MODEL-BASED HIGH-DIMENSIONAL NETWORK INFERENCE: THEORY & METHODS

A Dissertation

Submitted to the Faculty

of

Purdue University

by

Min Ren

In Partial Fulfillment of the

Requirements for the Degree

of

Doctor of Philosophy

December 2018

Purdue University

West Lafayette, Indiana

THE PURDUE UNIVERSITY GRADUATE SCHOOL
STATEMENT OF DISSERTATION APPROVAL

Dr. Dabao Zhang, Chair
Department of Statistics
Dr. Min Zhang
Department of Statistics
Dr. Bruce Craig
Department of Statistics
Dr. Lingsong Zhang
Department of Statistics

Approved by:
Dr. Hao Zhang
Head of the Department Graduate Program

To my parents.

ACKNOWLEDGMENTS

First of all, I would like to express my sincere appreciation to my advisor Professor Dabao Zhang, for his superb advice, his patience and his continuous and generous support of my Ph.D. study in the past five years. Every time I stopped by his office, he was ready to offer help. Every time I got stuck in research and felt really frustrated, he would always encourage and motivate me to overcome the difficulties and barriers. I am deeply grateful to him for the tremendous efforts and time he has given to me. I would also like to thank Professor Min Zhang for her generous and unending support in both my research and life. I also obtained a lot of experience through the collaboration projects Professor Min Zhang provided. I would like to thank Professor Lingsong Zhang and Professor Bruce Craig for serving as my committee member and providing me with many valuable and insightful suggestions on my projects. Without the incredible support and help from all of them, this dissertation may not have been completed.

I would also like to express my big gratitude to our departmental IT specialist Mr. Doug Crabill for his continuous and excellent assistance in Linux system and large scale parallel computing, which saved me hundreds of thousands of computing hours in my research. I would also like to extend my thanks to Prof. Jun Xie for giving me many heartful suggestions for my teaching work. Through multiple talks with her, I learned a lot of lessons and advice about how to get along with colleagues and work professionally. I would like to thank Professor Bruce Craig and Ce-Ce Furtner for providing me with the valuable opportunity and support to be a statistical consultant in the Statistical Consulting Service, which helped me accumulate a lot of experience on applying my statistical knowledge to solve real world problems. I would like to thank all the faculty for the impressive courses they offered and all the departmental staff for the service they provided.

I also want to express deep gratitude to my group member Dr. Chen Chen, for the great collaboration and inspiration in past several years. To my office mates, Shen Zhou, Eric Gerber, and Yucong Zhang, I really appreciate your constant support and encouragement. I will always remember the funny moments with you guys. I would also like to extend my thanks to many friends here, including but not limited to Hui Sun, Yuying Song, Yixuan Qiu, Botao Hao, Yun Lu, Jiexin Duan, Rongrong Zhang, Donglai Chen, Yumin Zhang, and Sarah Liu for their fabulous support of my life here.

Last but not least, I would like to express my tremendous gratitude to my beloved parents for their unconditional love.

ACKNOWLEDGMENTS TO THE GTEX PROJECT

The Genotype-Tissue Expression (GTEx) Project was supported by the Common Fund of the Office of the Director of the National Institutes of Health, and by NCI, NHGRI, NHLBI, NIDA, NIMH, and NINDS. The data used for the analysis described in this paper were obtained from dbGaP accession number phs000424.v7.p2 on 08/18/2017.

TABLE OF CONTENTS

Page

Page

LIST OF TABLES

Table Page

LIST OF FIGURES

Figure Page

MATHEMATICAL NOTATIONS

$\|\| \cdot \|\|_2$	ℓ_2 norm of a vector
$\|\| \cdot \|\|_1$	ℓ_1 norm of a vector
$\\| \cdot \\|_\infty$	maximal absolute value of the components of a vector
$\\| \cdot \\|_{-\infty}$	minimal absolute value of the component of a vector
$\| \cdot \|_1$	element-wise absolute values of a vector
$\\|A\\|_1$	$\max_{1\leq j\leq n} \sum_{i=1}^m \|a_{ij}\|$ for matrix $A = (a_{ij})_{m\times n}$, i.e., the maximal column sum of absolute values of its components
$\\|A\\|_\infty$	$\max_{1\leq i\leq m} \sum_{j=1}^n \|a_{ij}\|$ for matrix $A = (a_{ij})_{m\times n}$, i.e., the maximal row sum of absolute values of its components
a_i	the i-th entry of vector a
a_{-i}	the subvector excluding the i-th entry of vector a
$a_{\mathcal{S}}$	the sub-vector of vector a indexed by index set $\mathcal{S}$
$a_{\mathcal{S}_i}$	the sub-vector of vector a_i indexed by index set $\mathcal{S}_i$ for simplicity
A_i	the i-th column of matrix A
A_{-i}	the sub-matrix of A excluding its i-th column
$A_{\mathcal{S}_i}$	the submatrix of matrix A_i including columns indexed by index set $\mathcal{S}_i$ for simplicity
$a \vee b$	the maximum of a and b
$a \wedge b$	the minimum of a and b
$\lambda_{\min}(\cdot)$	the minimum eigenvalue of a matrix
$\lambda_{\max}(\cdot)$	the maximum eigenvalue of a matrix
$\mathbb{E}(\cdot)$	the expectation
$\mathbb{P}(\cdot)$	the probability of event
$\asymp$	two terms at the same order
$\mathrm{tr}(\cdot)$	the trace of a matrix

$\|S\|$	the number of elements in set S
$j\|p$	the remainder of j when divided by p,where j and p are positive integers
$a = O(b)$	implying that a/b is bounded by some constant
$a = o(b)$	implying that a/b goes to zero
$\mathrm{supp}(\cdot)$	the support set of a vector, i.e., its non-zero indices
$C_i, c_i, \tilde{c}_i, t_i$	positive constant numbers indexed by integer $i = 1, 2, 3, \cdots$

ABBREVIATIONS

2SLS	Two-Stage Least Squares
SEM	Structural Equation Model
GRN	Gene Regulatory Network
2SPLS	Two-Stage Penalized Least Squares
ReDNet	Reparametrization-Based Differential Analysis of Directed Network
eQTL	expression Quantitative Trait Locus
SNP	Single Nucleotide Polymorphism

ABSTRACT

Ren, M. Ph.D., Purdue University, December 2018. Model-Based High-Dimensional Network Inference: Theory & Methods. Major Professor: Dabao Zhang.

In the past several decades, the advent of high-throughput biotechnologies for genomics study provides appealing opportunities for us to understand the complex gene interaction inside biological systems, attracting many researches in constructing gene regulatory networks (GRNs). Motivated by the promise of the genetical genomics study, our research group has recently focused on representing gene regulatory networks using structural equation models and further revealing system-wide gene regulations. This dissertation presents two recent works along this direction.

Firstly, we conducted thorough theoretical analysis of the recently proposed Two-Stage Penalized Least Squares (2SPLS) method for constructing large systems of structural equation models. We establish the estimation and prediction error bounds for results at both stages of 2SPLS as well as its variable selection consistency. Specifically, a bounded eigenvalue assumption is imposed to ensure the consistency properties of the ℓ_2-penalized regressions at the first stage. At the second stage, the estimation and variable selection consistency of the ℓ_1-penalized regressions are obtained by assuming a restricted eigenvalue condition and a variant of irrepresentable condition, which are both commonly employed in the current literature. We will show that the 2SPLS estimator works not only for fixed dimensions but also diverging dimensions which can grow to infinity with the sample size but at an appropriate rate.

Secondly, we developed a novel statistical method to identify structural differences between two cognate networks characterized by structural equation models. We propose to reparameterize the model to separate the differential structures from common structures, and then design an algorithm with calibration and construction stages to identify these differential structures directly. The calibration stage serves to obtain consistent prediction by building the ℓ_2 regularized regression of each endogenous variables against pre-screened exogenous variables, correcting for potential endogeneity issue. The construction stage consistently selects and estimates both common and differential effects by undertaking ℓ_1 regularized regression of each endogenous variable against the predicts of other endogenous variables as well as its anchoring exogenous variables. Our method allows for easy parallel computation. Theoretical results are obtained to establish non-asymptotic error bounds of predictions and estimates at both stages. Our studies on simulated data demonstrated that the proposed method performed much better than independently constructing networks. A real data set was analyzed to illustrate the applicability of our method.

1. INTRODUCTION

1.1 Motivation From Understanding Gene Regulation

The past several decades witnessed the profound advent of high throughput genome, transcriptome, microbiome and more broadly “omics” sequencing technologies. Besides the wet lab experiments, the massive scale of data produced by these tools provide a multitude of ways for life science researchers and health-care scientists to uncover the complex interactions and relationships within organic systems. However, the vast scale and the heterogeneity of the generated data induce many challenges, calling for development of novel computational efficient and powerful statistical methodologies (Marx, 2013).

The gene expression data and DNA genotypic data are two fundamental blocks in the quantification of biological systems due to the central dogma, which formally illustrates the flow of inheritance information from DNA genotype to gene expression via transcription and finally to phenotypes via translation and other downstream processes. Due to the rapid advances in sequencing technologies, it is becoming more and more affordable and feasible to collect gene expression data or both of the whole genome gene expression and genotype information for each individual of a large population, i.e., genetical genomics data. In the past decades, numerous projects have been dedicated to obtain and curate these data, for instance, the Gene Expression Omnibus (GEO; Edgar et al., 2002), the Genotype-Tissue Expression (GTEx) project (Consortium and Others, 2015), and the 1001 Arabidopsis database (Kawakatsu et al., 2016). The rich sets of data have stimulated a myriad of developments of statistical methods by employing gene expression data, genotypic data or both of them in order to understand and model the biological systems in different fashions, for instance, the construction of gene regulatory networks (GRN) and cis/trans-eQTLs identification.

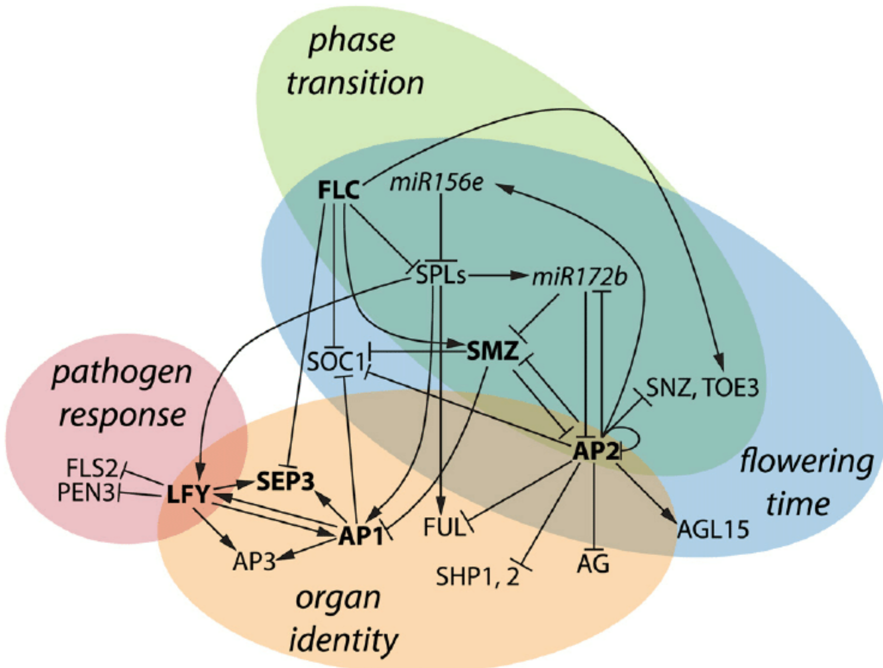


Figure 1.1. An Example of Constructed Gene Regulatory Network in Yant (2012) for a set of 17 Genes. Genes in different colored oval circles are grouped by their biological functionality. The regulation effects between genes are indicated by edges. The causal effects of genotypes are not shown.

Each species has a set of genes, for instance, human genome has around twenty thousand genes and fruit fly genome has around fifteen thousand genes. All the genes rarely act independently, but instead interact with each other in an orchestra fashion, whose interaction relationship can be naturally represented by a network, i.e., gene regulatory network (GRN). Figure 1.1 demonstrates an illustrative example of gene regulatory network for a set of 17 genes. The reconstruction of gene regulatory networks from data provides an important tool to explore and understand the organization and functionality of genes. Those novel information may further facilitate the characterization of the genetic profiles of many complicated diseases and biological traits. A great deal of approaches has been developed to construct gene regulatory networks by using only gene expression data, including partial correlation/graphical models (Dobra et al., 2004; Friedman, 2004; Yuan and Lin, 2007; Friedman et al., 2008; Menéndez et al., 2010), Bayesian and dynamic Bayesian networks (Friedman et al., 2000; Tamada et al., 2003; Zou and Conzen, 2004; Kim et al., 2004; Young et al., 2014), correlation based co-expression networks (Carter et al., 2004; Daub et al., 2004; Langfelder and Horvath, 2008; Teng and Huang, 2009). Furthermore, it would be very promising to improve the performance of gene regulatory network construction by combining gene expression data and whole genome genotypic data. There are also many proposed methods in this direction (Xiong et al., 2004; Liu et al., 2008; Logsdon and Mezey, 2010; Cai et al., 2013; Ni et al., 2016, 2018).

Gene regulatory network provides a concise representation of gene interactions for a single population. Sometimes, it is of more importance to compare or detect the differences between two cognate networks from different but related populations, such as healthy population and diseased population, or different tissues, such as heart cells and muscle cells (West et al., 2012). This network differences are also commonly referred to as **differential network**, which may help us gain critical insights into the deep mechanism of the development of complex diseases or the intriguing underlying biological processes of cell differentiation. Those insights may further assist us with developing more efficient and personalized drugs or therapies for different diseases or tissues. In view of the importance, many research efforts have been dedicated to this direction and a diversity of methods have been proposed, including correlation and entropy based methods (Fuller et al., 2007; Gill et al., 2010; de la Fuente, 2010; Gambardella et al., 2013) and similar F-statistic based methods (Lai et al., 2004; Ma et al., 2011), Gaussian graphical model based direct estimation and testing methods (Zhao et al., 2014; Xia et al., 2015; Liu et al., 2017b) and Fused lasso based methods (Zhang and Wang, 2012; Zhang et al., 2016; Liu et al., 2017a) and its related D-trace loss based methods (Yuan et al., 2017; Zhang et al., 2017).

1.2 Inferring Gene Regulation via Structural Equation Models

In this section, we first review structural equation model (SEM) and then introduce the use of SEM to characterize a gene regulatory network for one population, and then two cognate gene regulatory networks for two different but related populations.

1.2.1 Structural Equation Models and Simultaneous Equation Models

Structural equation models (SEMs) include a broad class of statistical models which provide a general framework for modeling complex dependence structures in multivariate data involving unobserved latent variables, observed variables or both of them (Jöreskog, 1970; Bollen and Noble, 2011). The flexibility of SEMs has resulted in widespread applications in a diverse of fields, including econometrics, social science, genetics, and behavioral science (Reiss and Wolak, 2007; Liu et al., 2008; Hoyle, 2012). There are also many well established statistical software packages specifically designed for the estimation and inference of SEMs, such as *LISREL* (Jöreskog, 1970),*Mplus* (Muthén and Muthén, 2017) and *AMOS* (Byrne, 2016). Although structural equation model is usually perceived to involve latent variables, observed variables only models are also commonly employed, such as simultaneous equation models which are popular in econometrics to model the dependence structures among a system of observable endogenous variables and exogenous variables (Nelson and Olson, 1978; Lee, 1982; Cai, 2010; Jeanty et al., 2010; Omri et al., 2014; Adewuyi and Awodumi, 2017). Its popularity in econometrics and related social sciences attracted many researches, for instance, the classical two-stage least square estimation method (Theil, 1953a,b, 1961; Basmann, 1957; Sargan, 1958) and the model estimation and its identification (Amemiya, 1977; Kai, 1998; Wilde, 2000; Imbens and Newey, 2009; Dijkstra and Henseler, 2015). As further pointed out in Bentler and Weeks (1980), Sánchez et al. (2005) and Bollen and Noble (2011), simultaneous equation models can be considered as special cases of a general formulation of structural equation models. Moreover, both structural equation model and simultaneous equation model enjoy nice interpretations

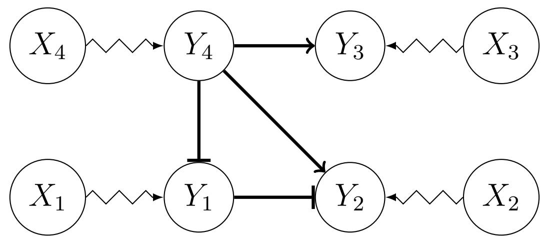


Figure 1.2. An Illustrative Example of Gene Regulatory Network. For $i = 1, 2, 3, 4$, X_i denotes the direct causal factors to Y_i. Y_i denotes the gene expression level. The solid lines refer to the regulation among gene expressions, while the zigzag lines refer to the causal effects on each gene, such as cis-eQTLs.

in causal inference setting (Koster et al., 1996; Pearl, 2003; Pearl et al., 2009). Thus, SEMs facilitate our understanding and interpretation of reconstructed gene regulatory networks in our study.

In this dissertation, we intend to model and study gene regulation structures using a system of linear equation models with observed endogenous and exogenous variables only. In this sense, our employed structure equation models refer to the simultaneous equations models in econometrics (Sánchez et al., 2005). Notwithstanding, following the current trend beyond econometrics and social sciences, we here choose to mainly use the term "structural equation models" as it is a more broad and general term.

1.2.2 Representation of Gene Regulatory Network

Structure equation models can readily characterize the dependence structures among a system of random variables. Thus, the interaction or regulation among genes can be naturally represented by the SEMs. Figure 1.2 provides an simple illustrative example of the gene regulatory network with four genes $Y_1, \ldots, Y_4$ and their causal factors $X_1, \ldots, X_4$. For a linear system with p genes, the gene regulatory network can be formulated by the linear structural equation models and each gene is regulated by effects from two major sources, i.e., two types of causal effects. One is the regulatory effects from other genes, and the other is the anchoring regulatory effects from its cis-expression Quantitative Trait Locus (cis-eQTLs) which are local genotypes inside the gene region and regulate the expression of the gene. Then, the formal model for each gene can be concisely formulated as follows.

$$\underbrace{\mathbf{Y}_i}_{\text{gene } i} = \underbrace{\mathbf{Y}_{-i}\boldsymbol{\gamma}_i}_{\substack{\text{regulatory effects}\\ \text{by other genes}}} + \underbrace{\mathbf{X}\boldsymbol{\phi}_i}_{\substack{\text{anchoring regulation}\\ \text{by cis-eQTLs}}} + \underbrace{\boldsymbol{\epsilon}_i}_{\text{error terms}}, \tag{1.1}$$

where $\mathbf{Y}_i$ and $\mathbf{Y}_{-i}$ denote the i-th column and the submatrix by excluding the i-th column of the $n \times p$ matrix $\mathbf{Y}$, $n \times q$ matrix $\mathbf{X}$ denotes the genotype matrix and $\boldsymbol{\gamma}_i$ and $\boldsymbol{\phi}_i$ encode the two types of causal effects for each term, respectively. In model (1.1), $\mathbf{Y}_{-i}$ and $\mathbf{X}$ are also commonly referred to as **endogenous** and **exogenous** variables, respectively (Fan and Liao, 2014), since $\mathbf{Y}_{-i}$ may not be independent of the error terms $\boldsymbol{\epsilon}_i$, while $\mathbf{X}_i$ and $\boldsymbol{\epsilon}_i$ are assumed to independent of each other.

The additional anchoring regulation plays an important role in revealing the directionality of the gene regulation. With proper identification assumption, say, each gene is assumed to have at least one unique direct causal factor as an anchoring variable, model (1.1) can not only identify the directionality of the regulation but also allow for both acyclic and cyclic or loop structures. As illustrated in Figure 1.3, we can not recover the directionality between node Y_1 and Y_2 without the extra information provided by the direct causal factors X_1 and X_2 because all four sub-networks consisting of Y_1 and Y_2 (without X_1 and X_2) will be Markov equivalent. Thus, the anchoring variables help reveal the directionality of the regulations between genes of the interest. The directionality and flexibility are crucial for researchers to understand complicated biological pathways.

For the analysis of genetical genomics data, $\mathbf{Y}_i$ usually represents the gene expression value of the i-th gene, which quantifies its degree of activity and can be measured by a microarray chip or RNA sequencing platform, while the matrix $\mathbf{X}$ consists of the cis-eQTLs of all genes in the study. In this dissertation, we just used the cis-eQTL data to construct the network for simplicity. In practice, some genes may not have significant anchoring factors. However, many other biological factors

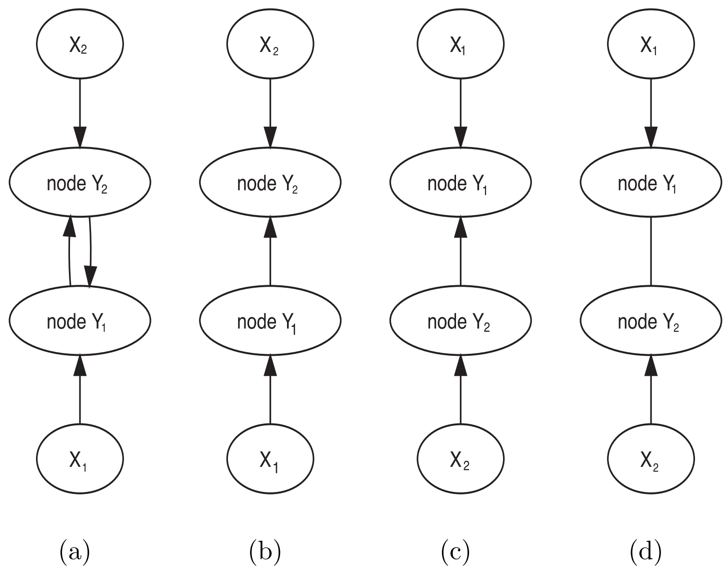


Figure 1.3. An Illustrative Example of Networks Which Are Not Markov Equivalent. However, without X_1 and X_2, sub-networks consisting of only node Y_1 and Y_2 will be Markov equivalent.

can be incorporated into the model as the direct causal factors, such as copy number variation, cis-acting methylation eQTLs or aggregated eQTLs from rare and low frequency SNPs introduced by Chen (2017). Model (1.1) has been employed in several studies to represent the gene regulatory networks, such as Xiong et al. (2004), Liu et al. (2008), Logsdon and Mezey (2010), and Cai et al. (2013). Most recently, Chen (2017) proposed to estimate model (1.1) in a two stage fashion coupled with penalization to achieve better performance, especially variable selection in a diverse of settings. We will investigate its theoretical properties in Chapter 2, in particular, for the case that dimensions p, q can grow to infinity with the sample size but at appropriate rate.

1.2.3 Differential Gene Regulatory Networks

As discussed in Section 1.1, it is oftentimes of more interest and importance to detect differences between two cognate gene regulatory networks, which is often referred to as differential network. Figure 1.4 demonstrates a simple illustrative example for two cognate or similar networks and their differential network. Both networks have four genes and their corresponding causal factors. The differences of these two networks are the disappearance of regulation effects from $\mathbf{Y}_4$ to $\mathbf{Y}_1$ and from $\mathbf{Y}_4$ to $\mathbf{Y}_3$. We omit the nodes of direct causal factors in the differential network for simplicity.

In this light, it is natural to represent each network by a structural equation model, which can be formulated as the model in the below,

$$\begin{cases} \mathbf{Y}_i^{(1)} &= \mathbf{Y}_{-i}^{(1)}\boldsymbol{\gamma}_i^{(1)} + \mathbf{X}^{(1)}\boldsymbol{\phi}_i^{(1)} + \boldsymbol{\epsilon}_i^{(1)}, \\ \mathbf{Y}_i^{(2)} &= \mathbf{Y}_{-i}^{(2)}\boldsymbol{\gamma}_i^{(2)} + \mathbf{X}^{(2)}\boldsymbol{\phi}_i^{(2)} + \boldsymbol{\epsilon}_i^{(2)}. \end{cases} \tag{1.2}$$

For $k \in \{1, 2\}$, the $n^{(k)} \times p$ matrix $\mathbf{Y}^{(k)} = [\mathbf{Y}_1^{(k)}, \ldots, \mathbf{Y}_p^{(k)}]$ denotes the gene expression matrix for p genes, and $n^{(k)} \times q$ matrix $\mathbf{X}^{(k)}$ denotes the direct casual factors of all the genes for each network. $\boldsymbol{\phi}_i^{(k)}$ consists of the effects of direct casual factors, such as cis-eQTLs and $\boldsymbol{\gamma}_i^{(k)}$ encodes the regulation structures.

Here, we mainly focus on the detection and the estimation of the non-zeros entries of the difference between $\boldsymbol{\gamma}_i^{(1)}$ and $\boldsymbol{\gamma}_i^{(2)}$, i.e., $\boldsymbol{\gamma}_i^{(1)} - \boldsymbol{\gamma}_i^{(2)}$. In practice, we assume that the two populations underlying the cognate networks are related, such as the healthy population and diseased population. Therefore, we assume that majority of the regulation structures of the two networks are similar to each other and the detection of the sparse differences in $\boldsymbol{\gamma}_i^{(1)} - \boldsymbol{\gamma}_i^{(2)}$ for each node i will be of the main interest.

Naively, we can construct the gene regulatory networks and estimate the regulatory effects $\boldsymbol{\gamma}_i^{(k)}$ independently. However, this approach will fail to take advantage of the similarities of the networks, which may result in high false positive rate or low power. In order to take account of this similarities or commonality, we propose to reparametrize the models and estimate the differential effects in a direct fashion, which could lead to much lower false discovery rate and better variable selection performance.

By virtue of the flexibility of structural equation models and the additional anchoring regulations, model (1.2) naturally reveals the directionality of regulations for both networks. Therefore, it can not only identify the change of effect sizes but also the change of the direction of regulations. Similar to the 2SPLS method, an slightly relaxed identification assumption will be imposed on the data for revealing the direction of regulations as well.

1.3 Challenges in Revealing Gene Regulation

In the current big data era, massive biological data from a variety of sequencing technologies offer a promising opportunity to infer the gene regulatory networks and differential gene regulatory networks, which can help us understand the gene-gene or gene-genotype interactions in a data driven manner. However, the promise is hindered by four major challenges: model flexibility, high dimensionality, computational burden, and differential analysis. Firstly, many current network inference methods focus on undirected or acyclic networks, such as graphical model and Bayesian network

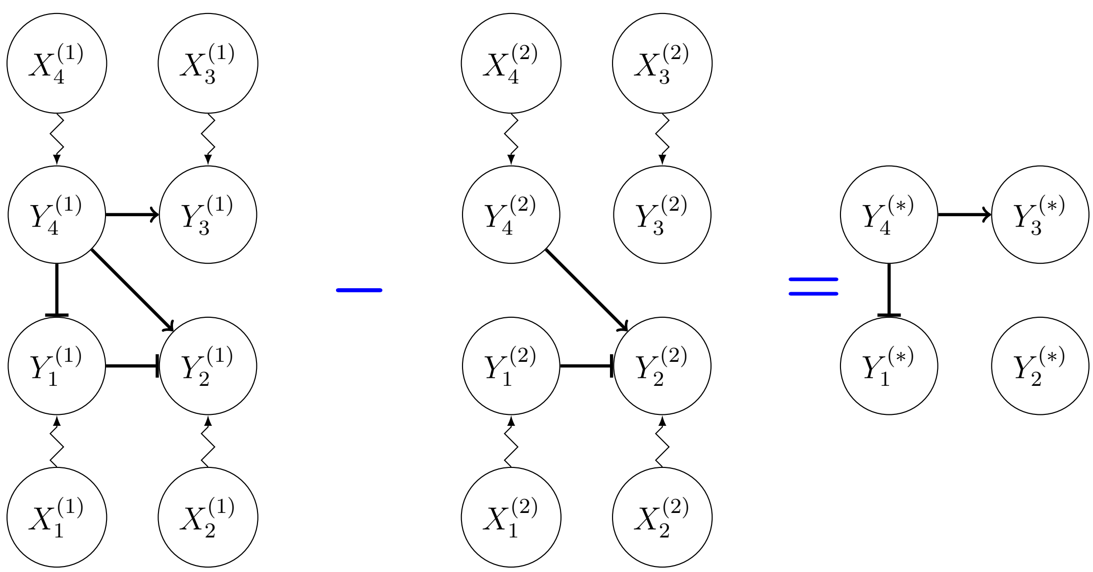


Figure 1.4. An Illustrative Example for Differential Gene Regulatory Network. The superscripts $1, 2$ are the indices for two networks. The differential network between the two networks is indicated by nodes with superscript $*$.

methods. In contrast, the flexible structural equation models can not only reveal the directionality of regulation but also allow for both acyclic and cyclic structures. As discussed before, the information of directionality is crucial for life science researchers to study complex biological pathways. Moreover, cyclic structures offer a promising way to understand important feedback structures in biological systems. Secondly, current large scale biological data commonly have high dimensionality, i.e., the number of predictors can be comparable or even larger than the sample size, which could result in inefficiency or even failure of traditional methods. Thus, we employ multiple penalization approaches to control the negative impact of high dimensionality. Thirdly, large scale data can pose significant demands on computation for even simple models. Most current methods focus on constructing networks as a whole, which is infeasible for parallel computation and may not be applicable for large scale data. By virtue of the node-wise representation of SEMs, we can fit one linear model for each node in a parallel fashion, which makes our method applicable for large datasets with hundreds of thousands of endogenous and exogenous variables. Fourthly, most studies on network construction focus on single network and the research on differential analysis of networks are relatively scarce due to its difficulty, especially in directed network setting. Moreover, combined with the other three aforementioned challenges, the differential analysis of directed networks with flexible structures is much more complicated. In conclusion, we hope our methods can facilitate the understanding of complex biological systems based on large scale data, which might eventually help us develop new drugs and therapies.

1.4 Sketch of the Research

The rest of the dissertation is organized as follows.

In Chapter 2, Section 2.2 and 2.3 first review the model setup and methodology of the Two-Stage Penalized Least Squares (2SPLS) method proposed in Chen (2017). Then, Section 2.4 presents the theoretical analysis of 2SPLS estimator for diverging dimensions. Our theoretical results show that 2SPLS estimator works not only for fixed dimensions but also for diverging dimensions, say $p \asymp q = o(n)$. The detailed technical proofs are displayed in Section 2.5.

In Chapter 3, Section 3.2 introduces the model and identifiability assumption of Reparametrization-Based Differential Analysis of Directed Networks (**ReDNet**) method, which detects the structural differences between two cognate networks characterized by structural equations models. Section 3.3 presents the methodology and comprehensive theoretical analysis of the **ReDNet** method. We will show the non-asymptotic error bounds for both the calibration and construction stages and discuss how the error bounds can be well controlled with proper sample sizes of both networks, the dimensions and other parameters. The simulation study is detailed in Section 3.4 to demonstrate the superior performance of our method in practice. Section 3.5 presents a real data analysis example to show the applicability of the proposed method. The technical proofs of all the theorems are relegated to Section 3.7.

Chapter 4 concludes the dissertation with a summary of all the works.

2. TWO-STAGE PENALIZED LEAST SQUARE (2SPLS) METHOD TO CONSTRUCT LARGE NETWORKS

2.1 Introduction

It is presumably expeditious to reveal gene regulation via genetical genomics data. However, the promise is far from realized due to the lack of a systematic and efficient approach to construct the networks, especially the directed networks that allow for both acyclic and cyclic or loop structures. Structural equation model (1.1) provides us an flexible and promising way to construct such networks and similar models were studied in many past literatures, such as Xiong et al. (2004), Liu et al. (2008), Logsdon and Mezey (2010), and Cai et al. (2013). Xiong et al. (2004) proposed the use of structural equation models to represent the gene regulatory networks and employed genetic algorithm to search for optimal network structure by minimizing the Akaike Information Criterion (AIC; Akaike, 1974). Similarly, Liu et al. (2008) also utilized the genetic algorithm to determine the network topology of the structural equations but select the model by minimizing Bayesian Information Criterion (BIC; Schwarz et al., 1978) or its variants (Broman and Speed, 2002) instead of AIC for the optimal genetic networks. All the aforementioned methods are only suitable for a small number of genes and genotype markers. In a large scale study, such as whole human genome network construction, there will be hundreds of thousands of endogenous variables (genes) and exgeonous variables (genotype markers). Thus, Cai et al. (2013) instead developed a regularized likelihood based approach to infer a sparse network as a whole. However this method is not easy for parallel computing and thus may not be able to scale for massive data.

In the current big data era, it is commonly computationally formidable to fit a large system based on the likelihood function of the complete model. In her dissertation, Chen (2017) instead proposed to identify a large network system via two-stage estimation approach on a set of limited information models, each for one endogenous variable in the system (Schmidt, 1976). Based on the instrumental variables (IVs) interpretation of the classical two-stage least squares method (2SLS; Theil, 1953a,b, 1961; Basmann, 1957), the estimation consistency of model parameters depends on the consistent estimation of the conditional expectations of the endogenous variables which are also referred to as the optimal instruments. Thus, Chen (2017) extended the classical 2SLS method and proposed a two-stage penalized least squares (2SPLS) method to fit penalized linear regression at both stages. At the first stage, ℓ_2 regularized linear model is employed to obtain the consistent prediction of the endogenous variables. Then, with the endogenous variables being replaced by its predicted values at the second stage, ℓ_1 regularized adaptive lasso step is utilized to identify the non-zero regulatory effects from a large pool of candidates.

The proposed 2SPLS method in Chen (2017) tackles two major challenging issues, i.e., flexibility and computational burden. Firstly, the structural equation model is very flexible for inferring the gene regulatory networks and allows for both acyclic and cyclic or loop structures. The current literature in machine learning and statistics mainly focus on the study of estimation of undirected and acyclic networks. The research on the identification of cyclic network structure are comparatively scarce due to its difficult nature. However, the loop or feedback regulation structures are indeed present in many species and of the great research interest as well (Boyer et al., 2005; Cooper et al., 2008; Chen and Wu, 2013; Lee et al., 2016). Thus, the proposed method may shed new light on this direction. Secondly, since the 2SPLS method identify the regulatory effects in a node-wise fashion, it's inherently very easy for parallel computation. Therefore, Many resampling methods, such as the bootstrap, are viable for evaluating the significance of detected regulatory effects even for large scale datasets.

In this chapter, we first review the methodology of 2SPLS method proposed in the dissertation by Chen (2017). We intend to analyze its theoretical properties of both stages in depth. Our derived theorems allow for diverging dimensions. We show that, when the numbers of endogenous and exogenous variables grow with the sample size at a polynomial rate, the established consistency properties hold for the 2SPLS estimator.

2.2 The Identifiable Structural Equation Models

For a system with p endogenous variables (genes) and q exogenous variables (genotype markers etc.), there will be p of model (1.1). Then, the p linear equations can be combined and rewritten in a systematic fashion as follows

$$\mathbf{Y} = \mathbf{Y}\boldsymbol{\Gamma} + \mathbf{X}\boldsymbol{\Phi} + \boldsymbol{\epsilon}, \tag{2.1}$$

where the $n \times p$ matrix $\mathbf{Y}$ denotes the n samples from the endogenous variables, the $n \times q$ matrix $\mathbf{X}$ denotes n sample from the exogenous variables and $p \times p$ matrix $\boldsymbol{\Gamma}$ and $q \times p$ matrix $\boldsymbol{\Phi}$ contain the regulatory effects and causal effects, respectively. Without loss of generality, each column of $\mathbf{X}$ is standardized to have ℓ_2 norm $\sqrt{n}$. In particular, the diagonal line of $\boldsymbol{\Gamma}$ are all zero. Each component of $\boldsymbol{\epsilon}$ is assumed to independently distributed as zero mean normal distribution and the matrix $\mathbf{X}$ is assumed to be independent of the error term $\boldsymbol{\epsilon}$.

The structural equation model (2.1) suffers from identifiability issue as other structure equation based models. Thus, proper identifiability assumption is needed. In this paper, we follow the common assumption in Logsdon and Mezey (2010) and Cai et al. (2013), which each endogenous variable is assumed to have a unique set of exogenous variables. In other words, the nonzero indices of the causal effects $\boldsymbol{\phi}_i$ are nonempty and can be pre-determined. This serves as "prior" information for the causal structure of the model and can be determined by domain knowledge. For example, each gene is directly affected by its local SNPs, a.k.a, cis-eQTLs, due to the central dogma. We emphasize that the true value of $\boldsymbol{\phi}_i$ still need to be estimated if necessary, though its nonzero indices are known. Denote the known set as $\mathcal{A}_i = \text{supp}(\boldsymbol{\phi}_i), \quad i = 1, 2, \ldots, p$. Then, the identifiability assumption below which has been considered in the dissertation Chen (2017) is formulated as follows,

Assumption 1. *For $i = 1, \cdots, p$, $\mathcal{A}_i \neq \emptyset$, but $\mathcal{A}_i \cap \mathcal{A}_j = \emptyset$ for $i \neq j$.*

Note that this assumption also satisfies the rank condition in Schmidt (1976) that is a sufficient model identifiability assumption. Since the support set of $\boldsymbol{\phi}_i$ is known, henceforth, the model for each node can be further rewritten as a limited information model,

$$\mathbf{Y}_i = \mathbf{Y}_{-i}\boldsymbol{\gamma}_i + \mathbf{X}_{\mathcal{A}_i}\boldsymbol{\phi}_{\mathcal{A}_i} + \boldsymbol{\epsilon}_i, \tag{2.2}$$

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\pi} + \boldsymbol{\xi}, \tag{2.3}$$

where the error term $\boldsymbol{\epsilon}_i \sim N(\mathbf{0}, \sigma_i^2\mathbf{I}_n)$. The equation (2.3) is also commonly referred to as reduced form equation and the effects of the reduced model $\boldsymbol{\pi} = \boldsymbol{\Phi}(\mathbf{I} - \boldsymbol{\Gamma})^{-1}$ and its error term $\boldsymbol{\xi} = \boldsymbol{\epsilon}(\mathbf{I} - \boldsymbol{\Gamma})^{-1}$. The reduced form equation (2.3) reveals that the direct causal factors in $\mathbf{X}$ serve as the instrumental variables for the full information model (2.1) in addition to being the anchoring regulation. This motivates the application of the instrumental variables based method (Reiersøl, 1941, 1945; Anderson, 2005).

2.3 Two-Stage Penalized Least Squares (2SPLS) Method

2.3.1 The Review of 2SPLS Method

In this section, we review the 2SPLS estimator proposed in Chen (2017). Each step of the 2SPLS method is detailed in Algorithm 1. In the first step, we aim to obtain consistent prediction $\hat{\mathbf{Z}}$ of $\mathbf{Z} = E(\mathbf{Y}|\boldsymbol{X})$ as the optimal instruments based on the reduced form equation (2.3). Variable selection is not necessary in this step. Therefore, we employ ridge regression to the model $\mathbf{Y}_i = \mathbf{X}\boldsymbol{\pi}_i + \boldsymbol{\xi}_i$ to obtain the estimates $\hat{\boldsymbol{\pi}}_i$ and predict $\mathbf{Z}_i$ with $\hat{\mathbf{Z}}_i = \mathbf{X}\hat{\boldsymbol{\pi}}_i$. Denote ρ_i as the ridge tuning parameter.

Algorithm 1 2SPLS Algorithm

Input: Gene expression matrix $\mathbf{Y}$, genotypic matrix $\mathbf{X}$, and known cis-eQTL sets $\mathcal{A}_i$ for $i = 1, 2, \cdots p$.

parallel for $i = 1, \ldots, p$

Step 1. Obtain prediction $\hat{\mathbf{Z}}_i$ of $\mathbf{Z}_i = E[\mathbf{Y}_i|\mathbf{X}]$ using the ridge regression;

parallel end

Collect the prediction from above loop and obtain prediction expression matrix $\hat{\mathbf{Z}}$;

parallel for $i = 1, \ldots, p$

Step 2. Use the adaptive lasso to estimate $\hat{\boldsymbol{\gamma}}_i$ by regressing $\mathbf{H}_i\mathbf{Y}_i$ against $\mathbf{H}_i\hat{\mathbf{Z}}_{-i}$;

parallel end

Output: The regulatory effects in $\hat{\gamma}_1, \hat{\gamma}_2, \cdots, \hat{\gamma}_p$.

Then, the ridge estimator $\hat{\boldsymbol{\pi}}_i$ can be obtained by minimizing the following objective function,

$$||\mathbf{Y}_i - \mathbf{X}\boldsymbol{\pi}_i||_2^2 + \rho_i||\boldsymbol{\pi}_i||_2^2.$$

Following the application of instrumental variables based methods, we replace the endogenous variables $\mathbf{Y}_{-i}$ with its prediction $\hat{\mathbf{Z}}_{-i}$ obtained in the first step and utilize the adaptive lasso regression to identify the regulatory effects $\hat{\boldsymbol{\gamma}}_i$ in the second step. The objective function can be formulated as below,

$$\frac{1}{n}||\mathbf{Y}_i - \hat{\mathbf{Z}}_{-i}\boldsymbol{\gamma}_i - \mathbf{X}_{\mathcal{A}_i}\boldsymbol{\phi}_{\mathcal{A}_i}||_2^2 + \lambda_i\boldsymbol{\omega}_i^T|\boldsymbol{\gamma}_i|_1, \tag{2.4}$$

where $\boldsymbol{\omega}_i$ is a prespecified weight vector inversely proportional to a initial estimator of $\boldsymbol{\gamma}_i$ and λ_i is the adaptive lasso tuning parameter. In order to keep consistent with Chapter 3, we employ factor $1/n$ in equation 2.4 instead of $1/2$ in Chen (2017). Since the objective functions in equation 2.4 and Chen (2017) are essentially equivalent, this change neither affects the solution of 2SPLS algorithm nor our theoretical results here. Only the adaptive tunning parameters differ by a multiplicative term.

Since there is no regularization imposed on $\boldsymbol{\phi}_{\mathcal{A}_i}$, we can first minimize the objective function (2.4) for $\boldsymbol{\phi}_{\mathcal{A}_i}$ and obtain its estimator,

$$\hat{\boldsymbol{\phi}}_{\mathcal{A}_i} = (\mathbf{X}_{\mathcal{A}_i}^T \mathbf{X}_{\mathcal{A}_i})^{-1} \mathbf{X}_{\mathcal{A}_i}^T (\mathbf{Y}_i - \hat{\mathbf{Z}}_{-i} \boldsymbol{\gamma}_i). \tag{2.5}$$

Plugging it back into the objective function, we have a simplified objective function at the second stage to obtain the estimate of regulatory effect $\hat{\boldsymbol{\gamma}}_i$,

$$\hat{\boldsymbol{\gamma}}_i = \arg\min_{\boldsymbol{\gamma}_i} \left\{ \frac{1}{n} ||\mathbf{H}_i \mathbf{Y}_i - \mathbf{H}_i \hat{\mathbf{Z}}_{-i} \boldsymbol{\gamma}_i||_2^2 + \lambda_i \boldsymbol{\omega}_i^T |\boldsymbol{\gamma}_i|_1 \right\}, \tag{2.6}$$

where $\mathbf{H}_i = \mathbf{I}_n - \mathbf{X}_{\mathcal{A}_i} (\mathbf{X}_{\mathcal{A}_i}^T \mathbf{X}_{\mathcal{A}_i})^{-1} \mathbf{X}_{\mathcal{A}_i}^T$ is a $n \times n$ projection matrix.

2.3.2 Motivation of Theoretical Analysis for Diverging Dimensions

Chen (2017) analyzed the theoretical properties of 2SPLS estimator for fixed dimensions p, q and showed that the 2SPLS estimator enjoys the consistency properties in both stages and the oracle variable selection property in the second stage. However, in current large scale genetical genomics data era, the dimensions of data can be quite large, which are usually comparable to the sample size or even large than the sample size. Moreover, the simulation study in Chen (2017) also demonstrates that the 2SPLS estimator can achieve good variable selection performance even when the dimensions are comparable to the sample size. Therefore, it is more practical and interesting to investigate whether the 2SPLS estimator still enjoys the consistency and variable selection properties even for diverging dimension case, i.e., the dimensions p, q can grow with the sample to infinity at an appropriate rate, e.g., polynomial rate.

2.4 Theoretical Properties

We are now ready to investigate the theoretical properties of 2SPLS estimator for diverging dimensions. That is, per Assumption 1, both p and q may grow with sample size n at the the same order $o(n)$. We will first introduce Assumption 2 below for the consistency property of the first stage. All the theoretical properties henceforth will be described by a prespecified sequence $f_n = o(n)$ but $f_n \to \infty$.

Assumption 2. The singular values of the matrix $\mathbf{I} - \mathbf{\Gamma}$ are positively bounded from below and there exist positive constants c_1 and c_2 such that, for any vector δ with $||\delta||_2 = 1$, $c_1 \geq n^{-1/2}||\mathbf{X}\delta||_2 \geq c_2$. Furthermore $r_{ni} \triangleq \rho_i^2||\boldsymbol{\pi}_i||_2^2/n = o(n)$

We have the following properties on the ridge regression estimator of $\boldsymbol{\pi}_i$ from the first stage.

Theorem 2.4.1 *Under Assumptions 1-2, for each ridge regression estimator $\hat{\boldsymbol{\pi}}_i$, there exist constants C_1 and C_2 such that, with probability at least $1 - e^{-f_n}$,*

(a) $||\hat{\boldsymbol{\pi}}_i - \boldsymbol{\pi}_i||_2^2 \leq C_1 \left(r_{ni} \vee q \vee f_n\right)/n;$

(b) $n^{-1}||\mathbf{X}(\hat{\boldsymbol{\pi}}_i - \boldsymbol{\pi}_i)||_2^2 \leq C_2 \left(r_{ni} \vee q \vee f_n\right)/n.$

Denote $r_{\max} = \max_{1\leq i\leq p} r_{ni}$. Then the system-wise losses in both $||\hat{\boldsymbol{\pi}}_i - \boldsymbol{\pi}_i||_2^2$ and $n^{-1}||\mathbf{X}(\hat{\boldsymbol{\pi}}_i - \boldsymbol{\pi}_i)||_2^2$ have upper bounds in the same order as $(r_{\max} \vee q \vee f_n)/n$, with probability at least $1 - e^{-f_n + \log(p)}$. With $q \asymp p \asymp n^c$ for some $c \in (0, 1)$, we can henceforth select $f_n = O(n^c)$ to dominate $\log(p) = O(1)$, i.e. $f_n - \log(p) \to \infty$, and note that we can choose $r_{\max} = O(n^c)$ as well, due to Assumption 2. Therefore the prediction and estimation losses over the whole system at the first stage can be well controlled.

Before we characterize the consistency of estimated regulatory effects $\hat{\boldsymbol{\gamma}}_i$ on the second stage, we first introduce the following concept of restricted eigenvalue which is used to present Assumption 3.

Definition 2.4.1 *The restricted eigenvalue of a matrix $\mathbf{A}$ on an index set $\mathcal{D}$ is defined as*

$$\phi_{re}(\mathbf{A}, \mathcal{D}) = \min_{||\delta_{\mathcal{D}^c}||_1 \leq 3||\delta_{\mathcal{D}}||_1} \frac{||\mathbf{A}\delta||_2}{\sqrt{n}||\delta_{\mathcal{D}}||_2}. \tag{2.7}$$

Denote $\mathcal{D}_i = \text{supp}(\boldsymbol{\gamma}_i)$. We make the following restricted eigenvalue assumption to pave the way for the estimation consistency in Theorem 2.4.2.

Assumption 3. There exists a constant $\phi_0 > 0$ such that $\phi_{\text{re}}(\mathbf{H}_i\mathbf{X}\boldsymbol{\pi}_{-i}, \mathcal{D}_i) \geq \phi_0$. Furthermore, $\|\boldsymbol{\omega}_{\mathcal{D}_i}\|_\infty \leq \|\boldsymbol{\omega}_{\mathcal{D}_i^c}\|_{-\infty}$.

We then have the consistency property of estimator $\hat{\boldsymbol{\gamma}}_i$.

Theorem 2.4.2 *(Estimation Consistency) Assume that, for each node i, the adaptive tuning parameter is chosen as* $\lambda_i \asymp \|\boldsymbol{\omega}_i\|_{-\infty}^{-1}||\boldsymbol{\Gamma}||_1\ ||\boldsymbol{\pi}||_1\sqrt{n^{-1}(r_{max} \vee q \vee f_n)\log p}$ *and* $\sqrt{(r_{max} \vee q \vee f_n)/n} + c_1||\boldsymbol{\pi}||_1 \ \leq \ \sqrt{c_1^2||\boldsymbol{\pi}||_1^2 + \phi_0^2/64C_2|\mathcal{D}_i|}$. *Denote* $h_n = (||\boldsymbol{\Gamma}||_1^2 \wedge 1)\left[(\frac{n}{q}||\boldsymbol{\pi}||_1^2) \wedge (r_{max} \vee q \vee f_n)\right]\log p$. *Under Assumptions 1-3, there exist constants* $C_3 > 0$ *and* $C_4 > 0$ *such that, with probability at least* $1 - e^{-C_3 h_n + \log(4pq)} - e^{-f_n + \log(p)}$, *each 2SPLS estimator* $\hat{\boldsymbol{\gamma}}_i$ *satisfies that*

1. $||\hat{\boldsymbol{\gamma}}_i - \boldsymbol{\gamma}_i||_1 \leq 8C_4 \frac{\|\boldsymbol{\omega}_{\mathcal{D}_i}\|_\infty ||\boldsymbol{\pi}||_1 ||\boldsymbol{\Gamma}||_1}{\phi_0^2 \|\boldsymbol{\omega}_i\|_{-\infty}} |\mathcal{D}_i| \sqrt{\frac{(r_{max} \vee q \vee f_n)\log p}{n}}$,
2. $n^{-1}||\mathbf{H}_i \hat{\mathbf{Z}}_{-i}(\hat{\boldsymbol{\gamma}}_i - \boldsymbol{\gamma}_i)||_2^2 \leq \frac{C_4^2 \|\boldsymbol{\omega}_{\mathcal{D}_i}\|_\infty^2 ||\boldsymbol{\pi}||_1^2 ||\boldsymbol{\Gamma}||_1^2}{\phi_0^2 \|\boldsymbol{\omega}_i\|_{-\infty}^2} |\mathcal{D}_i| \frac{(r_{max} \vee q \vee f_n)\log p}{n}$.

Note that the system-wide upper bounds, defined by replacing $|\mathcal{D}_i|$ with $\max_i |\mathcal{D}_i|$, can also be achieved with probability at least $1 - e^{-C_3 h_n + \log(4pq) + \log(p)} - e^{-f_n + 2\log(p)}$. The available anchoring effects required by the identifiability assumption implies that both $||\boldsymbol{\pi}||_1$ and $||\boldsymbol{\Gamma}||_1$ are positive. Therefore, we have $h_n/\log(p) \to \infty$. As discussed before, when the dimension $p \asymp q \asymp n^c$, we can still choose $f_n = O(n^c)$ to well control the two losses at a sufficiently large probability.

Let $W_i = diag\{\boldsymbol{\omega}_i\}$ and $\mathcal{V}_i = (v_{ij})_{(p-1)\times(p-1)} \triangleq \frac{1}{n}\boldsymbol{\pi}_{-i}^T \mathbf{X}^T \mathbf{H}_i \mathbf{X} \boldsymbol{\pi}_{-i}$. Further denote $W_{\mathcal{D}_i} = diag\{\boldsymbol{\omega}_{\mathcal{D}_i}\}$, $W_{\mathcal{D}_i^c} = diag\{\boldsymbol{\omega}_{\mathcal{D}_i^c}\}$, $\mathcal{V}_{i,21} = (v_{ij})_{i\in\mathcal{D}_i^c, j\in\mathcal{D}_i}$, $\mathcal{V}_{i,11} = (v_{ij})_{i\in\mathcal{D}_i, j\in\mathcal{D}_i}$ and $\theta_i = ||\mathcal{V}_{i,11}^{-1} W_{\mathcal{D}_i}||_\infty$. We then further introduce the Assumption 4 below for the selection consistency theorem.

Assumption 4. There exists a $\zeta \in (0,1)$ such that $||W_{\mathcal{D}_i^c}^{-1} \mathcal{V}_{i,21} \mathcal{V}_{i,11}^{-1} W_{\mathcal{D}_i}||_\infty < 1 - \zeta$.

Theorem 2.4.3 *(Selection Consistency) Suppose that $\mathcal{V}_{i,11}$ is invertible, for each i,* $\sqrt{(r_{max} \vee q \vee f_n)/n} + c_1||\boldsymbol{\pi}||_1 \leq \sqrt{c_1^2||\boldsymbol{\pi}||_1^2 + \min(\phi_0^2/64, \zeta(4-\zeta)^{-1}\|\boldsymbol{\omega}_i\|_{-\infty}/\theta_i)/(C_2|\mathcal{D}_i|)}$ *and* $\min_{j\in\mathcal{D}_i}|\boldsymbol{\gamma}_{ij}| > \frac{\lambda_i \theta_i}{(2-\zeta)}$. *Under Assumptions 1-4, there exists a 2SPLS estimator* $\hat{\boldsymbol{\gamma}}_i$ *satisfying that, with probability at least* $1 - e^{-C_5 h_n + \log(4pq)} - e^{-f_n + \log(p)}$ *for some constant* $C_5 > 0$, $\hat{\mathcal{D}}_i = \mathcal{D}_i$ *with* $\hat{\mathcal{D}}_i = supp(\hat{\boldsymbol{\gamma}}_i)$.

Theorem 2.4.3 states that the true set of signals can be recovered with a large probability approaching to one.

As shown by Hahn and Hausman (2002), the bias of traditional two-stage least squares (2SLS) estimator in fixed dimension setting is inversely proportional to the coefficient of determination R^2 in the first stage regression. In other words, weak instrumental effects may lead to large bias in parameter estimation or vice versa. For our theoretical results, if we take a closer look at the restricted eigenvalue lower bound ϕ_0, we can provide a similar but rather non-formal interpretation for such phenomena. Noting Assumption 3 and Definition 2.4.1 of restricted eigenvalue, we can know that weaker instrument effects in $\boldsymbol{\pi}_{-i}$ may lead to smaller lower bound value ϕ_0, and further larger loss for the bounds or bias in Theorem 2.4.2. Moreover, larger instrumental effects in $\boldsymbol{\pi}$ may result in larger h_n, and further larger probability for the selection consistency in Theorem 2.4.3. However, due to regularization in both stages of 2SPLS method, our bounds are intricate and exact interpretation of the strength of instrumental effects on error bounds is not very straightforward.

2.5 Proofs of Theoretical Properties

Denote ξ_{ji}, and ϵ_{ji} as the j-th row of $\boldsymbol{\xi}_i$ and $\boldsymbol{\epsilon}_i$, respectively. Note that $\boldsymbol{\xi} = \boldsymbol{\epsilon}(\mathbf{I}-\boldsymbol{\Gamma})^{-1}$. Following Assumption 2, the singular values of $(\mathbf{I}-\boldsymbol{\Gamma})$ are positively bounded from below by a constant c. Denote $\sigma_i^2 = \mathrm{var}(\epsilon_{ji})$ and $\tilde{\sigma}_i^2 = \mathrm{var}(\xi_{ji})$. Then $\tilde{\sigma}_i \leq \sigma_{p\max}/c = \max\limits_{1\leq i\leq p}(\sigma_i)/\mathrm{c}$.

2.5.1 Proof of Theorem 2.4.1

(a) From the ridge regression, we have the following closed form solution,

$$\hat{\boldsymbol{\pi}}_i = (\mathbf{X}^T\mathbf{X}+\rho_i I_q)^{-1}\mathbf{X}^T\mathbf{Y}_i = (\mathbf{X}^T\mathbf{X}+\rho_i I_q)^{-1}\mathbf{X}^T\mathbf{X}\boldsymbol{\pi}_i + (\mathbf{X}^T\mathbf{X}+\rho_i I_q)^{-1}\mathbf{X}^T\boldsymbol{\xi}_i.$$

Note that

$$\hat{\boldsymbol{\pi}}_i - \boldsymbol{\pi}_i = -\rho_i(\mathbf{X}^T\mathbf{X}+\rho_i I_q)^{-1}\boldsymbol{\pi}_i + (\mathbf{X}^T\mathbf{X}+\rho_i I_q)^{-1}\mathbf{X}^T\boldsymbol{\xi}_i = \boldsymbol{\mu} + A_i^T\boldsymbol{\xi}_i,$$

where $\boldsymbol{\mu} = -\rho_i(\mathbf{X}^T\mathbf{X} + \rho_i I_q)^{-1}\boldsymbol{\pi}_i$ and $A_i = \mathbf{X}(\mathbf{X}^T\mathbf{X} + \rho_i I_q)^{-1}$. Then we have

$$||\hat{\boldsymbol{\pi}}_i - \boldsymbol{\pi}_i||_2^2 = \underbrace{\boldsymbol{\mu}^T\boldsymbol{\mu}}_{T_1} + \underbrace{2\boldsymbol{\mu}^T A_i^T \boldsymbol{\xi}_i}_{T_2} + \underbrace{\boldsymbol{\xi}_i^T A_i A_i^T \boldsymbol{\xi}_i}_{T_3}. \tag{2.8}$$

Via the singular value decomposition of $\mathbf{X}$, we can have the decomposition $\mathbf{X}^T\mathbf{X} = \mathbf{P}^T\mathbf{U}\mathbf{P}$, where $\mathbf{P}$ is a unitary matrix and matrix $\mathbf{U}$ is a diagonal matrix with non-negative diagonal elements u_i. Therefore,

$$(\mathbf{X}^T\mathbf{X} + \rho_i I_q)^{-2} = \mathbf{P}^T(\mathbf{U} + \rho_i I_q)^{-2}\mathbf{P}.$$

Following Assumption 2, $\lambda_{min}(\mathbf{X}^T\mathbf{X}) > c_2^2 n$ and $\lambda_{max}(\mathbf{X}^T\mathbf{X}) < c_1^2 n$, which implies that $u_i \asymp n$ for all i. Therefore,

$$T_1 = \rho_i^2 \boldsymbol{\pi}_i^T \mathbf{P}^T(\mathbf{U} + \rho_i I_q)^{-2}\mathbf{P}\boldsymbol{\pi}_i = \sum_{i=1}^q \frac{\rho_i^2 a_{ik}^2}{(u_i + \rho_i)^2} = \mathcal{O}(\rho_i^2||\boldsymbol{\pi}_i||_2^2/n^2) = \mathcal{O}(r_{ni}/n), \tag{2.9}$$

where a_{ik} is the i-th element of $\mathbf{a}_i = \mathbf{P}\boldsymbol{\pi}_i$ with $||\mathbf{a}_i||_2 = ||\boldsymbol{\pi}_i||_2$.

For the term T_2, we have that

$$\mathbb{E}[T_2] = 0, \qquad \text{Var}(T_2) = 4\tilde{\sigma}_i^2 \boldsymbol{\mu}^T A_i^T A_i \boldsymbol{\mu}.$$

By the classical Gaussian tail probability, we have

$$\mathbb{P}\left(T_2 \le t\right) \ge 1 - \exp\left\{-t^2/\left(8\tilde{\sigma}_i^2 \boldsymbol{\mu}^T A_i^T A_i \boldsymbol{\mu}\right)\right\}.$$

Note that,

$$\boldsymbol{\mu}^T A_i^T A_i \boldsymbol{\mu} = \rho_i^2 \boldsymbol{\pi}_i^T \mathbf{P}^T(\mathbf{U}+\rho_i I_q)^{-2}\mathbf{U}(\mathbf{U}+\rho_i I_q)^{-2}\mathbf{P}\boldsymbol{\pi}_i = \sum_{i=1}^q \frac{\rho_i^2 u_i a_{ik}^2}{(u_i + \rho_i)^4} = \mathcal{O}(\rho_i^2||\boldsymbol{\pi}_i||_2^2/n^3).$$

Letting $t = \sqrt{8\tilde{\sigma}_i^2 \boldsymbol{\mu}^T A_i^T A_i \boldsymbol{\mu}(f_n + \log 2)}$, we have, with probability at least $1 - e^{-f_n}/2$,

$$T_2 = \mathcal{O}(\sqrt{r_{ni} f_n}/n). \tag{2.10}$$

For the term T_3, we can invoke the Hanson-Wright inequality (Rudelson et al., 2013) to have, for some constant $t_1 > 0$,

$$\mathbb{P}(T_3 \le \mathbb{E}[T_3] + t) \ge 1 - \exp\left\{-t_1 \min\left(\frac{t^2}{\tilde{\sigma}_i^4||A_i A_i^T||_F^2}, \frac{t}{\tilde{\sigma}_i^2||A_i A_i^T||_{op}}\right)\right\},$$

where $||\cdot||_{op} = \max_{x\neq 0}||\cdot x||_2/||x||_2$ is the operator norm and $||\cdot||_F$ is the Frobenius norm.

Since

$$A_iA_i^T = \mathbf{X}(\mathbf{X}^T\mathbf{X}+\rho_iI_q)^{-2}\mathbf{X}^T = \mathbf{X}\mathbf{P}^T(\mathbf{U}+\rho_iI_q)^{-2}\mathbf{P}\mathbf{X}^T,$$

we have

$$\begin{aligned}
\mathbb{E}[T_3] &= \tilde{\sigma}_i^2\mathrm{tr}(A_iA_i^T) = \tilde{\sigma}_i^2\mathrm{tr}(\mathbf{X}^T\mathbf{X}(\mathbf{X}^T\mathbf{X}+\rho_iI_q)^{-2}) \\
&= \tilde{\sigma}_i^2\mathrm{tr}(\mathbf{U}(\mathbf{U}+\rho_iI_q)^{-2}) = \sum_{i=1}^{q}\frac{\tilde{\sigma}_i^2u_i}{(u_i+\rho_i)^2} = \mathcal{O}(\tilde{\sigma}_i^2q/n), \\
||A_iA_i^T||_F^2 &= \mathrm{tr}(A_iA_i^TA_iA_i^T) = \mathrm{tr}(A_i^TA_iA_i^TA_i) \\
&= \mathrm{tr}(P^T\mathbf{U}(\mathbf{U}+\rho_iI_q)^{-2}U(U+\rho_iI_q)^{-2}) = \sum_{i=1}^{q}\frac{u_i^2}{(u_i+\rho_i)^4} = \mathcal{O}(q/n^2), \\
||A_iA_i^T||_{op} &= \mathcal{O}(\lambda_{max}\left(\mathbf{X}\mathbf{X}^T\right)/n^2) = \mathcal{O}(n^{-1}).
\end{aligned}$$

Letting $t = \max\left(\sqrt{\tilde{\sigma}_i^4||A_iA_i^T||_F^2(f_n+\log 2)/t_1}, \tilde{\sigma}_i^2||A_iA_i^T||_{op}(f_n+\log 2)/t_1\right)$, we obtain that, with probability at least $1-e^{-f_n}/2$,

$$T_3 = \mathcal{O}(q/n) + \mathcal{O}(\sqrt{f_nq}/n) + \mathcal{O}(f_n/n). \tag{2.11}$$

Collecting the bounds in (2.9), (2.10), and (2.11), we conclude that there exist a positive constant C_1 such that, with probability at least $1-e^{-f_n}$,

$$||\hat{\boldsymbol{\pi}}_i - \boldsymbol{\pi}_i||_2^2 \leq C_1(r_{ni}\vee q\vee f_n)/n.$$

(b) Similar to (2.8), we have

$$||\mathbf{X}(\hat{\boldsymbol{\pi}}_i-\boldsymbol{\pi}_i)||_2^2 = \underbrace{\boldsymbol{\mu}^T\mathbf{X}^T\mathbf{X}\boldsymbol{\mu}}_{T_4} + \underbrace{2\boldsymbol{\mu}^T\mathbf{X}^T\mathbf{X}A_i^T\boldsymbol{\xi}_i}_{T_5} + \underbrace{\boldsymbol{\xi}_i^TA_i\mathbf{X}^T\mathbf{X}A_i^T\boldsymbol{\xi}_i}_{T_6}.$$

For the term T_4, we have

$$\begin{aligned}
T_4 &= \rho_i^2\mathbf{a}_i^T\mathbf{U}(\mathbf{U}+\rho_iI_q)^{-1}\mathbf{U}(\mathbf{U}+\rho_iI_q)^{-1}\mathbf{a}_i \\
&= \rho_i^2\sum_{i=1}^{q}\frac{u_ia_{ik}^2}{(u_i+\rho_i)^2} = \mathcal{O}(\rho_i^2||\boldsymbol{\pi}_i||_2^2/n) = \mathcal{O}(r_{ni}).
\end{aligned} \tag{2.12}$$

For the term T_5, by the classical Gaussian tail inequality, we have

$$\mathbb{P}\left(T_5 \leq t\right) \geq 1 - \exp\left\{-t^2/(2\mathrm{Var}(T_5))\right\},$$

where

$$\begin{aligned}
\mathrm{Var}(T_5) &= 4\tilde{\sigma}_i^2 \boldsymbol{\mu}^T \mathbf{X}^T \mathbf{X} A_i^T A_i \mathbf{X}^T \mathbf{X} \boldsymbol{\mu} \\
&= 4\tilde{\sigma}_i^2 \rho_i^2 \mathbf{a}_i^T (\mathbf{U} + \rho_i I_q)^{-1} \mathbf{U} (\mathbf{U} + \rho_i I_q)^{-1} \mathbf{U} (\mathbf{U} + \rho_i I_q)^{-1} \mathbf{U} (\mathbf{U} + \rho_i I_q)^{-1} \mathbf{a}_i \\
&= 4\tilde{\sigma}_i^2 \rho_i^2 \sum_{i=1}^{q} \frac{u_i^3 a_{ik}^2}{(u_i + \rho_i)^4} = \mathcal{O}(\tilde{\sigma}_i^2 \rho_i^2 ||\boldsymbol{\pi}_i||_2^2/n).
\end{aligned}$$

Taking $t = \sqrt{2\mathrm{Var}(T_5)(f_n + \log 2)}$, we can obtain that, with probability at least $1 - e^{-f_n}/2$,

$$T_5 = \mathcal{O}(\sqrt{r_{ni} f_n}). \tag{2.13}$$

For the term T_6, by the Hanson-Wright inequality, we have, for some constant $t_2 > 0$,

$$\mathbb{P}(T_6 \leq \mathbb{E}(T_6) + t) \geq 1 - \exp\left\{-t_2 \min\left(\frac{t^2}{\tilde{\sigma}_i^4 ||A_i \mathbf{X}^T \mathbf{X} A_i^T||_F^2}, \frac{t}{\tilde{\sigma}_i^2 ||A_i \mathbf{X}^T \mathbf{X} A_i^T||_{op}}\right)\right\}.$$

Similar to managing the term T_3 in (a), we have

$$\begin{aligned}
\mathbb{E}[T_6] &= \tilde{\sigma}_i^2 \mathrm{tr}(A_i \mathbf{X}^T \mathbf{X} A_i^T) = \tilde{\sigma}_i^2 \mathrm{tr}(U(U + \rho_i I_q)^{-1} U (U + \rho_i I_q)^{-1}) \\
&= \tilde{\sigma}_i^2 \sum_{i=1}^{q} \frac{u_i^2}{(u_i + \rho_i)^2} = \mathcal{O}(\tilde{\sigma}_i^2 q), \\
||A_i \mathbf{X}^T \mathbf{X} A_i^T||_F^2 &= \mathrm{tr}(A_i \mathbf{X}^T \mathbf{X} A_i^T A_i \mathbf{X}^T \mathbf{X} A_i^T) = \mathrm{tr}(\mathbf{X}^T \mathbf{X} A_i^T A_i \mathbf{X}^T \mathbf{X} A_i^T A_i) \\
&= \mathrm{tr}(\mathbf{U}(\mathbf{U} + \rho_i I_q)^{-1} \mathbf{U} (\mathbf{U} + \rho_i I_q)^{-1} \mathbf{U} (\mathbf{U} + \rho_i I_q)^{-1} \mathbf{U} (\mathbf{U} + \rho_i I_q)^{-1}) \\
&= \sum_{i=1}^{q} \frac{u_i^4}{(u_i + \rho_i)^4} = \mathcal{O}(q), \\
||A_i \mathbf{X}^T \mathbf{X} A_i^T||_{op} &= ||\mathbf{X}(\mathbf{X}^T \mathbf{X} + \rho_i I_q)^{-1} \mathbf{X}^T \mathbf{X} (\mathbf{X}^T \mathbf{X} + \rho_i I_q)^{-1} \mathbf{X}^T||_{op} \\
&= \mathcal{O}(\lambda_{\max}\left(\mathbf{X}\mathbf{X}^T \mathbf{X}\mathbf{X}^T\right)/n^2) = \mathcal{O}(1).
\end{aligned}$$

Letting $t = \max\left(\sqrt{\tilde{\sigma}_i^4 ||A_i \mathbf{X}^T \mathbf{X} A_i^T||_F^2 (f_n + \log 2)/t_2}, \tilde{\sigma}_i^2 ||A_i \mathbf{X}^T \mathbf{X} A_i^T||_{op} (f_n + \log 2)/t_2\right)$, we have that, with probability at least $1 - e^{-f_n}/2$,

$$T_6 = \mathcal{O}(q) + \mathcal{O}(\sqrt{q\, f_n}) + \mathcal{O}(f_n). \tag{2.14}$$

Collecting the bounds in (2.12), (2.13), and (2.14), we conclude that there exists a positive constant C_2 such that, with probability at least $1-e^{-f_n}$,

$$n^{-1}||\mathbf{X}(\hat{\boldsymbol{\pi}}_i-\boldsymbol{\pi}_i)||_2^2 \leq C_2(r_{ni}\vee q\vee f_n)/n.$$

2.5.2 Proof of Theorem 2.4.2

Let's define the composite quantity g_n as

$$g_n = C_2\left(r_{\max}\vee q\vee f_n\right)/n + 2c_1C_2||\boldsymbol{\pi}||_1\sqrt{\left(r_{\max}\vee q\vee f_n\right)/n}.$$

We will first prove some lemmas, and then proceed to prove Theorem 2.4.2.

Lemma 2.5.1 *Assume that, for each node i, the following inequality holds,*

$$\sqrt{\frac{r_{max}\vee q\vee f_n}{n}} + c_1||\boldsymbol{\pi}||_1 \leq \sqrt{c_1^2||\boldsymbol{\pi}||_1^2 + \frac{\phi_0^2}{64C_2|\mathcal{D}_i|}}. \tag{2.15}$$

Under Assumptions 1-3, we have $\phi_{re}(\mathbf{H}_i\mathbf{X}\hat{\boldsymbol{\pi}}_{-i},\mathcal{D}_i)\geq\phi_0/2$ with probability at least $1-e^{-f_n+\log(p)}$.

Proof Note that the inequality (2.15) implies that $g_n\leq\frac{\phi_0^2}{64|\mathcal{D}_i|}$, then, for any index j_1 and j_2, we first investigate the bound of

$$\begin{aligned}
&(\mathbf{H}_i\mathbf{X}\hat{\boldsymbol{\pi}}_{j_1})^T(\mathbf{H}_i\mathbf{X}\hat{\boldsymbol{\pi}}_{j_2}) - (\mathbf{H}_i\mathbf{X}\boldsymbol{\pi}_{j_1})^T(\mathbf{H}_i\mathbf{X}\boldsymbol{\pi}_{j_2})\\
&= \underbrace{(\hat{\boldsymbol{\pi}}_{j_1}-\boldsymbol{\pi}_{j_1})^T\mathbf{X}^T\mathbf{H}_i\mathbf{X}(\hat{\boldsymbol{\pi}}_{j_2}-\boldsymbol{\pi}_{j_2})}_{T_7}\\
&\quad+\underbrace{(\hat{\boldsymbol{\pi}}_{j_1}-\boldsymbol{\pi}_{j_1})^T\mathbf{X}^T\mathbf{H}_i\mathbf{X}\boldsymbol{\pi}_{j_2}}_{T_8}+\underbrace{(\mathbf{X}\boldsymbol{\pi}_{j_1})^T\mathbf{H}_i\mathbf{X}(\hat{\boldsymbol{\pi}}_{j_2}-\boldsymbol{\pi}_{j_2})}_{T_9}.
\end{aligned}$$

Note that $\lambda_{max}(\mathbf{H}_i)=1$. By Theorem 2.4.1, we have, with probability at least $1-e^{-f_n}$,

$$\begin{aligned}
|T_7| &\leq ||\mathbf{H}_i\mathbf{X}(\hat{\boldsymbol{\pi}}_{j_1}-\boldsymbol{\pi}_{j_1})||_2\times||\mathbf{H}_i\mathbf{X}(\hat{\boldsymbol{\pi}}_{j_2}-\boldsymbol{\pi}_{j_2})||_2\\
&\leq \lambda_{max}(\mathbf{H}_i)\times||\mathbf{X}(\hat{\boldsymbol{\pi}}_{j_1}-\boldsymbol{\pi}_{j_1})||_2\times||\mathbf{X}(\hat{\boldsymbol{\pi}}_{j_2}-\boldsymbol{\pi}_{j_2})||_2\\
&= ||\mathbf{X}(\hat{\boldsymbol{\pi}}_{j_1}-\boldsymbol{\pi}_{j_1})||_2\times||\mathbf{X}(\hat{\boldsymbol{\pi}}_{j_2}-\boldsymbol{\pi}_{j_2})||_2\\
&\leq C_2\left(r_{\max}\vee q\vee f_n\right).
\end{aligned} \tag{2.16}$$

Following that $||\mathbf{X}\boldsymbol{\pi}_{j_2}||_2 \le c_1\sqrt{n}||\boldsymbol{\pi}_{j_2}||_2$, we have,

$$\begin{aligned} |T_8| &\le ||\mathbf{X}\boldsymbol{\pi}_{j_2}||_2 \times ||\mathbf{H}_{j_2}\mathbf{X}(\hat{\boldsymbol{\pi}}_{j_1} - \boldsymbol{\pi}_{j_1})||_2 \le c_1\sqrt{n}||\boldsymbol{\pi}_{j_2}||_2 \times ||\mathbf{X}(\hat{\boldsymbol{\pi}}_{j_1} - \boldsymbol{\pi}_{j_1})||_2 \\ &\le c_1 C_2 ||\boldsymbol{\pi}||_1 \sqrt{n\,(r_{\max} \vee q \vee f_n)}. \end{aligned} \tag{2.17}$$

Similarly, we have,

$$|T_9| \le c_1\sqrt{n}||\boldsymbol{\pi}_{j_1}||_2||\mathbf{X}(\hat{\boldsymbol{\pi}}_{j_2} - \boldsymbol{\pi}_{j_2})||_2 \le c_1 C_2||\boldsymbol{\pi}||_1\sqrt{n\,(r_{\max} \vee q \vee f_n)}. \tag{2.18}$$

Putting together the bounds in (2.16), (2.17), and (2.18), we have, with probability at least $1 - e^{-f_n}$,

$$|(\mathbf{H}_i\mathbf{X}\hat{\boldsymbol{\pi}}_{j_1})^T(\mathbf{H}_i\mathbf{X}\hat{\boldsymbol{\pi}}_{j_1}) - (\mathbf{H}_i\mathbf{X}\boldsymbol{\pi}_{j_2})^T(\mathbf{H}_i\mathbf{X}\boldsymbol{\pi}_{j_2})| \le n g_n. \tag{2.19}$$

By definition, for any set $\mathcal{D}_i$ and any vector δ, we have

$$||\delta||_1^2 \le (||\delta_{\mathcal{D}_i^c}||_1 + ||\delta_{\mathcal{D}_i}||_1)^2 \le (3\sqrt{|\mathcal{D}_i|}||\delta_{\mathcal{D}_i}||_2 + \sqrt{|\mathcal{D}_i|}||\delta_{\mathcal{D}_i}||_2)^2 = 16|\mathcal{D}_i|||\delta_{\mathcal{D}_i}||_2^2.$$

We then have, with probability at least $1 - e^{-f_n + log(p)}$,

$$\begin{aligned} &|\delta^T((\mathbf{H}_i\mathbf{X}\hat{\boldsymbol{\pi}}_{-i})^T(\mathbf{H}_i\mathbf{X}\hat{\boldsymbol{\pi}}_{-i}) - (\mathbf{H}_i\mathbf{X}\boldsymbol{\pi}_{-i})^T(\mathbf{H}_i\mathbf{X}\boldsymbol{\pi}_{-i}))\delta|/(n||\delta_{\mathcal{D}_i}||_2^2) \\ &\le ||\delta||_1^2||\delta_{\mathcal{D}_i}||_2^{-2}\max_{i,j}|(\mathbf{H}_i\mathbf{X}\hat{\boldsymbol{\pi}}_i)^T(\mathbf{H}_i\mathbf{X}\hat{\boldsymbol{\pi}}_j) - (\mathbf{H}_i\mathbf{X}\boldsymbol{\pi}_i)^T(\mathbf{H}_i\mathbf{X}\boldsymbol{\pi}_j)|/n \\ &\le 16|\mathcal{D}_i| \times g_n \le 16|\mathcal{D}_i| \times \phi_0^2/\left(64|\mathcal{D}_i|\right) = \phi_0^2/4, \end{aligned}$$

which, along with Assumption 3, implies that $\phi_{\mathrm{re}}(\mathbf{H}_i\mathbf{X}\hat{\boldsymbol{\pi}}_{-i}, \mathcal{D}_i) \ge \phi_0/2$. ■

Lemma 2.5.2 *(Basic Inequality) Let random vector*

$$\boldsymbol{J}_i = 2n^{-1}\hat{\mathbf{Z}}_{-i}^T\mathbf{H}_i\boldsymbol{\epsilon}_i - 2n^{-1}\hat{\mathbf{Z}}_{-i}^T\mathbf{H}_i(\hat{\mathbf{Z}}_{-i} - \mathbf{Y}_{-i})\boldsymbol{\gamma}_i.$$

Under Assumption 1-2, for the event $\mathcal{J}_i(\lambda_i) = \left\{||W_i^{-1}\boldsymbol{J}_i||_\infty \le \lambda_i/2\right\}$ *with* λ_i *and* h_n *being specified in Theorem 2.4.2, there exists a constant* $C_3 > 0$ *such that*

$$\mathbb{P}(\mathcal{J}_i(\lambda_i)) \ge 1 - e^{-C_3 h_n + \log(4qp)} - e^{-f_n + \log(p)},$$

Furthermore, concurring with the random vector $\boldsymbol{J}_i$*, we have the following basic inequality,*

$$n^{-1}||\mathbf{H}_i\hat{\mathbf{Z}}_{-i}(\hat{\gamma}_i - \boldsymbol{\gamma}_i)||_2^2 + \lambda_i\boldsymbol{\omega}_i^T|\hat{\gamma}_i| \le \lambda_i\boldsymbol{\omega}_i^T|\boldsymbol{\gamma}_i| + \boldsymbol{J}_i^T|\hat{\gamma}_i - \boldsymbol{\gamma}_i|. \tag{2.20}$$

Proof With $\mathbf{Y}_{-i} = \mathbf{X}\boldsymbol{\pi}_{-i} + \boldsymbol{\xi}_{-i}$ and $\hat{\mathbf{Z}}_{-i} = \mathbf{X}\hat{\boldsymbol{\pi}}_{-i}$, we have

$$
\begin{aligned}
\boldsymbol{J}_i &= 2n^{-1}\hat{\mathbf{Z}}_{-i}^T\mathbf{H}_i\boldsymbol{\epsilon}_i - 2n^{-1}\hat{\mathbf{Z}}_{-i}^T\mathbf{H}_i(\hat{\mathbf{Z}}_{-i} - \mathbf{Y}_{-i})\boldsymbol{\gamma}_i \\
&= 2n^{-1}\hat{\boldsymbol{\pi}}_{-i}^T\mathbf{X}^T\mathbf{H}_i\boldsymbol{\epsilon}_i - \frac{2}{n}\hat{\boldsymbol{\pi}}_{-i}^T\mathbf{X}^T\mathbf{H}_i(\mathbf{X}\hat{\boldsymbol{\pi}}_{-i} - \mathbf{X}\boldsymbol{\pi}_{-i} - \boldsymbol{\xi}_{-i})\boldsymbol{\gamma}_i \\
&= \underbrace{2n^{-1}(\hat{\boldsymbol{\pi}}_{-i} - \boldsymbol{\pi}_{-i})^T\mathbf{X}^T\mathbf{H}_i\boldsymbol{\epsilon}_i}_{T_{10}} + \underbrace{2n^{-1}\boldsymbol{\pi}_{-i}^T\mathbf{X}^T\mathbf{H}_i\boldsymbol{\epsilon}_i}_{T_{11}} + \underbrace{2n^{-1}(\hat{\boldsymbol{\pi}}_{-i} - \boldsymbol{\pi}_{-i})^T\mathbf{X}^T\mathbf{H}_i\boldsymbol{\xi}_{-i}\boldsymbol{\gamma}_i}_{T_{12}} \\
&\quad + \underbrace{2n^{-1}\boldsymbol{\pi}_{-i}^T\mathbf{X}^T\mathbf{H}_i\boldsymbol{\xi}_{-i}\boldsymbol{\gamma}_i}_{T_{13}} \underbrace{-2n^{-1}(\hat{\boldsymbol{\pi}}_{-i} - \boldsymbol{\pi}_{-i})^T\mathbf{X}^T\mathbf{H}_i\mathbf{X}(\hat{\boldsymbol{\pi}}_{-i} - \boldsymbol{\pi}_{-i})\boldsymbol{\gamma}_i}_{T_{14}} \\
&\quad \underbrace{-2n^{-1}\boldsymbol{\pi}_{-i}^T\mathbf{X}^T\mathbf{H}_i\mathbf{X}(\hat{\boldsymbol{\pi}}_{-i} - \boldsymbol{\pi}_{-i})\boldsymbol{\gamma}_i}_{T_{15}}.
\end{aligned}
$$

Denote $\mathbf{X} = (X_{\cdot 1}, X_{\cdot 2}, \ldots, X_{\cdot q})$, then $X_{\cdot j}^T X_{\cdot j} = n$ due to standardization. Let $\sigma_{pmax}^2 = \max_{1\le i\le p}\sigma_i^2$, then $\mathrm{Var}(X_{\cdot j}^T\mathbf{H}_i\boldsymbol{\epsilon}_i) = X_{\cdot j}^T\mathbf{H}_i X_{\cdot j}\sigma_i^2 \le n\sigma_i^2 \le n\sigma_{pmax}^2$. Further let, for some constant $t_\lambda > 0$,

$$
\lambda_i = t_\lambda\|\boldsymbol{\omega}_i\|_{-\infty}^{-1}||\boldsymbol{\Gamma}||_1||\boldsymbol{\pi}||_1\sqrt{n^{-1}(r_{\max} \vee q \vee f_n)\log p}.
$$

By the Gaussian tail inequality, we have

$$
\begin{aligned}
&\mathbb{P}\left(||W_i^{-1}T_{10}||_\infty \ge \lambda_i/12\right) \\
&\le \mathbb{P}\left(||T_{10}||_\infty \ge \lambda_i\|\boldsymbol{\omega}_i\|_{-\infty}/12\right) \\
&= \mathbb{P}\left(||2n^{-1}(\hat{\boldsymbol{\pi}}_{-i} - \boldsymbol{\pi}_{-i})^T\mathbf{X}^T\mathbf{H}_i\boldsymbol{\epsilon}_i||_\infty \ge \lambda_i\|\boldsymbol{\omega}_i\|_{-\infty}/12\right) \\
&\le \mathbb{P}\left(||(\hat{\boldsymbol{\pi}}_{-i} - \boldsymbol{\pi}_{-i})^T||_\infty \times ||2n^{-1}\mathbf{X}^T\mathbf{H}_i\boldsymbol{\epsilon}_i||_\infty \ge \lambda_i\|\boldsymbol{\omega}_i\|_{-\infty}/12\right) \\
&\le \mathbb{P}\left(||2n^{-1}\mathbf{X}^T\mathbf{H}_i\boldsymbol{\epsilon}_i||_\infty \ge \lambda_i\|\boldsymbol{\omega}_i\|_{-\infty}/(12\delta_\pi)\right) \\
&\le q\exp\left\{-n\lambda_i^2\|\boldsymbol{\omega}_i\|_{-\infty}^2/(1152\sigma_{pmax}^2\delta_\pi^2)\right\} \\
&= q \cdot p^{-\frac{n}{q}t_3||\Gamma||_1^2||\pi||_1^2} \\
&\le q \cdot p \cdot p^{-\frac{n}{q}t_3||\Gamma||_1^2||\pi||_1^2},
\end{aligned}
$$

where $t_3 = t_\lambda^2/\left(1152C_1\sigma_{pmax}^2\right)$ and

$$
\begin{aligned}
\delta_\pi &= \max_i||\hat{\boldsymbol{\pi}}_i - \boldsymbol{\pi}_i||_1 \\
&\le \max_i\sqrt{q}||\hat{\boldsymbol{\pi}}_i - \boldsymbol{\pi}_i||_2 \\
&= \sqrt{C_1 q(r_{\max} \vee q \vee f_n)/n}.
\end{aligned}
$$

Similarly, letting $t_4 = t_\lambda / \left(1152\sigma^2_{pmax}\right)$, we have

$$
\begin{aligned}
&\mathbb{P}\left(||W_i^{-1}T_{11}||_\infty \geq \lambda_i/12\right)\\
&\leq \ \mathbb{P}\left(||T_{11}||_\infty \geq \lambda_i\|\boldsymbol{\omega}_i\|_{-\infty}/12\right)\\
&= \ \mathbb{P}\left(||2n^{-1}\boldsymbol{\pi}^T_{-i}\mathbf{X}^T\mathbf{H}_i\boldsymbol{\epsilon}_i||_\infty \geq \lambda_i\|\boldsymbol{\omega}_i\|_{-\infty}/12\right)\\
&\leq \ \mathbb{P}\left(||\boldsymbol{\pi}^T_{-i}||_\infty||2n^{-1}\mathbf{X}^T\mathbf{H}_i\boldsymbol{\epsilon}_i||_\infty \geq \lambda_i\|\boldsymbol{\omega}_i\|_{-\infty}/12\right)\\
&\leq \ \mathbb{P}\left(||2n^{-1}\mathbf{X}^T\mathbf{H}_i\boldsymbol{\epsilon}_i||_\infty \geq \lambda_i\|\boldsymbol{\omega}_i\|_{-\infty}||\boldsymbol{\pi}^T_{-i}||^{-1}_\infty/12\right)\\
&\leq \ q\exp\left\{-n\lambda_i^2\|\boldsymbol{\omega}_i\|^2_{-\infty}||\boldsymbol{\pi}^T_{-i}||^{-2}_\infty/(1152\sigma^2_{pmax})\right\}\\
&= \ q\cdot p^{-t_4||\Gamma||_1^2(r_{\max}\vee q\vee f_n)}\\
&\leq \ q\cdot p\cdot p^{-t_4||\Gamma||_1^2(r_{\max}\vee q\vee f_n)}.
\end{aligned}
$$

Let $\tilde{\sigma}^2_{pmax} = \max_i(\tilde{\sigma}^2_i)$, then, $\mathrm{var}(X^T_{\cdot j_1}\mathbf{H}_i\boldsymbol{\xi}_{j_2}) \leq n\tilde{\sigma}^2_{pmax}$. Furthermore, denote $t_5 = t_\lambda/\left(1152C_1\tilde{\sigma}^2_{pmax}\right)$. For the term T_{12}, we have

$$
\begin{aligned}
&\mathbb{P}\left(||W_i^{-1}T_{12}||_\infty \geq \lambda_i/12\right)\\
&\leq \ \mathbb{P}\left(||T_{12}||_\infty \geq \lambda_i\|\boldsymbol{\omega}_i\|_{-\infty}/12\right)\\
&\leq \ \mathbb{P}\left(||(\hat{\boldsymbol{\pi}}_{-i}-\boldsymbol{\pi}_{-i})^T||_\infty||2n^{-1}\mathbf{X}^T\mathbf{H}_i\boldsymbol{\xi}_{-i}\boldsymbol{\gamma}_i||_1 \geq \lambda_i\|\boldsymbol{\omega}_i\|_{-\infty}/12\right)\\
&\leq \ \mathbb{P}\left(\delta_\pi\max_{j_1,j_2}\left|2n^{-1}X^T_{\cdot j_1}\mathbf{H}_i\boldsymbol{\xi}_{j_2}\right|||\boldsymbol{\gamma}_i||_1 \geq \lambda_i\|\boldsymbol{\omega}_i\|_{-\infty}/12\right)\\
&\leq \ \mathbb{P}\left(\max_{j_1,j_2}\left|2n^{-1}X^T_{\cdot j_1}\mathbf{H}_i\boldsymbol{\xi}_{j_2}\right| \geq \lambda_i\|\boldsymbol{\omega}_i\|_{-\infty}||\boldsymbol{\gamma}_i||_1^{-1}/(12\delta_\pi)\right)\\
&\leq \ qp\exp\left\{-n\lambda_i^2\|\boldsymbol{\omega}_i\|^2_{-\infty}\tilde{\sigma}^{-2}_{pmax}\delta_\pi^{-2}||\boldsymbol{\gamma}_i||_1^{-2}/1152\right\}\\
&= \ qp^{1-t_5||\pi||_1^2 n/q}.
\end{aligned}
$$

Letting $t_6 = t_\lambda/\left(1152\tilde{\sigma}^2_{pmax}\right)$, we similarly have

$$
\begin{aligned}
&\mathbb{P}\left(||W_i^{-1}T_{13}||_\infty \geq \lambda_i/12\right)\\
&\leq \ qp\exp\left\{-\lambda_i^2\tilde{\sigma}^{-2}_{pmax}||\boldsymbol{\pi}^T_{-i}||^{-2}_\infty||\boldsymbol{\gamma}_i||_1^{-2}/1152\right\}\\
&= \ qp^{1-t_6(r_{max}\vee q\vee f_n)}.
\end{aligned}
$$

When t_λ is sufficiently large, say $t_\lambda \geq 12C_2||\boldsymbol{\pi}||_1^{-1}\sqrt{(r_{max} \vee q \vee f_n)/(n \log p)}$, we have

$$\begin{aligned}
||W_i^{-1}T_{14}||_\infty &\leq n^{-1}\|\boldsymbol{\omega}_i\|_{-\infty}^{-1}||\boldsymbol{\gamma}_i||_1 \max_{j_1,j_2}|(\hat{\boldsymbol{\pi}}_{j_1} - \boldsymbol{\pi}_{j_1})^T \mathbf{X}^T \mathbf{H}_i \mathbf{X}(\hat{\boldsymbol{\pi}}_{j_2} - \boldsymbol{\pi}_{j_2})| \\
&\leq n^{-1}\|\boldsymbol{\omega}_i\|_{-\infty}^{-1}||\boldsymbol{\gamma}_i||_1 \max_{j_1,j_2}||\mathbf{H}_i \mathbf{X}(\hat{\boldsymbol{\pi}}_{j_1} - \boldsymbol{\pi}_{j_1})||_2||\mathbf{H}_i \mathbf{X}(\hat{\boldsymbol{\pi}}_{j_2} - \boldsymbol{\pi}_{j_2})||_2 \\
&\leq n^{-1}\|\boldsymbol{\omega}_i\|_{-\infty}^{-1}||\boldsymbol{\gamma}_i||_1 \max_{j_1,j_2}\lambda_{\max}(\mathbf{H}_i)||\mathbf{X}(\hat{\boldsymbol{\pi}}_{j_1} - \boldsymbol{\pi}_{j_1})||_2||\mathbf{X}(\hat{\boldsymbol{\pi}}_{j_2} - \boldsymbol{\pi}_{j_2})||_2 \\
&\leq n^{-1}\|\boldsymbol{\omega}_i\|_{-\infty}^{-1}||\boldsymbol{\gamma}_i||_1 \max_{i,j}||\mathbf{X}(\hat{\boldsymbol{\pi}}_i - \boldsymbol{\pi}_i)||_2||\mathbf{X}(\hat{\boldsymbol{\pi}}_j - \boldsymbol{\pi}_j)||_2 \\
&\leq C_2\|\boldsymbol{\omega}_i\|_{-\infty}^{-1}||\boldsymbol{\gamma}_i||_1 n^{-1}(r_{\max} \vee q \vee f_n) \\
&\leq \left\{\lambda_i/12\right\} \times \left\{12C_2 t_\lambda^{-1}||\boldsymbol{\pi}||_1^{-1}\sqrt{n^{-1}(\log p)^{-1}(r_{max} \vee q \vee f_n)}\right\} \leq \lambda_i/12.
\end{aligned}$$

Similarly, when $t_\lambda \geq 12\sqrt{C_2/\log p}$,

$$\begin{aligned}
||W_i^{-1}T_{15}||_\infty &\leq 2n^{-1}||\boldsymbol{\gamma}_i||_1||\boldsymbol{\pi}_{-i}^T||_\infty\|\boldsymbol{\omega}_i\|_{-\infty}^{-1}\max_{j_1,j_2}|X_{\cdot j_1}^T \mathbf{H}_i \mathbf{X}(\hat{\boldsymbol{\pi}}_{j_2} - \boldsymbol{\pi}_{j_2})| \\
&\leq 2n^{-1/2}||\boldsymbol{\gamma}_i||_1||\boldsymbol{\pi}_{-i}^T||_\infty\|\boldsymbol{\omega}_i\|_{-\infty}^{-1}\max_{j_2}||\mathbf{H}_i \mathbf{X}(\hat{\boldsymbol{\pi}}_{j_2} - \boldsymbol{\pi}_{j_2})||_2 \\
&\leq 2n^{-1/2}||\boldsymbol{\gamma}_i||_1||\boldsymbol{\pi}_{-i}^T||_\infty\|\boldsymbol{\omega}_i\|_{-\infty}^{-1}\max_{j_2}||\mathbf{X}(\hat{\boldsymbol{\pi}}_{j_2} - \boldsymbol{\pi}_{j_2})||_2 \\
&\leq \left\{\lambda_i/12\right\} \times \left\{12t_\lambda^{-1}\sqrt{C_2/\log p}\right\} \leq \lambda_i/12.
\end{aligned}$$

Putting together all the above results with union bounds, we have, for some constant $C_3 > 0$,

$$\mathbb{P}(\mathcal{J}_i(\lambda_i)) \geq 1 - e^{-C_3 h_n + \log(4qp)} - e^{-f_n + \log p}.$$

Concurring with the random vector $\boldsymbol{J}_i$, we have the following inequality based on the optimality of $\hat{\boldsymbol{\gamma}}_i$,

$$n^{-1}||\mathbf{H}_i \mathbf{Y}_i - \mathbf{H}_i \hat{\mathbf{Z}}_{-i}\hat{\boldsymbol{\gamma}}_i||_2 + \lambda_i \boldsymbol{\omega}_i^T|\hat{\boldsymbol{\gamma}}_i| \leq n^{-1}||\mathbf{H}_i \mathbf{Y}_i - \mathbf{H}_i \hat{\mathbf{Z}}_{-i}\boldsymbol{\gamma}_i||_2 + \lambda_i \boldsymbol{\omega}_i^T|\boldsymbol{\gamma}_i|. \quad (2.21)$$

With $\mathbf{H}_i \mathbf{Y}_i = \mathbf{H}_i \mathbf{Y}_{-i}\boldsymbol{\gamma}_i + \mathbf{H}_i \boldsymbol{\epsilon}_i$, we also have

$$\begin{aligned}
&||\mathbf{H}_i \mathbf{Y}_i - \mathbf{H}_i \hat{\mathbf{Z}}_{-i}\hat{\boldsymbol{\gamma}}_i||_2^2 \\
&= ||\mathbf{H}_i \mathbf{Y}_{-i}\boldsymbol{\gamma}_i + \mathbf{H}_i \boldsymbol{\epsilon}_i - \mathbf{H}_i \hat{\mathbf{Z}}_{-i}\hat{\boldsymbol{\gamma}}_i||_2^2 \\
&= ||\mathbf{H}_i \boldsymbol{\epsilon}_i||_2^2 - 2\boldsymbol{\epsilon}_i^T \mathbf{H}_i(\hat{\mathbf{Z}}_{-i}\hat{\boldsymbol{\gamma}}_i - \mathbf{Y}_{-i}\boldsymbol{\gamma}_i) + ||\mathbf{H}_i \hat{\mathbf{Z}}_{-i}\hat{\boldsymbol{\gamma}}_i - \mathbf{H}_i \mathbf{Y}_{-i}\boldsymbol{\gamma}_i||_2 \\
&= ||\mathbf{H}_i \boldsymbol{\epsilon}_i||_2^2 - 2\boldsymbol{\epsilon}_i^T \mathbf{H}_i(\hat{\mathbf{Z}}_{-i}\hat{\boldsymbol{\gamma}}_i - \mathbf{Y}_{-i}\boldsymbol{\gamma}_i) + ||\mathbf{H}_i \hat{\mathbf{Z}}_{-i}(\hat{\boldsymbol{\gamma}}_i - \boldsymbol{\gamma}_i)||_2^2 \\
&\quad + ||\mathbf{H}_i(\hat{\mathbf{Z}}_{-i} - \mathbf{Y}_{-i})\boldsymbol{\gamma}_i||_2^2 + 2\boldsymbol{\gamma}_i^T(\hat{\mathbf{Z}}_{-i} - \mathbf{Y}_{-i})^T \mathbf{H}_i \hat{\mathbf{Z}}_{-i}(\hat{\boldsymbol{\gamma}}_i - \boldsymbol{\gamma}_i), \qquad (2.22)
\end{aligned}$$

$$
\begin{aligned}
&||\mathbf{H}_i\mathbf{Y}_i - \mathbf{H}_i\hat{\mathbf{Z}}_{-i}\boldsymbol{\gamma}_i||_2^2 \\
&\quad = \; ||\mathbf{H}_i\mathbf{Y}_{-i}\boldsymbol{\gamma}_i + \mathbf{H}_i\boldsymbol{\epsilon}_i - \mathbf{H}_i\hat{\mathbf{Z}}_{-i}\boldsymbol{\gamma}_i||_2^2 \\
&\quad = \; ||\mathbf{H}_i\boldsymbol{\epsilon}_i||_2^2 + ||\mathbf{H}_i(\hat{\mathbf{Z}}_{-i} - \mathbf{Y}_{-i})\boldsymbol{\gamma}_i||_2^2 - 2\boldsymbol{\epsilon}_i^T\mathbf{H}_i(\hat{\mathbf{Z}}_{-i} - \mathbf{Y}_{-i})\boldsymbol{\gamma}_i. \qquad (2.23)
\end{aligned}
$$

Combining the equations (2.21), (2.22), and (2.23), we obtain that

$$
\begin{aligned}
&n^{-1}||\mathbf{H}_i\hat{\mathbf{Z}}_{-i}(\hat{\boldsymbol{\gamma}}_i - \boldsymbol{\gamma}_i)||_2^2 + \lambda_i\boldsymbol{\omega}_i^T|\hat{\boldsymbol{\gamma}}_i| \\
&\leq \; \lambda_i\boldsymbol{\omega}_i^T|\boldsymbol{\gamma}_i| + \left(\frac{2}{n}\hat{\mathbf{Z}}_{-i}^T\mathbf{H}_i\boldsymbol{\epsilon}_i - \frac{2}{n}\hat{\mathbf{Z}}_{-i}^T\mathbf{H}_i(\hat{\mathbf{Z}}_{-i} - \mathbf{Y}_{-i})\boldsymbol{\gamma}_i\right)^T(\hat{\boldsymbol{\gamma}}_i - \boldsymbol{\gamma}_i) \\
&= \; \lambda_i\boldsymbol{\omega}_i^T|\boldsymbol{\gamma}_i| + \boldsymbol{J}_i^T(\hat{\boldsymbol{\gamma}}_i - \boldsymbol{\gamma}_i),
\end{aligned}
$$

which concludes the proof. ■

By the basic inequality we just proved above and condition on the event $\mathcal{J}_i(\lambda_i)$, we have that

$$
\begin{aligned}
&n^{-1}||\mathbf{H}_i\hat{\mathbf{Z}}_{-i}(\hat{\boldsymbol{\gamma}}_i - \boldsymbol{\gamma}_i)||_2^2 \leq \lambda_i\boldsymbol{\omega}_i^T|\boldsymbol{\gamma}_i| - \lambda_i\boldsymbol{\omega}_i^T|\hat{\boldsymbol{\gamma}}_i| + \boldsymbol{J}_i^T(\hat{\boldsymbol{\gamma}}_i - \boldsymbol{\gamma}_i) \\
&\leq \; \lambda_i\boldsymbol{\omega}_{\mathcal{D}_i}^T|\boldsymbol{\gamma}_{\mathcal{D}_i}| - \lambda_i\boldsymbol{\omega}_{\mathcal{D}_i}^T|\hat{\boldsymbol{\gamma}}_{\mathcal{D}_i}| - \lambda_i\boldsymbol{\omega}_{\mathcal{D}_i^c}^T|\hat{\boldsymbol{\gamma}}_{\mathcal{D}_i^c}| \\
&\quad + \boldsymbol{J}_{\mathcal{D}_i^c}^T(\hat{\boldsymbol{\gamma}}_{\mathcal{D}_i^c}) + \boldsymbol{J}_{\mathcal{D}_i}^T(\hat{\boldsymbol{\gamma}}_{\mathcal{D}_i} - \boldsymbol{\gamma}_{\mathcal{D}_i}) \\
&\leq \; \lambda_i\boldsymbol{\omega}_{\mathcal{D}_i}^T|\hat{\boldsymbol{\gamma}}_{\mathcal{D}_i} - \boldsymbol{\gamma}_{\mathcal{D}_i}| - \lambda_i\boldsymbol{\omega}_{\mathcal{D}_i^c}^T|\hat{\boldsymbol{\gamma}}_{\mathcal{D}_i^c}| \\
&\quad + 2^{-1}\lambda_i\boldsymbol{\omega}_{\mathcal{D}_i^c}^T|\hat{\boldsymbol{\gamma}}_{\mathcal{D}_i^c}| + 2^{-1}\lambda_i\boldsymbol{\omega}_{\mathcal{D}_i}^T|\hat{\boldsymbol{\gamma}}_{\mathcal{D}_i} - \boldsymbol{\gamma}_{\mathcal{D}_i}| \\
&\leq \; \frac{3}{2}\lambda_i\boldsymbol{\omega}_{\mathcal{D}_i}^T|\hat{\boldsymbol{\gamma}}_{\mathcal{D}_i} - \boldsymbol{\gamma}_{\mathcal{D}_i}| - \frac{1}{2}\lambda_i\boldsymbol{\omega}_{\mathcal{D}_i^c}^T|\hat{\boldsymbol{\gamma}}_{\mathcal{D}_i^c}| \\
&\leq \; \frac{3}{2}\lambda_i\|\boldsymbol{\omega}_{\mathcal{D}_i}\|_\infty||\hat{\boldsymbol{\gamma}}_{\mathcal{D}_i} - \boldsymbol{\gamma}_{\mathcal{D}_i}||_1 - \frac{1}{2}\lambda_i\|\boldsymbol{\omega}_{\mathcal{D}_i^c}\|_{-\infty}||\hat{\boldsymbol{\gamma}}_{\mathcal{D}_i^c}||_1,
\end{aligned}
$$

which implies that

$$
\lambda_i\|\boldsymbol{\omega}_{\mathcal{D}_i^c}\|_{-\infty}||\hat{\boldsymbol{\gamma}}_{\mathcal{D}_i^c}||_1 \leq 3\lambda_i\|\boldsymbol{\omega}_{\mathcal{D}_i}\|_\infty||\hat{\boldsymbol{\gamma}}_{\mathcal{D}_i} - \boldsymbol{\gamma}_{\mathcal{D}_i}||_1. \qquad (2.24)
$$

Note that $\|\boldsymbol{\omega}_{\mathcal{D}_i}\|_\infty\|\boldsymbol{\omega}_{\mathcal{D}_i^c}\|_{-\infty}^{-1} \leq 1$ in Assumption 3, we have that

$$
\begin{aligned}
&||\hat{\boldsymbol{\gamma}}_{\mathcal{D}_i^c} - \boldsymbol{\gamma}_{\mathcal{D}_i^c}||_1 \\
&\quad \leq \; 3\|\boldsymbol{\omega}_{\mathcal{D}_i}\|_\infty\|\boldsymbol{\omega}_{\mathcal{D}_i^c}\|_{-\infty}^{-1}||\hat{\boldsymbol{\gamma}}_{\mathcal{D}_i} - \boldsymbol{\gamma}_{\mathcal{D}_i}||_1 \leq 3||\hat{\boldsymbol{\gamma}}_{\mathcal{D}_i} - \boldsymbol{\gamma}_{\mathcal{D}_i}||_1. \qquad (2.25)
\end{aligned}
$$

On the other hand, following Lemma 2.5.1, we have, with $C_4 = 3t_\lambda$,

$$
\begin{aligned}
n^{-1}||\mathbf{H}_i\hat{\mathbf{Z}}_{-i}(\hat{\boldsymbol{\gamma}}_i - \boldsymbol{\gamma}_i)||_2^2 &\leq \frac{3}{2}\lambda_i\|\boldsymbol{\omega}_{\mathcal{D}_i}\|_\infty\sqrt{|\mathcal{D}_i|}||\hat{\boldsymbol{\gamma}}_{\mathcal{D}_i} - \boldsymbol{\gamma}_{\mathcal{D}_i}||_2 \\
\leq\ & \frac{3}{2}\lambda_i\|\boldsymbol{\omega}_{\mathcal{D}_i}\|_\infty\sqrt{|\mathcal{D}_i|} \times 2n^{-1/2}\phi_0^{-1}||\mathbf{H}_i\hat{\mathbf{Z}}_{-i}(\hat{\boldsymbol{\gamma}}_i - \boldsymbol{\gamma}_i)||_2 \\
\leq\ & 9\phi_0^{-2}\|\boldsymbol{\omega}_{\mathcal{D}_i}\|_\infty^2|\mathcal{D}_i|\lambda_i^2 \\
=\ & C_4^2\phi_0^{-2}\|\boldsymbol{\omega}_i\|_{-\infty}^{-2}\|\boldsymbol{\omega}_{\mathcal{D}_i}\|_\infty^2||\boldsymbol{\pi}||_1^2||\boldsymbol{\Gamma}||_1^2|\mathcal{D}_i|(r_{\max} \vee q \vee f_n)\log p/n.
\end{aligned}
$$

Employing the inequality (2.24) and $\|\boldsymbol{\omega}_{\mathcal{D}_i}\|_\infty\|\boldsymbol{\omega}_{\mathcal{D}_i^c}\|_{-\infty}^{-1} \leq 1$ in Assumption 3, we have

$$
\begin{aligned}
||\hat{\boldsymbol{\gamma}}_i - \boldsymbol{\gamma}_i||_1 &\leq \left(3\|\boldsymbol{\omega}_{\mathcal{D}_i}\|_\infty\|\boldsymbol{\omega}_{\mathcal{D}_i^c}\|_{-\infty}^{-1} + 1\right)||\hat{\boldsymbol{\gamma}}_{\mathcal{D}_i} - \boldsymbol{\gamma}_{\mathcal{D}_i}||_1 \\
\leq\ & \left(3\|\boldsymbol{\omega}_{\mathcal{D}_i}\|_\infty\|\boldsymbol{\omega}_{\mathcal{D}_i^c}\|_{-\infty}^{-1} + 1\right)\sqrt{|\mathcal{D}_i|}||\hat{\boldsymbol{\gamma}}_{\mathcal{D}_i} - \boldsymbol{\gamma}_{\mathcal{D}_i}||_2 \\
\leq\ & \left(6\|\boldsymbol{\omega}_{\mathcal{D}_i}\|_\infty\|\boldsymbol{\omega}_{\mathcal{D}_i^c}\|_{-\infty}^{-1} + 2\right)\sqrt{|\mathcal{D}_i|} \times n^{-1/2}||\mathbf{H}_i\hat{\mathbf{Z}}_{-i}(\hat{\boldsymbol{\gamma}}_i - \boldsymbol{\gamma}_i)||_2\phi_0^{-1} \\
\leq\ & 8C_4 \times \|\boldsymbol{\omega}_{\mathcal{D}_i}\|_\infty||\boldsymbol{\pi}||_1||\boldsymbol{\Gamma}||_1\phi_0^{-2}\|\boldsymbol{\omega}_i\|_{-\infty}^{-1} \\
& \times|\mathcal{D}_i|\sqrt{(r_{\max} \vee q \vee f_n)\log p/n}.
\end{aligned}
$$

Since we condition on event $\mathcal{J}_i(\lambda_i)$, the above prediction and estimation bounds hold with probability at least $1 - e^{-C_3h_n + \log(4qp)} - e^{-f_n + \log p}$.

2.5.3 Proof of Theorem 2.4.3

Let $\hat{\mathcal{V}}_i = (\hat{v}_{ij})_{(p-1)\times(p-1)} \triangleq \frac{1}{n}\hat{\boldsymbol{\pi}}_{-i}^T\mathbf{X}^T\mathbf{H}_i\mathbf{X}\hat{\boldsymbol{\pi}}_{-i}$. Further denote $\hat{\mathcal{V}}_{i,21} = (\hat{v}_{ij})_{i\in\mathcal{D}_i^c, j\in\mathcal{D}_i}$, $\hat{\mathcal{V}}_{i,11} = (\hat{v}_{ij})_{i\in\mathcal{D}_i, j\in\mathcal{D}_i}$. Then, the proof of Theorem 2.4.3 will be presented after the following lemma.

Lemma 2.5.3 *Assume that, for each node i, the following inequality holds.*

$$
\begin{aligned}
&\sqrt{(r_{max} \vee q \vee f_n)/n} + c_1||\boldsymbol{\pi}||_1 \\
&\leq \sqrt{c_1^2||\boldsymbol{\pi}||_1^2 + \min(\phi_0^2/64, \zeta(4-\zeta)^{-1}\|\boldsymbol{\omega}_i\|_{-\infty}/\theta_i)/(C_2|\mathcal{D}_i|)}. \qquad (2.26)
\end{aligned}
$$

Under Assumptions 1-4, we have that, with probability at least $1 - e^{-f_n + \log(p)}$,

$$
||W_{\mathcal{D}_i^c}^{-1}\left(\hat{\mathcal{V}}_{i,21}\hat{\mathcal{V}}_{i,11}^{-1}\right)W_{\mathcal{D}_i}||_\infty \leq 1 - \zeta/2.
$$

Proof Following Theorem 2.4.1, we have, with probability at least $1 - e^{-f_n + \log(p)}$,

$$n^{-1}\max_{j_1,j_2}\left|(\mathbf{H}_i\mathbf{X}\hat{\boldsymbol{\pi}}_{j_1})^T(\mathbf{H}_i\mathbf{X}\hat{\boldsymbol{\pi}}_{j_1}) - (\mathbf{H}_i\mathbf{X}\boldsymbol{\pi}_{j_2})^T(\mathbf{H}_i\mathbf{X}\boldsymbol{\pi}_{j_2})\right| \le g_n.$$

The inequality (2.26) implies that $\theta_i\|\boldsymbol{\omega}_{\mathcal{D}_i}\|_{-\infty}^{-1}|\mathcal{D}_i|g_n \le \zeta/(4-\zeta)$, we have

$$||W_{\mathcal{D}_i}^{-1}(\hat{\mathcal{V}}_{i,11} - \mathcal{V}_{i,11})||_\infty \le \|\boldsymbol{\omega}_{\mathcal{D}_i}\|_{-\infty}^{-1}|\mathcal{D}_i|g_n \le \zeta/\{(4-\zeta)\theta_i\}.$$

Similarly we have that

$$||W_{\mathcal{D}_i^c}^{-1}(\hat{\mathcal{V}}_{i,21} - \mathcal{V}_{i,21})||_\infty \le \zeta/\{(4-\zeta)\theta_i\}.$$

Applying the matrix inversion error bound in Horn and Johnson (2012), we obtain

$$\begin{aligned} &||\hat{\mathcal{V}}_{i,11}^{-1}W_{\mathcal{D}_i}||_\infty \le ||\mathcal{V}_{i,11}^{-1}W_{\mathcal{D}_i}||_\infty + ||\hat{\mathcal{V}}_{i,11}^{-1}W_{\mathcal{D}_i} - \mathcal{V}_{i,11}^{-1}W_{\mathcal{D}_i}||_\infty \\ &\le \ \theta_i + \theta_i||W_{\mathcal{D}_i}^{-1}(\hat{\mathcal{V}}_{i,11} - \mathcal{V}_{i,11})||_\infty \left(1 - \theta_i||W_{\mathcal{D}_i}^{-1}(\hat{\mathcal{V}}_{i,11} - \mathcal{V}_{i,11})||_\infty\right)^{-1}\theta_i \\ &\le \ \theta_i(4-\zeta)/(4-2\zeta). \end{aligned}$$

Therefore,

$$\begin{aligned} &||W_{\mathcal{D}_i^c}^{-1}\left(\hat{\mathcal{V}}_{i,21}\hat{\mathcal{V}}_{i,11}^{-1} - \mathcal{V}_{i,21}\mathcal{V}_{i,11}^{-1}\right)W_{\mathcal{D}_i}||_\infty \\ &\le \ ||W_{\mathcal{D}_i^c}^{-1}\left(\hat{\mathcal{V}}_{i,21} - \mathcal{V}_{i,21}\right)(\hat{\mathcal{V}}_{i,11}^{-1})W_{\mathcal{D}_i}||_\infty \\ &\quad + ||W_{\mathcal{D}_i^c}^{-1}\mathcal{V}_{i,21}\mathcal{V}_{i,11}^{-1}W_{\mathcal{D}_i}W_{\mathcal{D}_i}^{-1}\left(\hat{\mathcal{V}}_{i,11} - \mathcal{V}_{i,11}\right)(\hat{\mathcal{V}}_{i,11}^{-1})W_{\mathcal{D}_i}||_\infty \\ &\le \ ||W_{\mathcal{D}_i^c}^{-1}\left(\hat{\mathcal{V}}_{i,21} - \mathcal{V}_{i,21}\right)||_\infty||(\hat{\mathcal{V}}_{i,11}^{-1})W_{\mathcal{D}_i}||_\infty \\ &\quad + ||W_{\mathcal{D}_i^c}^{-1}\mathcal{V}_{i,21}\mathcal{V}_{i,11}^{-1}W_{\mathcal{D}_i}||_\infty||W_{\mathcal{D}_i}^{-1}\left(\hat{\mathcal{V}}_{i,11} - \mathcal{V}_{i,11}\right)||_\infty||(\hat{\mathcal{V}}_{i,11}^{-1})W_{\mathcal{D}_i}||_\infty \\ &\le \ \zeta/2, \end{aligned}$$

which implies that $||W_{\mathcal{D}_i^c}^{-1}(\hat{\mathcal{V}}_{i,21}\hat{\mathcal{V}}_{i,11}^{-1})W_{\mathcal{D}_i}||_\infty \le 1 - \zeta/2$. ■

By the optimality of $\hat{\boldsymbol{\gamma}}_i$, it must satisfy the KKT condition as follows,

$$-2n^{-1}(\mathbf{H}_i\hat{\mathbf{Z}}_{-i})^T(\mathbf{H}_i\mathbf{Y}_i - \mathbf{H}_i\hat{\mathbf{Z}}_{-i}\hat{\boldsymbol{\gamma}}_i) + \lambda_i W_i\alpha_i = 0, \tag{2.27}$$

where $||\alpha_i||_\infty \le 1$ and $\alpha_{kj} I[\hat{\boldsymbol{\gamma}}_{kj} \neq 0] = sign(\hat{\boldsymbol{\gamma}}_{kj})$. Plug in the equation $\mathbf{H}_i\mathbf{Y}_i = \mathbf{H}_i\mathbf{Y}_{-i}\boldsymbol{\gamma}_i + \mathbf{H}_i\epsilon_i$, we can have that

$$\begin{aligned}\mathbf{H}_i\mathbf{Y}_i - \mathbf{H}_i\hat{\mathbf{Z}}_{-i}\hat{\boldsymbol{\gamma}}_i &= \mathbf{H}\mathbf{Y}_{-i}\boldsymbol{\gamma}_i + \mathbf{H}_i\epsilon_i - \mathbf{H}_i\hat{\mathbf{Z}}_{-i}\hat{\boldsymbol{\gamma}}_i \\ &= \mathbf{H}_i\epsilon_i + \mathbf{H}_i\mathbf{Y}_{-i}\boldsymbol{\gamma}_i - \mathbf{H}_i\hat{\mathbf{Z}}_{-i}\boldsymbol{\gamma}_i + \mathbf{H}_i\hat{\mathbf{Z}}_{-i}\boldsymbol{\gamma}_i - \mathbf{H}_i\hat{\mathbf{Z}}_{-i}\hat{\boldsymbol{\gamma}}_i \\ &= \mathbf{H}_i\epsilon_i - \mathbf{H}_i(\hat{\mathbf{Z}}_{-i} - \mathbf{Y}_{-i})\boldsymbol{\gamma}_i - \mathbf{H}_i\hat{\mathbf{Z}}_{-i}(\hat{\boldsymbol{\gamma}}_i - \boldsymbol{\gamma}_i). \qquad (2.28)\end{aligned}$$

Combining (2.27) and (2.28), we can get that

$$2\hat{\mathcal{V}}_i(\hat{\boldsymbol{\gamma}}_i - \boldsymbol{\gamma}_i) - \boldsymbol{J}_i = -\lambda_i W_i \alpha_i, \qquad (2.29)$$

where $\boldsymbol{J}_i = 2n^{-1}\hat{\mathbf{Z}}_{-i}^T\mathbf{H}_i\boldsymbol{\epsilon}_i - 2n^{-1}\hat{\mathbf{Z}}_{-i}^T\mathbf{H}_i(\hat{\mathbf{Z}}_{-i} - \mathbf{Y}_{-i})\boldsymbol{\gamma}_i$. For an estimator satisfying $\hat{\boldsymbol{\gamma}}_{\mathcal{D}_i^c} = \boldsymbol{\gamma}_{\mathcal{D}_i^c} = 0$, the above equation implies that

$$\begin{cases} 2\hat{\mathcal{V}}_{i,11}(\hat{\boldsymbol{\gamma}}_{\mathcal{D}_i} - \boldsymbol{\gamma}_{\mathcal{D}_i}) - \boldsymbol{J}_{\mathcal{D}_i} = -\lambda_i W_{\mathcal{D}_i}\alpha_{\mathcal{D}_i}, \\ 2\hat{\mathcal{V}}_{i,21}(\hat{\boldsymbol{\gamma}}_{\mathcal{D}_i} - \boldsymbol{\gamma}_{\mathcal{D}_i}) - \boldsymbol{J}_{\mathcal{D}_i^c} = -\lambda_i W_{\mathcal{D}_i^c}\alpha_{\mathcal{D}_i^c}. \end{cases} \qquad (2.30)$$

Manipulating the above equations, we have that

$$\begin{aligned}\hat{\boldsymbol{\gamma}}_{\mathcal{D}_i} - \boldsymbol{\gamma}_{\mathcal{D}_i} &= 2^{-1}\hat{\mathcal{V}}_{i,11}^{-1}(\boldsymbol{J}_{\mathcal{D}_i} - \lambda_i W_{\mathcal{D}_i}^T\alpha_{\mathcal{D}_i}) \\ &= 2^{-1}\hat{\mathcal{V}}_{i,11}^{-1}W_{\mathcal{D}_i}(W_{\mathcal{D}_i}^{-1}\boldsymbol{J}_{\mathcal{D}_i} - \lambda_i\alpha_{\mathcal{D}_i}). \qquad (2.31)\end{aligned}$$

Following the similar strategy in proving Lemma 2.5.2, we can prove that there exists a constant $C_5 > 0$ such that $||W_i^{-1}\boldsymbol{J}_i||_\infty \le \lambda_i\zeta/\{(4-\zeta)\}$ with probability at least $1 - e^{-C_5 h_n + \log(4qp)} - e^{-f_n + \log(p)}$. Therefore, with $||\alpha_{\mathcal{D}_i}||_\infty \le 1$, we have that

$$\begin{aligned}&||\hat{\boldsymbol{\gamma}}_{\mathcal{D}_i} - \boldsymbol{\gamma}_{\mathcal{D}_i}||_\infty \le 2^{-1}||\hat{\mathcal{V}}_{i,11}^{-1}W_{\mathcal{D}_i}||_\infty(||W_{\mathcal{D}_i}^{-1}\boldsymbol{J}_{\mathcal{D}_i}||_\infty + \lambda_i) \\ &\le \{\theta_i(4-\zeta)/(2-\zeta)\} \times \{4/(4-\zeta)\} \times \lambda_i = \lambda_i\theta_i/(2-\zeta) \le \min_{j\in\mathcal{D}_i}|\boldsymbol{\gamma}_{kj}|.\end{aligned}$$

The above inequality implies that $sign(\hat{\boldsymbol{\gamma}}_{\mathcal{D}_i}) = sign(\boldsymbol{\gamma}_{\mathcal{D}_i})$.

Combining (2.30) and (2.31), we can also verify that

$$\begin{aligned}&||W_{\mathcal{D}_i^c}^{-1}\hat{\mathcal{V}}_{i,21}(\hat{\mathcal{V}}_{i,11})^{-1}(\boldsymbol{J}_{\mathcal{D}_i} - \lambda_i W_{\mathcal{D}_i}\alpha_{\mathcal{D}_i}) - W_{\mathcal{D}_i^c}^{-1}J_{\mathcal{D}_i^c}||_\infty \\ &\le ||W_{\mathcal{D}_i^c}^{-1}\hat{\mathcal{V}}_{i,21}(\hat{\mathcal{V}}_{i,11})^{-1}W_{\mathcal{D}_i}||_\infty(||W_{\mathcal{D}_i}^{-1}\boldsymbol{J}_i||_\infty + \lambda_i) + ||W_{\mathcal{D}_i^c}^{-1}J_{\mathcal{D}_i^c}||_\infty \\ &\le (1-\zeta/2)(4/(4-\zeta))\lambda_i + \zeta/(4-\zeta)\lambda_i = \lambda_i.\end{aligned}$$

Therefore, there exists an estimator $\hat{\boldsymbol{\gamma}}_i$ satisfying the KKT condition (2.29) as well as $sign(\hat{\boldsymbol{\gamma}}_i) = sign(\boldsymbol{\gamma}_i)$ which implies $\hat{\mathcal{D}}_i = \mathcal{D}_i$.

3. DIFFERENTIAL ANALYSIS OF DIRECTED NETWORKS

3.1 Introduction

It is of great importance and interest to detect sparse structural differences or differential structures between two cognate networks. For instance, the gene regulatory networks of diseased and healthy individuals may differ slightly from each other (West et al., 2012), and identifying the subtle difference between them helps design specific drugs. Social networks evolve over times, and monitoring their abrupt changes may serve as surveillance to economic stability or disease epidemics (Pianese et al., 2013; Berkman and Syme, 1979). However, addressing such practical problems demands differential analysis of large networks, calling for development of efficient statistical method to infer and compare complex structures from high dimensional data. In this paper, we focus on differential analysis of directed acyclic or even cyclic networks which can be described by structural equation models (SEMs).

Many research efforts have been made towards construction of a single network via SEM. For example, both Xiong et al. (2004) and Liu et al. (2008) employed a genetic algorithm to search for the best SEM using different information criteria. Most recently, Ni et al. (2016, 2018) employed a hierarchical Bayes approach to construct the SEM based networks. However, these approaches were designed for small or medium scale networks. For large-scale networks that the number of endogenous variables p exceeds the sample size n, Cai et al. (2013) proposed a regularization approach to fit a sparse model. Because this method suffers from incapability of parallel computation, it may not be feasible for large networks. Logsdon and Mezey (2010) proposed another penalization approach to fit the model in a node-wise fashion which alleviates the computational burden. Most recently, Lin et al. (2015), Zhu (2018), and Chen (2017) together with Chapter 2 each proposed a two-stage approach to construct SEMs, with different algorithms designed at different stages. As shown by Chen (2017) and Chapter 2, such a two-stage estimation approach can have superior performance compared to other methods and enjoys good consistency and variable selection properties for both fixed and diverging dimensions.

To the best of our knowledge, no algorithm has been proposed to conduct differential analysis of directed networks characterized by SEM. While a naive approach would separately construct each individual network and identify common and differential structures, this approach fails to take advantage of the commonality as well as sparse differential structures of the paired networks, leading to higher false positive rate or lower power. In this light, we introduce a novel statistical method, specially in the directed network regime, to conduct differential analysis of two networks via appropriate reparameterization of the corresponding models. There are two major features of our method. Firstly, we jointly model the commonality and difference between two networks explicitly. This helps us to gain dramatic performance improvements over the naive construction method. Secondly, benefiting from the flexible framework of SEMs, we are able to conduct differential analysis of directed networks. Most importantly, our method allow for both acyclic and cyclic networks. Compared to the other methods, directionality and allowing for cyclicity are crucial for many network studies, especially in constructing gene regulatory networks. As far as we know, our method is the first work on differential analysis of directed networks that enjoys the two promising features.

The rest of this chapter is organized as follows. We first introduce the model and its identification assumption in Section 3.2.1 and Section 3.2.2, respectively. Then, we present our proposed method of **Re**parameterization-Based **D**ifferential analysis of directed **Net**works, termed as **ReDNet**, and its comprehensive theoretical justification in Section 3.3. Section 3.4 includes our studies on simulated data showing the superior performance of our method. Section 3.5 demonstrates a real data analysis using the Genotype-Tissue Expression (GTEx) data sets. We conclude our paper with brief discussion in Section 3.6.

3.2 Structural Equation Models and Their Identifiability

Here, we first introduce the use of structural equation model and its identifiability assumption, and then describe our proposed **ReDNet** method for identifying common and differential structures between two directed networks, followed with its theoretical justification.

3.2.1 The Model

We consider two networks, each describing the dependencies among a common set of variables or nodes in a unique population. For each node $i \in \{1, 2, \ldots, p\}$ in network $k \in \{1, 2\}$, its regulation structure can be represented by the following equation,

$$\underbrace{\mathbf{Y}_i^{(k)}}_{\text{node } i} = \underbrace{\mathbf{Y}_{-i}^{(k)}\boldsymbol{\gamma}_i^{(k)}}_{\text{regulation by others}} + \underbrace{\mathbf{X}^{(k)}\boldsymbol{\phi}_i^{(k)}}_{\text{anchoring regulation}} + \underbrace{\boldsymbol{\epsilon}_i^{(k)}}_{\text{error}}, \tag{3.1}$$

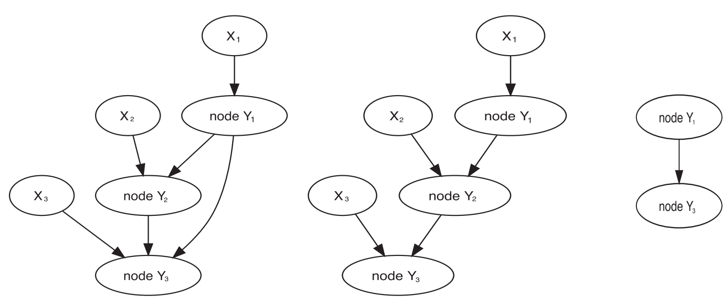
where $\mathbf{Y}_i^{(k)}$ is the i-th column of $\mathbf{Y}^{(k)}$ and $\mathbf{Y}_{-i}^{(k)}$ is the submatrix of $\mathbf{Y}^{(k)}$ by excluding $\mathbf{Y}_i^{(k)}$, with $\mathbf{Y}^{(k)}$ a $n^{(k)} \times p$ matrix. $\mathbf{X}^{(k)}$ is a $n^{(k)} \times q$ matrix with each column standardized to have ℓ_2 norm $\sqrt{n^{(k)}}$. The vectors $\boldsymbol{\gamma}_i^{(k)}$ and $\boldsymbol{\phi}_i^{(k)}$ encode the inter-nodes and anchoring regulatory effects, respectively. The index set of non-zeros of $\boldsymbol{\phi}_i^{(k)}$ is known and denoted by $\mathcal{A}_i^{(k)}$, in other words, $\mathcal{A}_i^{(k)} = \text{supp}(\boldsymbol{\phi}_i^{(k)})$. The support set $\mathcal{A}_i^{(k)}$ indexes the direct causal effects for the i-th node, and can be prespecified based on the domain knowledge. However, the size of nonzero effect $\boldsymbol{\phi}_i^{(k)}$ is unknown and can be estimated. Further property of $\mathcal{A}_i^{(k)}$ will be discussed in Section 3.2.2. All elements of the error term are independently distributed following a normal distribution with mean zero and standard deviation $\sigma_i^{(k)}$. We assume that the matrix $\mathbf{X}^{(k)}$ and the error term $\boldsymbol{\epsilon}_i^{(k)}$ are independent of each other. However $\mathbf{Y}_{-i}^{(k)}$ and $\boldsymbol{\epsilon}_i^{(k)}$ may correlate with each other. $\mathbf{Y}^{(k)}$ and $\mathbf{X}^{(k)}$ include observed endogenous variables and exogenous variables, respectively.

By combining the p linear equations in (3.1), we can rewrite the two sets of linear equations in a systematic fashion as two structural equation models below,

$$\begin{cases} \mathbf{Y}^{(1)} = \mathbf{Y}^{(1)}\mathbf{\Gamma}^{(1)} + \mathbf{X}^{(1)}\mathbf{\Phi}^{(1)} + \boldsymbol{\epsilon}^{(1)}, \\ \mathbf{Y}^{(2)} = \mathbf{Y}^{(2)}\mathbf{\Gamma}^{(2)} + \mathbf{X}^{(2)}\mathbf{\Phi}^{(2)} + \boldsymbol{\epsilon}^{(2)}, \end{cases} \tag{3.2}$$

where each matrix $\mathbf{\Gamma}^{(k)}$ is $p \times p$ with zero diagonal elements and represents the inter-nodes regulatory effects in the corresponding network. Specifically, excluding i-th element (which is zero) from the i-th column of $\mathbf{\Gamma}^{(k)}$ leads to $\boldsymbol{\gamma}_i^{(k)}$. The $q \times p$ matrix $\mathbf{\Phi}^{(k)}$ contains the anchoring regulatory effects and its i-th column is $\boldsymbol{\phi}_i^{(k)}$. Each error term $\boldsymbol{\epsilon}^{(k)}$ is $n^{(k)} \times p$ and has the error term $\boldsymbol{\epsilon}_i^{(k)}$ as its i-th column.

Figure 3.1 gives an illustrative example of networks with three nodes and one anchoring regulation per node for the structural equations in (3.2). For example, with anchoring regulation on nodes Y_1, X_1 has a direct effect on node Y_1 but indirect effects on node Y_2 and Y_3 via Y_1.



(a) Network I (b) Network II (c) Differential

Figure 3.1. An Illustrative Example of Differential Network Between Two Directed Networks. The error term for each node is not shown for simplicity.

For each network k, its full model in (3.2) can be further transformed into the reduced form as follows,

$$\mathbf{Y}^{(k)} = \mathbf{X}^{(k)}\boldsymbol{\pi}^{(k)} + \boldsymbol{\xi}^{(k)}, \tag{3.3}$$

where the $q \times p$ matrix $\boldsymbol{\pi}^{(k)} = \boldsymbol{\Phi}^{(k)}(\mathbf{I} - \boldsymbol{\Gamma}^{(k)})^{-1}$ and the transformed error term $\boldsymbol{\xi}^{(k)} = \boldsymbol{\epsilon}^{(k)}(\mathbf{I} - \boldsymbol{\Gamma}^{(k)})^{-1}$. The reduced model (3.3) reveals variables observed in $\mathbf{X}^{(k)}$ as instrumental variables which will be used later to correct for the endogeneity issue. Otherwise, directly applying any regularization based regression to equation (3.1) will result in non-consistent or suboptimal estimation of model parameters (Fan and Liao, 2014; Chen, 2017; Lin et al., 2015; Zhu, 2018).

3.2.2 The Model Identifiability

Here we introduce an identifiability assumption which helps to infer an identifiable system (3.2) from available data. We assume that each endogenous variable is directly regulated by a unique set of exogenous variables as long as it regulates other endogenous variables. That is, any regulatory node needs at least one anchoring exogenous variable to distinguish the corresponding regulatory effects from association. Explicitly let $\mathcal{M}_{i0}^{(k)}$ denote the index set of endogenous variables which either directly or indirectly regulate the i-th endogenous variable in the k-th network. Thus, $\mathcal{A}_i^{(k)} \subseteq \mathcal{M}_{i0}^{(k)}$. The model identification assumption can be stated in the below.

Assumption 1. For any $i = 1, \cdots, p$, $\mathcal{A}_i^{(k)} \neq \emptyset$ if there exists j such that $i \in \mathcal{M}_{j0}^{(k)}$. Furthermore, $\mathcal{A}_i^{(k)} \cap \mathcal{A}_j^{(k)} = \emptyset$ as long as $i \neq j$.

This assumption is slightly less restrictive than the one employed by Chen (2017), and is a sufficient condition for model identifiability as it satisfies the rank condition in Schmidt (1976). It can be further relaxed to allow nonempty $\mathcal{A}_i^{(k)} \cap \mathcal{A}_j^{(k)}$ as long as each regulatory node has its own unique anchoring exogenous variables.

The above identifiability assumption not only identifies $\boldsymbol{\gamma}_i^{(k)}$ in model (3.1) from $\boldsymbol{\pi}^{(k)}$ in model (3.3) but also helps reveal regulatory directionality of the networks. As illustrated in Figure 1.3 of Section 1.2.2, the additional anchoring variables break the "Markov Equivalence" and recover the directionality between nodes. In other words, the known set $\mathcal{A}_j^{(k)}$ serves as external prior knowledge which helps recover the directionality. In our two-stage construction of the differential network, the additional anchoring variables serve as instrumental variables in the calibration stage, since they are independent of the error terms. The present direct causal effects from $\mathbf{X}^{(k)}$ together with Assumption 1 differentiates our approach from the classical graphical models (Meinshausen and Bühlmann, 2006; Yuan and Lin, 2007) or the PC algorithm approaches (Spirtes et al., 2000; Kalisch and Bühlmann, 2007), since those methods either cannot recover edge directions or do not allow for cyclic structure due to lack of additional direct causal effects from $\mathbf{X}^{(k)}$.

3.3 Two-Stage Differential Analysis of Networks

Here we intend to develop a regularized version of the two-stage least squares. We first screen for exogenous variables and conduct ℓ_2 regularized regression of each endogenous variable against screened exogenous variables to obtain its good prediction which helps address the endogeneity issue in the following stage. At the second stage, we reparametrize the model to explicitly model the common and differential regulatory effects and identify them via the adaptive lasso method.

3.3.1 The Calibration Stage

To address the endogeneity issue, we aim for good prediction of each endogenous variable following the reduced model in (3.3). However, in the high-dimensional setting, the dimension q of $\mathbf{X}^{(k)}$ can be much larger than the sample size $n^{(k)}$, and any direct prediction with all exogenous variables may not produce consistent prediction. Note that both Lin et al. (2015) and Zhu (2018) proposed to conduct variable selection with lasso or its variants and predict with selected exogenous variables. We here instead propose to first screen for exogenous variables with ISIS (Fan and Lv, 2008), and then apply ridge regression to predict the endogenous variables with screened exogenous variables. While variable screening is more robust and provides higher coverage of true variables than variable selection, its combination with ridge regression puts less computational burden. Furthermore, as shown by Chen (2017), ridge regression performs well in predicting the endogenous variables.

Let $\mathcal{M}_i^{(k)}$ denotes the selected index set for the i-th node in the k-th network from the variable screening which reduces the dimension from q to $d = |\mathcal{M}_i^{(k)}|$. The *Sure Independence Screening Property* in Fan and Lv (2008) can be directly applied in our case to guarantee that $\mathcal{M}_i^{(k)}$ covers the true set $\mathcal{M}_{i0}^{(k)}$ with a large probability. Here, We state Assumption 2 and 3 by recollecting the conditions in Fan and Lv (2008) to pave that way for Theorem 3.3.1 for sure screening.

Denote $Y_{ji}^{(k)}$, $X_{jl}^{(k)}$, $\xi_{ji}^{(k)}$, and $\pi_{ji}^{(k)}$ as the j-th row of $\mathbf{Y}_i^{(k)}$, $\mathbf{X}_l^{(k)}$, $\boldsymbol{\xi}_i^{(k)}$, and $\boldsymbol{\pi}_i^{(k)}$, respectively. Further denote $\Sigma^{(k)}$ the variance-covariance matrix of the q random variables in observing $\mathbf{X}^{(k)}$. For any index subet $\mathcal{M} \subset \{1, 2, \cdots, q\}$, denote $\Sigma_{\mathcal{M}}^{(k)}$ the variance-covariance matrix of the random variables in observing $\mathbf{X}_{\mathcal{M}}^{(k)}$.

Assumption 2. $n^{(1)}$ and $n^{(2)}$ are at the same order, i.e., $n_{\min} = \min(n^{(1)}, n^{(2)}) \asymp n^{(1)} \asymp n^{(2)}$, and $p \asymp q$.

Assumption 3. For each node i in network $k \in \{1, 2\}$,

(a) Each $\xi_{ji}^{(k)}$ is normally distributed with mean zero. $(\Sigma^{(k)})^{-1/2}\mathbf{X}^{(k)T}$ is observed from a spherically symmetric distribution, and has the concentration property: there exist some constants $\tilde{c}_1^{(k)}, \tilde{c}_2^{(k)} > 1$ and $\tilde{c}_3^{(k)} > 0$ such that, for any index subset $\mathcal{M} \subset \{1, 2, \cdots, q\}$ with $|\mathcal{M}| \geq \tilde{c}_1^{(k)} n^{(k)}$, the eigenvalues of $|\mathcal{M}|^{-1}\mathbf{X}_{\mathcal{M}}^{(k)}(\Sigma_{\mathcal{M}}^{(k)})^{-1/2}(\Sigma_{\mathcal{M}}^{(k)T})^{-1/2}\mathbf{X}_{\mathcal{M}}^{(k)T}$ are bounded either from above by $\tilde{c}_2^{(k)}$ or from below by $1/\tilde{c}_2^{(k)}$ with probability at least $1 - \exp(-\tilde{c}_3^{(k)} n^{(k)})$.

(b) $\mathrm{var}(Y_{ji}^{(k)}) = O(1)$. For some $\kappa^{(k)} \geq 0$, $\tilde{c}_4^{(k)} > 0$, and $\tilde{c}_5^{(k)} > 0$,

$$\min_{l \in \mathcal{M}_{i0}^{(k)}} \left|\pi_{li}^{(k)}\right| \geq \frac{\tilde{c}_4^{(k)}}{(n^{(k)})^{\kappa^{(k)}}} \quad \text{and} \quad \min_{l \in \mathcal{M}_{i0}^{(k)}} \left|\mathrm{cov}\left((\pi_{li}^{(k)})^{-1} Y_{ji}^{(k)}, X_{jl}^{(k)}\right)\right| \geq \tilde{c}_5^{(k)}.$$

(c) $\log(q) = O((n^{(k)})^{\tilde{c}})$ for some $\tilde{c} \in (0, 1 - 2\kappa^{(k)})$.

(d) There are some $\tau^{(k)} \geq 0$ and $\tilde{c}_6^{(k)} > 0$ such that $\lambda_{\max}(\Sigma^{(k)}) \leq \tilde{c}_6^{(k)} (n^{(k)})^{\tau^{(k)}}$.

Theorem 3.3.1 *Under Assumption 1, 2 and 3 which restrict the positive pairs $\tau^{(k)}$ and $\kappa^{(k)}$. Denote $\tilde{\tau} = \max\{\tau^{(1)}, \tau^{(2)}\}$ and $\tilde{\kappa} = \max\{\kappa^{(1)}, \kappa^{(2)}\}$, then there exists some*

$\theta \in (0, 1-2\tilde{\kappa}-\tilde{\tau})$ *such that, when* $d = |\mathcal{M}_i^{(k)}| = O((n_{\min})^{1-\theta})$*, we have, for some constant* $C > 0$*,*

$$\mathbb{P}(\mathcal{M}_{i0}^{(k)} \subseteq \mathcal{M}_i^{(k)}) = 1 - \mathcal{O}\left(\exp\left\{-\frac{C(n^{(k)})^{1-2\tilde{\kappa}}}{\log(n^{(k)})}\right\}\right).$$

Hereafter we assume that $\mathcal{M}_i^{(k)}$ successfully covers the true set $\mathcal{M}_{i0}^{(k)}$ for convenience of stating the following assumptions and theorems. That is, the probability of successful screening is not incorporated into our assumptions or theorems in the below.

For node i in network k, let $\mathbf{X}_{\mathcal{M}_i^{(k)}}^{(k)}$ denotes the submatrix of $\mathbf{X}^{(k)}$ with prescreened columns which are indexed by $\mathcal{M}_i^{(k)}$. With $\boldsymbol{\pi}_i^{(k)}$ denoting the i-th column of $\boldsymbol{\pi}^{(k)}$, the subvector of $\boldsymbol{\pi}_i^{(k)}$ indexed by $\mathcal{M}_i^{(k)}$ will be simply denoted by $\boldsymbol{\pi}_{\mathcal{M}_i^{(k)}}^{(k)}$ without confusion. Such simplified notations will apply to other vectors and matrices in the rest of this paper.

With d pre-screened exogenous variables, we can apply ridge regression to the model

$$\mathbf{Y}_i^{(k)} = \mathbf{X}_{\mathcal{M}_i^{(k)}}^{(k)} \boldsymbol{\pi}_{\mathcal{M}_i^{(k)}}^{(k)} + \boldsymbol{\xi}_i^{(k)}, \tag{3.4}$$

to obtain the estimates $\hat{\boldsymbol{\pi}}_{\mathcal{M}_i^{(k)}}^{(k)}$ of $\boldsymbol{\pi}_{\mathcal{M}_i^{(k)}}^{(k)}$, and predict $\mathbf{Y}_i^{(k)}$ with $\hat{\mathbf{Y}}_i^{(k)} = \mathbf{X}_{\mathcal{M}_i^{(k)}}^{(k)} \hat{\boldsymbol{\pi}}_{\mathcal{M}_i^{(k)}}^{(k)}$.

3.3.2 The Construction Stage

With known $\mathcal{A}_i^{(k)}$, we can rewrite model (3.1) as,

$$\mathbf{Y}_i^{(k)} = \mathbf{Y}_{-i}^{(k)}\boldsymbol{\gamma}_i^{(k)} + \mathbf{X}_{\mathcal{A}_i^{(k)}}^{(k)}\boldsymbol{\phi}_{\mathcal{A}_i^{(k)}}^{(k)} + \boldsymbol{\epsilon}_i^{(k)}. \tag{3.5}$$

Before we use the predicted $\mathbf{Y}^{(k)}$ to identify both common and differential regulatory effects across the two networks, we first reparametrize the model so as to define differential regulatory effects explicitly,

$$\boldsymbol{\beta}_i^- = \frac{\boldsymbol{\gamma}_i^{(1)} - \boldsymbol{\gamma}_i^{(2)}}{2},\ \boldsymbol{\beta}_i^+ = \frac{\boldsymbol{\gamma}_i^{(1)} + \boldsymbol{\gamma}_i^{(2)}}{2}. \tag{3.6}$$

Here $\boldsymbol{\beta}_i^-$ and $\boldsymbol{\beta}_i^+$ represent the **differential** and **average regulatory effects** between the two networks, respectively. We need compare $\boldsymbol{\beta}_i^+$ with $\boldsymbol{\beta}_i^-$ to identify the **common**

regulatory effects, that is, effects of all regulations with nonzero values in $\boldsymbol{\beta}_i^{+}$ but zero values in $\boldsymbol{\beta}_i^{-}$.

Note that other differential analysis of networks may suggest a different reparametrization to identify common and differential regulatory effects. For example, in a typical case-control study, we may expect few structures in the case network mutated from the control network. While we are interested in identifying differential structures in the case network, we may be also interested in identifying baseline network structures in the control network. Therefore we may reparametrize the model with the regulatory effects in the control network, as well as the differential regulatory effects defined as the difference of regulatory effects between case and control networks. We want to point out that the method described here still applies and we can also derive similar theoretical results as follows.

Following the reparametrization in (3.6), we can rewrite model (3.5) as follows,

$$\begin{pmatrix} \mathbf{Y}_i^{(1)} \\ \mathbf{Y}_i^{(2)} \end{pmatrix} = \begin{pmatrix} \mathbf{Y}_{-i}^{(1)} & \mathbf{Y}_{-i}^{(1)} \\ \mathbf{Y}_{-i}^{(2)} & -\mathbf{Y}_{-i}^{(2)} \end{pmatrix} \begin{pmatrix} \boldsymbol{\beta}_i^{+} \\ \boldsymbol{\beta}_i^{-} \end{pmatrix} + \begin{pmatrix} \mathbf{X}_{\mathcal{A}_i^{(1)}}^{(1)} & 0 \\ 0 & \mathbf{X}_{\mathcal{A}_i^{(2)}}^{(2)} \end{pmatrix} \begin{pmatrix} \boldsymbol{\phi}_{\mathcal{A}_i^{(1)}}^{(1)} \\ \boldsymbol{\phi}_{\mathcal{A}_i^{(2)}}^{(2)} \end{pmatrix} + \begin{pmatrix} \boldsymbol{\epsilon}_i^{(1)} \\ \boldsymbol{\epsilon}_i^{(2)} \end{pmatrix}. \tag{3.7}$$

Denote

$$\mathbf{Y}_i = \begin{pmatrix} \mathbf{Y}_i^{(1)} \\ \mathbf{Y}_i^{(2)} \end{pmatrix}, \quad \mathbf{Z}_{-i} = \begin{pmatrix} \mathbf{Y}_{-i}^{(1)} & \mathbf{Y}_{-i}^{(1)} \\ \mathbf{Y}_{-i}^{(2)} & -\mathbf{Y}_{-i}^{(2)} \end{pmatrix},$$
$$\boldsymbol{\beta}_i = \begin{pmatrix} \boldsymbol{\beta}_i^{+} \\ \boldsymbol{\beta}_i^{-} \end{pmatrix}, \quad \boldsymbol{\epsilon}_i = \begin{pmatrix} \boldsymbol{\epsilon}_i^{(1)} \\ \boldsymbol{\epsilon}_i^{(2)} \end{pmatrix}.$$

Further define the projection matrix for each network,

$$\mathbf{H}_i^{(k)} = \mathbf{I}_{n^{(k)}} - \mathbf{X}_{\mathcal{A}_i^{(k)}}^{(k)} \left(\mathbf{X}_{\mathcal{A}_i^{(k)}}^{(k)T} \mathbf{X}_{\mathcal{A}_i^{(k)}}^{(k)} \right)^{-1} \mathbf{X}_{\mathcal{A}_i^{(k)}}^{(k)T}.$$

Applying the projection matrix $\mathbf{H}_i = \mathrm{diag}\{\mathbf{H}_i^{(1)}, \mathbf{H}_i^{(2)}\}$ to both sides of model (3.7), we can remove the exogenous variables from the model and obtain,

$$\mathbf{H}_i \mathbf{Y}_i = \mathbf{H}_i \mathbf{Z}_{-i} \boldsymbol{\beta}_i + \mathbf{H}_i \boldsymbol{\epsilon}_i. \tag{3.8}$$

Algorithm 2 Reparameterization-Based Differential Analysis of Network (ReDNet)

Input: For $k \in \{1,2\}$, $\mathbf{Y}^{(k)}$, $\mathbf{X}^{(k)}$, index set $\mathcal{A}_i^{(k)}$ for each $i \in \{1,2,\ldots,p\}$. Set $d = O(n_{\min}^{1-\theta})$.

for $i \to 1$ **to** p **do**

Stage 1.a. Screen for a sub-matrix $\mathbf{X}^{(k)}_{\mathcal{M}_i^{(k)}}$ of $\mathbf{X}^{(k)}$ for model $\mathbf{Y}_i^{(k)}$ versus $\mathbf{X}^{(k)}$ and set $\mathbf{X}^{(k)}_{\mathcal{M}_i^{(k)}} = \mathbf{X}^{(k)}$ if $q \leq n^{(k)}$.

Stage 1.b. Apply ridge regression to regress $\mathbf{Y}_i^{(k)}$ against $\mathbf{X}^{(k)}_{\mathcal{M}_i^{(k)}}$ to obtain prediction $\hat{\mathbf{Y}}_i^{(k)}$.

end for

for $i \to 1$ **to** p **do**

Stage 2. Apply adaptive lasso to regress $\mathbf{H}_i\mathbf{Y}_i$ against $\mathbf{H}_i\hat{\mathbf{Z}}_{-i}$ to obtain coefficients estimate $\hat{\boldsymbol{\beta}}_i$.

end for

Output: The common and differential regulatory effects in $\hat{\boldsymbol{\beta}}_1, \ldots, \hat{\boldsymbol{\beta}}_p$.

To address the endogeneity issue, we predict $\mathbf{Z}_{-i}$ by replacing its component $\mathbf{Y}_{-i}^{(k)}$ with the predicted value $\hat{\mathbf{Y}}_{-i}^{(k)}$ from the previous stage, and then regressing $\mathbf{H}_i\mathbf{Y}_i$ against $\mathbf{H}_i\hat{\mathbf{Z}}_{-i}$ with the adaptive lasso to consistently estimate $\boldsymbol{\beta}_i$. That is, an optimal $\boldsymbol{\beta}_i$ can be obtained as,

$$\hat{\boldsymbol{\beta}}_i = \arg\min_{\boldsymbol{\beta}_i} \left\{ \frac{1}{n} ||\mathbf{H}_i\mathbf{Y}_i - \mathbf{H}_i\hat{\mathbf{Z}}_{-i}\boldsymbol{\beta}_i||_2^2 + \lambda_i \boldsymbol{\omega}_i^T |\boldsymbol{\beta}_i|_1 \right\},$$

where $|\boldsymbol{\beta}_i|_1$ is a vector taking element-wise absolute values of $\boldsymbol{\beta}_i$ and $\boldsymbol{\omega}_i$ is the adaptive weights whose components are inversely proportional to the components of an initial estimator of $\boldsymbol{\beta}_i$, and λ_i is the adaptive tuning parameter.

The two-stage algorithm is summarized in Algorithm 2. With the estimator $\hat{\boldsymbol{\beta}}_i$ from the second stage, we can accordingly obtain estimators $\hat{\boldsymbol{\gamma}}_i^{(1)} = \hat{\boldsymbol{\beta}}_i^+ + \hat{\boldsymbol{\beta}}_i^-$ and $\hat{\boldsymbol{\gamma}}_i^{(2)} = \hat{\boldsymbol{\beta}}_i^+ - \hat{\boldsymbol{\beta}}_i^-$.

As shown in Theorem 3.3.1, a screening method like ISIS (Fan and Lv, 2008) can identify $\mathcal{M}_i^{(k)}$ with size $d = O(n_{\min}^{1-\theta})$ which covers the true set $\mathcal{M}_{i0}^{(k)}$ with a sufficiently large probability. For the sake of simplicity and without loss of generality, in the following we assume $\mathcal{M}_{i0}^{(k)} \subseteq \mathcal{M}_i^{(k)}$.

We first investigate the consistency of predictions from the first stage. The consistency properties will be characterized by prespecified sequences $f^{(k)} = o(n^{(k)})$ but $f^{(k)} \to \infty$ as $n^{(k)} \to \infty$. We also denote $f_{\max} = f^{(1)} \vee f^{(2)}$, i.e., $\max\{f^{(1)}, f^{(2)}\}$.

The following assumption is required for the consistency properties.

Assumption 4. For each network k, the singular values of $\mathbf{I} - \mathbf{\Gamma}^{(k)}$ are positively bounded from below, and there exist constants $c_1^{(k)}, c_2^{(k)} > 0$ such that, for each node i, $\max_{||\delta||_2=1}(n^{(k)})^{-1/2}||\mathbf{X}^{(k)}_{\mathcal{M}_i^{(k)}}\delta||_2 \le c_1^{(k)}$ and $\min_{||\delta||_2=1}(n^{(k)})^{-1/2}||\mathbf{X}^{(k)}_{\mathcal{M}_i^{(k)}}\delta||_2 \ge c_2^{(k)}$. Furthermore, the ridge parameter $\lambda_i^{(k)} = o(n_{\min})$.

For the ease of exposition, we will omit the subscript $\mathcal{M}_i^{(k)}$ from $\mathbf{X}^{(k)}_{\mathcal{M}_i^{(k)}}$ henceforth, and accordingly use $\boldsymbol{\pi}_i^{(k)}$ and $\hat{\boldsymbol{\pi}}_i^{(k)}$ which include the zero components of excluded predictors.

Denote $\mathbf{X} = \mathrm{diag}\{\mathbf{X}^{(1)}, \mathbf{X}^{(2)}\}$, and

$$\mathbf{Z} = \begin{pmatrix} \mathbf{Y}^{(1)} & \mathbf{Y}^{(1)} \\ \mathbf{Y}^{(2)} & -\mathbf{Y}^{(2)} \end{pmatrix}, \qquad \mathbf{\Pi} = \begin{pmatrix} \boldsymbol{\pi}^{(1)} & \boldsymbol{\pi}^{(1)} \\ \boldsymbol{\pi}^{(2)} & -\boldsymbol{\pi}^{(2)} \end{pmatrix}.$$

We use $\mathbf{\Pi}_j$ to denote the j-th column of the matrix $\mathbf{\Pi}$ and $\boldsymbol{\pi}_j^{(k)}$ to denote the j-th column of the matrix $\boldsymbol{\pi}^{(k)}$. We also use $\hat{\mathbf{Z}}$ and $\hat{\mathbf{\Pi}}$ to denote the prediction of $\mathbf{Z}$ and estimate of $\mathbf{\Pi}$, respectively. Note that, with the ridge parameter $\lambda_i^{(k)}$ for the ridge regression taken on node i in network k, we have $r_i^{(k)} = (\lambda_i^{(k)})^2||\boldsymbol{\pi}_i^{(k)}||_2^2/n^{(k)}$ and hence define $r_{\max} = \max\limits_{1\le i\le p}[r_i^{(1)} \vee r_i^{(2)}]$. Then the estimation and prediction losses at the first stage can be summarized in the following theorem.

Theorem 3.3.2 *Under Assumptions 1-4, for each $j \in \{1, 2, \ldots, 2p\}$, there will exist some constant C_1 and C_2 such that, with probability at least $1 - e^{-f^{(1)}} - e^{-f^{(2)}}$,*

1. $||\hat{\mathbf{\Pi}}_j - \mathbf{\Pi}_j||_2^2 \le C_1\left(d \vee r_{\max} \vee f_{\max}\right)/n_{\min}$;

2. $||\mathbf{X}(\hat{\mathbf{\Pi}}_j - \mathbf{\Pi}_j)||_2^2 \le C_2\left(d \vee r_{\max} \vee f_{\max}\right)$.

Note that these two sets of losses can be controlled by the same upper bounds across the two networks with probability at least $1 - e^{-f^{(1)}+\log(p)} - e^{-f^{(2)}+\log(p)}$. Therefore, $f^{(k)}$ can be selected such that $f^{(k)} - log(p) \to \infty$, which will provide a probability approaching one to have the network-wide losses approaching zero.

Furthermore, the dimension p can be divergent up to an exponential order, say $p = e^{n_{\min}^c}$ for some $c \in (0, 1)$. We can set $f^{(1)} = f^{(2)} = n_{\min}^{(1+c)/2}$ and, apparently, $f^{(k)} = o(n_{\min})$ but $f^{(k)} - log(p) = n_{\min}^{(1+c)/2} - n_{\min}^c \to \infty$.

Since the ridge parameter $\lambda_i^{(k)} = o(n_{\min})$, $r_i^{(k)} = ||\boldsymbol{\pi}_i^{(k)}||_2^2 \times o(n_{\min})$. Therefore, when all $||\boldsymbol{\pi}_i^{(k)}||_2$ are uniformly bounded, we have $r_{\max} = o(n_{\min})$. Otherwise, the ridge parameter $\lambda_i^{(k)}$ should be adjusted accordingly to control both estimation and prediction losses. The proof is detailed in Section 3.7.

For the i-th node, we use $\mathcal{S}_i$ to denote the non-zero indices of $\boldsymbol{\beta}_i$, i.e., $\mathcal{S}_i = \mathrm{supp}(\boldsymbol{\beta}_i)$. Further denote

$$\mathbf{\Pi}_{-i} = \begin{pmatrix} \boldsymbol{\pi}_{-i}^{(1)} & \boldsymbol{\pi}_{-i}^{(1)} \\ \boldsymbol{\pi}_{-i}^{(2)} & -\boldsymbol{\pi}_{-i}^{(2)} \end{pmatrix}.$$

As in Bickel et al. (2009), we utilize again the restricted eigenvalue defined in Definition 2.4.1 to impose the following restricted eigenvalue condition on the design matrix in (3.8).

Assumption 5. There exists a constant $\boldsymbol{\phi}_0 > 0$ such that $\phi_{\mathrm{re}}(\mathbf{H}_i\mathbf{X}\mathbf{\Pi}_{-i}, \mathcal{S}_i) \geq \boldsymbol{\phi}_0$. Furthermore, $\|\boldsymbol{\omega}_{\mathcal{S}_i}\|_\infty \leq \|\boldsymbol{\omega}_{\mathcal{S}_i^c}\|_{-\infty}$.

Let $n = n^{(1)} + n^{(2)}$, $c_{\max} = c_1^{(1)} \vee c_1^{(2)}$, and $\mathbf{B} = [\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \ldots, \boldsymbol{\beta}_p]$. The matrix norms $||\cdot||_1$ and $\|\cdot\|_\infty$ are the maximum of column and row sums of absolute values of the matrix, respectively. For a vector, we define $\|\cdot\|_\infty$ and $\|\cdot\|_{-\infty}$ to be the maximum and minimum absolute values of its components. Then, we can derive the following loss bounds for the estimation and prediction at the second stage on the basis of Theorem 3.3.2.

Theorem 3.3.3 *Suppose that, for node i, the adaptive lasso at the second stage takes the tuning parameter* $\lambda_i \asymp \|\boldsymbol{\omega}_i\|_{-\infty}^{-1}||\mathbf{B}||_1||\mathbf{\Pi}||_1\sqrt{(d \vee r_{\max} \vee f_{\max})\log(p)/n_{\min}}$, *and* $\sqrt{(d \vee r_{\max} \vee f_{\max})/n} + c_{\max}||\mathbf{\Pi}||_1 \leq \sqrt{c_{\max}^2||\mathbf{\Pi}||_1^2 + \phi_0^2/(64C_2|\mathcal{S}_i|)}$. *Let* $h_n = (||\mathbf{B}||_1^2 \wedge 1) \times ((n||\mathbf{\Pi}||_1^2/d) \wedge (d \vee r_{\max} \vee f_{\max})) \log(p)$. *Under Assumptions 1-5, there exist positive constants* C_3 *and* C_4 *such that, with probability at least* $1 - 3e^{-C_3 h_n + \log(4pq)} - e^{-f^{(1)} + \log(p)} - e^{-f^{(2)} + \log(p)}$,

1. *Estimation Loss:*

$$||\hat{\boldsymbol{\beta}}_i - \boldsymbol{\beta}_i||_1 \leq 8C_4|\mathcal{S}_i| \times \frac{\|\boldsymbol{\omega}_{\mathcal{S}_i}\|_\infty ||\mathbf{B}||_1 ||\mathbf{\Pi}||_1}{\boldsymbol{\phi}_0^2 \|\boldsymbol{\omega}_i\|_{-\infty}} \sqrt{\frac{(d \vee r_{max} \vee f_{\max}) \log(p)}{n_{\min}}};$$

2. *Prediction Loss:*

$$\frac{1}{n}||\mathbf{H}_i \hat{\mathbf{Z}}_{-i}(\hat{\boldsymbol{\beta}}_i - \boldsymbol{\beta}_i)||_2^2 \leq C_4^2|\mathcal{S}_i| \times \frac{\|\boldsymbol{\omega}_{\mathcal{S}_i}\|_\infty^2 ||\mathbf{B}||_1^2 ||\mathbf{\Pi}||_1^2}{\boldsymbol{\phi}_0^2 \|\boldsymbol{\omega}_i\|_{-\infty}^2} \frac{(d \vee r_{max} \vee f_{\max}) \log(p)}{n_{\min}}.$$

The main idea of the proof is to take advantage of the commonly used restricted eigenvalue condition and irrepresentable condition for lasso-type estimator. However, the design matrix in our case includes predicted values instead of the original one, which complicates the proof. We claim that the restricted eigenvalue and irrepresentable condition still hold for the predicted design matrix as long as the estimation and prediction losses are well controlled at the calibration stage. The proof is detailed in Section 3.7.

The available anchoring regulators as required by Assumption 1 implies that both $||\mathbf{B}||_1 > 0$ and $||\mathbf{\Pi}||_1 > 0$, so $h_n/\log(p) \to \infty$. That is, these loss bounds hold with a sufficient large probability with properly chosen $f^{(k)}$.

The two sets of losses in Theorem 3.3.3 can also be controlled across the whole system by the same upper bounds defined by replacing $|\mathcal{S}_i|$ with $s_{\max} = \max_i |\mathcal{S}_i|$, with probability at least $1 - 3e^{-C_3 h_n + \log(4q) + 2\log(p)} - e^{-f^{(1)} + 2\log(p)} - e^{-f^{(2)} + 2\log(p)}$. When both p and q are divergent up to an exponential order, say $p \asymp q \asymp e^{n_{\min}^c}$ for some $c \in (0, 1)$, we can set $f^{(1)} = f^{(2)} = n_{\min}^{(1+c)/2}$ to guarantee the bounds at a sufficient large probability. However, the bounds are determined by $(d \vee r_{\max} \vee f_{\max})\log(p)$ which is $o(n_{\min})$ only when $c < \min(1/3, \theta)$. Therefore, if $s_{\max}$ also diverges up to $n_{\min}^{\tilde{c}}$ with $\tilde{c} < \min(1/4, \theta/2, 1-\theta)$, the losses can be well controlled for $c < \min((1-4\tilde{c})/3, \theta-2\tilde{c})$.

Note that, with properly chosen $f^{(1)}$ and $f^{(2)}$, these losses are well controlled at $o(n_{\min})$, revealing the fact that we need to have sufficient observations for each network for consistent differential analysis of the two networks.

Let $W_i = diag\{\boldsymbol{\omega}_i\}$. Denote $\mathcal{I}_i = n^{-1}\mathbf{\Pi}_{-i}^T\mathbf{X}^T\mathbf{H}_i\mathbf{X}\mathbf{\Pi}_{-i}$ and $\hat{\mathcal{I}}_i = n^{-1}\hat{\mathbf{\Pi}}_{-i}^T\mathbf{X}^T\mathbf{H}_i\mathbf{X}\hat{\mathbf{\Pi}}_{-i}$. Let $\mathcal{I}_{i,11}$ be a submatrix of $\mathcal{I}_i$ with rows and columns both indexed by $\mathcal{S}_i$, and $\mathcal{I}_{i,21}$ be a submatrix of $\mathcal{I}_i$ with rows and columns indexed by $\mathcal{S}_i^c$ and $\mathcal{S}_i$, respectively. $\hat{\mathcal{I}}_{i,11}$ and $\hat{\mathcal{I}}_{i,21}$ are similarly defined from $\hat{\mathcal{I}}_i$. We further define the minimal signal strength $b_i = \max_{j\in\mathcal{S}_i}|\boldsymbol{\beta}_{ij}|$ and $\psi_i = ||\mathcal{I}_{i,11}^{-1}W_{\mathcal{S}_i}||_\infty$.

The following assumption, reminiscent of the *adaptive irrepresentable condition* in Huang et al. (2008), helps investigate the selection consistency of regulatory effects.

Assumption 6. (Weighted Irrepresentable Condition) There exists a constant $\tau \in (0,1)$ such that $||W_{\mathcal{S}_i^c}^{-1}\mathcal{I}_{i,21}\mathcal{I}_{i,11}^{-1}W_{\mathcal{S}_i}||_\infty < 1-\tau$.

Theorem 3.3.4 *(Variable Selection Consistency) Denote $\hat{\mathcal{S}}_i = supp(\hat{\boldsymbol{\beta}}_i)$. Suppose that, for node i, $\hat{\mathcal{I}}_{i,11}$ is invertible, $b_i > \lambda_i\psi_i/(2-\tau)$, and $\sqrt{(d \vee r_{\max} \vee f_{\max})/n} + c_{\max}||\mathbf{\Pi}||_1 \le \sqrt{c_{\max}^2||\mathbf{\Pi}||_1^2 + \min(\phi_0^2/64, \tau(4-\tau)^{-1}\|\boldsymbol{\omega}_i\|_{-\infty}/\psi_i)/(C_2|\mathcal{S}_i|)}$. Under Assumptions 1-6, there exists some constant $C_5 > 0$ such that $\hat{\mathcal{S}}_i = \mathcal{S}_i$ with probability at least $1 - 3e^{-C_5h_n+\log(4pq)} - e^{-f^{(1)}+\log(p)} - e^{-f^{(2)}+\log(p)}$.*

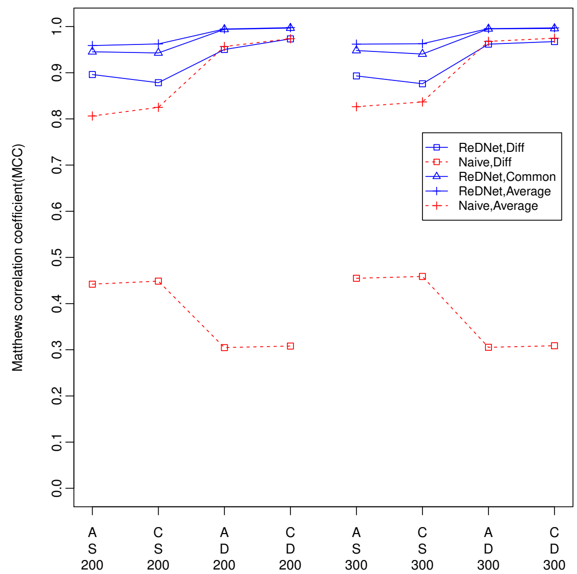
The above theorem implies that our proposed method can identify both common and differential regulatory effects between the two networks with a sufficiently large probability. On the other hand, the assumed weighted irrepresentable condition means that the true signal should not correlate too much with irrelevant covariates so as to conduct a successful differential analysis of networks. The corresponding proof is displayed in Section 3.7.

3.4 Simulation Study

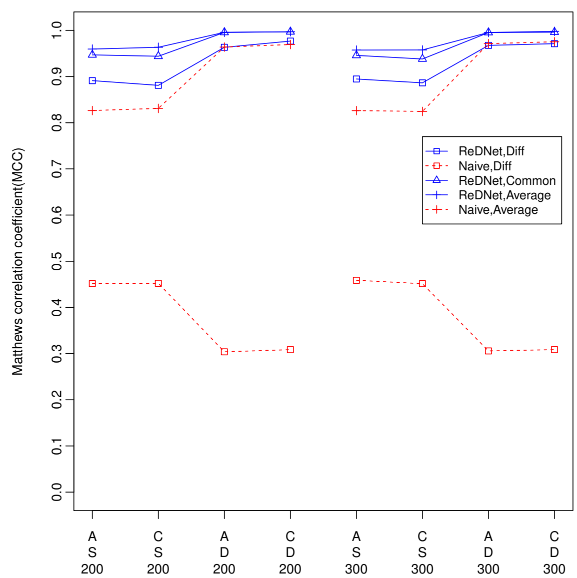
Here we report on experiments with synthetic data to show the superior performance of our method. We compare the method **ReDNet** to a naive differential analysis which employs the 2SPLS method proposed by Chen (2017) to construct each network separately. We refer to this method as **Naive**. Note that the 2SPLS method is modified here by applying ISIS to screen exogenous variables before conducting ridge regression to predict endogenous variables, making the naive differential analysis comparable to **ReDNet**.

Synthetic data are generated from both acyclic and cyclic networks involving 1000 endogenous variables, with the sample size from 200 to 300. Each network includes a subnetwork of 50 endogenous variables, whose shared and differential structures will be investigated against its pair. On average, each endogenous variable has one regulatory effect in a sparse subnetwork, and three regulatory effects on average in a dense network. While each pair of subnetworks in comparison share many identical regulatory effects, they also share five regulatory effects but with opposite signs, and each network has five unique regulatory effects (so the total number of differential regulatory effects is 15). The nonzero regulatory effects were independently sampled from a uniform distribution over the range $[-0.8, -0.3] \cup [0.3, 8]$. While assuming each node is directly regulated by one exogenous variable, each exogenous variable was sampled from discrete values 0, 1 and 2 with probabilities 0.25, 0.5 and 0.25, respectively. All of the noise terms were independently sampled from the normal distribution $N(0, 0.1^2)$. We also conducted differential analysis between two networks with both $\mathbf{X}^{(1)} \neq \mathbf{X}^{(2)}$ and $\mathbf{X}^{(1)} = \mathbf{X}^{(2)}$ as in practice the paired networks may or may not share identically valued exogenous variables.

We evaluate the the performance in terms of the false discovery rate (FDR), power and Matthews correlation coefficient (MCC) (Matthews, 1975). Let TP, TN, FP and

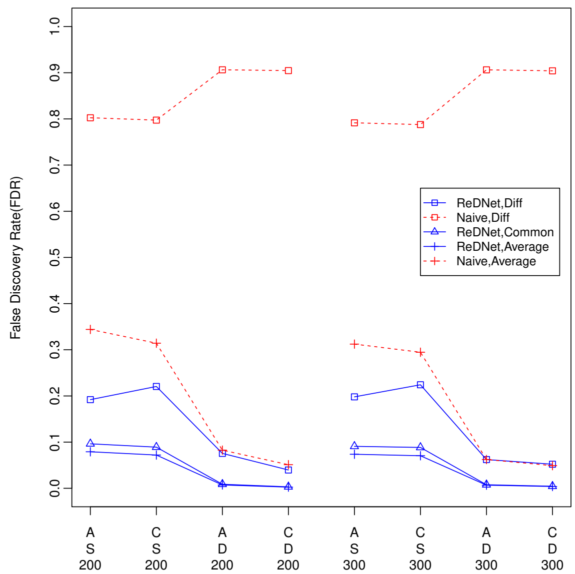


(a) $X_1 \neq X_2$, MCC

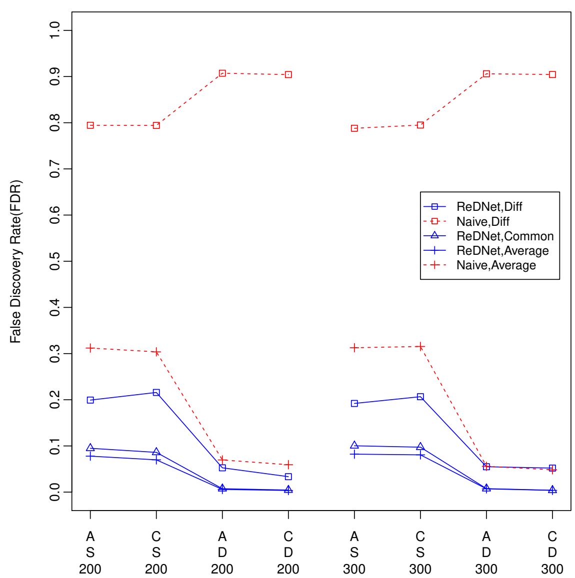


(b) $X_1 = X_2$, MCC

Figure 3.2. Performance of ReDNet Versus the Naive Approach which Independently Constructs Two Networks. The results average over 100 synthetic data sets for different types of networks, with letters A, C, S, D in the x-axis denoting Acyclic, Cyclic, Sparse and Dense networks, respectively. “Diff”, “Common” and “Average” summarize the performance on differential, common and average regulatory effects, respectively. MCC of the naive approach are undefined due to its failure to identify common effects. The sample size $n^{(2)} = n^{(2)}$ is either 200 or 300.

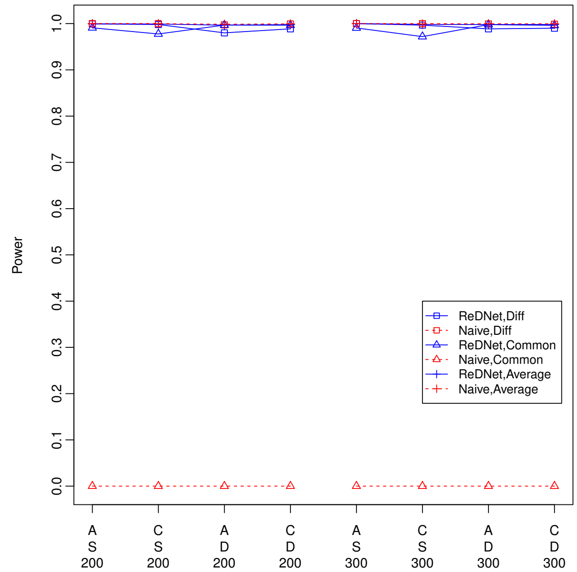


(a) $X_1 \neq X_2$, FDR

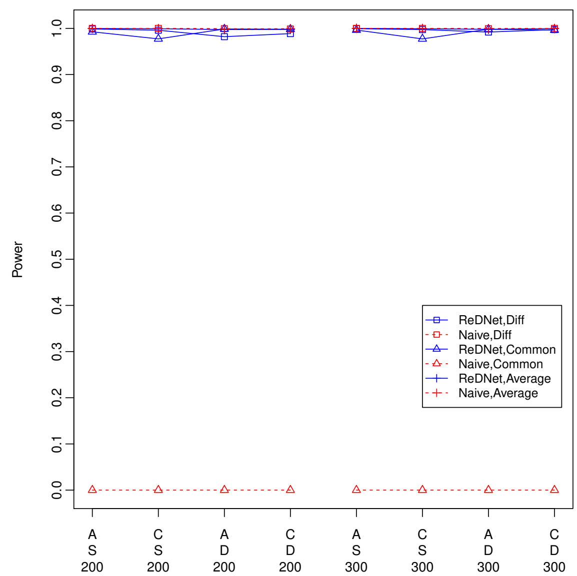


(b) $X_1 = X_2$, FDR

Figure 3.3. Performance of ReDNet Versus the Naive Approach which Independently Constructs Two Networks. The results average over 100 synthetic data sets for different types of networks, with letters A, C, S, D in the x-axis denoting Acyclic, Cyclic, Sparse and Dense networks, respectively. "Diff", "Common" and "Average" summarize the performance on differential, common and average regulatory effects, respectively. FDR of the naive approach are undefined due to its failure to identify common effects. The sample size $n^{(2)} = n^{(2)}$ is either 200 or 300.



(a) $X_1 \neq X_2$, Power



(b) $X_1 = X_2$, Power

Figure 3.4. Performance of ReDNet Versus the Naive Approach which Independently Constructs Two Networks.. The results average over 100 synthetic data sets for different types of networks, with letters A, C, S, D in the x-axis denoting Acyclic, Cyclic, Sparse and Dense networks, respectively. "Diff", "Common" and "Average" summarize the performance on differential, common and average regulatory effects, respectively. Power of the naive approach are always zero due to its failure to identify common effects. The sample size $n^{(2)} = n^{(2)}$ is either 200 or 300.

FN denote the numbers of true positives, true negatives, false positives and false negatives, respectively. MCC is defined as,

$$\text{MCC} = \frac{\text{TP} \times \text{TN} - \text{FP} \times \text{FN}}{\sqrt{(\text{TP}+\text{FP})(\text{TP}+\text{FN})(\text{TN}+\text{FP})(\text{TN}+\text{FN})}}.$$

Here we refer nonzero effects as positives and zero effects as negatives. The MCC varies from 0 to 1 with larger values implying better variable selection.

In each differential analysis, the ridge regression employed the generalized cross validation (Golub et al., 1979) to select the ridge parameter, and the adaptive lasso used 10-fold cross-validation to choose its tuning parameter. Following the recommendation by Fan and Lv (2008), $(n^{(k)})^{0.9}$ variables are screened by ISIS. The algorithm is implemented in R based on packages *SIS* (Saldana and Feng, 2018) and *parcor* (Kraemer et al., 2009).

For each type of networks, 100 synthetic data sets were generated, and the differential analysis results are summarized in Figure 3.2, Figure 3.3 and Figure 3.4. Overall, both **ReDNet** and the naive approach maintain high power in identifying differential regulatory effects. However, the naive approach fails to identify common regulatory effects and tends to report FDR over 80% on differential regulatory effects. Such a tendency to report false positives by the naive approach results in lower MCC, with dramatic decrease in identifying differential regulatory effects.

While both methods performed stably across networks with $\mathbf{X}^{(1)} \neq \mathbf{X}^{(2)}$ and $\mathbf{X}^{(1)} = \mathbf{X}^{(2)}$, **ReDNet** performed better in identifying differential regulatory effects from dense networks than sparse networks in terms of FDR and MCC. However, the naive approach tends to report even higher FDR and so much lower MCC when identifying differential regulatory effects from dense networks. Nonetheless, the naive approach fails to identify common regulatory effects for each type of networks so the corresponding FDR and MCC are undefined.

We also calculated the standard errors (SE) of the reported FDR, power, and MCC over 100 synthetic data sets. They are all small with most at the scale of thousandth and others at the scale of hundredth as shown in Figure 3.5 and 3.6. Therefore,

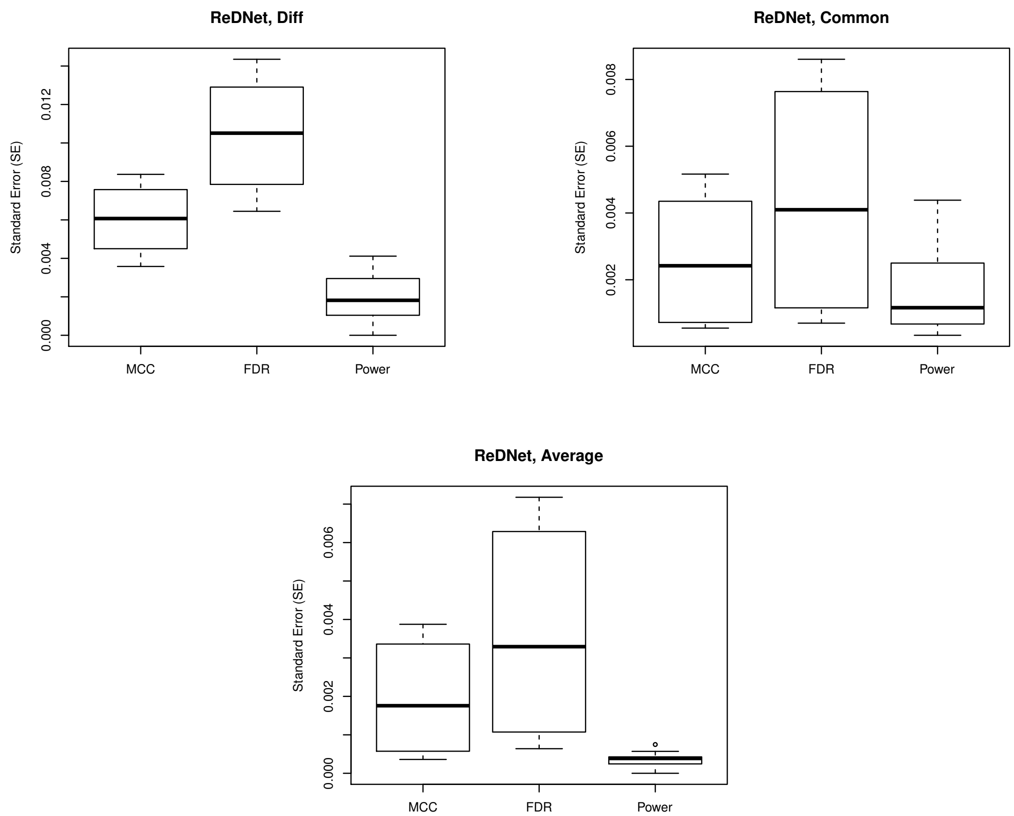


Figure 3.5. Boxplots of the Stardard Errors (SE) of the Reported FDR, Power and MCC for **ReDNet** Across Different Settings as Stated in Figure 3.2, 3.3 and 3.4.

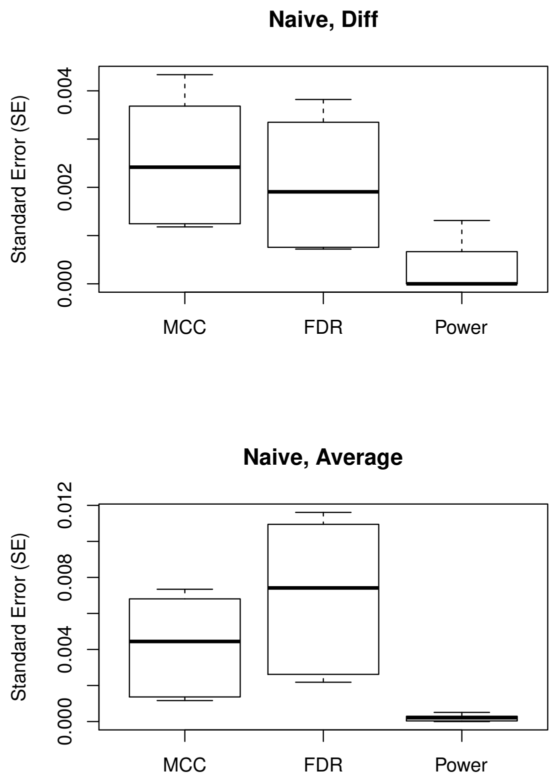


Figure 3.6. Boxplots of the Stardard Errors (SE) of the Reported FDR, Power and MCC for **Naive** Methods Across Different Settings as Stated in Figure 3.2, 3.3 and 3.4.

ReDNet performed robustly in differential analysis of networks, and the 2SPLS approach by Chen (2017) performed also robustly in constructing single networks.

3.5 The Genotype-Tissue Expression (GTEx) Data

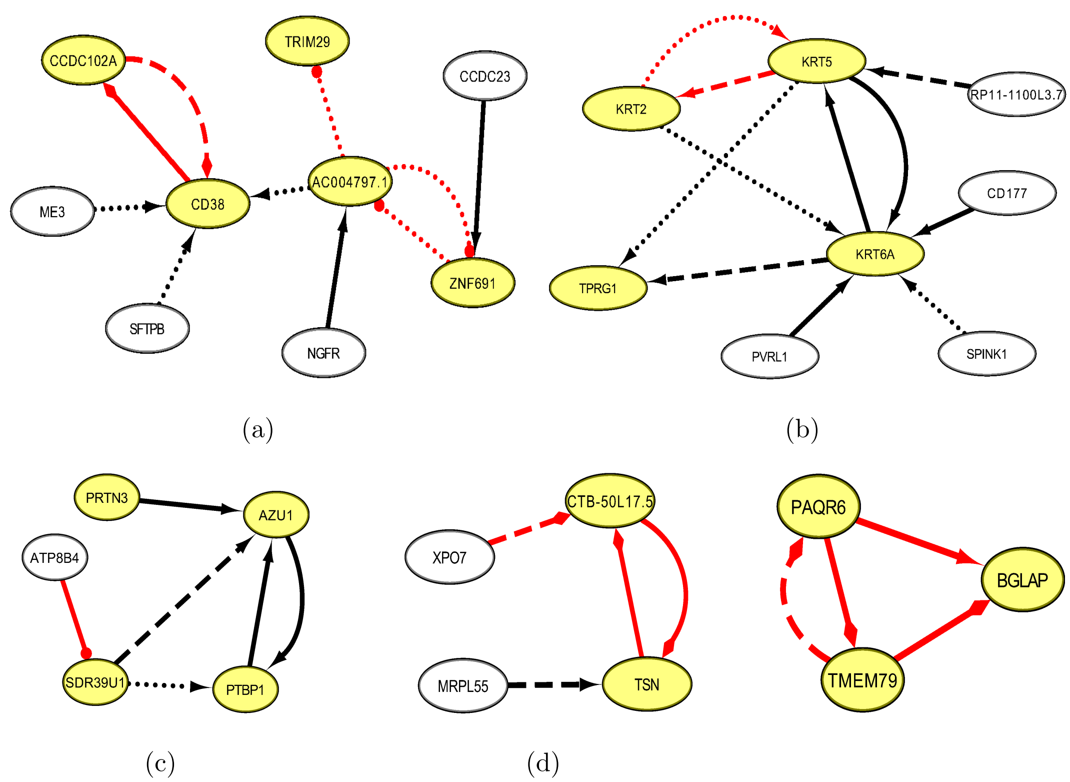


Figure 3.7. The Top Five Differential Subnetworks of Gene Regulation Identified by **ReDNet** from GTEx Data. The dotted, dashed, and solid lines imply regulations constructed in over 70%, 80%, and 90% of the bootstrap data sets, respectively. Highlighted in yellow are the target genes whose regulatory genes are focused in this study. The differential regulations are in red while common regulations are in black. The arrow head implies up regulation in both networks or no regulation in at most one network; the circle head implies down regulation in the whole blood but up regulation in muscle skeletal; and the diamond head implies up regulation in whole blood but down regulation muscle skeletal.

We performed differential analysis of gene regulatory networks on two sets of genetic genomics data from the Genotype-Tissue Expression(GTEx) project (Carithers et al., 2015), with one collected from human whole blood (WB) and another one from human muscle skeletal (MS). The WB and MS data included genome-wide genetic and genotypic values from 350 and 367 healthy subjects, respectively. The flowchart of the analysis of the GTEx data is shown in Figure 3.8. Both data sets were preprocessed following Carithers et al. (2015) and Stegle et al. (2010), resulting in a total of 15,899 genes and 1,083,917 single nucleotide polymorphisms (SNPs) being shared by WB and MS.

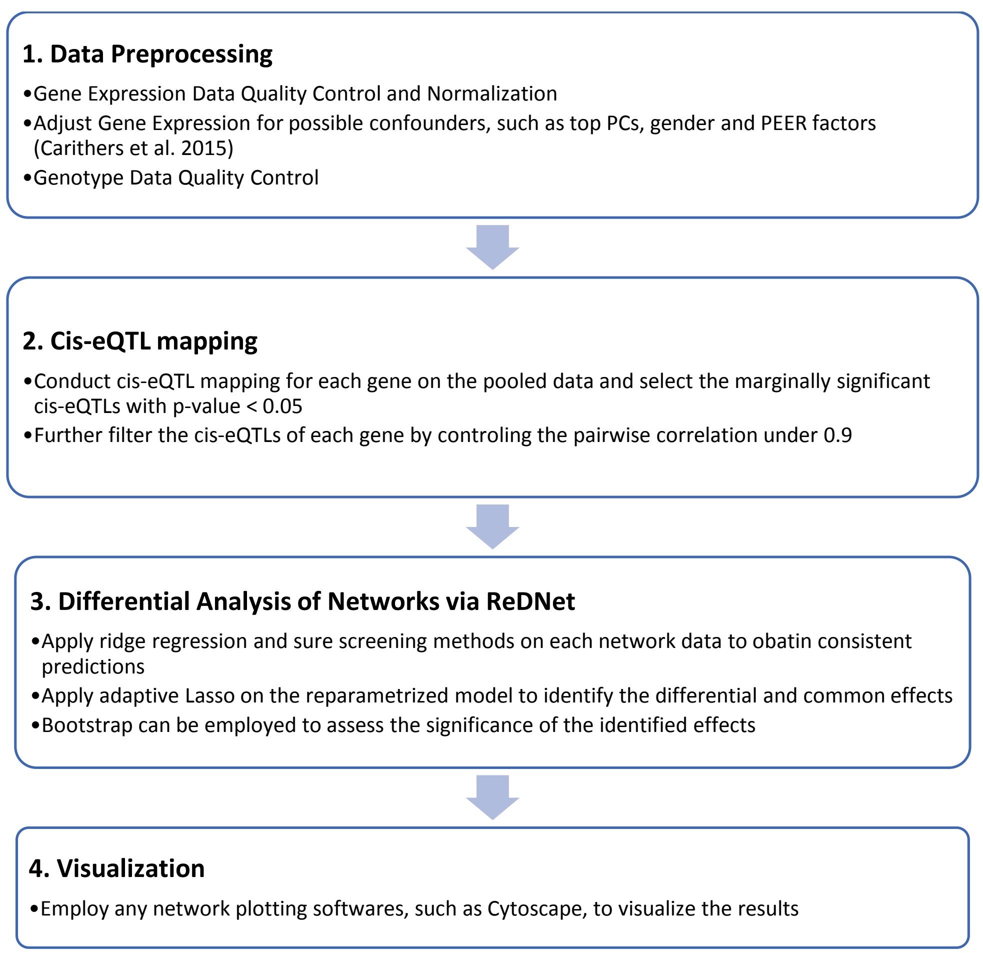


Figure 3.8. The Flowchart for The GTEx Data Analysis.

Expression quantitative trait loci (eQTL) mapping (Gilad et al., 2008) was conducted and identified 9875 genes with at least one marginally significant cis-eQTL (with p-value< 0.05). For each gene, we further filtered its set of cis-eQTL by controlling the pairwise correlation under 0.9 and keeping up to three cis-eQTL which have the strongest association with the corresponding gene expression. These cis-eQTL serve as anchoring exogenous variables for the genes, and expression levels of different genes are endogenous variables. At completion of preprocessing data, we have 9,875 endogenous variables and 23,920 exogenous variables.

We applied **ReDNet** to infer the differential gene regulation on a set of eighty target genes, which had largest changes on gene-gene correlation between the two tissues. We identified a total of 640 common and 572 differential regulations on the eighty target genes. To evaluate the significance of identified regulations, we bootstrapped 100 data sets, and conducted differential analysis on each bootstrap data set. As summarized in Table 3.1, 50, 43 and 34 differential regulatory effects were identified in over 70%, 80% and 90% of the bootstrap data sets, respectively.

Table 3.1.
Summary of Regulations Identified in Over 70%, 80%, 90% of the Bootstrap Data Sets by **ReDNet** From the GTEx Data. Shown under "Original" are for those identified from the original data.

	Original	70%	80%	90%
Common	640	49	40	34
Differential	572	50	43	34

The top five subnetworks bearing differential regulations on some of the eighty target genes were shown in Figure 3.7. We also constructed the differential networks using the naive approach and reported more differential regulations which cover the reported ones by **ReDNet** as shown in Table 3.2. This concurs with our observation in the synthetic data evaluation that the naive approach tends to report higher false positives, especially for differential regulatory effects.

Table 3.2.
Summary of Regulations Identified in Over 70%, 80%, 90% of the Bootstrap Data Sets by **Naive** Method From the GTEx Data. Shown under "Original" are for those identified from the original data.

	Original	70%	80%	90%
Differential	1516	151	129	109
Overlap with **ReDNet**	183	50	43	34

3.6 Discussion

We have developed a novel two-stage differential analysis method named **ReDNet**. The first stage, i.e., the calibration stage, aims for good prediction of the endogenous variables, and the second stage, i.e., the construction stage, identifies both common and differential network structures in a node-wise fashion. The key idea of **ReDNet** method is to appropriately re-parametrize the independent models into a joint model so as to estimate differential and common effects directly. This approach can dramatically reduce the false discovery rate. In the experiments with synthetic data, we demonstrated the effectiveness of our method, which outperformed the naive approach with a large margin. Note that **ReDNet** allows independently conducting all ℓ_2 regularized regressions at the same time at the first stage, and all ℓ_1 regularized regressions at the same time at the second stage. Therefore, **ReDNet** not only permits parallel computation but also allows for fast subnetwork construction to avoid potential huge computational demands from differential analysis of large networks.

There are some interesting directions for future research. Firstly, it is worthwhile to explore other re-parametrization approaches such as baseline reparametrizaiton in a case-control study. Secondly, while we only consider differential analysis of two networks, it is possible to generalize our method to compare multiple networks, demanding more complex reparametrization. Finally, applying the proposed methods for fully differential analysis of 53 tissues presented in the GTEx project still provides challenging computational and methodological issues.

3.7 Technical Details in Theoretical Analysis

3.7.1 Proof of Theorem 3.3.1

Proof Following the *Sure Independence Screening Property* by Fan and Lv [2008], there exists some $\theta^{(k)} \in (0, 1 - 2\kappa^{(k)} - \tau^{(k)})$ such that, when $d^{(k)} = |\mathcal{M}_i^{(k)}| = O((n^{(k)})^{1-\theta^{(k)}})$, we have, for some constant $C > 0$,

$$\mathbb{P}(\mathcal{M}_{i0}^{(k)} \subseteq \mathcal{M}_i^{(k)}) = 1 - \mathcal{O}\left(\exp\left\{-\frac{C(n^{(k)})^{1-2\kappa^{(k)}}}{\log(n^{(k)})}\right\}\right).$$

Let $\theta = \min(\theta^{(1)}, \theta^{(2)})$, then for $d^{(k)} = |\mathcal{M}_i^{(k)}| \equiv d = O(n_{\min}^{1-\theta})$, we have

$$\mathbb{P}(\mathcal{M}_{i0}^{(k)} \subseteq \mathcal{M}_i^{(k)}) = 1 - \mathcal{O}\left(\exp\left\{-\frac{C(n^{(k)})^{1-2\tilde{\kappa}}}{\log(n^{(k)})}\right\}\right).$$

■

3.7.2 Proof of Theorem 3.3.2

Note that $\boldsymbol{\xi}^{(k)} = \boldsymbol{\epsilon}^{(k)}(\mathbf{I} - \boldsymbol{\Gamma}^{(k)})^{-1}$ for $k \in \{1, 2\}$. Following Assumption 4, the singular values of both $(\mathbf{I} - \boldsymbol{\Gamma}^{(k)})$ are positively bounded from below by a constant c. Denote $\sigma_i^{(k)2} = \mathrm{var}(\epsilon_{ji}^{(k)})$ and $\tilde{\sigma}_i^{(k)2} = \mathrm{var}(\xi_{ji}^{(k)})$. Then $\tilde{\sigma}_i^{(k)} \leq \sigma_{p\max}/c = \max_{1\leq i\leq p}(\sigma_i^{(1)} \vee \sigma_i^{(2)})/\mathrm{c}$.

Lemma 3.7.1 *Under Assumptions 1-4, for each network $k \in \{1, 2\}$ in the calibration step, there exist positive constants $C_1^{(k)}$ and $C_2^{(k)}$ such that, with probability at least $1 - e^{-f^{(k)}}$,*

1. *(Estimation Loss)* $||\hat{\boldsymbol{\pi}}_i^{(k)} - \boldsymbol{\pi}_i^{(k)}||_2^2 \leq C_1^{(k)} \left(r_i^{(k)} \vee d \vee f^{(k)}\right) / n^{(k)}$;

2. (Prediction Loss) $||\mathbf{X}^{(k)}(\hat{\boldsymbol{\pi}}_i^{(k)} - \boldsymbol{\pi}_i^{(k)})||_2^2/n^{(k)} \leq C_2^{(k)}\left(r_i^{(k)} \vee d \vee f^{(k)}\right)/n^{(k)}$.

Proof We have the closed form ridge estimator $\hat{\boldsymbol{\pi}}^{(k)}_{\mathcal{M}_i^{(k)}}$ for the linear model $\mathbf{Y}_i^{(k)} = \mathbf{X}^{(k)}_{\mathcal{M}_i^{(k)}}\boldsymbol{\pi}^{(k)}_{\mathcal{M}_i^{(k)}} + \boldsymbol{\xi}_i^{(k)}$,

$$\hat{\boldsymbol{\pi}}^{(k)}_{\mathcal{M}_i^{(k)}} = \left(\mathbf{X}^{(k)T}_{\mathcal{M}_i^{(k)}}\mathbf{X}^{(k)}_{\mathcal{M}_i^{(k)}} + \lambda_i^{(k)}I_d\right)^{-1}\mathbf{X}^{(k)T}_{\mathcal{M}_i^{(k)}}\mathbf{Y}_i^{(k)},$$

where $\lambda_i^{(k)}$ is the ridge tuning parameter. Plugging in the equation $\mathbf{Y}_i^{(k)} = \mathbf{X}^{(k)}_{\mathcal{M}_i^{(k)}}\boldsymbol{\pi}^{(k)}_{\mathcal{M}_i^{(k)}} + \boldsymbol{\xi}_i^{(k)}$, we have

$$\begin{aligned}\hat{\boldsymbol{\pi}}^{(k)}_{\mathcal{M}_i^{(k)}} = &\left\{\left(\mathbf{X}^{(k)T}_{\mathcal{M}_i^{(k)}}\mathbf{X}^{(k)}_{\mathcal{M}_i^{(k)}} + \lambda_i^{(k)}I_d\right)^{-1}\mathbf{X}^{(k)T}_{\mathcal{M}_i^{(k)}}\mathbf{X}^{(k)}_{\mathcal{M}_i^{(k)}}\boldsymbol{\pi}^{(k)}_{\mathcal{M}_i^{(k)}}\right\} \\ &+ \left\{\left(\mathbf{X}^{(k)T}_{\mathcal{M}_i^{(k)}}\mathbf{X}^{(k)}_{\mathcal{M}_i^{(k)}} + \lambda_i^{(k)}I_d\right)^{-1}\mathbf{X}^{(k)T}_{\mathcal{M}_i^{(k)}}\boldsymbol{\xi}_i^{(k)}\right\}.\end{aligned}$$

The difference between the ridge estimator $\hat{\boldsymbol{\pi}}^{(k)}_{\mathcal{M}_i^{(k)}}$ and the true $\boldsymbol{\pi}^{(k)}_{\mathcal{M}_i^{(k)}}$ can be written as

$$\begin{aligned}\hat{\boldsymbol{\pi}}^{(k)}_{\mathcal{M}_i^{(k)}} - \boldsymbol{\pi}^{(k)}_{\mathcal{M}_i^{(k)}} = &-\lambda_i^{(k)}\left(\mathbf{X}^{(k)T}_{\mathcal{M}_i^{(k)}}\mathbf{X}^{(k)}_{\mathcal{M}_i^{(k)}} + \lambda_i^{(k)}I_d\right)^{-1}\boldsymbol{\pi}^{(k)}_{\mathcal{M}_i^{(k)}} \\ &+ \left(\mathbf{X}^{(k)T}_{\mathcal{M}_i^{(k)}}\mathbf{X}^{(k)}_{\mathcal{M}_i^{(k)}} + \lambda_i^{(k)}I_d\right)^{-1}\mathbf{X}^{(k)T}_{\mathcal{M}_i^{(k)}}\boldsymbol{\xi}_i^{(k)}.\end{aligned}$$

For simplicity, we denote the composite forms of $\boldsymbol{\pi}^{(k)}_{\mathcal{M}_i^{(k)}}$ and $\mathbf{X}^{(k)}_{\mathcal{M}_i^{(k)}}$ as follows,

$$\begin{aligned}\tilde{\boldsymbol{\pi}}^{(k)}_{\mathcal{M}_i^{(k)}} &= -\lambda_i^{(k)}\left(\mathbf{X}^{(k)T}_{\mathcal{M}_i^{(k)}}\mathbf{X}^{(k)}_{\mathcal{M}_i^{(k)}} + \lambda_i^{(k)}I_d\right)^{-1}\boldsymbol{\pi}^{(k)}_{\mathcal{M}_i^{(k)}}; \\ \tilde{\mathbf{X}}^{(k)}_{\mathcal{M}_i^{(k)}} &= \mathbf{X}^{(k)}_{\mathcal{M}_i^{(k)}}\left(\mathbf{X}^{(k)T}_{\mathcal{M}_i^{(k)}}\mathbf{X}^{(k)}_{\mathcal{M}_i^{(k)}} + \lambda_i^{(k)}I_d\right)^{-1}.\end{aligned}$$

Then we have the following simplified form of the difference,

$$\hat{\boldsymbol{\pi}}^{(k)}_{\mathcal{M}_i^{(k)}} - \boldsymbol{\pi}^{(k)}_{\mathcal{M}_i^{(k)}} = \tilde{\boldsymbol{\pi}}^{(k)}_{\mathcal{M}_i^{(k)}} + \tilde{\mathbf{X}}^{(k)T}_{\mathcal{M}_i^{(k)}}\boldsymbol{\xi}_i^{(k)}.$$

To obtain the ℓ_2 norm losses of estimation and prediction, we write

$$\begin{aligned}&||\hat{\boldsymbol{\pi}}^{(k)}_{\mathcal{M}_i^{(k)}} - \boldsymbol{\pi}^{(k)}_{\mathcal{M}_i^{(k)}}||_2^2 \\ &= \underbrace{\tilde{\boldsymbol{\pi}}^{(k)T}_{\mathcal{M}_i^{(k)}}\tilde{\boldsymbol{\pi}}^{(k)}_{\mathcal{M}_i^{(k)}}}_{T_{21}} + \underbrace{2\tilde{\boldsymbol{\pi}}^{(k)T}_{\mathcal{M}_i^{(k)}}\tilde{\mathbf{X}}^{(k)T}_{\mathcal{M}_i^{(k)}}\boldsymbol{\xi}_i^{(k)}}_{T_{22}} + \underbrace{\boldsymbol{\xi}_i^{(k)T}\tilde{\mathbf{X}}^{(k)}_{\mathcal{M}_i^{(k)}}\tilde{\mathbf{X}}^{(k)T}_{\mathcal{M}_i^{(k)}}\boldsymbol{\xi}_i^{(k)}}_{T_{23}},\end{aligned}$$

$$
\begin{aligned}
&||\mathbf{X}^{(k)}_{\mathcal{M}_i^{(k)}}\big(\hat{\boldsymbol{\pi}}^{(k)}_{\mathcal{M}_i^{(k)}} - \boldsymbol{\pi}^{(k)}_{\mathcal{M}_i^{(k)}}\big)||_2^2 \\
&= \underbrace{\boldsymbol{\pi}^{(k)}_{\mathcal{M}_i^{(k)}}\big(\mathbf{X}^{(k)T}_{\mathcal{M}_i^{(k)}}\mathbf{X}^{(k)}_{\mathcal{M}_i^{(k)}}\big)\tilde{\boldsymbol{\pi}}^{(k)}_{\mathcal{M}_i^{(k)}}}_{T_{24}} + \underbrace{2\tilde{\boldsymbol{\pi}}^{(k)T}_{\mathcal{M}_i^{(k)}}\big(\mathbf{X}^{(k)T}_{\mathcal{M}_i^{(k)}}\mathbf{X}^{(k)}_{\mathcal{M}_i^{(k)}}\big)\tilde{\mathbf{X}}^{(k)T}_{\mathcal{M}_i^{(k)}}\boldsymbol{\xi}_i^{(k)}}_{T_{25}} \\
&\quad + \underbrace{\boldsymbol{\xi}_i^{(k)T}\tilde{\mathbf{X}}^{(k)}_{\mathcal{M}_i^{(k)}}\big(\mathbf{X}^{(k)T}_{\mathcal{M}_i^{(k)}}\mathbf{X}^{(k)}_{\mathcal{M}_i^{(k)}}\big)\tilde{\mathbf{X}}^{(k)T}_{\mathcal{M}_i^{(k)}}\boldsymbol{\xi}_i^{(k)}}_{T_{26}}.
\end{aligned}
$$

Firstly, we will derive the bound for T_{24}, T_{25} and T_{26} terms, then we can obtain similar results for term T_{21}, T_{22} and T_{23} by simply removing the matrix $\mathbf{X}^{(k)T}_{\mathcal{M}_i^{(k)}}\mathbf{X}^{(k)}_{\mathcal{M}_i^{(k)}}$. Denote the singular value decomposition $\mathbf{X}^{(k)T}_{\mathcal{M}_i^{(k)}}\mathbf{X}^{(k)}_{\mathcal{M}_i^{(k)}} = U_i^{(k)T}V_i^{(k)}U_i^{(k)}$, where $U_i^{(k)}$ is a unitary matrix, $V_i^{(k)}$ is a diagonal matrix with eigenvalues v_i. Therefore, the shared component of $\tilde{\boldsymbol{\pi}}^{(k)}_{\mathcal{M}_i^{(k)}}$ and $\tilde{\mathbf{X}}^{(k)}_{\mathcal{M}_i^{(k)}}$ can be rewritten as

$$
\big(\mathbf{X}^{(k)T}_{\mathcal{M}_i^{(k)}}\mathbf{X}^{(k)}_{\mathcal{M}_i^{(k)}} + \lambda_i^{(k)}I_d\big)^{-1} = U_i^{(k)T}\big(V_i^{(k)} + \lambda_i^{(k)}I_d\big)^{-1}U_i^{(k)}.
$$

By Assumption 4, there are some constants c_1, c_2 such that $\max_{||\delta||_2=1}(n^{(k)})^{-1/2}||\mathbf{X}^{(k)}_{\mathcal{M}_i^{(k)}}\delta||_2 \le c_1$ and $\min_{||\delta||_2=1}(n^{(k)})^{-1/2}||\mathbf{X}^{(k)}_{\mathcal{M}_i^{(k)}}\delta||_2 \ge c_2$. Thus, $\lambda_{\max}\big(\mathbf{X}^{(k)T}_{\mathcal{M}_i^{(k)}}\mathbf{X}^{(k)}_{\mathcal{M}_i^{(k)}}\big) < c_1^2 n^{(k)}$ and $\lambda_{\min}\big(\mathbf{X}^{(k)T}_{\mathcal{M}_i^{(k)}}\mathbf{X}^{(k)}_{\mathcal{M}_i^{(k)}}\big) > c_2^2 n^{(k)}$. That is, $v_j \asymp n^{(k)}$ for each eigenvalue. Let $b = U_i^{(k)}\boldsymbol{\pi}^{(k)}_{\mathcal{M}_i^{(k)}}$, then $||b||_2 = ||\boldsymbol{\pi}^{(k)}_{\mathcal{M}_i^{(k)}}||_2$. Noting that $\lambda_i^{(k)} = o(n^{(k)})$ in Assumption 4, we can bound the term T_{24} as follows,

$$
\begin{aligned}
T_{24} &= \tilde{\boldsymbol{\pi}}^{(k)T}_{\mathcal{M}_i^{(k)}}\big(\mathbf{X}^{(k)T}_{\mathcal{M}_i^{(k)}}\mathbf{X}^{(k)}_{\mathcal{M}_i^{(k)}}\big)\tilde{\boldsymbol{\pi}}^{(k)}_{\mathcal{M}_i^{(k)}} = \lambda_i^{(k)2}b^T V_i^{(k)}\big(V_i^{(k)} + \lambda_i^{(k)}I_d\big)^{-1}V_i^{(k)}\big(V_i^{(k)} + \lambda_i^{(k)}I_d\big)^{-1}b \\
&= \lambda_i^{(k)2}\sum_{j=1}^{d}\frac{v_j b_{ij}^2}{\big(v_j + \lambda_i^{(k)}\big)^2} = \mathcal{O}(\lambda_i^{(k)2}||\boldsymbol{\pi}^{(k)}_{\mathcal{M}_i^{(k)}}||_2^2/n^{(k)}) = \mathcal{O}(r_i^{(k)}).
\end{aligned}
\tag{3.9}
$$

Similarly, removing the term $\mathbf{X}^{(k)T}_{\mathcal{M}_i^{(k)}}\mathbf{X}^{(k)}_{\mathcal{M}_i^{(k)}}$, we have

$$
T_{21} = \mathcal{O}(\lambda_i^{(k)2}||\boldsymbol{\pi}^{(k)}_{\mathcal{M}_i^{(k)}}||_2^2/n^{(k)}) = \mathcal{O}(r_i^{(k)}/n^{(k)}). \tag{3.10}
$$

Noting that T_{25} follows a Gaussian distribution, we can write the probability of deviation of T_{25} with the classical Gaussian tail inequality, for any positive number t,

$$
\mathbb{P}\left(T_{25} \le t\right) \ge 1 - \exp\left(-\frac{1}{2}t^2/\mathrm{var}(T_{25})\right).
$$

Furthermore,

$$\begin{aligned}
\mathrm{var}(T_{25}) &= 4\tilde{\sigma}_i^{(k)2}\tilde{\boldsymbol{\pi}}_{(i)}^{(k)T}\big(\mathbf{X}_{\mathcal{M}_i^{(k)}}^{(k)T}\mathbf{X}_{\mathcal{M}_i^{(k)}}^{(k)}\big)\tilde{\mathbf{X}}_{\mathcal{M}_i^{(k)}}^{(k)T}\tilde{\mathbf{X}}_{\mathcal{M}_i^{(k)}}^{(k)}\big(\mathbf{X}_{\mathcal{M}_i^{(k)}}^{(k)T}\mathbf{X}_{\mathcal{M}_i^{(k)}}^{(k)}\big)\tilde{\boldsymbol{\pi}}_{\mathcal{M}_i^{(k)}}^{(k)} \\
&= 4\tilde{\sigma}_i^{(k)2}\lambda_i^{(k)2}b^T(V+\lambda_i^{(k)}I_d)^{-1}V_i^{(k)}(V_i^{(k)}+\lambda_i^{(k)}I_d)^{-1} \\
&\quad \times V_i^{(k)}\big(V_i^{(k)}+\lambda_i^{(k)}I_d\big)^{-1}V_i^{(k)}\big(V_i^{(k)}+\lambda_i^{(k)}I_d\big)^{-1}b \\
&= 4\tilde{\sigma}_i^{(k)2}\lambda_i^{(k)2}\sum_{j=1}^{d}\frac{v_j^3 b_{ij}^2}{(v_j+\lambda_i^{(k)})^4} \\
&= \mathcal{O}(\tilde{\sigma}_i^{(k)2}\lambda_i^{(k)2}||\boldsymbol{\pi}_{\mathcal{M}_i^{(k)}}^{(k)}||_2^2/n^{(k)}) \\
&= \mathcal{O}(\tilde{\sigma}_i^{(k)2}r_i^{(k)}).
\end{aligned}$$

Letting $t = \sqrt{2\mathrm{var}(T_{25})(f^{(k)}+\log 2)}$, we obtain that, with probability at least $1 - e^{-f^{(k)}}/2$,

$$T_{25} = \mathcal{O}(\sqrt{r_i^{(k)}f^{(k)}}). \tag{3.11}$$

Similarly, removing $\mathbf{X}_{\mathcal{M}_i^{(k)}}^{(k)T}\mathbf{X}_{\mathcal{M}_i^{(k)}}^{(k)}$, we can obtain that, concurring with (3.11),

$$T_{22} = \mathcal{O}(\sqrt{r_i^{(k)}f^{(k)}}/n^{(k)}). \tag{3.12}$$

The term T_{26} follows a non-central χ^2 distribution. We can invoke the Hanson-Wright inequality (Rudelson et al., 2013) to bound the probability of its extreme deviation, for some constant $t_2 > 0$,

$$\begin{aligned}
&\mathbb{P}(T_{26} \leq \mathbb{E}(T_{26}) + t) \\
&\quad \geq \; 1 - \exp\left\{\frac{-t^2 t_2}{\tilde{\sigma}_i^{(k)4}||\tilde{\mathbf{X}}_{\mathcal{M}_i^{(k)}}^{(k)}\mathbf{X}_{\mathcal{M}_i^{(k)}}^{(k)T}\mathbf{X}_{\mathcal{M}_i^{(k)}}^{(k)}\tilde{\mathbf{X}}_{\mathcal{M}_i^{(k)}}^{(k)T}||_F^2}\right\} \\
&\qquad \wedge \exp\left\{\frac{-t t_2}{\tilde{\sigma}_i^{(k)2}||\tilde{\mathbf{X}}_{\mathcal{M}_i^{(k)}}^{(k)}\mathbf{X}_{\mathcal{M}_i^{(k)}}^{(k)T}\mathbf{X}_{\mathcal{M}_i^{(k)}}^{(k)}\tilde{\mathbf{X}}_{\mathcal{M}_i^{(k)}}^{(k)T}||_{op}}\right\}.
\end{aligned} \tag{3.13}$$

To understand this probabilistic bound, we need to calculate $\mathbb{E}(T_{26})$ and the two involved norms. Firstly,

$$\begin{aligned}
\mathbb{E}(T_{26}) &= \tilde{\sigma}_i^{(k)2}\mathrm{tr}\left(\tilde{\mathbf{X}}^{(k)}_{\mathcal{M}_i^{(k)}}\mathbf{X}^{(k)T}_{\mathcal{M}_i^{(k)}}\mathbf{X}^{(k)}_{\mathcal{M}_i^{(k)}}\tilde{\mathbf{X}}^{(k)T}_{\mathcal{M}_i^{(k)}}\right)\\
&= \tilde{\sigma}_i^{(k)2}\mathrm{tr}\left(V_i^{(k)}\big(V_i^{(k)}+\lambda_i^{(k)}I_d\big)^{-1}V_i^{(k)}\big(V_i^{(k)}+\lambda_i^{(k)}I_d\big)^{-1}\right)\\
&= \tilde{\sigma}_i^{(k)2}\sum_{j=1}^{d}\frac{v_j^2}{(v_j+\lambda_i^{(k)})^2}=\mathcal{O}(d\tilde{\sigma}_i^{(k)2}). \qquad (3.14)
\end{aligned}$$

The Frobenius norm can be simplified as follows,

$$\begin{aligned}
&||\tilde{\mathbf{X}}^{(k)}_{\mathcal{M}_i^{(k)}}\mathbf{X}^{(k)T}_{\mathcal{M}_i^{(k)}}\mathbf{X}^{(k)}_{\mathcal{M}_i^{(k)}}\tilde{\mathbf{X}}^{(k)T}_{\mathcal{M}_i^{(k)}}||_F^2\\
&= \mathrm{tr}\left(\tilde{\mathbf{X}}^{(k)}_{\mathcal{M}_i^{(k)}}\mathbf{X}^{(k)T}_{\mathcal{M}_i^{(k)}}\mathbf{X}^{(k)}_{\mathcal{M}_i^{(k)}}\tilde{\mathbf{X}}^{(k)T}_{\mathcal{M}_i^{(k)}}\tilde{\mathbf{X}}^{(k)}_{\mathcal{M}_i^{(k)}}\mathbf{X}^{(k)T}_{\mathcal{M}_i^{(k)}}\mathbf{X}^{(k)}_{\mathcal{M}_i^{(k)}}\tilde{\mathbf{X}}^{(k)T}_{\mathcal{M}_i^{(k)}}\right)\\
&= \mathrm{tr}\left(((\mathbf{X}^{(k)}_{\mathcal{M}_i^{(k)}})^T\mathbf{X}^{(k)}_{\mathcal{M}_i^{(k)}})(\tilde{\mathbf{X}}^{(k)}_{\mathcal{M}_i^{(k)}})^T\tilde{\mathbf{X}}^{(k)}_{\mathcal{M}_i^{(k)}}((\mathbf{X}^{(k)}_{\mathcal{M}_i^{(k)}})^T\mathbf{X}^{(k)}_{\mathcal{M}_i^{(k)}})(\tilde{\mathbf{X}}^{(k)}_{\mathcal{M}_i^{(k)}})^T\tilde{\mathbf{X}}^{(k)}_{\mathcal{M}_i^{(k)}}\right)\\
&= \mathrm{tr}\Big(V_i^{(k)}(V_i^{(k)}+\lambda_i^{(k)}I_d)^{-1}V_i^{(k)}(V_i^{(k)}+\lambda_i^{(k)}I_d)^{-1}V_i^{(k)}\\
&\quad\times(V_i^{(k)}+\lambda_i^{(k)}I_d)^{-1}V_i^{(k)}(V_i^{(k)}+\lambda_i^{(k)}I_d)^{-1}\Big)\\
&= \sum_{j=1}^{d}\frac{v_j^4}{(v_j+\lambda_i^{(k)})^4}=\mathcal{O}(d). \qquad (3.15)
\end{aligned}$$

Note that $\lambda_{\max}(\mathbf{X}^{(k)}_{\mathcal{M}_i^{(k)}}\mathbf{X}^{(k)T}_{\mathcal{M}_i^{(k)}}\mathbf{X}^{(k)}_{\mathcal{M}_i^{(k)}}\mathbf{X}^{(k)T}_{\mathcal{M}_i^{(k)}})\asymp n^{(k)}$, then, the operator norm can be simplified as follows,

$$\begin{aligned}
&||\tilde{\mathbf{X}}^{(k)}_{\mathcal{M}_i^{(k)}}\mathbf{X}^{(k)T}_{\mathcal{M}_i^{(k)}}\mathbf{X}^{(k)}_{\mathcal{M}_i^{(k)}}\tilde{\mathbf{X}}^{(k)T}_{\mathcal{M}_i^{(k)}}||_{op}\\
&= ||\mathbf{X}^{(k)}_{\mathcal{M}_i^{(k)}}\big(\mathbf{X}^{(k)T}_{\mathcal{M}_i^{(k)}}\mathbf{X}^{(k)}_{\mathcal{M}_i^{(k)}}+\lambda_i^{(k)}I_d\big)^{-1}\mathbf{X}^{(k)T}_{\mathcal{M}_i^{(k)}}\mathbf{X}^{(k)}_{\mathcal{M}_i^{(k)}}\big(\mathbf{X}^{(k)T}_{\mathcal{M}_i^{(k)}}\mathbf{X}^{(k)}_{\mathcal{M}_i^{(k)}}+\lambda_i^{(k)}I_d\big)^{-1}\mathbf{X}^{(k)T}_{\mathcal{M}_i^{(k)}}||_{op}\\
&= \mathcal{O}(\lambda_{\max}\big(\mathbf{X}^{(k)}_{\mathcal{M}_i^{(k)}}\mathbf{X}^{(k)T}_{\mathcal{M}_i^{(k)}}\mathbf{X}^{(k)}_{\mathcal{M}_i^{(k)}}\mathbf{X}^{(k)T}_{\mathcal{M}_i^{(k)}}\big)/n^{(k)2})=\mathcal{O}(1). \qquad (3.16)
\end{aligned}$$

Letting

$$\begin{aligned}
t=&\sqrt{\tilde{\sigma}_i^{(k)4}||\tilde{\mathbf{X}}^{(k)}_{\mathcal{M}_i^{(k)}}\mathbf{X}^{(k)T}_{\mathcal{M}_i^{(k)}}\mathbf{X}^{(k)}_{\mathcal{M}_i^{(k)}}\tilde{\mathbf{X}}^{(k)T}_{\mathcal{M}_i^{(k)}}||_F^2\times(f^{(k)}+\log 2)/t_2}\\
&\vee\left(\tilde{\sigma}_i^{(k)2}||\tilde{\mathbf{X}}^{(k)}_{\mathcal{M}_i^{(k)}}\mathbf{X}^{(k)T}_{\mathcal{M}_i^{(k)}}\mathbf{X}^{(k)}_{\mathcal{M}_i^{(k)}}\tilde{\mathbf{X}}^{(k)T}_{\mathcal{M}_i^{(k)}}||_{op}\times(f^{(k)}+\log 2)/t_2\right),
\end{aligned}$$

and combining (3.13), (3.14), (3.15), and (3.16), we obtain that, with probability at least $1-e^{-f^{(k)}}/2$,

$$T_{26}=\mathcal{O}(d\vee\sqrt{df^{(k)}}\vee f^{(k)}). \qquad (3.17)$$

Similarly, removing $\mathbf{X}^{(k)T}_{\mathcal{M}_i^{(k)}}\mathbf{X}^{(k)}_{\mathcal{M}_i^{(k)}}$, we can obtain that, concurring with (3.17),

$$T_{23} = \mathcal{O}\big(\big(d \vee \sqrt{d\, f^{(k)}} \vee f^{(k)}\big)/n^{(k)}\big). \tag{3.18}$$

Collecting the bounds (3.9), (3.11), (3.17) and noting the definition of $\mathbf{X}^{(k)}_{\mathcal{M}_i^{(k)}}$ and $\boldsymbol{\pi}^{(k)}_{\mathcal{M}_i^{(k)}}$, we conclude there exists some constant $C_2^{(k)} > 0$ such that, with probability at least $1 - e^{-f^{(k)}}$,

$$\frac{1}{n^{(k)}}||\mathbf{X}^{(k)}(\hat{\boldsymbol{\pi}}_i^{(k)} - \boldsymbol{\pi}_i^{(k)})||_2^2 = \frac{1}{n^{(k)}}||\mathbf{X}^{(k)}_{\mathcal{M}_i^{(k)}}(\hat{\boldsymbol{\pi}}^{(k)}_{\mathcal{M}_i^{(k)}} - \boldsymbol{\pi}^{(k)}_{\mathcal{M}_i^{(k)}})||_2^2 \le C_2^{(k)}\frac{r_i^{(k)} \vee d \vee f^{(k)}}{n^{(k)}}.$$

Similarly, collecting the bound (3.10), (3.12) and (3.18), we conclude there exists some constant $C_1^{(k)} > 0$ such that, with probability at least $1 - e^{-f^{(k)}}$,

$$||\hat{\boldsymbol{\pi}}_i^{(k)} - \boldsymbol{\pi}_i^{(k)}||_2^2 = ||\hat{\boldsymbol{\pi}}^{(k)}_{\mathcal{M}_i^{(k)}} - \boldsymbol{\pi}^{(k)}_{\mathcal{M}_i^{(k)}}||_2^2 \le C_1^{(k)}\frac{r_i^{(k)} \vee d \vee f^{(k)}}{n^{(k)}}.$$

This concludes the proof of Lemma 3.7.1. ∎

To bound the estimation loss, we write

$$||\hat{\mathbf{\Pi}}_j - \mathbf{\Pi}_j||_2^2 = ||\hat{\boldsymbol{\pi}}^{(1)}_{j|p} - \boldsymbol{\pi}^{(1)}_{j|p}||_2^2 + ||\hat{\boldsymbol{\pi}}^{(2)}_{j|p} - \boldsymbol{\pi}^{(2)}_{j|p}||_2^2,$$

where $\boldsymbol{\pi}^{(k)}_{j|p}$ and $\hat{\boldsymbol{\pi}}^{(k)}_{j|p}$ are the $j|p$ columns of $\boldsymbol{\pi}^{(k)}$ and $\hat{\boldsymbol{\pi}}^{(k)}$, respectively. Following the bounds in Lemma 3.7.1 for both networks, we obtain the overall estimation bound as, with probability at least $1 - e^{-f^{(1)}} - e^{-f^{(2)}}$,

$$\begin{aligned}
||\hat{\mathbf{\Pi}}_j - \mathbf{\Pi}_j||_2^2 &\le C_1^{(1)}\frac{r^{(1)}_{j|p} \vee d \vee f^{(1)}}{n^{(1)}} + C_1^{(2)}\frac{r^{(2)}_{j|p} \vee d \vee f^{(2)}}{n^{(2)}} \\
&\le \big(C_1^{(1)} + C_1^{(2)}\big)\frac{\big(r^{(2)}_{j|p} \vee d \vee f^{(2)}\big) \vee \big(r^{(2)}_{j|p} \vee d \vee f^{(2)}\big)}{n^{(1)} \wedge n^{(2)}} \\
&= C_1\frac{d \vee \big(r^{(1)}_{j|p} \vee r^{(2)}_{j|p}\big) \vee \big(f^{(1)} \vee f^{(2)}\big)}{n^{(1)} \wedge n^{(2)}} \le C_1\frac{d \vee r_{\max} \vee f_{\max}}{n^{(1)} \wedge n^{(2)}},
\end{aligned}$$

where $C_1 = C_1^{(1)} + C_1^{(2)}$. Similarly, we write the prediction bound as, with probability at least $1 - e^{-f^{(1)}} - e^{-f^{(2)}}$,

$$\begin{aligned}
||\mathbf{X}(\hat{\mathbf{\Pi}}_j - \mathbf{\Pi}_j)||_2^2 &= ||X^{(1)}(\hat{\boldsymbol{\pi}}^{(1)}_{j|p} - \boldsymbol{\pi}^{(1)}_{j|p})||_2^2 + ||X^{(2)}(\hat{\boldsymbol{\pi}}^{(2)}_{j|p} - \boldsymbol{\pi}^{(2)}_{j|p})||_2^2 \\
&\le C_2^{(1)}\left\{r^{(1)}_{j|p} \vee d \vee f^{(1)} + C_2^{(2)} r^{(2)}_{j|p} \vee d \vee f^{(2)}\right\} \\
&\le C_2\left\{d \vee \big(r^{(1)}_{j|p} \vee r^{(2)}_{j|p}\big) \vee \big(f^{(1)} \vee f^{(2)}\big)\right\} \le C_2\left\{d \vee r_{\max} \vee f_{\max}\right\},
\end{aligned}$$

where $C_2 = C_2^{(1)} + C_2^{(2)}$ and $r_{\max} = \max_{1\le i\le p}(r_i^{(1)} \vee r_i^{(2)})$. This concludes the proof of Theorem 3.3.2.

3.7.3 Proof of Theorem 3.3.3

Let $c_{\max} = c_1^{(1)} \vee c_1^{(2)}$, and further denote

$$g_n = C_2 \frac{d \vee r_{\max} \vee f_{\max}}{n} + 2c_{\max} C_2 ||\mathbf{\Pi}||_1 \sqrt{\frac{d \vee r_{\max} \vee f_{\max}}{n}}.$$

Lemma 3.7.2 *Suppose that, for node i,*

$$\sqrt{(d \vee r_{\max} \vee f_{\max})/n} + c_{\max}||\mathbf{\Pi}||_1 \le \sqrt{c_{\max}^2 ||\mathbf{\Pi}||_1^2 + \phi_0^2/(64 C_2 |\mathcal{S}_i|)}. \qquad (3.19)$$

Under Assumptions 1-5, we have $\phi_{re}(\mathbf{H}_i \mathbf{X} \hat{\mathbf{\Pi}}_{-i}, \mathcal{S}_i) \ge \boldsymbol{\phi}_0/2$ with probability at least $1 - e^{-f^{(1)} + \log p} - e^{-f^{(2)} + \log p}$.

Proof The inequality (3.19) implies that $g_n \le \phi_0^2/(64|\mathcal{S}_i|)$.

For any index set $\mathcal{S}_i$ and vector δ, note the definition of $\boldsymbol{\phi}_{\text{re}}(\cdot)$, then, we have that $||\delta||_1^2 \le (||\delta_{\mathcal{S}_i^c}||_1 + ||\delta_{\mathcal{S}_i}||_1)^2 \le (3\sqrt{|\mathcal{S}_i|}||\delta_{\mathcal{S}_i}||_2 + \sqrt{|\mathcal{S}_i|}||\delta_{\mathcal{S}_i}||_2)^2 = 16|\mathcal{S}_i|||\delta_{\mathcal{S}_i}||_2^2$. we also have

$$\begin{aligned}
&\frac{\delta^T((\mathbf{H}_i \mathbf{X}\hat{\mathbf{\Pi}}_{-i})^T(\mathbf{H}_i \mathbf{X}\hat{\mathbf{\Pi}}_{-i}) - (\mathbf{H}_i \mathbf{X}\mathbf{\Pi}_{-i})^T(\mathbf{H}_i \mathbf{X}\mathbf{\Pi}_{-i}))\delta}{n||\delta_{\mathcal{S}_i}||_2^2} \\
&\le \frac{||\delta||_1^2}{n||\delta_{\mathcal{S}_i}||_2^2} \max_{j_1,j_2} |(\mathbf{H}_i \mathbf{X}\hat{\mathbf{\Pi}}_{j_1})^T(\mathbf{H}_i \mathbf{X}\hat{\mathbf{\Pi}}_{j_2}) - (\mathbf{H}_i \mathbf{X}\mathbf{\Pi}_{j_1})^T(\mathbf{H}_i \mathbf{X}\mathbf{\Pi}_{j_2})| \\
&\le \frac{16|\mathcal{S}_i|}{n} \max_{j_1,j_2} |(\mathbf{H}_i \mathbf{X}\hat{\mathbf{\Pi}}_{j_1})^T(\mathbf{H}_i \mathbf{X}\hat{\mathbf{\Pi}}_{j_2}) - (\mathbf{H}_i \mathbf{X}\mathbf{\Pi}_{j_1})^T(\mathbf{H}_i \mathbf{X}\mathbf{\Pi}_{j_2})|. \qquad (3.20)
\end{aligned}$$

Note that,

$$\begin{aligned}
&(\mathbf{H}_i \mathbf{X}\hat{\mathbf{\Pi}}_{j_1})^T(\mathbf{H}_i \mathbf{X}\hat{\mathbf{\Pi}}_{j_2}) - (\mathbf{H}_i \mathbf{X}\mathbf{\Pi}_{j_1})^T(\mathbf{H}_i \mathbf{X}\mathbf{\Pi}_{j_2}) \\
&= \underbrace{(\hat{\mathbf{\Pi}}_{j_1} - \mathbf{\Pi}_{j_1})^T \mathbf{X}^T \mathbf{H}_i \mathbf{X}(\hat{\mathbf{\Pi}}_{j_2} - \mathbf{\Pi}_{j_2})}_{T_{31}} \\
&\quad + \underbrace{(\hat{\mathbf{\Pi}}_{j_1} - \mathbf{\Pi}_{j_1})^T \mathbf{X}^T \mathbf{H}_i \mathbf{X}\mathbf{\Pi}_{j_2}}_{T_{32}} + \underbrace{(\mathbf{X}\mathbf{\Pi}_{j_1})^T \mathbf{H}_i \mathbf{X}(\hat{\mathbf{\Pi}}_{j_2} - \mathbf{\Pi}_{j_2})}_{T_{33}}.
\end{aligned}$$

We will derive the bounds for each of these three terms separately. With $\mathbf{H}_i$ a projection matrix, we have $\lambda_{max}(\mathbf{H}_i) = 1$. We can obtain that

$$\begin{aligned} |T_{31}| &\leq ||\mathbf{H}_i\mathbf{X}(\hat{\mathbf{\Pi}}_{j_1} - \mathbf{\Pi}_{j_1})||_2 \times ||\mathbf{H}_i\mathbf{X}(\hat{\mathbf{\Pi}}_{j_2} - \mathbf{\Pi}_{j_2})||_2 \\ &\leq \lambda_{max}(\mathbf{H}_i)||\mathbf{X}(\hat{\mathbf{\Pi}}_{j_1} - \mathbf{\Pi}_{j_1})||_2 \times ||\mathbf{X}(\hat{\mathbf{\Pi}}_{j_2} - \mathbf{\Pi}_{j_2})||_2 \\ &= ||\mathbf{X}(\hat{\mathbf{\Pi}}_{j_1} - \mathbf{\Pi}_{j_1})||_2 \times ||\mathbf{X}(\hat{\mathbf{\Pi}}_{j_2} - \mathbf{\Pi}_{j_2})||_2. \end{aligned}$$

Note that $|T_{32}| \leq ||\mathbf{X}\ \mathbf{\Pi}_{j_2}||_2||\mathbf{H}_i\mathbf{X}(\hat{\mathbf{\Pi}}_{j_1} - \mathbf{\Pi}_{j_1})||_2$, and following Assumption 4, we have that

$$\begin{aligned} ||\mathbf{X}\ \mathbf{\Pi}_{j_2}||_2^2 &= ||X^{(1)}\boldsymbol{\pi}_{j|p}^{(1)}||_2^2 + ||X^{(2)}\boldsymbol{\pi}_{j|p}^{(2)}||_2^2 \\ &\leq (c_1^{(1)})^2 n^{(1)}||\boldsymbol{\pi}_{j|p}^{(1)}||_2^2 + (c_1^{(2)})^2 n^{(2)}||\boldsymbol{\pi}_{j|p}^{(2)}||_2^2 \\ &\leq c_{\max}^2 n(||\boldsymbol{\pi}_{j|p}^{(1)}||_2^2 + ||\boldsymbol{\pi}_{j|p}^{(2)}||_2^2) \\ &\leq c_{\max}^2 n\left(||\boldsymbol{\pi}_{j|p}^{(1)}||_2 + ||\boldsymbol{\pi}_{j|p}^{(2)}||_2\right)^2 \\ &\leq c_{\max}^2 n||\mathbf{\Pi}||_1^2. \end{aligned}$$

Therefore,

$$|T_{32}| \leq ||\mathbf{X}\mathbf{\Pi}_{j_2}||_2||\mathbf{H}_i\mathbf{X}(\hat{\mathbf{\Pi}}_{j_1} - \mathbf{\Pi}_{j_1})||_2 \leq c_{\max}\sqrt{n}||\mathbf{\Pi}||_1||\mathbf{X}(\hat{\mathbf{\Pi}}_{j_1} - \mathbf{\Pi}_{j_1})||_2. \quad (3.21)$$

Similarly, we can have

$$|T_{33}| \leq c_{\max}\sqrt{n}||\mathbf{\Pi}||_1||\mathbf{X}(\hat{\mathbf{\Pi}}_{j_2} - \mathbf{\Pi}_{j_2})||_2. \quad (3.22)$$

Theorem 3.3.2 leads to the following, with probability at least $1 - e^{-f^{(1)}+\log(p)} - e^{-f^{(2)}+\log(p)}$,

$$\begin{cases} \dfrac{|T_{31}|}{n} \leq C_2\dfrac{d \vee r_{\max} \vee f_{\max}}{n}, \\ \dfrac{|T_{32}|}{n} \leq c_{\max}C_2||\mathbf{\Pi}||_1\sqrt{\dfrac{d \vee r_{\max} \vee f_{\max}}{n}}, \\ \dfrac{|T_{33}|}{n} \leq c_{\max}C_2||\mathbf{\Pi}||_1\sqrt{\dfrac{d \vee r_{\max} \vee f_{\max}}{n}}. \end{cases} \quad (3.23)$$

Putting the above three inequalities together, we have,

$$\begin{aligned} &\frac{\delta^T((\mathbf{H}_i\mathbf{X}\hat{\mathbf{\Pi}}_{-i})^T(\mathbf{H}_i\mathbf{X}\hat{\mathbf{\Pi}}_{-i}) - (\mathbf{H}_i\mathbf{X}\mathbf{\Pi}_{-i})^T(\mathbf{H}_i\mathbf{X}\mathbf{\Pi}_{-i}))\delta}{n||\delta_{\mathcal{S}_i}||_2^2} \\ \leq\ & 16|\mathcal{S}_i| \times \frac{|T_{31}| + |T_{32}| + |T_{33}|}{n} = 16|\mathcal{S}_i|g_n \leq 16|\mathcal{S}_i|\frac{\phi_0^2}{64|\mathcal{S}_i|} = \phi_0^2/4. \quad (3.24) \end{aligned}$$

Together with Assumption 5, we have $\phi_{\text{re}}(\mathbf{H}_i\mathbf{X}\hat{\mathbf{\Pi}}_{-k}, \mathcal{S}_k) \geq \boldsymbol{\phi}_0/2$. This concludes the proof of Lemma 3.7.2. ■

Lemma 3.7.3 *(Basic Inequality) Let* $\boldsymbol{\eta}_i = 2n^{-1}\hat{\mathbf{Z}}_{-i}^T\mathbf{H}_i\boldsymbol{\epsilon}_i - 2n^{-1}\hat{\mathbf{Z}}_{-i}^T\mathbf{H}_i(\hat{\mathbf{Z}}_{-i} - \mathbf{Z}_{-i})\boldsymbol{\beta}_i$ *and*

$$\mathscr{E}(\lambda_i) = \left\{||W_i^{-1}\boldsymbol{\eta}_i||_\infty \leq \lambda_i/2\right\},$$

for λ_i *specified in Theorem 3.3.3. Under Assumptions 1-4, with* h_n *defined in Theorem 3.3.3, there exit a positive constant* $C_3 > 0$ *such that*

$$\mathbb{P}(\mathscr{E}(\lambda_i)) \geq 1 - e^{-C_3 h_n + \log(4q)} - e^{-f^{(1)} + \log(p)} - e^{-f^{(2)} + \log(p)}.$$

Concurring with event $\mathscr{E}(\lambda_i)$*, we have the following basic inequality,*

$$n^{-1}||\mathbf{H}_i\hat{\mathbf{Z}}_{-i}(\hat{\boldsymbol{\beta}}_i - \boldsymbol{\beta}_i)||_2^2 + \lambda_i\boldsymbol{\omega}_i^T|\hat{\boldsymbol{\beta}}_i|_1 \leq \lambda_i\boldsymbol{\omega}_i^T|\boldsymbol{\beta}_i|_1 + \boldsymbol{\eta}_i^T(\hat{\boldsymbol{\beta}}_i - \boldsymbol{\beta}_i). \tag{3.25}$$

Proof Letting

$$\boldsymbol{\xi}_{-i} = \begin{pmatrix} \boldsymbol{\xi}_{-i}^{(1)} & \boldsymbol{\xi}_{-i}^{(1)} \\ \boldsymbol{\xi}_{-i}^{(2)} & -\boldsymbol{\xi}_{-i}^{(2)} \end{pmatrix}, \tag{3.26}$$

we have $\mathbf{Z}_{-i} = \mathbf{X}\mathbf{\Pi}_{-i} + \boldsymbol{\xi}_{-i}$. With $\hat{\mathbf{Z}}_{-i} = \mathbf{X}\hat{\mathbf{\Pi}}_{-i}$, we get

$$\begin{aligned}
\boldsymbol{\eta}_i &= \frac{2}{n}\hat{\mathbf{\Pi}}_{-i}^T\mathbf{X}^T\mathbf{H}_i\boldsymbol{\epsilon}_i - \frac{2}{n}\hat{\mathbf{\Pi}}_{-i}^T\mathbf{X}^T\mathbf{H}_i(\mathbf{X}\hat{\mathbf{\Pi}}_{-i} - \mathbf{X}\mathbf{\Pi}_{-i} - \boldsymbol{\xi}_{-i})\boldsymbol{\beta}_i \\
&= \underbrace{\frac{2}{n}(\hat{\mathbf{\Pi}}_{-i} - \mathbf{\Pi}_{-i})^T\mathbf{X}^T\mathbf{H}_i\boldsymbol{\epsilon}_i}_{T_{34}} + \underbrace{\frac{2}{n}\mathbf{\Pi}_{-i}^T\mathbf{X}^T\mathbf{H}_i\boldsymbol{\epsilon}_i}_{T_{35}} \\
&\quad + \underbrace{\frac{2}{n}(\hat{\mathbf{\Pi}}_{-i} - \mathbf{\Pi}_{-i})^T\mathbf{X}^T\mathbf{H}_i\boldsymbol{\xi}_{-i}\boldsymbol{\beta}_i}_{T_{36}} + \underbrace{\frac{2}{n}\mathbf{\Pi}_{-i}^T\mathbf{X}^T\mathbf{H}_i\boldsymbol{\xi}_{-i}\boldsymbol{\beta}_i}_{T_{37}} \\
&\quad - \underbrace{\frac{2}{n}(\hat{\mathbf{\Pi}}_{-i} - \mathbf{\Pi}_{-i})^T\mathbf{X}^T\mathbf{H}_i\mathbf{X}(\hat{\mathbf{\Pi}}_{-i} - \mathbf{\Pi}_{-i})\boldsymbol{\beta}_i}_{T_{38}} \\
&\quad - \underbrace{\frac{2}{n}\mathbf{\Pi}_{-i}^T\mathbf{X}^T\mathbf{H}_i\mathbf{X}(\hat{\mathbf{\Pi}}_{-i} - \mathbf{\Pi}_{-i})\boldsymbol{\beta}_i}_{T_{39}}.
\end{aligned}$$

We aim to bound each of these six terms by $\lambda_i/12$ either probabilistically or deterministically.

Firstly, for some constant $t_\lambda > 0$, we choose the adaptive lasso tuning parameter as below,

$$\lambda_i = t_\lambda \|\boldsymbol{\omega}_i\|_{-\infty}^{-1} ||\mathbf{B}||_1 ||\boldsymbol{\Pi}||_1 \sqrt{\frac{(d \vee r_{\max} \vee f_{\max}) \log(p)}{n_{\min}}}. \tag{3.27}$$

Denoting the j-th column of $\mathbf{X}$ by $X_{\cdot j}$, we have $X_{\cdot j}^T X_{\cdot j} = n^{(k)}$ for $k \in \{1, 2\}$ due to standardization. Furthermore,

$$\operatorname{var}\left(\frac{1}{n} X_{\cdot j}^T \mathbf{H}_i \boldsymbol{\epsilon}_i\right) \leq \frac{1}{n^2} X_{\cdot j}^T \mathbf{H}_i X_{\cdot j} \sigma_{p\max}^2 \leq \frac{n^{(k)}}{n^2} \sigma_{p\max}^2 \leq \frac{1}{n} \sigma_{p\max}^2.$$

For T_{34}, via the classical Gaussian tail inequality, we have

$$\begin{aligned}
&\mathbb{P}\left(||W_i^{-1} T_{34}||_\infty \geq \frac{\lambda_i}{12}\right) \\
&\leq \ \mathbb{P}\left(||\frac{2}{n}(\hat{\boldsymbol{\Pi}}_{-i} - \boldsymbol{\Pi}_{-i})^T \mathbf{X}^T \mathbf{H}_i \boldsymbol{\epsilon}_i||_\infty \geq \frac{\lambda_i \|\boldsymbol{\omega}_i\|_{-\infty}}{12}\right) \\
&\leq \ \mathbb{P}\left(||(\hat{\boldsymbol{\Pi}}_{-i} - \boldsymbol{\Pi}_{-i})^T||_\infty ||\frac{2}{n} \mathbf{X}^T \mathbf{H}_i \boldsymbol{\epsilon}_i||_\infty \geq \frac{\lambda_i \|\boldsymbol{\omega}_i\|_{-\infty}}{12}\right) \\
&\leq \ \mathbb{P}\left(||\frac{2}{n} \mathbf{X}^T \mathbf{H}_i \boldsymbol{\epsilon}_i||_\infty \geq \frac{\lambda_i \|\boldsymbol{\omega}_i\|_{-\infty}}{12 \delta_\Pi}\right) \leq 2q \exp\left\{-\frac{n \lambda_i^2 \|\boldsymbol{\omega}_i\|_{-\infty}^2}{1152 \sigma_{p\max}^2 \delta_\Pi^2}\right\} \\
&\leq \ 2q \cdot p^{-\frac{n}{d} t_1 ||\mathbf{B}||_1^2 ||\boldsymbol{\Pi}||_1^2} \leq 2q \cdot p \cdot p^{-t_1 ||\mathbf{B}||_1^2 \frac{n}{d} ||\boldsymbol{\Pi}||_1^2},
\end{aligned} \tag{3.28}$$

where $t_1 = t_\lambda^2 / (2304 C_1 \sigma_{p\max}^2)$, and δ_Π is the maximum estimation loss of the first stage. The last inequality is obtained based on the following bound of δ_Π. Following Theorem 3.3.2, δ_Π satisfies the following inequality with probability at least $1 - e^{-f^{(1)} + \log(p)} - e^{-f^{(2)} + \log(p)}$,

$$\begin{aligned}
\delta_\Pi^2 &= \max_{1 \leq j \leq 2p} ||\hat{\boldsymbol{\Pi}}_j - \boldsymbol{\Pi}_j||_1^2 \\
&\leq \max_{1 \leq j \leq 2p} \left(2d ||\hat{\boldsymbol{\Pi}}_j - \boldsymbol{\Pi}_j||_2^2\right) \\
&\leq 2C_1 d \left\{\frac{d \vee r_{\max} \vee f_{\max}}{n_{\min}}\right\}.
\end{aligned} \tag{3.29}$$

Note that the first inequality of (3.29) holds, since $\hat{\boldsymbol{\Pi}}$ and $\boldsymbol{\Pi}$ have at most $2d$ non-zeros based on our assumptions and the screening in the calibration step.

Similarly, for the second term T_{35}, we have that, with $t_2 = \frac{(t_\lambda)^2}{1152\sigma^2_{p\max}}$,

$$\begin{aligned}
&\mathbb{P}\left(||W_i^{-1}T_{35}||_\infty \geq \frac{\lambda_i}{12}\right)\\
\leq\ & \mathbb{P}\left(||\frac{2}{n}\mathbf{\Pi}_{-i}^T\mathbf{X}^T\mathbf{H}_i\boldsymbol{\epsilon}_i||_\infty \geq \frac{\lambda_i\|\boldsymbol{\omega}_i\|_{-\infty}}{12}\right)\\
\leq\ & \mathbb{P}\left(||\mathbf{\Pi}_{-i}^T||_\infty||\frac{2}{n}\mathbf{X}^T\mathbf{H}_i\boldsymbol{\epsilon}_i||_\infty \geq \frac{\lambda_i\|\boldsymbol{\omega}_i\|_{-\infty}}{12}\right)\\
\leq\ & \mathbb{P}\left(||\frac{2}{n}\mathbf{X}^T\mathbf{H}_i\boldsymbol{\epsilon}_i||_\infty \geq \frac{\lambda_i\|\boldsymbol{\omega}_i\|_{-\infty}}{12||\mathbf{\Pi}_{-i}^T||_\infty}\right)\\
\leq\ & 2q\exp\left\{-\frac{n\lambda_i^2\|\boldsymbol{\omega}_i\|_{-\infty}^2}{1152\sigma^2_{p\max}||\mathbf{\Pi}_{-i}^T||_\infty^2}\right\}\\
=\ & 2q\cdot p^{-t_2||\mathbf{B}||_1^2(d\vee r_{\max}\vee f_{\max})n/n_{\min}}\\
\leq\ & 2q\cdot p\cdot p^{-t_2||\mathbf{B}||_1^2(d\vee r_{\max}\vee f_{\max})n/n_{\min}}. \qquad (3.30)
\end{aligned}$$

For the third term T_{36}, we write

$$\begin{aligned}
&\mathbb{P}\left(||W_i^{-1}T_{36}||_\infty \geq \frac{\lambda_i}{12}\right)\\
\leq\ & \mathbb{P}\left(||(\hat{\mathbf{\Pi}}_{-i}-\mathbf{\Pi}_{-i})^T||_\infty||\frac{2}{n}\mathbf{X}^T\mathbf{H}_i\boldsymbol{\xi}_{-i}\boldsymbol{\beta}_i||_1 \geq \frac{\lambda_i\|\boldsymbol{\omega}_i\|_{-\infty}}{12}\right)\\
\leq\ & \mathbb{P}\left(\delta_\Pi\times\max_{j_1,j_2}|\frac{2}{n}X_{\cdot j_1}^T\mathbf{H}_i\boldsymbol{\xi}_{j_2}|\times||\boldsymbol{\beta}_i||_1 \geq \frac{\lambda_i\|\boldsymbol{\omega}_i\|_{-\infty}}{12}\right)\\
\leq\ & \mathbb{P}\left(\max_{j_1,j_2}|\frac{2}{n}X_{\cdot j_1}^T\mathbf{H}_i\boldsymbol{\xi}_{j_2}| \geq \frac{\lambda_i\|\boldsymbol{\omega}_i\|_{-\infty}}{12\delta_\Pi||\boldsymbol{\beta}_i||_1}\right)\\
\leq\ & 2q\cdot 2p\exp\left\{-\frac{n\lambda_i^2\|\boldsymbol{\omega}_i\|_{-\infty}^2}{1152\tilde{\sigma}^2_{p\max}\delta_\Pi^2||\boldsymbol{\beta}_i||_1^2}\right\}\\
=\ & 4q\cdot p\cdot p^{-t_3||\mathbf{\Pi}||_1^2 n/d}, \qquad (3.31)
\end{aligned}$$

where $\tilde{\sigma}^2_{p\max} = \max_i(\tilde{\sigma}_i^{(1)}\vee\tilde{\sigma}_i^{(2)})$, $\mathrm{var}(\frac{1}{n}X_{\cdot j_1}^T\mathbf{H}_i\boldsymbol{\xi}_{j_2}) \leq \tilde{\sigma}^2_{p\max}/n$ and $t_3 = \frac{t_\lambda^2}{2304C_1\tilde{\sigma}^2_{p\max}}$. Similarly, with $t_4 = \frac{t_\lambda^2}{1152\tilde{\sigma}^2_{p\max}}$, we write T_{37} term as

$$\begin{aligned}
&\mathbb{P}\left(||W_i^{-1}T_{37}||_\infty \geq \frac{\lambda_i}{12}\right)\\
\leq\ & 2q\cdot 2p\cdot\exp\left\{-\frac{n\lambda_i^2\|\boldsymbol{\omega}_i\|_{-\infty}^2}{1152\tilde{\sigma}^2_{p\max}||\mathbf{\Pi}_{-i}^T||_\infty^2||\boldsymbol{\beta}_i||_1^2}\right\}\\
=\ & 4q\cdot p\cdot p^{-t_4(d\vee r_{\max}\vee f_{\max})n/n_{\min}}. \qquad (3.32)
\end{aligned}$$

For the deterministic term T_{38}, choosing $t_\lambda \geq 12C_2||\mathbf{\Pi}||_1^{-1}\sqrt{(d \vee r_{\max} \vee f_{\max})/(n\log(p))}$, along with *Cauchy-Schwarz Inequality*, we have

$$\begin{aligned}
||W_i^{-1}T_{38}||_\infty &\leq \frac{||\boldsymbol{\beta}_i||_1\|\boldsymbol{\omega}_i\|_{-\infty}^{-1}}{n}\max_{j_1,j_2}|(\hat{\mathbf{\Pi}}_{j_1}-\mathbf{\Pi}_{j_1})^T\mathbf{X}^T\mathbf{H}_i\mathbf{X}(\hat{\mathbf{\Pi}}_{j_2}-\mathbf{\Pi}_{j_2})| \\
&\leq \frac{||\boldsymbol{\beta}_i||_1\|\boldsymbol{\omega}_i\|_{-\infty}^{-1}}{n}\max_{j_1,j_2}\left\{||\mathbf{H}_i\mathbf{X}(\hat{\mathbf{\Pi}}_{j_1}-\mathbf{\Pi}_{j_1})||_2||\mathbf{H}_i\mathbf{X}(\hat{\mathbf{\Pi}}_{j_2}-\mathbf{\Pi}_{j_2})||_2\right\} \\
&\leq \frac{||\boldsymbol{\beta}_i||_1\|\boldsymbol{\omega}_i\|_{-\infty}^{-1}}{n}\max_{j_1,j_2}\left\{\lambda_{\max}(\mathbf{H}_i)||\mathbf{X}(\hat{\mathbf{\Pi}}_{j_1}-\mathbf{\Pi}_{i_1})||_2||\mathbf{X}(\hat{\mathbf{\Pi}}_{j_2}-\mathbf{\Pi}_{j_2})||_2\right\} \\
&\leq \frac{||\boldsymbol{\beta}_i||_1\|\boldsymbol{\omega}_i\|_{-\infty}^{-1}}{n}\max_{j_1,j_2}\left\{||\mathbf{X}(\hat{\mathbf{\Pi}}_{j_1}-\mathbf{\Pi}_{j_1})||_2||\mathbf{X}(\hat{\mathbf{\Pi}}_{j_2}-\mathbf{\Pi}_{j_2})||_2\right\} \\
&\leq ||\boldsymbol{\beta}_i||_1\|\boldsymbol{\omega}_i\|_{-\infty}^{-1}C_2\frac{d \vee r_{\max} \vee f_{\max}}{n} \\
&\leq \frac{\lambda_i}{12} \times \left(\frac{12C_2}{t_\lambda||\mathbf{\Pi}||_1}\sqrt{\frac{d \vee r_{\max} \vee f_{\max}}{n\log(p)}}\right) \leq \frac{\lambda_i}{12}.
\end{aligned}$$

Similarly, we choose $t_\lambda \geq 24\sqrt{C_2 n_{\min}/(n\log(p))}$, and take Theorem 3.3.2 to obtain

$$\begin{aligned}
||W_i^{-1}T_{39}||_\infty &\leq 2\frac{||\boldsymbol{\beta}_i||_1||\mathbf{\Pi}_{-i}^T||_\infty\|\boldsymbol{\omega}_i\|_{-\infty}^{-1}}{n}\max_{j_1,j_2}|X_{\cdot j_1}^T\mathbf{H}_i\mathbf{X}(\hat{\mathbf{\Pi}}_{j_2}-\mathbf{\Pi}_{j_2})| \\
&\leq 2\frac{||\boldsymbol{\beta}_i||_1||\mathbf{\Pi}_{-i}^T||_\infty\|\boldsymbol{\omega}_i\|_{-\infty}^{-1}}{\sqrt{n}}\max_{j_2}||\mathbf{H}_i\mathbf{X}(\hat{\mathbf{\Pi}}_{j_2}-\mathbf{\Pi}_{j_2})||_2 \\
&\leq 2\frac{||\boldsymbol{\beta}_i||_1||\mathbf{\Pi}_{-i}^T||_\infty\|\boldsymbol{\omega}_i\|_{-\infty}^{-1}}{\sqrt{n}}\max_{j_2}||\mathbf{X}(\hat{\mathbf{\Pi}}_{j_2}-\mathbf{\Pi}_{j_2})||_2 \\
&\leq \frac{\lambda_i}{12} \times \left(\frac{24}{t_\lambda}\sqrt{\frac{C_2 n_{\min}}{n\log(p)}}\right) \leq \frac{\lambda_i}{12}.
\end{aligned}$$

Note that $n \geq n_{\min}$. Putting together the probabilistic bounds (3.28), (3.29), (3.30), (3.31) and (3.32), along with union bound, there exist a constant $C_3 > 0$ such that

$$\mathbb{P}(\mathscr{E}(\lambda_i)) \geq 1 - 3e^{-C_3 h_n + \log(4pq)} - e^{-f^{(1)}+\log(p)} - e^{-f^{(2)}+\log(p)}.$$

Next we will establish the basic inequality, concurring with the event $\mathscr{E}(\lambda_i)$.

Since the estimator $\hat{\boldsymbol{\beta}}_i$ from the adaptive lasso minimizes the corresponding objective function, we have

$$\frac{1}{n}||\mathbf{H}_i\mathbf{Y}_i - \mathbf{H}_i\hat{\mathbf{Z}}_{-i}\hat{\boldsymbol{\beta}}_i||_2 + \lambda_i\boldsymbol{\omega}_i^T|\hat{\boldsymbol{\beta}}_i|_1 \leq \frac{1}{n}||\mathbf{H}_i\mathbf{Y}_i - \mathbf{H}_i\hat{\mathbf{Z}}_{-i}\boldsymbol{\beta}_i||_2 + \lambda_i\boldsymbol{\omega}_i^T|\boldsymbol{\beta}_i|_1. \quad (3.33)$$

Because $\mathbf{H}_i\mathbf{Y}_i = \mathbf{H}_i\mathbf{Z}_{-i}\boldsymbol{\beta}_i + \mathbf{H}_i\boldsymbol{\epsilon}_i$, we can rewrite

$$\begin{aligned}
&||\mathbf{H}_i\mathbf{Y}_i - \mathbf{H}_i\hat{\mathbf{Z}}_{-i}\hat{\boldsymbol{\beta}}_i||_2^2 \\
&\quad= ||\mathbf{H}_i\mathbf{Z}_{-i}\boldsymbol{\beta}_i + \mathbf{H}_i\boldsymbol{\epsilon}_i - \mathbf{H}_i\hat{\mathbf{Z}}_{-i}\hat{\boldsymbol{\beta}}_i||_2^2 \\
&\quad= ||\mathbf{H}_i\boldsymbol{\epsilon}_i||_2^2 - 2\boldsymbol{\epsilon}_i^T\mathbf{H}_i(\hat{\mathbf{Z}}_{-i}\hat{\boldsymbol{\beta}}_i - \mathbf{Z}_{-i}\boldsymbol{\beta}_i) + ||\mathbf{H}_i\hat{\mathbf{Z}}_{-i}\hat{\boldsymbol{\beta}}_i - \mathbf{H}_i\hat{\mathbf{Z}}_{-i}\boldsymbol{\beta}_i + \mathbf{H}_i\hat{\mathbf{Z}}_{-i}\boldsymbol{\beta}_i - \mathbf{H}_i\mathbf{Z}_{-i}\boldsymbol{\beta}_i||_2^2 \\
&\quad= ||\mathbf{H}_i\boldsymbol{\epsilon}_i||_2^2 - 2\boldsymbol{\epsilon}_i^T\mathbf{H}_i(\hat{\mathbf{Z}}_{-i}\hat{\boldsymbol{\beta}}_i - \mathbf{Z}_{-i}\boldsymbol{\beta}_i) \\
&\qquad+||\mathbf{H}_i\hat{\mathbf{Z}}_{-i}(\hat{\boldsymbol{\beta}}_i - \boldsymbol{\beta}_i)||_2^2 + ||\mathbf{H}_i(\hat{\mathbf{Z}}_{-i} - \mathbf{Z}_{-i})\boldsymbol{\beta}_i||_2^2 \\
&\qquad+2\boldsymbol{\beta}_i^T(\hat{\mathbf{Z}}_{-i} - \mathbf{Z}_{-i})^T\mathbf{H}_i\hat{\mathbf{Z}}_{-i}(\hat{\boldsymbol{\beta}}_i - \boldsymbol{\beta}_i).
\end{aligned} \tag{3.34}$$

Similarly we can rewrite

$$\begin{aligned}
&||\mathbf{H}_i\mathbf{Y}_i - \mathbf{H}_i\hat{\mathbf{Z}}_{-i}\boldsymbol{\beta}_i||_2^2 \\
&\quad= ||\mathbf{H}_i\mathbf{Z}_{-i}\boldsymbol{\beta}_i + \mathbf{H}_i\boldsymbol{\epsilon}_i - \mathbf{H}_i\hat{\mathbf{Z}}_{-i}\boldsymbol{\beta}_i||_2^2 \\
&\quad= ||\mathbf{H}_i\boldsymbol{\epsilon}_i||_2^2 + ||\mathbf{H}_i(\hat{\mathbf{Z}}_{-i} - \mathbf{Z}_{-i})\boldsymbol{\beta}_i||_2^2 - 2\boldsymbol{\epsilon}_i^T\mathbf{H}_i(\hat{\mathbf{Z}}_{-i} - \mathbf{Z}_{-i})\boldsymbol{\beta}_i.
\end{aligned} \tag{3.35}$$

Plugging equations (3.34) and (3.35) into (3.33), we then have

$$\begin{aligned}
&\frac{1}{n}||\mathbf{H}_i\hat{\mathbf{Z}}_{-i}(\hat{\boldsymbol{\beta}}_i - \boldsymbol{\beta}_i)||_2^2 + \lambda_i\boldsymbol{\omega}_i^T|\hat{\boldsymbol{\beta}}_i|_1 \\
&\quad\leq \lambda_i\boldsymbol{\omega}_i^T|\boldsymbol{\beta}_i|_1 + \left(\frac{2}{n}\hat{\mathbf{Z}}_{-i}^T\mathbf{H}_i\boldsymbol{\epsilon}_i - \frac{2}{n}\hat{\mathbf{Z}}_{-i}^T\mathbf{H}_i(\hat{\mathbf{Z}}_{-i} - \mathbf{Z}_{-i})\boldsymbol{\beta}_i\right)^T(\hat{\boldsymbol{\beta}}_i - \boldsymbol{\beta}_i) \\
&\quad= \lambda_i\boldsymbol{\omega}_i^T|\boldsymbol{\beta}_i|_1 + \boldsymbol{\eta}_i^T(\hat{\boldsymbol{\beta}}_i - \boldsymbol{\beta}_i).
\end{aligned}$$

Thus, the basic inequality is established. This concludes the proof of Lemma 3.7.3. ■

Conditioning on the event $\mathscr{E}(\lambda_i)$, we remove the random term $\boldsymbol{\eta}_i$ from the basic inequality as

$$\begin{aligned}
&\frac{1}{n}||\mathbf{H}_i\hat{\mathbf{Z}}_{-i}(\hat{\boldsymbol{\beta}}_i - \boldsymbol{\beta}_i)||_2^2 \\
&\quad\leq \lambda_i\boldsymbol{\omega}_i^T|\boldsymbol{\beta}_i|_1 - \lambda_i\boldsymbol{\omega}_i^T|\hat{\boldsymbol{\beta}}_i|_1 + \boldsymbol{\eta}_i^T(\hat{\boldsymbol{\beta}}_i - \boldsymbol{\beta}_i) \\
&\quad\leq \lambda_i\boldsymbol{\omega}_{\mathcal{S}_i}^T|\boldsymbol{\beta}_{\mathcal{S}_i}|_1 - \lambda_i\boldsymbol{\omega}_{\mathcal{S}_i}^T|\hat{\boldsymbol{\beta}}_{\mathcal{S}_i}|_1 - \lambda_i\boldsymbol{\omega}_{\mathcal{S}_i^c}^T|\hat{\boldsymbol{\beta}}_{\mathcal{S}_i^c}|_1 + \boldsymbol{\eta}_{\mathcal{S}_i^c}^T(\hat{\boldsymbol{\beta}}_{\mathcal{S}_i^c}) + \boldsymbol{\eta}_{\mathcal{S}_i}^T(\hat{\boldsymbol{\beta}}_{\mathcal{S}_i} - \boldsymbol{\beta}_{\mathcal{S}_i}) \\
&\quad\leq \lambda_i\boldsymbol{\omega}_{\mathcal{S}_i}^T|\hat{\boldsymbol{\beta}}_{\mathcal{S}_i} - \boldsymbol{\beta}_{\mathcal{S}_i}|_1 - \lambda_i\boldsymbol{\omega}_{\mathcal{S}_i^c}^T|\hat{\boldsymbol{\beta}}_{\mathcal{S}_i^c}|_1 + \frac{\lambda_i}{2}\boldsymbol{\omega}_{\mathcal{S}_i^c}^T|\hat{\boldsymbol{\beta}}_{\mathcal{S}_i^c}|_1 + \frac{\lambda_i}{2}\boldsymbol{\omega}_{\mathcal{S}_i}^T|\hat{\boldsymbol{\beta}}_{\mathcal{S}_i} - \boldsymbol{\beta}_{\mathcal{S}_i}|_1 \\
&\quad\leq \frac{3}{2}\lambda_i\boldsymbol{\omega}_{\mathcal{S}_i}^T|\hat{\boldsymbol{\beta}}_{\mathcal{S}_i} - \boldsymbol{\beta}_{\mathcal{S}_i}|_1 - \frac{1}{2}\lambda_i\boldsymbol{\omega}_{\mathcal{S}_i^c}^T|\hat{\boldsymbol{\beta}}_{\mathcal{S}_i^c}|_1 \\
&\quad\leq \frac{3}{2}\lambda_i\|\boldsymbol{\omega}_{\mathcal{S}_i}\|_\infty||\hat{\boldsymbol{\beta}}_{\mathcal{S}_i} - \boldsymbol{\beta}_{\mathcal{S}_i}||_1 - \frac{1}{2}\lambda_i\|\boldsymbol{\omega}_{\mathcal{S}_i^c}\|_{-\infty}||\hat{\boldsymbol{\beta}}_{\mathcal{S}_i^c}||_1.
\end{aligned} \tag{3.36}$$

The fact that $||\mathbf{H}_i\hat{\mathbf{Z}}_{-i}(\hat{\boldsymbol{\beta}}_i - \boldsymbol{\beta}_i)||_2^2$ is always positive leads to

$$\|\boldsymbol{\omega}_{\mathcal{S}_i^c}\|_{-\infty}||\hat{\boldsymbol{\beta}}_{\mathcal{S}_i^c}||_1 \leq 3\|\boldsymbol{\omega}_{\mathcal{S}_i}\|_\infty||\hat{\boldsymbol{\beta}}_{\mathcal{S}_i} - \boldsymbol{\beta}_{\mathcal{S}_i}||_1, \tag{3.37}$$

which, following Assumption 5, further implies that

$$||\hat{\boldsymbol{\beta}}_{\mathcal{S}_i^c} - \boldsymbol{\beta}_{\mathcal{S}_i^c}||_1 \leq 3||\hat{\boldsymbol{\beta}}_{\mathcal{S}_i} - \boldsymbol{\beta}_{\mathcal{S}_i}||_1. \tag{3.38}$$

The above inequality, as well as the last inequality in (3.36), implies that

$$\begin{aligned}
&\frac{1}{n}||\mathbf{H}_i\hat{\mathbf{Z}}_{-i}(\hat{\boldsymbol{\beta}}_i - \boldsymbol{\beta}_i)||_2^2 \\
&\leq \frac{3}{2}\lambda_i\|\boldsymbol{\omega}_{\mathcal{S}_i}\|_\infty||\hat{\boldsymbol{\beta}}_{\mathcal{S}_i} - \boldsymbol{\beta}_{\mathcal{S}_i}||_1 \leq \frac{3}{2}\lambda_i\|\boldsymbol{\omega}_{\mathcal{S}_i}\|_\infty\sqrt{|\mathcal{S}_i|}||\hat{\boldsymbol{\beta}}_{\mathcal{S}_i} - \boldsymbol{\beta}_{\mathcal{S}_i}||_2 \\
&\leq \frac{3}{2}\lambda_i\|\boldsymbol{\omega}_{\mathcal{S}_i}\|_\infty\sqrt{|\mathcal{S}_i|}\frac{2||\mathbf{H}_i\hat{\mathbf{Z}}_{-i}(\hat{\boldsymbol{\beta}}_i - \boldsymbol{\beta}_i)||_2}{\sqrt{n}\boldsymbol{\phi}_0},
\end{aligned} \tag{3.39}$$

where the last inequality follows Assumption 5 and Lemma 3.7.2. The above inequality leads to that,

$$\frac{1}{n}||\mathbf{H}_i\hat{\mathbf{Z}}_{-i}(\hat{\boldsymbol{\beta}}_i - \boldsymbol{\beta}_i)||_2^2 \leq \frac{9(\|\boldsymbol{\omega}_{\mathcal{S}_i}\|_\infty)^2}{\boldsymbol{\phi}_0^2}|\mathcal{S}_i|\lambda_i^2.$$

Plugging in (3.27), and letting $C_4 = 3t_\lambda$, we obtain that

$$\frac{1}{n}||\mathbf{H}_i\hat{\mathbf{Z}}_{-i}(\hat{\boldsymbol{\beta}}_i - \boldsymbol{\beta}_i)||_2^2 \leq \frac{C_4^2\|\boldsymbol{\omega}_{\mathcal{S}_i}\|_\infty^2||\mathbf{B}||_1^2||\boldsymbol{\Pi}||_1^2}{\boldsymbol{\phi}_0^2\|\boldsymbol{\omega}_i\|_{-\infty}^2}|\mathcal{S}_i|\frac{(d \vee r_{\max} \vee f_{\max})\log(p)}{n_{\min}}. \tag{3.40}$$

Taking this inequality, we can follow Assumption 5 and Lemma 3.7.2 to derive that

$$\begin{aligned}
&||\hat{\boldsymbol{\beta}}_i - \boldsymbol{\beta}_i||_1 \\
&\leq ||\hat{\boldsymbol{\beta}}_{\mathcal{S}_i^c}||_1 + ||\hat{\boldsymbol{\beta}}_{\mathcal{S}_i} - \boldsymbol{\beta}_{\mathcal{S}_i}||_1 \leq \left(3\frac{\|\boldsymbol{\omega}_{\mathcal{S}_i}\|_\infty}{\|\boldsymbol{\omega}_{\mathcal{S}_i^c}\|_{-\infty}} + 1\right)||\hat{\boldsymbol{\beta}}_{\mathcal{S}_i} - \boldsymbol{\beta}_{\mathcal{S}_i}||_1 \\
&\leq \left(3\frac{\|\boldsymbol{\omega}_{\mathcal{S}_i}\|_\infty}{\|\boldsymbol{\omega}_{\mathcal{S}_i^c}\|_{-\infty}} + 1\right)\sqrt{|\mathcal{S}_i|}\frac{2||\mathbf{H}_i\hat{\mathbf{Z}}_{-i}(\hat{\boldsymbol{\beta}}_i - \boldsymbol{\beta}_i)||_2}{\sqrt{n}\boldsymbol{\phi}_0} \\
&\leq \left(3\frac{\|\boldsymbol{\omega}_{\mathcal{S}_i}\|_\infty}{\|\boldsymbol{\omega}_{\mathcal{S}_i^c}\|_{-\infty}} + 1\right)\sqrt{|\mathcal{S}_i|}\frac{2C_4\|\boldsymbol{\omega}_{\mathcal{S}_i}\|_\infty||\mathbf{B}||_1||\boldsymbol{\Pi}||_1}{\boldsymbol{\phi}_0^2\|\boldsymbol{\omega}_i\|_{-\infty}}\sqrt{|\mathcal{S}_i|}\sqrt{\frac{(d \vee r_{\max} \vee f_{\max})\log(p)}{n_{\min}}} \\
&\leq 8C_4\frac{\|\boldsymbol{\omega}_{\mathcal{S}_i}\|_\infty||\mathbf{B}||_1||\boldsymbol{\Pi}||_1}{\boldsymbol{\phi}_0^2\|\boldsymbol{\omega}_i\|_{-\infty}}|\mathcal{S}_i|\sqrt{\frac{(d \vee r_{\max} \vee f_{\max})\log(p)}{n_{\min}}},
\end{aligned} \tag{3.41}$$

where the last inequality follows Assumption 5. Since the inequality (3.36) concurs with the event $\mathscr{E}(\lambda_i)$, the above prediction and estimation bounds hold with probability at least $1 - 3e^{-C_3h_n+\log(4pq)} - e^{-f^{(1)}+\log(p)} - e^{-f^{(2)}+\log(p)}$. This completes the proof of Theorem 3.3.3.

3.7.4 Proof of Theorem 3.3.4

Lemma 3.7.4 *Suppose that, for node i,*

$$\begin{aligned}&\sqrt{(d \vee r_{\max} \vee f_{\max})/n} + c_{\max}||\mathbf{\Pi}||_1 \\ &\leq \sqrt{c_{\max}^2||\mathbf{\Pi}||_1^2 + \min(\phi_0^2/64, \tau(4-\tau)^{-1}\|\boldsymbol{\omega}_i\|_{-\infty}/\psi_i)/(C_2|\mathcal{S}_i|)}. \quad (3.42)\end{aligned}$$

Under Assumptions 1-6, we have $||W_{\mathcal{S}_i^c}^{-1}(\hat{\mathcal{I}}_{i,21}\hat{\mathcal{I}}_{i,11}^{-1})W_{\mathcal{S}_i}||_\infty \leq 1-\tau/2$ *with the probability at least* $1 - e^{-f^{(1)}+\log(p)} - e^{-f^{(2)}+\log(p)}$.

Proof The inequality (3.42) implies that

$$\psi_i\|\boldsymbol{\omega}_i\|_{-\infty}^{-1}|\mathcal{S}_i|g_n \leq \frac{\tau}{4-\tau}.$$

By the inequalities (3.23) and (3.24) in the proof of Lemma 3.7.2 and union bound, we have that, with probability at least $1 - e^{-f^{(1)}+\log(p)} - e^{-f^{(2)}+\log(p)}$,

$$\max_{j_1,j_2}\left\{\frac{1}{n}|(\mathbf{H}_i\mathbf{X}\hat{\mathbf{\Pi}}_{j_1})^T(\mathbf{H}_i\mathbf{X}\hat{\mathbf{\Pi}}_{j_2}) - (\mathbf{H}_i\mathbf{X}\mathbf{\Pi}_{j_1})^T(\mathbf{H}_i\mathbf{X}\mathbf{\Pi}_{j_2})|\right\} \leq g_n.$$

With the definitions of infinity norm $||\cdot||_\infty$, $\hat{\mathcal{I}}_{i,11}$, and $\mathcal{I}_{i,11}$, we can obtain the following inequality indexed by set $\mathcal{S}_i$,

$$\begin{aligned}&\psi_i||W_{\mathcal{S}_i}^{-1}(\hat{\mathcal{I}}_{i,11} - \mathcal{I}_{i,11})||_\infty \\ &\leq \psi_i\|\boldsymbol{\omega}_{\mathcal{S}_i}\|_{-\infty}^{-1}||\hat{\mathcal{I}}_{i,11} - \mathcal{I}_{i,11}||_\infty \\ &\leq \psi_i\|\boldsymbol{\omega}_{\mathcal{S}_i}\|_{-\infty}^{-1}|\mathcal{S}_i|g_n \leq \frac{\tau}{4-\tau}. \quad (3.43)\end{aligned}$$

Similarly we can obtain the following bound indexed by the complement set $\mathcal{S}_i^c$,

$$\psi_i||W_{\mathcal{S}_i^c}^{-1}(\hat{\mathcal{I}}_{i,21} - \mathcal{I}_{i,21})||_\infty \leq \psi_i\|\boldsymbol{\omega}_{\mathcal{S}_i^c}\|_{-\infty}^{-1}|\mathcal{S}_i|g_n \leq \frac{\tau}{4-\tau}. \quad (3.44)$$

Applying the matrix inversion error bound in Horn and Johnson (2012) and the triangular inequality, we have that

$$\begin{aligned}||\hat{\mathcal{I}}_{i,11}^{-1}W_{\mathcal{S}_i}||_\infty &\leq ||\mathcal{I}_{i,11}^{-1}W_{\mathcal{S}_i}||_\infty + ||\hat{\mathcal{I}}_{i,11}^{-1}W_{\mathcal{S}_i} - \mathcal{I}_{i,11}^{-1}W_{\mathcal{S}_i}||_\infty \\ &\leq \psi_i + \frac{\psi_i||W_{\mathcal{S}_i}^{-1}(\hat{\mathcal{I}}_{i,11} - \mathcal{I}_{i,11})||_\infty}{1 - \psi_i||W_{\mathcal{S}_i}^{-1}(\hat{\mathcal{I}}_{i,11} - \mathcal{I}_{i,11})||_\infty}\psi_i \\ &\leq \psi_i + \frac{\tau}{4-2\tau}\psi_i = \frac{4-\tau}{4-2\tau}\psi_i. \quad (3.45)\end{aligned}$$

Also note that we can rewrite

$$\begin{aligned}
& W_{\mathcal{S}_i^c}^{-1}\left(\hat{\mathcal{I}}_{i,21}\hat{\mathcal{I}}_{i,11}^{-1}-\mathcal{I}_{i,21}\mathcal{I}_{i,11}^{-1}\right)W_{\mathcal{S}_i} \\
= \ & W_{\mathcal{S}_i^c}^{-1}\left(\hat{\mathcal{I}}_{i,21}-\mathcal{I}_{i,21}\right)\hat{\mathcal{I}}_{i,11}^{-1}W_{\mathcal{S}_i} \\
& +W_{\mathcal{S}_i^c}^{-1}\mathcal{I}_{i,21}\mathcal{I}_{i,11}^{-1}W_{\mathcal{S}_i}W_{\mathcal{S}_i}^{-1}\left(\hat{\mathcal{I}}_{i,11}-\mathcal{I}_{i,11}\right)\hat{\mathcal{I}}_{i,11}^{-1}W_{\mathcal{S}_i}.
\end{aligned}$$

Then, it follows from (3.43), (3.44), (3.45) and Assumption 6 that

$$\begin{aligned}
& ||W_{\mathcal{S}_i^c}^{-1}\left(\hat{\mathcal{I}}_{i,21}\hat{\mathcal{I}}_{i,11}^{-1}-\mathcal{I}_{i,21}\mathcal{I}_{i,11}^{-1}\right)W_{\mathcal{S}_i}||_\infty \\
\leq \ & ||W_{\mathcal{S}_i^c}^{-1}\left(\hat{\mathcal{I}}_{i,21}-\mathcal{I}_{i,21}\right)||_\infty||\hat{\mathcal{I}}_{i,11}^{-1}W_{\mathcal{S}_i}||_\infty \\
& +||W_{\mathcal{S}_i^c}^{-1}\mathcal{I}_{i,21}\mathcal{I}_{i,11}^{-1}W_{\mathcal{S}_i}||_\infty||W_{\mathcal{S}_i}^{-1}\left(\hat{\mathcal{I}}_{i,11}-\mathcal{I}_{i,11}\right)||_\infty||\hat{\mathcal{I}}_{i,11}^{-1}W_{\mathcal{S}_i}||_\infty \leq \tau/2.
\end{aligned}$$

Therefore, together with Assumption 6 again, we can conclude that $||W_{\mathcal{S}_i^c}^{-1}(\hat{\mathcal{I}}_{i,21}\hat{\mathcal{I}}_{i,11}^{-1})W_{\mathcal{S}_i}||_\infty \leq 1-\tau/2$. This concludes the proof of Lemma 3.7.4. ■

The optimality of $\hat{\boldsymbol{\beta}}_i$ in the adaptive lasso step and KKT condition lead to

$$-\frac{2}{n}(\mathbf{H}_i\hat{\mathbf{Z}}_{-i})^T(\mathbf{H}_i\mathbf{Y}_i-\mathbf{H}_i\hat{\mathbf{Z}}_{-i}\hat{\boldsymbol{\beta}}_i)+\lambda_iW_i\alpha_i=0, \tag{3.46}$$

where $\alpha_i\in\mathbb{R}^{2p-2}$, satisfying that $||\alpha_i||_\infty\leq 1$ and $\alpha_{ij}I(\hat{\boldsymbol{\beta}}_{ij}\neq 0)=sign(\hat{\boldsymbol{\beta}}_{ij})$.

Plug in the equation $\mathbf{H}_i\mathbf{Y}_i=\mathbf{H}_i\mathbf{Z}_{-i}\boldsymbol{\beta}_i+\mathbf{H}_i\boldsymbol{\epsilon}_i$, we can have that

$$\begin{aligned}
& \mathbf{H}_i\mathbf{Y}_i-\mathbf{H}_i\hat{\mathbf{Z}}_{-i}\hat{\boldsymbol{\beta}}_i \\
= \ & \mathbf{H}\mathbf{Z}_{-i}\boldsymbol{\beta}_i+\mathbf{H}_i\boldsymbol{\epsilon}_i-\mathbf{H}_i\hat{\mathbf{Z}}_{-i}\hat{\boldsymbol{\beta}}_i \\
= \ & \mathbf{H}_i\boldsymbol{\epsilon}_i+\mathbf{H}_i\mathbf{Z}_{-i}\boldsymbol{\beta}_i-\mathbf{H}_i\hat{\mathbf{Z}}_{-i}\boldsymbol{\beta}_i+\mathbf{H}_i\hat{\mathbf{Z}}_{-i}\boldsymbol{\beta}_i-\mathbf{H}_i\hat{\mathbf{Z}}_{-i}\hat{\boldsymbol{\beta}}_i \\
= \ & \mathbf{H}_i\boldsymbol{\epsilon}_i-\mathbf{H}_i(\hat{\mathbf{Z}}_{-i}-\mathbf{Z}_{-i})\boldsymbol{\beta}_i-\mathbf{H}_i\hat{\mathbf{Z}}_{-i}(\hat{\boldsymbol{\beta}}_i-\boldsymbol{\beta}_i).
\end{aligned} \tag{3.47}$$

This, along with KKT condition (3.46), leads to

$$2\hat{\mathcal{I}}_i(\hat{\boldsymbol{\beta}}_i-\boldsymbol{\beta}_i)-\boldsymbol{\eta}_i=-\lambda_iW_i\alpha_i, \tag{3.48}$$

where $\boldsymbol{\eta}_i$ is defined in Lemma 3.7.3.

Letting $\hat{\boldsymbol{\beta}}_{\mathcal{S}_i^c}=\boldsymbol{\beta}_{\mathcal{S}_i^c}=0$, equation (3.48) can be decomposed as

$$\begin{cases}
2\hat{\mathcal{I}}_{i,11}(\hat{\boldsymbol{\beta}}_{\mathcal{S}_i}-\boldsymbol{\beta}_{\mathcal{S}_i})-\boldsymbol{\eta}_{\mathcal{S}_i}=-\lambda_iW_{\mathcal{S}_i}\alpha_{\mathcal{S}_i}, \\
2\hat{\mathcal{I}}_{i,21}(\hat{\boldsymbol{\beta}}_{\mathcal{S}_i}-\boldsymbol{\beta}_{\mathcal{S}_i})-\boldsymbol{\eta}_{\mathcal{S}_i^c}=-\lambda_iW_{\mathcal{S}_i^c}\alpha_{\mathcal{S}_i^c}.
\end{cases} \tag{3.49}$$

We can solve for $\hat{\boldsymbol{\beta}}_{\mathcal{S}_i}$ from the first equation of (3.49) as

$$\begin{aligned}
&\hat{\boldsymbol{\beta}}_{\mathcal{S}_i} - \boldsymbol{\beta}_{\mathcal{S}_i} \\
&= 2^{-1}\hat{\mathcal{I}}_{i,11}^{-1}(\boldsymbol{\eta}_{\mathcal{S}_i} - \lambda_i W_{\mathcal{S}_i}^T \alpha_{\mathcal{S}_i}) \\
&= 2^{-1}\hat{\mathcal{I}}_{i,11}^{-1} W_{\mathcal{S}_i}(W_{\mathcal{S}_i}^{-1}\boldsymbol{\eta}_{\mathcal{S}_i} - \lambda_i \alpha_{\mathcal{S}_i}). \qquad (3.50)
\end{aligned}$$

Following the similar strategy in the proof of Lemma 3.7.3, we can prove that there exists a constant $C_5 > 0$ such that $||W_i^{-1}\boldsymbol{\eta}_i||_\infty \leq \frac{\tau}{4-\tau}\lambda_i$ with probability at least $1 - 3e^{-C_5\, h_n + \log(4q) + \log(p)} - e^{-f^{(1)} + \log(p)} - e^{-f^{(2)} + \log(p)}$. Thus, together with $||\alpha_{\mathcal{S}_i}||_\infty \leq 1$, we obtain the infinity norm estimation loss on the true support set $\mathcal{S}_i$

$$\begin{aligned}
||\hat{\boldsymbol{\beta}}_{\mathcal{S}_i} - \boldsymbol{\beta}_{\mathcal{S}_i}||_\infty &\leq 2^{-1}||\hat{\mathcal{I}}_{i,11}^{-1} W_{\mathcal{S}_i}||_\infty(||W_{\mathcal{S}_i}^{-1}\boldsymbol{\eta}_{\mathcal{S}_i}||_\infty + \lambda_i) \\
&\leq 2^{-1}\frac{4-\tau}{4-2\tau}\psi_i \frac{4}{4-\tau}\lambda_i = \frac{\lambda_i \psi_i}{2-\tau} \leq \min_{j \in \mathcal{S}_i}|\boldsymbol{\beta}_{ij}| = b_i,
\end{aligned}$$

where the last inequality comes from the condition on the minimal signal strength b_i. The above inequality implies $sign(\hat{\boldsymbol{\beta}}_{\mathcal{S}_i}) = sign(\boldsymbol{\beta}_{\mathcal{S}_i})$.

Plugging (3.50) into the left hand side of the second equation in (3.49), we can verify that

$$\begin{aligned}
&||W_{\mathcal{S}_i^c}^{-1}\hat{\mathcal{I}}_{i,21}(\hat{\mathcal{I}}_{i,11})^{-1}(\boldsymbol{\eta}_{\mathcal{S}_i} - \lambda_i W_{\mathcal{S}_i}\alpha_{\mathcal{S}_i}) - W_{\mathcal{S}_i^c}^{-1}\boldsymbol{\eta}_{\mathcal{S}_i^c}||_\infty \\
&\leq ||W_{\mathcal{S}_i^c}^{-1}\hat{\mathcal{I}}_{i,21}\hat{\mathcal{I}}_{i,11}^{-1}W_{\mathcal{S}_i}||_\infty(||W_{\mathcal{S}_i}^{-1}\boldsymbol{\eta}_{\mathcal{S}_i}||_\infty + \lambda_i) + ||W_{\mathcal{S}_i^c}^{-1}\boldsymbol{\eta}_{\mathcal{S}_i^c}||_\infty \\
&\leq (1-\tau/2)(4/(4-\tau))\lambda_i + \tau/(4-\tau)\lambda_i = \lambda_i.
\end{aligned}$$

Therefore, we have constructed a solution $\hat{\boldsymbol{\beta}}_i$ which satisfies the KKT condition (3.48) and $sign(\hat{\boldsymbol{\beta}}_i) = sign(\boldsymbol{\beta}_i)$, that is, $\hat{\mathcal{S}}_i = \mathcal{S}_i$. This completes the proof of Theorem 3.3.4.

4. SUMMARY

In the current big data era, large scale genetical genomics data provide promising opportunities for understanding complex biological systems. However, many traditional analysis methods suffer from inefficiency or even failure in the big data setting. Thus, it is important to develop new powerful and computational efficient statistical methods. Motivated by this practical needs, in this dissertation, we presented two recent works for efficiently modeling large scale systems or networks from different perspectives.

In the first part, we introduce and review the Two-Stage Penalized Least Square (2SPLS) Method for inferring casual networks from large scale data using structural equation models. We analyzed its theoretical properties for the diverging dimension case. We showed that if the dimensions grow with the sample size up to some polynomial order, i.e., $O(n^c)$ for some $0 < c < 1$. The estimation error bounds can be well controlled and the set of true signals can be recovered as well. In particular, our results mainly depend on the restricted eigenvalue condition and a variant of irrepresentible condition, which are widely employed in current literature.

It will be interesting to further extend the 2SPLS method to partially linear or even non-linear structural equation model. This direction will make the model more general for the real data. Notwithstanding, the extensions may require careful specification and identifiability assumption of the model and the resulting estimation procedure may induce higher computational burden.

In the second part, we propose the Reparametrization-Based Differential Analysis of Directed Network (**ReDNet**) method to directly detect the sparse differences between two cognate networks from related populations. Both of the networks are characterized via structural equation models and the model estimation is designed in two stages fashion similar to the 2SPLS method. In the first stage, we incorporate additional sure independence screening step to fast screen for important instrument variables to obtain consistent predictions for the second stage. In the second stage, in order to take advantage of the commonality between two networks, we reparametrize the two structural equations to directly estimate the differential and common effects between two networks. We show that the newly proposed method can achieve much better performance, especially for the detection of differential effects, than that of estimating the networks independently. We also analyzed the theoretical properties of **ReDNet** for diverging dimension case comprehensively. Our main theorems indicate that the proposed method allows the dimensions to grow with the minimum sample size up to some exponential order, i.e., $O(e^{n_{\min}^c})$ for some $0 < c < 1$. A real data was also analyzed to demonstrate the applicability of our method in practice.

The **ReDNet** is designed for detecting structural differences between two networks. If the data from multiple related populations are available, **ReDNet** can be naturally extended to jointly modeling multiple networks. A possible approach for this extension may be further reparametrizing the structural equations by employing "contrast" idea in ANONA method.

In conclusion, we hope our study and the novel methods in this dissertation can assist us to understand and model large scale systems or networks represented by structural equation models. Though our work was motivated by modeling gene regulatory network from genetical genomics data, We believe our models and proposed methods can also be employed in other fields, such as the modeling of social networks and stock interaction networks.

REFERENCES

LIST OF REFERENCES

Adewuyi, A. O. and Awodumi, O. B. (2017). Biomass energy consumption, economic growth and carbon emissions: Fresh evidence from west africa using a simultaneous equation model. *Energy*, 119: 453–471.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6): 716–723.

Amemiya, T. (1977). The maximum likelihood and the nonlinear three-stage least squares estimator in the general nonlinear simultaneous equation model. *Econometrica*, 45(4): 955–968.

Anderson, T. (2005). Origins of the limited information maximum likelihood and two-stage least squares estimators. *Journal of Econometrics*, 127(1): 1–16.

Basmann, R. L. (1957). A generalized classical method of linear estimation of coefficients in a structural equation. *Econometrica*, 25(1): 77–83.

Bentler, P. M. and Weeks, D. G. (1980). Linear structural equations with latent variables. *Psychometrika*, 45(3): 289–308.

Berkman, L. F. and Syme, S. L. (1979). Social networks, host resistance, and mortality: a nine-year follow-up study of alameda county residents. *American Journal of Epidemiology*, 109(2): 186–204.

Bickel, P. J., Ritov, Y., Tsybakov, A. B., et al. (2009). Simultaneous analysis of lasso and dantzig selector. *The Annals of Statistics*, 37(4): 1705–1732.

Bollen, K. A. and Noble, M. D. (2011). Structural equation models and the quantification of behavior. *Proceedings of the National Academy of Sciences*, 108(Supplement 3): 15639–15646.

Boyer, L. A., Lee, T. I., Cole, M. F., Johnstone, S. E., Levine, S. S., Zucker, J. P., Guenther, M. G., Kumar, R. M., Murray, H. L., Jenner, R. G., et al. (2005). Core transcriptional regulatory circuitry in human embryonic stem cells. *Cell*, 122(6): 947–956.

Broman, K. W. and Speed, T. P. (2002). A model selection approach for the identification of quantitative trait loci in experimental crosses. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(4): 641–656.

Byrne, B. M. (2016). *Structural Equation Modeling with AMOS: Basic Concepts, Applications, and Programming*. Routledge.

Cai, L. (2010). The relationship between health and labour force participation: Evidence from a panel data simultaneous equation model. *Labour Economics*, 17(1): 77–90.

Cai, X., Bazerque, J. A., and Giannakis, G. B. (2013). Inference of gene regulatory networks with sparse structural equation models exploiting genetic perturbations. *PLoS Computational Biology*, 9(5): e1003068.

Carithers, L. J., Ardlie, K., Barcus, M., Branton, P. A., Britton, A., Buia, S. A., Compton, C. C., DeLuca, D. S., Peter-Demchok, J., Gelfand, E. T., et al. (2015). A novel approach to high-quality postmortem tissue procurement: the GTEx project. *Biopreservation and Biobanking*, 13(5): 311–319.

Carter, S. L., Brechbühler, C. M., Griffin, M., and Bond, A. T. (2004). Gene co-expression network topology provides a framework for molecular characterization of cellular state. *Bioinformatics*, 20(14): 2242–2250.

Chen, B.-S. and Wu, C.-C. (2013). Systems biology as an integrated platform for bioinformatics, systems synthetic biology, and systems metabolic engineering. *Cells*, 2(4): 635–688.

Chen, C. (2017). *Parallel Construction of Large-Scale Gene Regulatory Networks.* Ph.D. Dissertation, Department of Statistics, Purdue University.

Consortium, G. and Others (2015). The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans. *Science*, 348(6235): 648–660.

Cooper, M. B., Loose, M., and Brookfield, J. F. (2008). Evolutionary modelling of feed forward loops in gene regulatory networks. *Biosystems*, 91(1): 231–244.

Daub, C. O., Steuer, R., Selbig, J., and Kloska, S. (2004). Estimating mutual information using b-spline functions–an improved similarity measure for analysing gene expression data. *BMC Bioinformatics*, 5(1): 118.

de la Fuente, A. (2010). From differential expressionto differential networking–identification of dysfunctional regulatory networks in diseases. *Trends in Genetics*, 26(7): 326–333.

Dijkstra, T. K. and Henseler, J. (2015). Consistent and asymptotically normal pls estimators for linear structural equations. *Computational Statistics & Data Analysis*, 81: 10–23.

Dobra, A., Hans, C., Jones, B., Nevins, J. R., Yao, G., and West, M. (2004). Sparse graphical models for exploring gene expression data. *Journal of Multivariate Analysis*, 90(1): 196–212.

Edgar, R., Domrachev, M., and Lash, A. E. (2002). Gene expression omnibus: Ncbi gene expression and hybridization array data repository. *Nucleic Acids Research*, 30(1): 207–210.

Fan, J. and Liao, Y. (2014). Endogeneity in high dimensions. *The Annals of Statistics*, 42(3): 872–917.

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

Friedman, J., Hastie, T., and Tibshirani, R. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3): 432–441.

Friedman, N. (2004). Inferring cellular networks using probabilistic graphical models. *Science*, 303(5659): 799–805.

Friedman, N., Linial, M., Nachman, I., and Pe'er, D. (2000). Using bayesian networks to analyze expression data. *Journal of Computational Biology*, 7(3-4): 601–620.

Fuller, T. F., Ghazalpour, A., Aten, J. E., Drake, T. A., Lusis, A. J., and Horvath, S. (2007). Weighted gene coexpression network analysis strategies applied to mouse weight. *Mammalian Genome*, 18(6-7): 463–472.

Gambardella, G., Moretti, M. N., De Cegli, R., Cardone, L., Peron, A., and Di Bernardo, D. (2013). Differential network analysis for the identification of condition-specific pathway activity and regulation. *Bioinformatics*, 29(14): 1776–1785.

Gilad, Y., Rifkin, S. A., and Pritchard, J. K. (2008). Revealing the architecture of gene regulation: the promise of eqtl studies. *Trends in Genetics*, 24(8): 408–415.

Gill, R., Datta, S., and Datta, S. (2010). A statistical framework for differential network analysis from microarray data. *BMC Bioinformatics*, 11(1): 95.

Golub, G. H., Heath, M., and Wahba, G. (1979). Generalized cross-validation as a method for choosing a good ridge parameter. *Technometrics*, 21(2): 215–223.

Hahn, J. and Hausman, J. (2002). Notes on bias in estimators for simultaneous equation models. *Economics Letters*, 75(2): 237–241.

Horn, R. A. and Johnson, C. R. (2012). *Matrix Analysis*. Cambridge University Press.

Hoyle, R. H. (2012). *Handbook of Structural Equation Modeling*. The Guilford Press.

Huang, J., Ma, S., and Zhang, C.-H. (2008). Adaptive lasso for sparse high-dimensional regression models. *Statistica Sinica*, 18(4): 1603–1618.

Imbens, G. W. and Newey, W. K. (2009). Identification and estimation of triangular simultaneous equations models without additivity. *Econometrica*, 77(5): 1481–1512.

Jeanty, P. W., Partridge, M., and Irwin, E. (2010). Estimation of a spatial simultaneous equation model of population migration and housing price dynamics. *Regional Science and Urban Economics*, 40(5): 343–352.

Jöreskog, K. G. (1970). A general method for estimating a linear structural equation system. *ETS Research Bulletin Series*, 1970(2):i–41.

Kai, L. (1998). Bayesian inference in a simultaneous equation model with limited dependent variables. *Journal of Econometrics*, 85(2): 387–400.

Kalisch, M. and Bühlmann, P. (2007). Estimating high-dimensional directed acyclic graphs with the pc-algorithm. *Journal of Machine Learning Research*, 8: 613–636.

Kawakatsu, T., Huang, S.-s. C., Jupe, F., Sasaki, E., Schmitz, R. J., Urich, M. A., Castanon, R., Nery, J. R., Barragan, C., He, Y., et al. (2016). Epigenomic diversity in a global collection of arabidopsis thaliana accessions. *Cell*, 166(2): 492–505.

Kim, S., Imoto, S., and Miyano, S. (2004). Dynamic bayesian network and nonparametric regression for nonlinear modeling of gene networks from time series gene expression data. *Biosystems*, 75(1-3): 57–65.

Koster, J. T. et al. (1996). Markov properties of nonrecursive causal models. *The Annals of Statistics*, 24(5): 2148–2177.

Kraemer, N., Schaefer, J., and Boulesteix, A.-L. (2009). Regularized estimation of large-scale gene regulatory networks using Gaussian graphical models. *BMC Bioinformatics*, 10(384).

Lai, Y., Wu, B., Chen, L., and Zhao, H. (2004). A statistical method for identifying differential gene–gene co-expression patterns. *Bioinformatics*, 20(17): 3146–3155.

Langfelder, P. and Horvath, S. (2008). WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics*, 9(1): 559.

Lee, L.-F. (1982). Health and wage: a simultaneous equation model with multiple discrete indicators. *International Economic Review*, 23(1): 199–221.

Lee, S.-H., Chen, T.-Y., Dhar, S. S., Gu, B., Chen, K., Kim, Y. Z., Li, W., and Lee, M. G. (2016). A feedback loop comprising prmt7 and mir-24-2 interplays with oct4, nanog, klf4 and c-myc to regulate stemness. *Nucleic Acids Research*, 44(22): 10603–10618.

Lin, W., Feng, R., and Li, H. (2015). Regularization methods for high-dimensional instrumental variables regression with an application to genetical genomics. *Journal of the American Statistical Association*, 110(509): 270–288.

Liu, B., de La Fuente, A., and Hoeschele, I. (2008). Gene network inference via structural equation modeling in genetical genomics experiments. *Genetics*, 178(3): 1763–1776.

Liu, S., Suzuki, T., Relator, R., Sese, J., Sugiyama, M., Fukumizu, K., et al. (2017a). Support consistency of direct sparse-change learning in markov networks. *The Annals of Statistics*, 45(3): 959–990.

Liu, W. et al. (2017b). Structural similarity and difference testing on multiple sparse Gaussian graphical models. *The Annals of Statistics*, 45(6): 2680–2707.

Logsdon, B. A. and Mezey, J. (2010). Gene expression network reconstruction by convex feature selection when incorporating genetic perturbations. *PLoS Computational Biology*, 6(12): e1001014.

Ma, H., Schadt, E. E., Kaplan, L. M., and Zhao, H. (2011). Cosine: Condition-specific sub-network identification using a global optimization method. *Bioinformatics*, 27(9): 1290–1298.

Marx, V. (2013). Biology: The big challenges of big data. *Nature*, 498: 255–260.

Matthews, B. W. (1975). Comparison of the predicted and observed secondary structure of t4 phage lysozyme. *Biochimica et Biophysica Acta (BBA)-Protein Structure*, 405(2): 442–451.

Meinshausen, N. and Bühlmann, P. (2006). High-dimensional graphs and variable selection with the lasso. *The Annals of Statistics*, 34: 1436–1462.

Menéndez, P., Kourmpetis, Y. A., ter Braak, C. J., and van Eeuwijk, F. A. (2010). Gene regulatory networks from multifactorial perturbations using graphical lasso: application to the DREAM4 challenge. *PloS One*, 5(12): e14147.

Muthén, L. and Muthén, B. (1998-2017). *Mplus Users Guide*. Los Angeles, CA: Muthén & Muthén.

Nelson, F. and Olson, L. (1978). Specification and estimation of a simultaneous-equation model with limited dependent variables. *International Economic Review*, 19(3): 695–709.

Ni, Y., Ji, Y., Müller, P., et al. (2016). Reciprocal graphical models for integrative gene regulatory network analysis. *arXiv:1607.06849v1*.

Ni, Y., Mller, P., Zhu, Y., and Ji, Y. (2018). Heterogeneous reciprocal graphical models. *Biometrics*, 74(2): 606–615.

Omri, A., Nguyen, D. K., and Rault, C. (2014). Causal interactions between CO2 emissions, FDI, and economic growth: Evidence from dynamic simultaneous-equation models. *Economic Modelling*, 42: 382–389.

Pearl, J. (2003). *Causality: Models, Reasoning, and Inference: 2nd Edition*. Cambridge University Press.

Pearl, J. et al. (2009). Causal inference in statistics: An overview. *Statistics Surveys*, 3: 96–146.

Pianese, F., An, X., Kawsar, F., and Ishizuka, H. (2013). Discovering and predicting user routines by differential analysis of social network traces. In *2013 IEEE 14th International Symposium on "A World of Wireless, Mobile and Multimedia Networks" (WoWMoM)*, pages 1–9. IEEE.

Reiersøl, O. (1941). Confluence analysis by means of lag moments and other methods of confluence analysis. *Econometrica*, 9(1): 1–24.

Reiersøl, O. (1945). Confluence analysis by means of instrumental sets of variables. *Arkiv for Matematik, Astronomi Och Fysik*, 32A(4): 1–24.

Reiss, P. C. and Wolak, F. A. (2007). Structural econometric modeling: Rationales and examples from industrial organization. *Handbook of Econometrics*, 6: 4277–4415.

Rudelson, M., Vershynin, R., et al. (2013). Hanson-Wright inequality and sub-Gaussian concentration. *Electronic Communications in Probability*, 18(82): 1–9.

Saldana, D. F. and Feng, Y. (2018). SIS: An R package for sure independence screening in ultrahigh dimensional statistical models. *Journal of Statistical Software*, 83(2): 1–25.

Sánchez, B. N., Budtz-Jørgensen, E., Ryan, L. M., and Hu, H. (2005). Structural equation models: a review with applications to environmental epidemiology. *Journal of the American Statistical Association*, 100(472): 1443–1455.

Sargan, J. D. (1958). The estimation of economic relationships using instrumental variables. *Econometrica*, 26(3): 393–415.

Schmidt, P. (1976). *Econometrics.* New York: Marcel Dekker.

Schwarz, G. et al. (1978). Estimating the dimension of a model. *The Annals of Statistics*, 6(2): 461–464.

Spirtes, P., Glymour, C. N., and Scheines, R. (2000). *Causation, Prediction, and Search.* MIT Press.

Stegle, O., Parts, L., Durbin, R., and Winn, J. (2010). A Bayesian framework to account for complex non-genetic factors in gene expression levels greatly increases power in eqtl studies. *PLoS Computational Biology*, 6(5): e1000770.

Tamada, Y., Kim, S., Bannai, H., Imoto, S., Tashiro, K., Kuhara, S., and Miyano, S. (2003). Estimating gene networks from gene expression data by combining bayesian network model with promoter element detection. *Bioinformatics*, 19(suppl_2): ii227–ii236.

Teng, S. L. and Huang, H. (2009). A statistical framework to infer functional gene relationships from biologically interrelated microarray experiments. *Journal of the American Statistical Association*, 104(486): 465–473.

Theil, H. (1953a). *Estimating and Simultaneous Correlation in Complete Equation Systems*. Mimeo. The Hague: Central Planning Bureau.

Theil, H. (1953b). *Repeated Least-Squares Applied to Complete Equation Systems*. Mimeo. The Hague: Central Planning Bureau.

Theil, H. (1961). *Economic Forecasts and Policy*. Amsterdam: North Holland.

West, J., Bianconi, G., Severini, S., and Teschendorff, A. E. (2012). Differential network entropy reveals cancer system hallmarks. *Scientific Reports*, 2(802).

Wilde, J. (2000). Identification of multiple equation probit models with endogenous dummy regressors. *Economics Letters*, 69(3): 309–312.

Xia, Y., Cai, T., and Cai, T. T. (2015). Testing differential networks with applications to the detection of gene-gene interactions. *Biometrika*, 102(2): 247–266.

Xiong, M., Li, J., and Fang, X. (2004). Identification of genetic networks. *Genetics*, 166(2): 1037–1052.

Yant, L. (2012). Genome-wide mapping of transcription factor binding reveals developmental process integration and a fresh look at evolutionary dynamics. *American Journal of Botany*, 99(2): 277–290.

Young, W. C., Raftery, A. E., and Yeung, K. Y. (2014). Fast Bayesian inference for gene regulatory networks using ScanBMA. *BMC Systems Biology*, 8: 47.

Yuan, H., Xi, R., Chen, C., and Deng, M. (2017). Differential network analysis via lasso penalized d-trace loss. *Biometrika*, 104(4): 755–770.

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

Zhang, B. and Wang, Y. (2012). Learning structural changes of Gaussian graphical models in controlled experiments. *arXiv:1203.3532*.

Zhang, X.-F., Ou-Yang, L., and Yan, H. (2017). Node-based differential network analysis in genomics. *Computational biology and chemistry*, 69: 194–201.

Zhang, X.-F., Ou-Yang, L., Zhao, X.-M., and Yan, H. (2016). Differential network analysis from cross-platform gene expression data. *Scientific Reports*, 6(34112).

Zhao, S. D., Cai, T. T., and Li, H. (2014). Direct estimation of differential networks. *Biometrika*, 101(2): 253–268.

Zhu, Y. (2018). Sparse linear models and ℓ_1-regularized 2SLS with high-dimensional endogenous regressors and instruments. *Journal of Econometrics*, 202(2): 196–213.

Zou, M. and Conzen, S. D. (2004). A new dynamic bayesian network (DBN) approach for identifying gene regulatory networks from time course microarray data. *Bioinformatics*, 21(1): 71–79.

VITA

VITA

Min Ren was born in Xiangan, Anhui, China. He received his Bachelor's Degree in Mathematics and Applied Mathematics from Mathematics Elite Class of Nankai University in May 2013. He entered the doctorate program of the Department of Statistics at Purdue University in August 2013 and received a Master's Degree in Mathematical Statistics from Purdue University in December 2015. He is a PhD candidate in the Department of Statistics at Purdue University, working under the supervision of Dr. Dabao Zhang. After graduation, He will join Capital One as a Principal Quantitative Modeler in New York City.

The list of Min Ren's papers and publications below summarize his research work at Purdue University:

- **Ren M**, Zhang D. Differential Analysis of directed networks. *Proceedings of the 34rd Conference on Uncertainty in Artificial Intelligence (UAI)*, 2018.

- Chen C*, **Ren M***, Zhang M, and Zhang D. A two-stage penalized least squares method for constructing large systems of structural equations. *Journal of Machine Learning Research*, 19(2): 1-34, 2018. (* co-author).

- Pauli D, Ziegler G, **Ren M**, Jenks M, Hunsaker D, Zhang M, Baxter I, and Gore M. (2018). Multivariate analysis of the cotton seed ionome reveals integrated genetic signatures of abiotic stress response. *G3: Genes|Genomes|Genetics*, 8(4): 1147-1160.

- Park J, Salmi ML, Wan Salim WWA, Rademacher A, Wickizer B, Schooley A, Benton J, Cantero A, Argote P, **Ren M**, Zhang M, Porterfield DM, Ricco AJ, Roux SJ, and Rickus JL. (2017). An autonomous lab on a chip for space flight calibration of gravity-induced transcellular calcium polarization in single-cell fern spores. *Lab on a Chip, Royal Society of Chemistry*, 17(6): 1095-1103.

- Wang X*, **Ren M***, Zhang D, Zhang C, Lang Z, Macho A, Zhang M, and Zhu J-K. Large-scale eQTL identification in Arabidopsis reveals novel candidate regulators of immune responses and other processes. *In Preparation*, (* co-author).

- **Ren M**, Zhang M, and Zhang D. QTLBayes: Mapping via Bayesian classification of main and epistatic effects. *In Preparation*.