|_F^2 - \frac{8\sqrt{2}c_1\sqrt{\phi_{\max}}}{C_{\text{bias}}\kappa(s_0)} r_n \left(\sum_k \|\Delta^k\|_F^2 \right)^{1/2} \\ &\quad - \frac{r_n}{(1 - \tau_1)d_0^2} \left(\sum_k \|\Delta^k\|_F^2 \right)^{1/2} \\ &\geq Mr_n^2 \left\{ \frac{1}{2(c_0 + \tau_1 d_0 + \tau_2 d_0)^2} - \frac{8c_1\sqrt{2\phi_{\max}}}{C_{\text{bias}}\kappa(s_0)} \frac{1}{\sqrt{M}} - \frac{1}{(1 - \tau_1)d_0^2\sqrt{M}} \right\} > 0, \end{aligned}$$

for M sufficiently large. □

3.8 Proof of Theorem III.2

Consider the group lasso estimator $\hat{\Theta}_i$ defined in (3.2). Since the problem (3.2) is a special case of the generic group lasso in *Basu et al.* (2012), we adapt their results in Theorem 4.1 to our design.

Proof of Theorem III.2. Let \mathcal{X}_i be the block diagonal matrix composed of all variables but \mathbf{X}_i^k ($k = 1, \dots, K$). Without loss of generality, suppose $\mathcal{X}_i = (\mathcal{X}_{i,(1)}, \mathcal{X}_{i,(2)})$ such that

$$\mathcal{X}_{i,(1)} = \text{diag}(\mathbf{X}_{I_1}^1, \dots, \mathbf{X}_{I_K}^K)$$

is the sub-matrix consisting of all relevant variables. Denote the Gram matrix

$$C = \frac{1}{n} \mathcal{X}_i^T \mathcal{X}_i = \begin{pmatrix} C_{11} & C_{12} \\ C_{21} & C_{22} \end{pmatrix}$$

with $C_{11} = \mathcal{X}_{i,(1)}^T \mathcal{X}_{i,(1)}/n$ and $C_{22} = \mathcal{X}_{i,(2)}^T \mathcal{X}_{i,(2)}/n$. C_{12} and C_{21} are also defined accordingly.

Now consider interchanging the columns of \mathcal{X}_i such that

$$\tilde{\mathcal{X}}_i = \mathcal{X}_i \text{diag}(R_1, R_2) = (\mathcal{X}_{i,(1)}R_1, \mathcal{X}_{i,(2)}R_2) = (\tilde{\mathcal{X}}_{i,(1)}, \tilde{\mathcal{X}}_{i,(2)}),$$

where the columns of $\tilde{\mathcal{X}}_{i,(1)}$ and $\tilde{\mathcal{X}}_{i,(2)}$ are ordered in groups of variables. Here R_l is the product of elementary column switching matrices and satisfies $R_l^{-1} = R_l^T$ ($l = 1, 2$). Note $R_1 \in \mathbb{M}(\sum_k |I_k|, \sum_k |I_k|)$. Based on $\tilde{\mathcal{X}}_i$, we can define \tilde{C}_{11} , \tilde{C}_{21} and \tilde{C}_{22} similarly as above. The advantage of using $\tilde{\mathcal{X}}_i$ as the design matrix is that it orders the variables based on the grouping structures, and is in the form of the generic group lasso design in *Basu et al.* (2012). It is thus more straightforward to adapt their results using $\tilde{\mathcal{X}}_i$.

With the above notations, the Uniform IC in Assumption A3 is equivalent to saying for all $\boldsymbol{\xi} = ((\boldsymbol{\xi}^1)^T, \dots, (\boldsymbol{\xi}^K)^T)^T \in \mathbb{R}^{\sum_k |I_k|}$ with $\max_{(j,g) \in J(\Theta_{0,i})} \|\boldsymbol{\xi}_j^{[g]}\| \leq 1$ and all $(j, g) \notin J(\Theta_{0,i})$

$$\|[\tilde{C}_{21}(\tilde{C}_{11})^{-1}\tilde{\boldsymbol{\xi}}]_{[j,g]}\| \leq 1 - \eta, \quad (3.36)$$

where $\tilde{\boldsymbol{\xi}} = R_1^T \boldsymbol{\xi}$. It remains to select λ and α_n to ensure that the direction consistency results hold simultaneously for all i with probability tending to 1. For any $(j, g) \in J(\Theta_{0,i})$, denote $(\tilde{C}_{11})_{[j,g]}^{-1}$ the diagonal block in \tilde{C}_{11}^{-1} corresponding to the group (j, g) . By Theorem 4.1 of *Basu et al.* (2012), it suffices to find the upper bounds for $\|\tilde{C}_{11}^{-1}\|$, $\|(\tilde{C}_{11})_{[j,g]}^{-1}\|$, $\|(\tilde{C}_{22})_{[j,g]}\|$ and substitute the constant variance σ with the appropriate bound for $\text{Var}(X_i^k | X_{-i}^k) = 1/\omega_{0,ii}^k$ ($k = 1, \dots, K$).

By definition and the fact that \mathbf{X}^k are centered and standardized, $(\tilde{C}_{11})_{[j,g]}$ is the identity matrix of size $|g| \times |g|$. It follows that

$$1 = \phi_{\min}^{-1}((\tilde{C}_{11})_{[j,g]}) \leq \phi_{\max}((\tilde{C}_{11})_{[j,g]}^{-1}) = \|(\tilde{C}_{11})_{[j,g]}^{-1}\| \leq \|(\tilde{C}_{11})^{-1}\|, \quad (3.37)$$

where the last step is obtained by applying Courant minimax principle since $0 \prec (\tilde{C}_{11})_{[j,g]}^{-1} \preceq (\tilde{C}_{11})^{-1}$. Similarly, for any $(j, g) \notin J(\Theta_{0,i})$, $(\tilde{C}_{22})_{[j,g]}$ is the identity matrix and

$$\|(\tilde{C}_{22})_{[j,g]}\| = 1. \quad (3.38)$$

Moreover, the variance for the random design in our problem

$$\text{Var}(X_i^k | X_{-i}^k) = 1/\omega_{0,ii}^k \leq 1/d_0, \quad \forall k \quad (3.39)$$

by Assumption A2.

It remains to find an upper bound for $\|\tilde{C}_{11}^{-1}\|$. Under Assumption A1 with $s = s_0$,

if we set $\Delta \in \mathcal{F}$ such that $\delta_j^{[g]} = \mathbf{0}$ for any $(j, g) \notin J(\Theta_{0,i})$, then

$$\frac{\sum_k \|\mathbf{X}^k \boldsymbol{\delta}^k\|^2/n}{\|\Delta_{J(\Theta_{0,i})}\|_F^2} = \frac{\boldsymbol{\xi}^T C_{11} \boldsymbol{\xi}}{\boldsymbol{\xi}^T \boldsymbol{\xi}},$$

where $\boldsymbol{\xi} = ((\boldsymbol{\xi}^1)^T, \dots, (\boldsymbol{\xi}^K)^T)^T \in \mathbb{R}^{\sum_k |I_k|}$ such that each $\boldsymbol{\xi}^k$ corresponds to the nonzero part of $\boldsymbol{\delta}^k$. If we choose Δ such that $\boldsymbol{\xi}$ is the eigenvector corresponding to the smallest eigenvalue of C_{11} , then

$$\kappa^2(s_0) \leq \frac{\sum_k \|\mathbf{X}^k \boldsymbol{\delta}^k\|^2/n}{\|\Delta_{J(\Theta_{0,i})}\|_F^2} = \frac{\boldsymbol{\xi}^T C_{11} \boldsymbol{\xi}}{\boldsymbol{\xi}^T \boldsymbol{\xi}} = \phi_{\min}(C_{11}).$$

Since $R_1^{-1} = R_1^T$, C_{11} and \tilde{C}_{11} are similar (i.e. there exists a non-singular matrix P such that $P^{-1}C_{11}P = \tilde{C}_{11}$) and thus share the same set of eigenvalues. Therefore $\phi_{\min}(\tilde{C}_{11}) \geq \kappa^2(s_0)$ and

$$\|\tilde{C}_{11}^{-1}\| \leq \kappa^{-2}(s_0). \quad (3.40)$$

Combining the upper bounds in (3.37), (3.38), (3.39) and (3.40), Theorem 4.1 of *Basu et al.* (2012) implies that if we select λ and α_n as in (3.8) and (3.9), respectively, the direction consistency results follow by considering the union bound on all probabilities made across $i = 1, \dots, p$.

Further, if $\alpha_n < 1$, the direction consistency property of $\hat{\Theta}_i$ implies exact recovery of all nonzero entries in the inverse covariance matrices, provided that the sparsity pattern \mathcal{G} is correctly specified. In other words, the set in (3.10) estimates correctly the true edge set E_0^k for all k .

This completes the proof. □

BIBLIOGRAPHY

BIBLIOGRAPHY

- Al-Shahrour, F., R. Díaz-Uriarte, and J. Dopazo (2005), Discovering molecular functions significantly related to phenotypes by combining gene expression data and biological information, *Bioinformatics*, *21*(13), 2988–2993.
- Banerjee, O., L. E. Ghaoui, and A. d’Aspremont (2008), Model selection through sparse maximum likelihood estimation for multivariate gaussian or binary data, *The Journal of Machine Learning Research*, *9*, 485–516.
- Basu, S., A. Shojaie, and G. Michailidis (2012), Network granger causality with inherent grouping structure, *arXiv preprint arXiv:1210.3711*.
- Baur, J. A., et al. (2006), Resveratrol improves health and survival of mice on a high-calorie diet, *Nature*, *444*(7117), 337–342.
- Beißbarth, T., and T. P. Speed (2004), Gostat: find statistically overrepresented gene ontologies within a group of genes, *Bioinformatics*, *20*(9), 1464–1465.
- Benjamini, Y., and Y. Hochberg (1995), Controlling the false discovery rate: A practical and powerful approach to multiple testing, *Journal of the Royal Statistical Society. Series B (Methodological)*, *57*(1), 289–300.
- Bickel, P. J., and E. Levina (2008), Regularized estimation of large covariance matrices, *The Annals of Statistics*, *36*(1), 199–227.
- Bickel, P. J., Y. Ritov, and A. B. Tsybakov (2009), Simultaneous analysis of lasso and dantzig selector, *The Annals of Statistics*, *37*(4), 1705–1732.
- Boyd, S., and L. Vandenberghe (2004), *Convex optimization*, Cambridge University Press.
- Boyd, S., N. Parikh, E. Chu, B. Peleato, and J. Eckstein (2011), Distributed optimization and statistical learning via the alternating direction method of multipliers, *Foundations and Trends® in Machine Learning*, *3*(1), 1–122.
- Breheeny, P., and J. Huang (2009), Penalized methods for bi-level variable selection, *Stat Interface*, *2*(3), 369–380.
- Byrd, R. H., P. Lu, J. Nocedal, and C. Zhu (1995), A limited memory algorithm for bound constrained optimization, *SIAM Journal on Scientific Computing*, *16*(5), 1190–1208.

- Cai, J.-F., E. J. Candès, and Z. Shen (2010), A singular value thresholding algorithm for matrix completion, *SIAM Journal on Optimization*, *20*(4), 1956–1982.
- Candes, E. J., and B. Recht (2009), Exact matrix completion via convex optimization, *Foundations of Computational Mathematics*, *9*, 717–772.
- Chiquet, J., Y. Grandvalet, and C. Ambroise (2011), Inferring multiple graphical structures, *Statistics and Computing*, *21*(4), 537–553.
- Chuang, H.-Y., L. Rassenti, M. Salcedo, K. Licon, A. Kohlmann, T. Haferlach, R. Foà, T. Ideker, and T. J. Kipps (2012), Subnetwork-based analysis of chronic lymphocytic leukemia identifies pathways that associate with disease progression, *Blood*, *120*(13), 2639–2649.
- Cui, L., H. Jeong, F. Borovecki, C. N. Parkhurst, N. Tanese, and D. Krainc (2006), Transcriptional repression of pgc-1 α by mutant huntingtin leads to mitochondrial dysfunction and neurodegeneration, *Cell*, *127*(1), 59–69.
- Danaher, P., P. Wang, and D. M. Witten (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.
- Dehmer, M., and F. Emmert-Streib (2008), *Analysis of microarray data: a network-based approach*, John Wiley & Sons.
- Dempster, A. P. (1972a), Covariance selection, *Biometrics*, *28*(1), 157–175.
- Dempster, A. P. (1972b), Covariance selection, *Biometrics*, pp. 157–175.
- Drton, M., and M. D. Perlman (2004), Model selection for gaussian concentration graph, *Biometrika*, *91*(3), 591–602.
- Efron, B., and R. Tibshirani (2007), On testing the significance of sets of genes, *The Annals of Applied Statistics*, *1*(1), 107–129.
- Friedman, J., T. Hastie, and R. Tibshirani (2008), Sparse inverse covariance estimation with the graphical lasso, *Biostatistics*, *9*(3), 432–441.
- Gottwein, E., et al. (2007), A viral microRNA functions as an orthologue of cellular mir-155, *Nature*, *450*(7172), 1096–1099.
- Green, M. R., et al. (2011), Signatures of murine b-cell development implicate yy1 as a regulator of the germinal center-specific program, *Proceedings of the National Academy of Sciences*, *108*(7), 2873–2878.
- Guo, J., E. Levina, G. Michailidis, and J. Zhu (2011), Joint estimation of multiple graphical models, *Biometrika*, *98*(1), 1–15.
- Haberman, S. J. (1989), Concavity and estimation, *The Annals of Statistics*, *17*(4), 1631–1661.

- Harris, I., P. Jones, T. Osborn, and D. Lister (2014), Updated high-resolution grids of monthly climatic observations—the cru ts3. 10 dataset, *International Journal of Climatology*, *34*(3), 623–642.
- Houstis, N., E. D. Rosen, and E. S. Lander (2006), Reactive oxygen species have a causal role in multiple forms of insulin resistance, *Nature*, *440*(7086), 944–948.
- Huang, D. W., B. T. Sherman, and R. A. Lempicki (2008), Systematic and integrative analysis of large gene lists using david bioinformatics resources, *Nature protocols*, *4*(1), 44–57.
- Huarte, M., et al. (2010), A large intergenic noncoding rna induced by p53 mediates global gene repression in the p53 response, *Cell*, *142*(3), 409–419.
- Huerta, A. M., H. Salgado, D. Thieffry, and J. Collado-Vides (1998), Regulondb: A database on transcriptional regulation in escherichia coli, *Nucleic Acids Research*, *26*(1), 55–59, doi:10.1093/nar/26.1.55.
- Ideker, T., and N. J. Krogan (2012), Differential network biology, *Molecular systems biology*, *8*(1).
- Ideker, T., J. Dutkowski, and L. Hood (2011), Boosting signal-to-noise in complex biology: prior knowledge is power, *Cell*, *144*(6), 860–863.
- Ischenko, I., J. Liu, O. Petrenko, and M. Hayman (2014), Transforming growth factor-beta signaling network regulates plasticity and lineage commitment of lung cancer cells, *Cell Death & Differentiation*.
- Joshi-Tope, G., et al. (2003), The genome knowledgebase: A resource for biologists and bioinformaticists, *Cold Spring Harbor Symposia on Quantitative Biology*, *68*, 237–244, doi:10.1101/sqb.2003.68.237.
- Kanehisa, M., and S. Goto (2000), Kegg: Kyoto encyclopedia of genes and genomes, *Nucleic Acids Research*, *28*(1), 27–30, doi:10.1093/nar/28.1.27.
- Khatri, P., M. Sirota, and A. J. Butte (2012), Ten years of pathway analysis: current approaches and outstanding challenges, *PLoS Comput Biol*, *8*(2), e1002375, doi:10.1371/journal.pcbi.1002375.
- Kottek, M., J. Grieser, C. Beck, B. Rudolf, and F. Rubel (2006), World map of the köppen-geiger climate classification updated, *Meteorologische Zeitschrift*, *15*(3), 259–263.
- Lauritzen, S. L. (1996), *Graphical models*, Oxford University Press.
- Lounici, K., M. Pontil, S. van de Geer, and A. B. Tsybakov (2011), Oracle inequalities and optimal inference under group sparsity, *The Annals of Statistics*, *39*(4), 2164–2204.

- Lozano, A. C., H. Li, A. Niculescu-Mizil, Y. Liu, C. Perlich, J. Hosking, and N. Abe (2009), Spatial-temporal causal modeling for climate change attribution, in *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 587–596, ACM.
- Meinshausen, N., and P. Bühlmann (2006), High dimensional graphs and variable selection with the lasso, *The Annals of Statistics*, *34*(3), 1436–1462.
- Meinshausen, N., and P. Bühlmann (2010), Stability selection, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *72*(4), 417–473.
- Mirsky, L. (1975), A trace inequality of john von neumann, *Monatshefte für Mathematik*, *79*(4), 303–306.
- Mohan, K., P. London, M. Fazel, D. Witten, and S.-I. Lee (2014), Node-based learning of multiple gaussian graphical models, *The Journal of Machine Learning Research*, *15*(1), 445–488.
- Nishimura, D. (2001), Biocarta, *Biotech Software & Internet Report: The Computer Software Journal for Scient*, *2*(3), 117–120.
- Palomero, T., et al. (2006), Notch1 directly regulates c-myc and activates a feed-forward-loop transcriptional network promoting leukemic cell growth, *Proceedings of the National Academy of Sciences*, *103*(48), 18,261–18,266.
- Peng, J., P. Wang, N. Zhou, and J. Zhu (2009), Partial correlation estimation by joint sparse regression model, *Journal of the American Statistical Association*, *104*(486), 735–746.
- Perroud, B., J. Lee, N. Valkova, A. Dhirapong, P.-Y. Lin, O. Fiehn, D. Kültz, and R. H. Weiss (2006), Pathway analysis of kidney cancer using proteomics and metabolic profiling, *Molecular Cancer*, *5*(1), 64.
- Peterson, C., F. Stingo, and M. Vannucci (2014), Bayesian inference of multiple gaussian graphical models, *Journal of the American Statistical Association*, (just-accepted), 00–00.
- Pujana, M. A., et al. (2007), Network modeling links breast cancer susceptibility and centrosome dysfunction, *Nature Genetics*, *39*(11), 1338 – 1349.
- Putluri, N., et al. (2011), Metabolomic profiling reveals potential markers and bioprocesses altered in bladder cancer progression, *Cancer Research*, *71*(24), 7376–7386.
- Raskutti, G., B. Yu, M. J. Wainwright, and P. K. Ravikumar (2009), Model selection in gaussian graphical models: High-dimensional consistency of ℓ_1 -regularized mle, in *Advances in Neural Information Processing Systems*, pp. 1329–1336.
- Rothman, A. J., P. J. Bickel, E. Levina, and J. Zhu (2008), Sparse permutation invariant covariance estimation, *Electronic Journal of Statistics*, *2*, 494–515.

- Searle, S. (1971), *Linear models*, New York [etc.]: Wiley [etc.], doi:10.1002/9781118491782.
- Shah, R. D., and R. J. Samworth (2012), Variable selection with error control: Another look at stability selection, *Journal of the Royal Statistical Society*.
- Shojaie, A., and G. Michailidis (2009), Analysis of gene sets based on the underlying regulatory network, *Journal of Computational Biology*, 16(3), 407–426.
- Shojaie, A., and G. Michailidis (2010), Network enrichment analysis in complex experiments, *Statistical Applications in Genetics and Molecular Biology*, 9(1).
- Stocker, T. F., et al. (2013), Climate change 2013: The physical science basis. contribution of working group i to the fifth assessment report of the intergovernmental panel on climate change, *Tech. rep.*, IPCC.
- Subramanian, A., et al. (2005), Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles, *Proceedings of the National Academy of Sciences of the United States of America*, 102(43), 15,545–15,550.
- Wille, A., et al. (2004), Sparse graphical gaussian modeling of the isoprenoid gene network in arabidopsis thaliana, *Genome Biol*, 5(11), R92.
- Wilson, B. G., et al. (2010), Epigenetic antagonism between polycomb and swi/snf complexes during oncogenic transformation, *Cancer cell*, 18(4), 316–328.
- Yuan, M., and Y. Lin (2007), Model selection and estimation in the gaussian graphical model, *Biometrika*, 94(1), 19–35.
- Zaki, N., D. Efimov, and J. Berenguères (2013), Protein complex detection using interaction reliability assessment and weighted clustering coefficient, *BMC Bioinformatics*, 14(1), 163, doi:10.1186/1471-2105-14-163.
- Zhao, P., and B. Yu (2006), On model selection consistency of lasso, *Journal of Machine Learning Research*, 7, 2541–2563.
- Zhou, L., A. Dai, Y. Dai, R. S. Vose, C.-Z. Zou, Y. Tian, and H. Chen (2009), Spatial dependence of diurnal temperature range trends on precipitation from 1950 to 2004, *Climate Dynamics*, 32(2-3), 429–440, doi:10.1007/s00382-008-0387-5.
- Zhou, S. (2010), Thresholded lasso for high dimensional variable selection and statistical estimation, *arXiv preprint arXiv:1002.1583*.
- Zhou, S., P. Rutimann, M. Xu, and P. Bühlmann (2011), High-dimensional covariance estimation based on gaussian graphical models, *The Journal of Machine Learning Research*, 12, 2975–3026.
- Zhu, Y., X. Shen, and W. Pan (2014), Structural pursuit over multiple undirected graphs, *Journal of the American Statistical Association*, 109(508), 1683–1696.