Regularized Learning of High-dimensional Sparse Graphical Models

A DISSERTATION
SUBMITTED TO THE FACULTY OF THE GRADUATE SCHOOL
OF THE UNIVERSITY OF MINNESOTA
BY

LINGZHOU XUE

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

HUI ZOU, ADVISER

July 2012

Acknowledgements

First and foremost, I wish to take this opportunity to express my great appreciation to my advisor Professor Hui Zou. Over the last four years, I have been truly fortunate to have Professor Hui Zou as my advisor, mentor and friend. I am extremely grateful to Professor Hui Zou for influencing me with his enthusiasm, insights and professionalism, for supporting me with patience and encouragement in pursuing my interests and dreams, and for the numerous conversations about research, career and life. Thanks, Professor Hui Zou, for everything.

I am very grateful to Professor Charles Geyer, Professor Glen Meeden and Professor Fadil Santosa for serving on my thesis committee and providing me many insightful suggestions. My special thanks go to Professor Fadil Santosa for his encouraging words and for generously introducing me to the Institute for Mathematics and its Applications and its 2012 Annual Program on Mathematics of Information.

I have been truly lucky to have met many excellent researchers and collaborators through my doctoral studies. I would like to thank Professor Runze Li, Professor Tianxi Cai, Professor Han Liu, Yu (David) Mao, Shiqian Ma, Teng Zhang and Qing Mai for their constant inspiration, stimulating discussions and close collaborations.

I wish to thank other faculty members in the School of Statistics for their support, availability and encouragement. I am thankful to all my friends for making my life at Minnesota a very enjoyable experience, and I will always cherish our friendship.

Last but not least, I want to thank my family for their love, support and encouragement. I would like to dedicate this thesis to my parents and my girlfriend for their boundless love.

Dedication

Dedicated to my parents Yiyu Xue, Yaohui Lin, and my girlfriend Qian Chen.

Abstract

High-dimensional graphical models are important tools for characterizing complex interactions within a large-scale system. In this thesis, our emphasis is to utilize the increasingly popular regularization technique to learn sparse graphical models, and our focus is on two types of graphs: Ising model for binary data and nonparanormal graphical model for continuous data. In the first part, we propose an efficient procedure for learning a sparse Ising model based on a non-concave penalized composite likelihood, which extends the methodology and theory of non-concave penalized likelihood. An efficient solution path algorithm is devised by using a novel coordinate-minorization-ascent algorithm. Asymptotic oracle properties of our proposed estimator are established with NP-dimensionality. We demonstrate its finite sample performance via simulation studies and real applications to study the Human Immunodeficiency Virus type 1 protease structure. In the second part, we study the nonparanormal graphical model that is much more robust than the Gaussian graphical model while retains the good interpretability of the latter. In this thesis we show that the nonparanormal graphical model can be efficiently estimated by using a unified regularized rank estimation scheme which does not require estimating those unknown transformation functions in the nonparanormal graphical model. In particular, we study the rank-based Graphical LASSO, the rank-based Dantzig selector and the rank-based CLIME. We establish their theoretical properties in the setting where the dimension is nearly exponentially large relative to the sample size. It is shown that the proposed rank-based estimators work as well as their oracle counterparts in both simulated and real data.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Nowadays, massive high-dimensional data are being routinely generated in various research fields, such as image processing, web mining, computational biology, climate studies, risk management and so on. How to succinctly characterizing complex interactions within a large-scale system is still an extremely challenging problem rooted in many real-world applications throughout science and engineering.

- One of our motivating examples is to study the associated viral mutations in the drug resistance of HIV-1-infected patients by investigating the inter-residue contacts between HIV-1 protease mutations and susceptibility to HIV antiretroviral therapy drugs. The viral mutation data are binary, and the Ising models from the statistical physics are usually used to study the underlying interactions of binary networks in practice. However, estimating the high-dimensional Ising model requires extremely intensive computation due to the intractable partition function. Thus, it is still a complicated task to discover the interaction patterns in the large scale binary network.

- The other motivating example is to analyze microarray gene expression measurements for recovering the isoprenoid genetic regulatory network in Arabidposis thaliana, which is one popular model organism in plant biology and genetics. The gene expression values are continuous data, and the common approach is

to study the underlying interactions (i.e. conditional dependencies) under the Gaussian graphical model. However, as evidenced in the normality test, the normal assumption usually appears to be inappropriate for the gene expression values with or without the log-transformation preprocessing. How to estimate the complex network effectively and efficiently from the high-dimensional non-normal data is still a difficult task.

In addition to these two examples, similar problems to construct large-scale binary or non-normal networks also arise in the real-world analysis of biological, geological, social or financial networks.

In this thesis, we emphasize the desirable parsimony in the estimation of graphical models, which offers an accurate predictive graph with a concise representation of complex interactions within a large-scale system. In particular, we utilize the well-developed regularization technique to learn sparse graphical models, and our focus is on two types of graphs: Ising model for binary data and nonparanormal graphical model for continuous data. In what follows, we outline the main results of this thesis. The details of Chapter 2 and Chapter 3 correspond to Xue et al. (2012) and Xue and Zou (2011) respectively.

In Chapter 2, we consider the problem of estimating a sparse Ising model for the binary data. To this end, we propose an efficient procedure for learning a sparse Ising model based on a non-concave penalized composite likelihood, which extends the methodology and theory of non-concave penalized likelihood. An efficient solution path algorithm is devised by using a novel coordinate-minorization-ascent algorithm. Asymptotic oracle properties of our proposed estimator are established with NP-dimensionality. Technical proofs are given in the Appendix A.

In Chapter 3, we consider the problem of estimating a sparse nonparanormal graphical model for the non-normal data. The nonparanormal graphical model is introduced as a robust alternative to the Gaussian graphical model, while it retains

the good interpretability of conditional independencies as in the latter model. We show that the nonparanormal graphical model can be efficiently estimated by using a unified regularized rank estimation scheme which does not require estimating these unknown transformation functions. In particular, we study the rate of convergence and the graphical model selection consistency of the rank-based Graphical LASSO, the rank-based Dantzig selector and the rank-based CLIME in the high-dimensional setting where the dimension is nearly exponentially large relative to the sample size. Simulated and real data demonstrate that the proposed rank-based estimators work as well as their oracle counterparts. Technical proofs are given in the Appendix B.

Chapter 2

Estimating Sparse Ising Models

2.1 Introduction

The Ising model was first introduced in statistical physics (Ising, 1925) as a mathematical model for describing magnetic interactions and the structures of ferromagnetic substances. Although rooted in physics, the Ising model has been successfully exploited to simplify complex interactions for network exploration in various research fields such as social-economics (Stauffer, 2008), protein modeling (Irbäck et al., 1996), and statistical genetics (Majewski et al., 2001). Following the terminology in physics, consider an Ising model with K magnetic dipoles denoted by X_j, $1 \leq j \leq K$. Each X_j equals $+1$ or -1, corresponding to the up or down spin state of the j-th magnetic dipole. The energy function is defined as $E = -\frac{1}{4}(\sum_{i \neq j} \beta_{ij} X_i X_j)$, where the coupling coefficient β_{ij} describes the physical interactions between dipoles X_i and X_j under the external magnetic field, $\beta_{ii} = 0$ and $\beta_{ij} = \beta_{ji}$ for any (i, j). According to Boltzmann's law, the joint distribution of $\boldsymbol{X} = (X_1, \cdots, X_K)$ should be

$$\Pr(X_1 = x_1, \cdots, X_K = x_K) = \frac{1}{Z(\boldsymbol{\beta})} \exp(\sum_{(i,j)} \frac{\beta_{ij} x_j x_i}{4}), \tag{2.1}$$

where $Z(\boldsymbol{\beta})$ is the partition function.

In this chapter we focus on learning sparse Ising models, i.e., many coupling coefficients are zero. Our research is motivated by the HIV drug resistance study where understanding the inter-residue couplings (interactions) could potentially shed light on the mechanisms of drug resistance. A suitable statistical learning method is to fit a sparse Ising model to the binary data, in order to discover the inter-residue couplings. More details are given in Section 2.5. In the recent statistical literature penalized likelihood estimation has become a standard tool for sparse estimation. See a recent review paper by Fan and Lv (2010). In principle we can follow the penalized likelihood estimation paradigm to derive a sparse penalized estimator of the Ising model. Unfortunately, the penalized likelihood estimation method is very difficult to compute under the Ising model because the partition function $Z(\boldsymbol{\beta})$ is computationally intractable when the number of dipoles is relatively large. On the other hand, the composite likelihood idea (Lindsay, 1988; Varin et al., 2011) offers a nice alternative. To elaborate, suppose we have N independent identically distributed (iid) realizations of $\boldsymbol{X}$ from the Ising model, denoted by $\{(x_{1n}, ..., x_{Kn}), n = 1, ..., N\}$. Let $\theta_j = P(X_i = x_j|\boldsymbol{X}_{(-j)})$, describing the conditional distribution of the jth dipole given the remaining dipoles, where $\boldsymbol{X}_{(-j)}$ denotes $\boldsymbol{X}$ with the j-th element removed. By (2.1), it is to easy see that for the n-th observation

$$\theta_{jn} = \frac{\exp(\sum_{k:k\neq j} \beta_{jk} x_{jn} x_{kn})}{\exp(\sum_{k:k\neq j} \beta_{jk} x_{jn} x_{kn}) + 1}.$$

Note that θ_{jn} does not involves the partition function. The conditional log-likelihood of the j-th dipole given the remaining dipoles is given by

$$\ell^{(j)} = \frac{1}{N} \sum_{n=1}^{N} \log(\theta_{jn}).$$

As in Lindsay (1988) a composite log-likelihood function can be defined as

$$\ell_c = \sum_{j=1}^{K} \ell^{(j)}.$$

This kind of composite conditional likelihood was also called pseudo-likelihood in Besag (1974). Another popular type of composite likelihood is composite marginal likelihood (Varin, 2008). Maximum composite likelihood is especially useful when the full likelihood is intractable. Such an approach has important applications in many areas including spatial statistics, clustered and longitudinal data and time series models. A nice review on the recent developments in composite likelihood can be found in Varin et al. (2011).

To estimate a high-dimensional sparse Ising model, we consider the following penalized composite likelihood estimator

$$\widehat{\boldsymbol{\beta}} = \arg\max_{\boldsymbol{\beta}} \{\ell_c(\boldsymbol{\beta}) - \sum_{j=1}^{K} \sum_{k=j+1}^{K} P_\lambda(|\beta_{jk}|)\} \tag{2.2}$$

where $P_\lambda(t)$ is a positive penalty function defined on $[0, \infty)$. In this work we focus primarily on the LASSO penalty (Tibshirani, 1996) and Smoothly Clipped Absolute Deviation (SCAD) penalty (Fan and Li, 2001). The LASSO penalty is $P_\lambda(t) = \lambda t$. The SCAD penalty is defined by

$$P'_\lambda(t) = \lambda \left\{ I(t \leq \lambda) + \frac{(a\lambda - t)_+}{(a-1)\lambda} I(t > \lambda) \right\}, \quad t \geq 0; \quad a > 2.$$

Following Fan and Li (2001) we set $a = 3.7$. We should make it clear that when $P_\lambda(t)$ is non-concave, $\widehat{\boldsymbol{\beta}}$ should be understood as a good local maximizer of (2.2). See discussions in Section 2.2.

The optimization problem in (2.2) is very challenging because of two major issues:

(1) the number of unknown parameters is $\frac{1}{2}K(K-1)$ and hence the optimization problem is high-dimensional in nature; and (2) the penalty function is concave and non-differentiable at zero, although ℓ_c is a smooth concave function. We propose to combine both strengthes of the coordinate-ascent and minorization-maximization principles, which results in two new algorithms, CMA and LLA-CMA, for computing a local solution of the non-concave penalized composite likelihood. See Section 2.2 for details. With the aid of the new algorithms, the SCAD penalized estimators are able to enjoy computational efficiency comparable to that of the LASSO penalized estimator.

Fan and Li (2001) advocated the oracle properties of the non-concave penalized likelihood estimator in the sense that it performs as well as the oracle estimator which is the hypothetical maximum likelihood estimator knowing the true submodel. Zhang (2010a) and Lv and Fan (2009) were among the first to study the concave penalized least-squares estimator with NP-dimensionality (i.e. p can grow faster than any polynomial function of n). Fan and Lv (2011) studied the asymptotic properties of non-concave penalized likelihood for generalized linear models with NP-dimensionality. In this chapter we show that the oracle model selection theory remains to hold nicely for non-concave penalized composite likelihood methods with NP-dimensionality. Furthermore, we show that under certain regularity conditions the oracle estimator can be attained asymptotically via the LLA-CMA algorithm.

There is some related work in the literature. Ravikumar et al. (2010) viewed the Ising model as a binary Markov graph and used a neighborhood LASSO-penalized logistic regression algorithm to select the edges. Their idea is an extension of neighborhood selection by LASSO regression proposed by Meinshausen and Bühlmann (2006) for estimating Gaussian graphical models. Höfling and Tibshirani (2009) suggested using the LASSO-penalized pseudo-likelihood to estimate binary Markov graphs. However, they did not provide any theoretical result nor application. In this

chapter we compare the LASSO and the SCAD penalized composite likelihood estimators and show the latter has substantial advantages with respect to both numerical and theoretical properties.

The rest of this chapter is organized as follows. In Section 2.2, we introduce the CMA and LLA-CMA algorithms. The statistical theory is presented in Section 2.3. Monte Carlo simulation results are shown in Section 2.4. In Section 2.5 we present a real application of the proposed method to study the network structure of the amino-acid sequences of retroviral proteases using data from the Stanford HIV drug resistance database. Technical proofs are relegated to the Appendix A.

2.2 Computing Algorithms

In this section we discuss how to efficiently implement the penalized composite likelihood estimators. As mentioned before, the computational challenges come from (1) penalizing the concave composite likelihood with a non-concave penalty which is not differentiable at zero; (2) the intrinsically high dimension of the unknown parameters. Zou and Li (2008) proposed the local linear approximation (LLA) algorithm to derive an iterative ℓ_1 optimization procedure for computing non-concave penalized estimators. The basic idea behind LLA is the minorization-maximization principle (Lange et al., 2000; Hunter and Lange, 2004). Coordinate-ascent (or descent) algorithms (Tseng, 1988) have been successfully used for solving penalized estimators with LASSO-type penalties. See, e.g., Fu (1998), Daubechies et al. (2004), Genkin et al. (2007), Yuan and Lin (2006), Meier et al. (2008), Wu and Lange (2008) and Friedman et al. (2010). In this work we combine the strengthes of minorization-maximization and coordinatewise optimization to overcome the computational challenges.

2.2.1 The CMA algorithm

Let $\widetilde{\boldsymbol{\beta}}$ be the current estimate. The coordinate-ascent algorithm sequentially updates $\tilde{\beta}_{ij}$ by solving the following univariate optimization problem

$$\tilde{\beta}_{jk} \Leftarrow \arg\max_{\beta_{jk}} \left\{ \ell_c(\beta_{jk}; \beta_{j'k'} = \tilde{\beta}_{j'k'}, (j',k') \neq (j,k)) - P_\lambda(|\beta_{jk}|) \right\}. \tag{2.3}$$

However, we do not have a closed-form solution for the maximizer of (2.3). The exact maximization has to be conducted by some numerical optimization routine, which may not be a good choice in the coordinate-ascent algorithm because the maximization routine needs to be repeated many times to reach convergence. On the other hand, one can find an update to increase rather than maximize the objective function in (2.3), maintaining the crucial ascent property of the coordinate-ascent algorithm. This idea is in line with the generalized E-M algorithm (Dempster et al., 1977) in which one seeks to increase the expected log likelihood in the M-step.

First, we observe that for any β_{ij}

$$\frac{\partial^2 \ell_c(\boldsymbol{\beta})}{\partial \beta_{jk}^2} = -\frac{1}{N} \sum_{n=1}^{N} (\theta_{kn}(1-\theta_{kn}) + \theta_{jn}(1-\theta_{jn})) \geq -\frac{1}{2}. \tag{2.4}$$

Thus, by Taylor's expansion we have

$$\ell_c(\beta_{jk}; \beta_{j'k'} = \tilde{\beta}_{j'k'}, (j',k') \neq (j,k)) \geq Q(\beta_{jk})$$

where

$$\begin{aligned} Q(\beta_{jk}) \;\equiv\; & \ell_c(\beta_{jk} = \tilde{\beta}_{jk}; \beta_{j'k'} = \tilde{\beta}_{j'k'}, (j',k') \neq (j,k)) \\ & + \tilde{z}_{jk}(\beta_{jk} - \widetilde{\beta}_{jk}) - \frac{1}{4}(\beta_{jk} - \tilde{\beta}_{jk})^2 \end{aligned} \tag{2.5}$$

$$\tilde{z}_{jk} = \frac{\partial \ell_c(\boldsymbol{\beta})}{\partial \beta_{jk}}\Big|_{\boldsymbol{\beta}=\widetilde{\boldsymbol{\beta}}} = \frac{1}{N}\sum_{n=1}^{N} x_{kn}x_{jn}(2 - \theta_{kn}(\widetilde{\beta}) - \theta_{jn}(\widetilde{\beta})). \tag{2.6}$$

Next, Zou and Li (2008) showed that

$$P_\lambda(|\beta_{jk}|) \leq P_\lambda(|\tilde{\beta}_{jk}|) + P'_\lambda(|\tilde{\beta}_{jk}|) \cdot (|\beta_{jk}| - |\tilde{\beta}_{jk}|) \equiv L(|\beta_{jk}|). \tag{2.7}$$

Combining (2.5)–(2.7) we see that $Q(\beta_{jk}) - L(|\beta_{jk}|)$ is a minorization function of the objective function in (2.3). We update $\tilde{\beta}_{jk}$ by

$$\tilde{\beta}_{jk}^{\text{new}} = \arg\max_{\beta_{jk}} \left\{ Q(\beta_{jk}) - L(|\beta_{jk}|) \right\}, \tag{2.8}$$

whose solution is given by

$$\tilde{\beta}_{jk}^{\text{new}} = S(\tilde{\beta}_{jk} + 2\tilde{z}_{jk}, 2P'_\lambda(|\tilde{\beta}_{jk}|))$$

with $S(r,t) = \text{sgn}(r)(|r| - t)_+$ being the soft-thresholding operator (Tibshirani, 1996). The above arguments lead to the following Algorithm 1 which we call the coordinate-minorization-ascent (CMA) algorithm.

Algorithm 1 The CMA algorithm

1. Initialization of $\widetilde{\boldsymbol{\beta}}$.
2. Cyclic coordinate-minorization-ascent: sequentially update $\tilde{\beta}_{ij}$ $(1 \leq j < k \leq K)$ via soft-thresholding $\tilde{\beta}_{jk} \Leftarrow S(\tilde{\beta}_{jk} + 2\tilde{z}_{jk}, 2P'_\lambda(|\tilde{\beta}_{jk}|))$.
3. Repeat the above cycle till convergence.

Remark 3.1. It is easy to prove that Algorithm 1 has a nice ascent property which is a direct consequence of the minorization-maximizaton principle. Note that Algorithm 1 can be directly used to compute the LASSO-penalized composite likelihood

estimator. We simply modify the coordinate-wise updating formula as

$$\tilde{\beta}_{jk} \Leftarrow S(\tilde{\beta}_{jk} + 2\tilde{z}_{jk}, 2\lambda).$$

In practice we need to specify the λ value. BIC has been shown to perform very well for selecting the tuning parameter of the penalized likelihood estimator (Wang, Li and Tsai, 2007). The BIC score is defined as

$$\hat{\lambda} = \arg\max_{\lambda}\{2\ell_c(\widehat{\boldsymbol{\beta}}(\lambda)) - \log(n) \cdot \sum_{(j,k)} I(\hat{\beta}_{jk}(\lambda) \neq 0)\}. \tag{2.9}$$

The above BIC score is used to tune all methods considered in this work. We use SCAD1 to denote the SCAD solution computed by Algorithm 1 with the BIC tuned LASSO solution being the starting value.

For computational efficiency considerations, we implement Algorithm 1 by using the path-following idea and some other tricks including warm-starts and active-set-cycling (Friedman et al., 2010). We have implemented the algorithm in R language functions. The core cyclic coordinate-wise soft-thresholding operations were carried out in C.

2.2.2 Issues of local solution and the LLA-CMA algorithm

The objective function in (2.2) is generally non-concave if a non-concave penalty function is used. Using Algorithm 1 we find a local solution to (2.2) but there is no guarantee that it is the global solution. A similar case is Schelldorfer et al. (2011) where the objective function is the LASSO-penalized maximum likelihood of a high-dimensional linear mixed-effects model and the authors derived a coordinate-wise gradient descent algorithm to find a local solution.

The local solution issue should not be considered as a special weakness of Algo-

rithm 1 or other coordinate-wise descent algorithm as in Schelldorfer et al. (2011) that the algorithm can only find a local solution, because in the current literature there is no algorithm that can guarantee to find the global solution of non-concave maximization (or non-convex minimization) problems, especially when the dimension is huge. Consider, for example, the E-M algorithm which is perhaps the most famous algorithm in statistical literature. The E-M algorithm often offers an elegant way to fit some statistical models that are formulated as non-concave maximization problems. However, the E-M algorithm provides a local solution in general. A recent application of the E-M algorithm to high-dimensional modeling can be found in Städler et al. (2010) who considered a LASSO-penalized maximum likelihood estimator of a high-dimensional linear regression model with inhomogeneous errors that are modeled by a finite mixture of Gaussians. To handle the computational challenges in their problem, Städler et al. (2010) proposed a generalized E-M algorithm in which a coordinate descent loop is used in the M-step and showed that the obtained solution is a local solution.

Our numerical results show that in the penalized composite likelihood estimation problem the SCAD performs much better than the LASSO. To offer theoretical understanding of their differences, it is important to show that the obtained local solution of the SCAD-penalized likelihood has better theoretical properties than the LASSO estimator. In Section 2.3 we establish the asymptotic properties of the LASSO estimator and a local solution of (2.2) with the SCAD penalty. However, a general technical difficulty in non-concave maximization problems is to show that the computed local solution is the one local solution with proven theoretical properties. In Städler et al. (2010) and Schelldorfer et al. (2011), nice asymptotic properties are established for their proposed methods but it is not clear whether the computed local solutions could have those theoretical properties. The same issue exists in Fan and Lv (2011).

To circumvent the technical difficulty, we can consider combining the LLA idea (Zou and Li, 2008) and Algorithm 1 to solve (2.2) with a non-concave penalty. The LLA algorithm turns a non-concave penalization problem into a sequence of weighted LASSO penalization problems. Similar ideas of iterative LLA convex relaxation have been used in Candes et al. (2008), Zhang (2010b) and Bradic et al. (2011). Applying the LLA algorithm to (2.2), we need to iteratively solve

$$\widehat{\boldsymbol{\beta}}^{(m+1)} = \arg\max_{\boldsymbol{\beta}} \{\ell_c(\boldsymbol{\beta}) - \sum_{j=1}^{K} \sum_{k=j+1}^{K} w_{jk} \cdot |\beta_{jk}|\} \tag{2.10}$$

for $m = 0, 1, 2, \ldots$ where $w_{jk} = P'_\lambda(|\tilde{\beta}_{jk}^{(m)}|)$. Note that Algorithm 1 can be used to solve (2.10) by simply modifying the coordinate-wise updating formula as

$$\tilde{\beta}_{jk} \Leftarrow S(\tilde{\beta}_{jk} + 2\tilde{z}_{jk}, 2w_{jk}).$$

Therefore, we have the following LLA-CMA algorithm for computing a local solution of (2.2).

Algorithm 2 The LLA-CMA algorithm

1. Initialize $\widetilde{\boldsymbol{\beta}}^{(0)}$ and compute $w_{jk} = P'_\lambda(|\tilde{\beta}_{jk}^{(0)}|)$.
2. For $m = 0, 1, 2, 3, \ldots$, repeat the LLA iteration.
   (2.a) Use Algorithm 1 to solve $\widehat{\boldsymbol{\beta}}^{(m+1)}$ defined in (2.10).
   (2.b) Update the weights w_{jk} by $P'_\lambda(|\tilde{\beta}_{jk}^{(m+1)}|)$.

In Section 2.3 we show that if the LASSO estimator is chosen as the initial solution $\widetilde{\boldsymbol{\beta}}^{(0)}$ then under certain regularity conditions the LLA-CMA algorithm finds the oracle estimator with an overwhelming probability, which is the maximum likelihood estimator over the true support set and will be formally introduced in Section 2.3.

In fact, after two LLA iterations, the LLA-CMA algorithm converges to the oracle estimator with probability tending to 1. These results suggest us to take the following steps to compute the SCAD solution by the LLA-CMA algorithm.

Proposed three-step procedure for computing a SCAD estimator

Step 1. Use Algorithm 1 to compute the LASSO solution path and find the LASSO estimator by BIC.

Step 2. Use the LLA-CMA algorithm to compute the two-step LLA-CMA solution path and find the two-step LLA-CMA solution by BIC.

Step 2a. Use the LASSO estimator as $\widetilde{\boldsymbol{\beta}}^{(0)}$ in the LLA-CMA algorithm to compute the solution path. Choose the best λ by BIC and find the corresponding $\widetilde{\boldsymbol{\beta}}^{(1)}$ as the one-step LLA-CMA estimator.

Step 2b. Use the one-step LLA-CMA estimator as $\widetilde{\boldsymbol{\beta}}^{(0)}$ in the LLA-CMA algorithm to compute the solution path. Choose the best λ by BIC and the corresponding $\widetilde{\boldsymbol{\beta}}^{(1)}$ is referred as the two-step LLA-CMA solution in what follows and we denote it by SCAD2.

Step 3. For the chosen λ in Step 2b, use Algorithm 2 to compute the fully converged SCAD solution with SCAD2 being the starting value. Denote this SCAD solution by SCAD2**.

Note the fully converged LLA-CMA solution (i.e. SCAD2**) is essentially computed by Algorithm 2 with the BIC tuned one-step LLA-CMA estimator being the starting value. Based on our experience, the fully converged SCAD2** works slightly better than the two-step LLA-CMA solution (i.e. SCAD2) but these two solutions are generally very close. Both SCAD2 and SCAD2** perform better than SCAD1, which is the solution by Algorithm 1. The tradeoff is that SCAD2 and SCAD2** require

almost another half computing time of that for SCAD1. Generally we recommend using SCAD2** in the real-world applications. These numerical results are consistent with the numerical evidence provided by Bühlmann and Meier (2008) to advocate the two-step adaptive LASSO for the non-convex penalized regression problem.

2.3 Theoretical Properties

In this section we establish the statistical theory for the penalized composite conditional likelihood estimator using the SCAD and the LASSO penalty, respectively. Such results allow us to compare the SCAD and the LASSO estimators theoretically.

In order to present the theory we need some necessary notation. For a matrix $\boldsymbol{A} = (a_{ij})$, we define the following matrix norms: the Frobenius norm $\|\boldsymbol{A}\|_F = \sqrt{\sum_{(i,j)} a_{ij}^2}$, the entry-wise ℓ_∞ norm $\|\boldsymbol{A}\|_{\max} = \max_{(i,j)} |a_{ij}|$ and the matrix ℓ_∞ norm $\|\boldsymbol{A}\|_\infty = \max_i \sum_j |a_{ij}|$. Let $\boldsymbol{\beta}^* = \{\beta_{jk}^* : j < k\}$ denote the true coefficients. Define the true support set

$$\mathcal{A} = \{(j,k) : \beta_{jk}^* \neq 0, j < k\}$$

and its cardinality $s = |\mathcal{A}|$. Define $\rho(s,N) = \min_{(j,k)\in\mathcal{A}} |\beta_{jk}^*|$ which represents the weakness of the signal. Let H be the Hessian matrix of ℓ_c such that

$$H_{(j_1k_1),(j_2k_2)} = -\frac{\partial^2 \ell_c(\boldsymbol{\beta})}{\partial\beta_{j_1k_1}\partial\beta_{j_2k_2}}$$

for $1 \le j_1 < k_1 \le K$ and $1 \le j_2 < k_2 \le K$. For simplicity we use $H^* = H(\boldsymbol{\beta}^*)$. We partition H and $\boldsymbol{\beta}$ according to $\mathcal{A}$ as $\begin{pmatrix} H_{\mathcal{A}\mathcal{A}} & H_{\mathcal{A}\mathcal{A}^c} \\ H_{\mathcal{A}^c\mathcal{A}} & H_{\mathcal{A}^c\mathcal{A}^c} \end{pmatrix}$ and $\boldsymbol{\beta} = (\boldsymbol{\beta}_{\mathcal{A}}^T, \boldsymbol{\beta}_{\mathcal{A}^c}^T)^T$,

respectively. We let

$$\boldsymbol{X}_{\mathcal{A}} = (X_j : (j,k) \text{ or } (k,j) \in \mathcal{A} \text{ for some } k)$$

and

$$\boldsymbol{x}_{\mathcal{A}n} = (x_{jn} : (j,k) \text{ or } (k,j) \in \mathcal{A} \text{ for some } k).$$

Finally, we define

$$b = \lambda_{\min}(E[H^*_{\mathcal{A}\mathcal{A}}]),$$

$$B = \lambda_{\max}(E[\boldsymbol{X}_{\mathcal{A}}\boldsymbol{X}^T_{\mathcal{A}}])$$

and

$$\phi = \|E[H^*_{\mathcal{A}^c\mathcal{A}}](E[H^*_{\mathcal{A}\mathcal{A}}])^{-1}\|_\infty.$$

Define the oracle estimator as

$$\widehat{\boldsymbol{\beta}}^{oracle} = (\widetilde{\boldsymbol{\beta}}^{hmle}_{\mathcal{A}}, 0)$$

where

$$\widetilde{\boldsymbol{\beta}}^{hmle}_{\mathcal{A}} = \arg\max_{\boldsymbol{\beta}_{\mathcal{A}}} \ell_c((\boldsymbol{\beta}_{\mathcal{A}}, 0)).$$

If we knew the true submodel, then we would use the oracle estimator to estimate the Ising model.

Theorem 2.1 Consider the SCAD-penalized composite likelihood defined in (2.2). We have the following two conclusions.

(1) For any $R < \frac{b}{3B}\frac{\sqrt{N}}{s}$, we have

$$\Pr\left(\|\widetilde{\boldsymbol{\beta}}_{\mathcal{A}}^{hmle} - \boldsymbol{\beta}_{\mathcal{A}}^{*}\|_2 \leq \sqrt{\frac{s}{N}}R\right) \geq 1 - \tau_1 \tag{2.11}$$

with $\tau_1 = \exp(-R^2\frac{b^2}{8^3}) + 2s^2\exp(-\frac{N}{s^2}\frac{b^2}{2}) + 2s^2\exp(-\frac{N}{s^2}\frac{B^2}{8})$.

(2) Pick a λ satisfying

$$\lambda < \min(\frac{\rho(s,N)}{2a}, \frac{(2\phi+1)b^2}{3sB}).$$

With probability at least $1-\tau_2$, $\widehat{\boldsymbol{\beta}}^{oracle}$ is a local maximizer of the SCAD-penalized composite likelihood estimator where

$$\begin{aligned}\tau_2 \;=\;& \exp(-R_*^2\frac{b^2}{8^3}) + K^2\exp(-\frac{N\lambda^2}{32(2\phi+1)^2}) \\ &+ \exp(-\frac{N\lambda}{3B(2\phi+1)s}\frac{b^2}{8^3}) + K^2 s\exp(-\frac{Nb^2}{2s^3}) + 2s^2\exp(-\frac{b^2N}{8s^3}) \\ &+4s^2[\exp(-\frac{N}{s^2}\frac{b^2}{2}) + \exp(-\frac{N}{s^2}\frac{B^2}{8})]\end{aligned} \tag{2.12}$$

and

$$R_* = \min(\frac{1}{2}\sqrt{\frac{N}{s}}\rho(s,N), \frac{b}{3B}\frac{\sqrt{N}}{s}). \qquad \square$$

We also analyzed the theoretical properties of the LASSO estimator. If the LASSO can consistently select the true model, it must equal to the hypothetically oracle

LASSO estimator $(\widetilde{\boldsymbol{\beta}}_{\mathcal{A}}, 0)$ where

$$\widetilde{\boldsymbol{\beta}}_{\mathcal{A}} = \arg\max_{\boldsymbol{\beta}_{\mathcal{A}}} \{\ell_c((\boldsymbol{\beta}_{\mathcal{A}}, 0)) - \lambda \sum_{(j,k)\in\mathcal{A}} |\beta_{jk}|\}.$$

Theorem 2.2 Consider the LASSO-penalized composite likelihood estimator.

(1) Choose λ such that $\lambda s < \frac{8b^2}{3B}$, and then we have

$$\Pr\left(\|\widetilde{\boldsymbol{\beta}}_{\mathcal{A}} - \boldsymbol{\beta}^*_{\mathcal{A}}\|_2 \leq \frac{16\lambda\sqrt{s}}{b}\right) \geq 1 - \tau_1'$$

with

$$\tau_1' = e^{-N\lambda^2/2} + 2s^2[\exp(\frac{-Nb^2}{2s^2}) + \exp(\frac{-NB^2}{8s^2})].$$

(2) Assume the ir-representable condition that $\phi \leq 1 - \eta < 1$. Choose λ such that

$$\lambda s < \min(\frac{b^2}{16^2 B}\frac{\eta/3}{4-\eta}, \frac{8b^2}{3B}).$$

Then $(\widetilde{\boldsymbol{\beta}}_{\mathcal{A}}, 0)$ is the LASSO-penalized composite likelihood estimator with probability at least $1 - \tau_2'$, where

$$\begin{aligned}\tau_2' &= e^{-N\lambda^2/2} + K^2 s\exp(-\frac{Nb^2\eta^2}{8s^3}) + K^2\exp(-\frac{N\lambda^2\eta^2}{32(4-\eta)^2}) \\ &\quad +2s^2[\exp(-\frac{Nb^2\eta^2}{2s^3(2-\eta)^2}) + \exp(\frac{-Nb^2}{2s^2}) + \exp(\frac{-NB^2}{8s^2})].\end{aligned}$$

□

In Theorems 2.1 and 2.2 the three quantities b, B and ϕ do not need to be constants. We can obtain a more straightforward understanding of the properties of the penalized composite likelihood estimators by considering the asymptotic consequences of these probability bounds. To highlight the main point, we consider b, B

and ϕ are fixed constants and derive the following asymptotic results.

Corollary 2.1 Suppose that b, B and ϕ are fixed constants and further assume

$$N \gg s^3 \log(K)$$

and

$$\rho(s, N) \gg \sqrt{\frac{\log(K)}{N}}.$$

(1) Pick the SCAD penalty parameter λ^{scad} satisfying

$$\lambda^{scad} < \min(\frac{\rho(s,N)}{2a}, \frac{(2\phi+1)b^2}{3sB}), \quad \lambda^{scad} \gg \sqrt{\frac{\log(K)}{N}}.$$

With probability tending to 1, the oracle estimator is a local maximizer of the SCAD-penalized estimator and

$$\|\widehat{\boldsymbol{\beta}}_{\mathcal{A}}^{oracle} - \boldsymbol{\beta}_{\mathcal{A}}^*\|_2 = O_P(\sqrt{\frac{s}{N}}).$$

(2) Assume the ir-representable condition in Theorem 2.2. Pick the LASSO penalty parameter λ^{lasso} satisfying

$$\min(\frac{1}{\sqrt{s}}\rho(s,N), \frac{1}{s}) \gg \lambda^{lasso} \gg \frac{1}{\sqrt{N}}$$

then the LASSO estimator consistently selects the true model and

$$\|\widehat{\boldsymbol{\beta}}_{\mathcal{A}}^{lasso} - \boldsymbol{\beta}_{\mathcal{A}}^*\|_2 = O_P(\lambda^{lasso}\sqrt{s}).$$ □

Remark 3.2. For the LASSO-penalized least squares, it is now known that the model selection consistency critically depends on the ir-representable condition Meinshausen and Bühlmann (2006); Zou (2006); Zhao and Yu (2007). A similar condition is again needed in the LASSO-penalized composite likelihood. Furthermore, Corollary 2.1 shows that even when it is possible for the LASSO to achieve consistent selection, λ^{lasso} should be much greater than $\sqrt{\frac{1}{N}}$, which means that

$$\lambda^{lasso}\sqrt{s} \gg \sqrt{\frac{s}{N}}.$$

So the LASSO yields larger bias than the SCAD.

Remark 3.3. We have shown that asymptotically speaking the oracle estimator is in fact a local solution of the SCAD-penalized composite likelihood model. This property is stronger than the oracle properties defined in Fan and Li (2001). Our result is the first to show that the oracle model selection theory holds nicely for non-concave penalized composite conditional likelihood models with NP-dimensionality. The usual composite likelihood theory in the literature is only applied to the fixed-dimension setting. Our result fills a long-standing gap in the composite likelihood literature.

What we have shown so far is the existence of a SCAD-penalized estimator that is superior to the LASSO-penalized estimator. Moreover, we would like to show that the computed SCAD estimator is equal to the oracle estimator. As discussed earlier in Section 2.2, such a result is very difficult to prove due to the non-concavity of the penalized likelihood function. See also Fan and Lv (2011), Städler et al. (2010) and Schelldorfer et al. (2011).

If one can prove that the objective function has only one maximizer, then the computed solution and the theoretically proven solution must be the same. This idea has been used in Fan and Lv (2011) to study the non-concave penalized generalized

linear models and Bradic et al. (2012) to study the non-concave penalized Cox's proportional hazards models. Their arguments are based on the observation that the SCAD penalty function has a finite maximum concavity (Zhang, 2010; Lv and Fan, 2009). Hence, if the smallest eigenvalue of the Hessian matrix of the negative log-likelihood is sufficiently large, the overall penalized likelihood function is concave and hence has a unique global maximizer. This argument requires the sample size is greater than the dimension, otherwise the Hessian matrix does not have full rank. To deal with the high-dimensional case, Fan and Lv (2011) further refined their arguments by considering a subspace denoted by $\mathbb{S}_s$ which is the union of all s-dimensional coordinate subspaces. Under some regularity conditions, Fan and Lv (2011) showed that the oracle estimator is the unique global maximizer in $\mathbb{S}_s$, which was referred to as restricted global optimality. Then by assuming that the computed solution has exactly s nonzero elements, it can be concluded that the computed solution is in $\mathbb{S}_s$ and hence equals the oracle estimator. See Proposition 3.b of Fan and Lv (2011). However, a fundamental problem with these arguments is that we have no idea whether the computed solution selects s nonzero coefficients, because s is unknown.

Here we take a different route to tackle the local solution issue. Instead of trying to prove the uniqueness of maximizer, we directly analyze the local solution by the LLA-CMA algorithm and discuss under which regularity conditions the LLA-CMA algorithm can actually find the oracle estimator.

Theorem 2.3 Consider the SCAD-penalized composite likelihood estimator in (2.2). Let $\widehat{\boldsymbol{\beta}}^{scad}$ be the local solution computed by Algorithm 2 (the LLA-CMA algorithm) with $\widetilde{\boldsymbol{\beta}}^{(0)}$ being the initial value. Pick a regularization parameter λ satisfying

$$\lambda < \min\left(\frac{\rho(s,N)}{2a}, \frac{(2\phi+1)b^2}{3sB}\right).$$

Write $\tau_0 = \Pr(\|\widetilde{\boldsymbol{\beta}}^{(0)} - \boldsymbol{\beta}^*\|_\infty > \lambda)$.

(1) The LLA-CMA algorithm finds the oracle estimator after one LLA iteration with probability at least $1 - \tau_0 - \tau_3$ where

$$\begin{aligned}\tau_3 &= K^2 \exp(\frac{-N\lambda^2}{32(2\phi+1)^2}) + \exp(\frac{-N\lambda}{3B(2\phi+1)s}\frac{b^2}{8^3}) + K^2 s \exp(\frac{-Nb^2}{2s^3}) \\ &\quad +2s^2[\exp(-\frac{Nb^2}{8s^3}) + \exp(-\frac{N}{s^2}\frac{b^2}{2}) + \exp(-\frac{N}{s^2}\frac{B^2}{8})].\end{aligned}$$ □

(2) The LLA-CMA algorithm converges after two LLA iterations and $\widehat{\boldsymbol{\beta}}^{scad}$ equals the oracle estimator with probability at least $1 - \tau_0 - \tau_2$, where τ_2 is defined in (2.12).

Theorem 2.3 can be used to drive the following asymptotic result.

Corollary 2.2 Suppose that b, B and ϕ are fixed constants and further assume

$$N \gg s^3 \log(K)$$

and

$$\rho(s, N) \gg \frac{\max(\sqrt{\log(K)}, 16\sqrt{s}/b)}{\sqrt{N}}.$$

Consider the SCAD-penalized composite likelihood estimator with the SCAD penalty parameter λ^{scad} satisfying

$$\lambda^{scad} < \min(\frac{\rho(s,N)}{2a}, \frac{(2\phi+1)b^2}{3sB}), \quad \lambda^{scad} \gg \sqrt{\frac{\log(K)}{N}}.$$

(1) If $\tau_0 \to 0$, then with probability tending to one, the LLA-CMA algorithm converges after two LLA iterations and the LLA-CMA solution (or its one-step

version) is equal to the oracle estimator.

(2) Consider using the LASSO estimator as $\widetilde{\boldsymbol{\beta}}^{(0)}$. Assume the ir-representable condition in Theorem 2.2 and pick the LASSO penalty parameter λ^{lasso} satisfying

$$\frac{1}{\sqrt{N}} \ll \lambda^{lasso} \ll \min(\frac{1}{\sqrt{s}}\rho(s,N), \frac{1}{s}), \quad \lambda^{lasso} < \frac{\lambda^{scad}}{\sqrt{s}}\frac{b}{16}.$$

Then $\tau_0 \to 0$ and the conclusion in (1) holds. □

Remark 3.4. Part (1) of Corollary 2.2 basically says that any estimator that converges to $\boldsymbol{\beta}^*$ in probability at a rate faster than λ^{scad} can be used as the starting value in the LLA-CMA algorithm to find the oracle estimator with high probability. Note that such a condition is not very restrictive. Part (2) of Corollary 2.2 shows that the LASSO estimator satisfies that condition. We could also consider using other estimators as the starting value in the LLA-CMA algorithm. For example, we can use the neighborhood selection estimator as $\widetilde{\boldsymbol{\beta}}^{(0)}$. Following Ravikumar et al. (2010) we assume an ir-representable condition for each of the K neighborhood LASSO-penalized logistic regression and some other regularity conditions. Then it is not hard to show that the neighborhood selection estimator is also a qualified starting value. In this work, we would like to faithfully follow the composite likelihood idea and hence prefer to use the LASSO-penalized composite likelihood estimator as the starting value in the LLA-CMA algorithm.

2.4 Numerical Properties

2.4.1 Monte-Carlo simulations

In this section we use simulation to study the finite sample performance of the SCAD-penalized composite likelihood estimator. For comparison, we also include other two

methods: neighborhood selection by LASSO-penalized logistic regression (Ravikumar et al., 2010) and the LASSO-penalized composite likelihood estimator.

For each coupling coefficient β_{jk}, the LASSO-penalized logistic method provides two estimates: $\hat{\beta}_{j\mapsto k}$ based on the model for the jth dipole and $\hat{\beta}_{k\mapsto j}$ based on the model for the kth dipole. Then we carry out two types of neighborhood selections: (i) aggregation by intersection (NSAI) based on $\hat{\beta}_{jk}^{\text{NSAI}}$, and (ii) aggregation by union (NSAU) based on $\hat{\beta}_{jk}^{\text{NSAU}}$, where

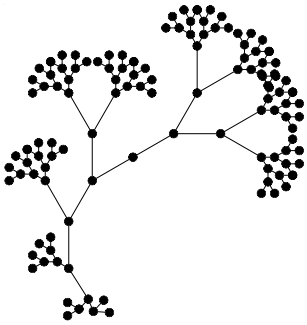
$$\hat{\beta}_{jk}^{\text{NSAI}} = \begin{cases} 0 & \text{if } \hat{\beta}_{j\mapsto k}\hat{\beta}_{k\mapsto j} = 0 \\ \frac{\hat{\beta}_{j\mapsto k}+\hat{\beta}_{k\mapsto j}}{2} & \text{otherwise} \end{cases}$$

and

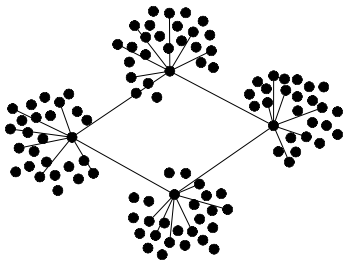
$$\hat{\beta}_{jk}^{\text{NSAU}} = \begin{cases} 0 & \text{if } \hat{\beta}_{j\mapsto k} = 0 \text{ and } \hat{\beta}_{k\mapsto j} = 0 \\ \hat{\beta}_{j\mapsto k} & \text{if } \hat{\beta}_{j\mapsto k} \neq 0 \text{ and } \hat{\beta}_{k\mapsto j} = 0 \\ \hat{\beta}_{k\mapsto j} & \text{if } \hat{\beta}_{j\mapsto k} = 0 \text{ and } \hat{\beta}_{k\mapsto j} \neq 0 \\ \frac{\hat{\beta}_{j\mapsto k}+\hat{\beta}_{k\mapsto j}}{2} & \text{if } \hat{\beta}_{j\mapsto k}\hat{\beta}_{k\mapsto j} \neq 0 \end{cases}.$$

BIC has been shown to perform very well for selecting the tuning parameter of the penalized likelihood estimator (Wang et al., 2007; Städler et al., 2010; Schelldorfer et al., 2011). We used BIC to tune all competitors.

Two sparse Ising models were considered in our simulation. Their graphical structure is displayed in Figure 1 where solid dots represent the dipoles and two dipoles are connected if and only if their coupling coefficient is non-zero. We generated the nonzero coupling coefficients as follows. If dipoles i and j are connected, we let β_{ij} be $t_{ij}s_{ij}$ where t_{ij} is a random variable following the uniform distribution on $[1, 2]$ and s_{ij} is a Bernoulli variable with $\Pr(s_{ij} = 1) = \Pr(s_{ij} = -1) = 0.5$. For each model,



(A) Model 1: 127 dipoles and 126 nonzero coupling coefficients.



(B) Model 2: 104 dipoles and 24 nonzero couplingcoefficients.

Figure 2.1: Plots of two simulated Ising models.

we used Gibbs sampling to generate 100 independent datasets consisting 300 observations. For comparison we use three measurements: the total number of discovered edges (NDE), the false discovery rate (FDR) and mean square errors (MSE).

(N, p)	Neighborhood Selection	LASSO	SCAD1	SCAD2	SCAD2**
Model 1 (300,7875)	51.1	32.7	67.9	84.7	95.1
Model 2 (300,5356)	29.8	16.0	34.8	42.6	51.2

Table 2.1: Total time (in seconds) for computing solutions at 100 penalization parameters, averaged over 3 replications. Timing was carried out on a laptop with an Intel Core 1.60GHz processor. The timing of SCAD1, SCAD2 and SCAD2** includes the timing for computing the starting value.

In Table 2.1 we compare the run times of the three methods. Compared to the LASSO case, the run time for fitting the SCAD model is doubled or tripled, but it is still very manageable for the high-dimensional data.

Based on the simulation results summarized in Table 2.2, we make the following interesting observations:

	Model 1			Model 2		
	MSE	NDE	FDR	MSE	NDE	FDR
NSAI	22.96	138.9	0.09	8.16	26.8	0.16
	(0.18)	(0.4)	(0.01)	(0.12)	(0.2)	(0.01)
NSAU	17.34	197.3	0.36	6.38	39.7	0.39
	(0.14)	(1.0)	(0.01)	(0.16)	(0.5)	(0.01)
LASSO	21.33	332.5	0.62	12.19	117.1	0.79
	(0.13)	(3.8)	(0.04)	(0.12)	(3.0)	(0.05)
SCAD1	2.86	145.0	0.12	5.64	30.0	0.22
	(0.10)	(2.4)	(0.01)	(0.17)	(1.8)	(0.02)
SCAD2	2.43	129.2	0.07	4.41	26.1	0.17
	(0.05)	(0.5)	(0.01)	(0.13)	(0.7)	(0.02)
SCAD2**	2.42	128.6	0.06	4.39	25.7	0.16
	(0.05)	(0.5)	(0.01)	(0.13)	(0.6)	(0.02)

Table 2.2: Comparing different estimators using simulation models 1 and 2 with standard errors in the bracket.

- NSAU, while selecting larger models than NSAI, provides more accurate estimation. Neighborhood selection outperforms the LASSO-penalized composite likelihood estimator.

- Note that SCAD2** has the smallest MSE in both models. SCAD2** and SCAD2 gave almost identical results and their improvement over SCAD1 is statistically significant. All three SCAD solutions perform much better than the LASSO for fitting penalized composite likelihood in terms of estimation and selection.

- The SCAD solutions and NSAI have similar model selection performance, but the SCAD is substantial better in estimation. Using the relaxed LASSO can improve the estimation accuracy of neighborhood selection methods, but their improved MSEs are still significantly higher than those of SCAD2 and SCAD2**.

2.4.2 Applications to the HIV drug resistance data

We illustrate our methods in a real example using a HIV antiretroviral therapy (ART) susceptibility dataset obtained from the Stanford HIV Drug Resistance Database (Rhee et al., 2004). The data for analysis consists of virus mutation information at 99 protease residues for $N = 702$ isolates from the plasma of HIV-1-infected patients. This dataset has been previously used in Rhee et al. (2006) and Wu et al. (2010).

A well recognized problem with current ART treatment such as PIs for treating HIV is that individuals who initially respond to therapy may develop resistance to it due to viral mutations. HIV-1 protease plays a key role in the late stage of viral replication and its ability to rapidly acquire a variety of mutations in response to various PIs confers the enzyme with high resistance to ARTs. A high cooperativity has been observed among drug-resistant mutations in HIV-1 protease (Ohtaka et al., 2003). The sequence data retrieved from treated patients is likely to include mutations that reflect cooperative effects originating from late functional constraints, rather than stochastic evolutionary noise (Atchley et al., 2000). However, the molecular mechanisms of drug resistance is yet to be elucidated. It is thus of great interest to study inter-residue couplings which might be relevant to protein structure or function and thus could potentially shed light on the mechanisms of drug resistance. We apply the proposed method to the protease sequence data to investigate such inter-residue contacts. Our analysis included $K = 79$ of the 99 residues that contain mutations.

We split the data into a training set with 500 data and a test set with 202 data. Model fitting and selection were done on the training set and the test data were used to compare the model errors. For a given estimate $\widehat{\boldsymbol{\beta}}$ obtained from the training set, its model error is gauged by the composite likelihood evaluated on the test set, i.e.,

$$\mathrm{ME}(\widehat{\boldsymbol{\beta}}) = -\ell_c^{\mathrm{test}}(\widehat{\boldsymbol{\beta}}) = -\frac{1}{202}\sum_{n=1}^{202}\sum_{j=1}^{79}\log(\theta_{jn}(\widehat{\boldsymbol{\beta}})).$$

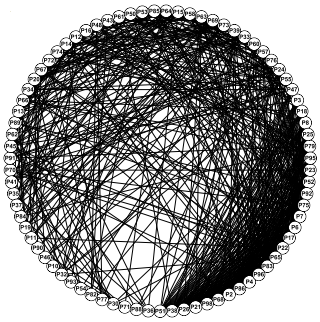
	NSAI	NSAU	LASSO	SCAD1	SCAD2	SCAD2**
NDE	57	305	631	101	141	132
ME	26.38	36.34	18.35	18.30	16.76	16.74
Stability selection	NSAI	NSAU	LASSO	SCAD1	SCAD 2	SCAD2**
NSE ($\pi_{thr} = 0.9$)	15	63	160	17	20	20
E[V]	≤ 3.2	≤ 48	≤ 147.5	≤ 4.3	≤ 8.0	≤ 7.2

Table 2.3: Application to HIVRT data. NSE is the number of "stable edges". $E[V]$ is the expected number of falsely selected edges. Its upper bounds were computed by Theorem 1 in Meinshausen and Bühlmann (2009).

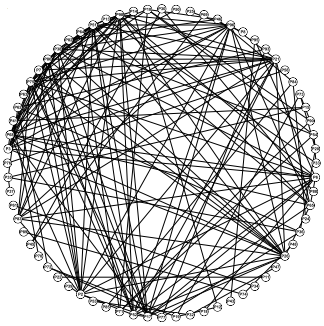
We report the analysis results in Table 2.3. There are total 3081 coupling coefficients to be estimated. Graphical presentations of the selected models are shown in Figure 3.2. Note that SCAD2 and SCAD2** again gave almost identical results and performed better SCAD1. We also performed stability selection (Meinshausen and Bühlmann, 2010) on each method to find "stable edges". A remarkable property of stability selection is that under some suitable conditions stability selection achieves finite sample control over the expected number of false discoveries in the set of "stable edges". We use the SCAD selector to explain the stability selection procedure. We took a random subsample of size 250 and fitted the SCAD model. The process was repeated 100 times. On average, SCAD1 selected 103.1 edges, SCAD2 selected 140.7 edges and SCAD2** chose 133.4 edges. For each coefficient β_{jk} we computed its frequency of being selected, denoted by $\hat{\Pi}_{jk}$. The set of "stable edges" is defined as $\{(k, j) : \hat{\Pi}_{kj} > \pi_{thr}\}$. In Table 3, we report the results using the threshold $\pi_{thr} = 0.9$, as suggested by Meinshausen and Bühlmann (2009). Stability selection found 17 edges in the SCAD1. SCAD2 and SCAD2** selected the same 20 stable edges. By theorem 1 in Meinshausen and Bühlmann (2009), among these 17 stable edges selected by SCAD1, the expected number of false discoveries is no greater than 4.3, and among the 20 stable edges selected by SCAD2 or SCAD2**, the expected

number of false discoveries is at most 7.2. Likewise, we did stability selection with the LASSO selector and neighborhood selection and the results are reported in Table 3 as well. Figure 3 shows the "stable edges" by stability selection. We see that the computed upper bounds are very useful for the SCAD selector and NSAI and not so informative for the LASSO selector and NSAU. Interestingly, both NSAI and SCAD suggest there are about 12 true discoveries by stability selection. In fact, we found that NSAI and SCAD have 11 "stable edges" in common, and NSAI and SCAD2 (or SCAD2**) have 12 "stable edges" in common.

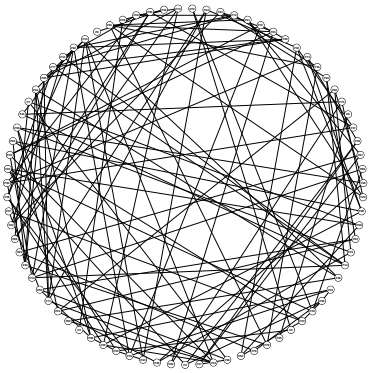
These results are consistent with some of the previous findings. For example, it has long been known that co-substitutions at residues 30 and 88 are most effective in reducing the susceptibility of nelfinavir (Liu et al., 2008). Among the top 30 most common drug resistance mutations (Rhee et al, 2004), 7 of those had a joint mutation at residues 54 and 82, the joint mutation at residues 88 and 30 was the second most common mutations among all. A co-mutation at residues 54, 82 and 90 was associated with high resistance to multiple drugs and an additional co-mutation at 46 was associated with an even higher level of resistance. It is interesting to note that using a larger set of isolates from treated HIV patients, Wu et al. (2003) reported (54, 82), (32, 47), (73, 90) as the three most highly correlated pairs. All these three pairs showed up as the stable edges in our analysis. Mutation at residue 71, often described as a compensatory or accessory mutation, has been reported as a critical mutation which appears to improve virus growth and contribute to resistance phenotype (Markowitz et al., 1995; Tisdale et al., 1995; Muzammil et al., 2003). Accessory mutations contribute to resistance only when present with a mutation in the substrate cleft or flap or at residue 90 (Wu et al., 2003). The stable edges connect this accessory mutation with residues 90 and 54 (a flap residue), as well as with another flap residue at 46 through residue 10.



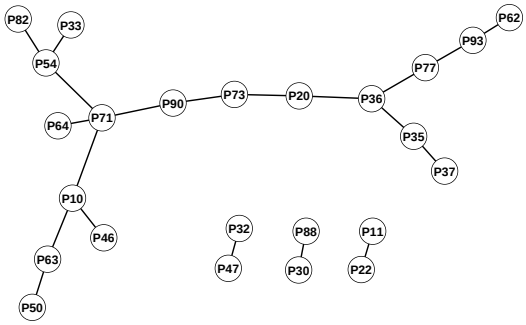
(A1) The LASSO model.



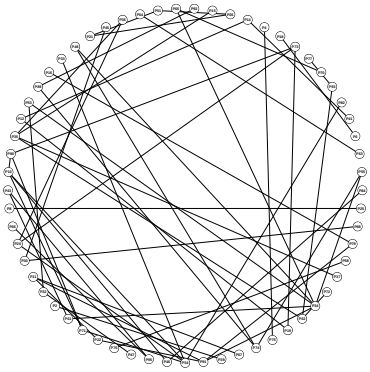
(A2) "Stable edges" in the LASSO model.



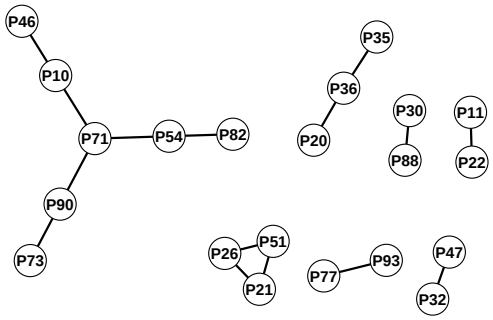
(B1) The SCAD2** model.



(B2) "Stable edges" in the SCAD2** model.



(C1) The model by neighborhood selection.



(C2) "Stable edges" in neighborhood selection.

Figure 2.2: Shown in the left three panels (A1,B1,C1) are the selected models by BIC. The right three panels (A2,B2,C2) show the stability selection results using $\pi_{thr} = 0.9$. SCAD2 and SCAD2** select the same stable edges. Neighborhood selection and SCAD1 have 11 common stable edges. Neighborhood selection and SCAD2 (or SCAD2**) select 12 common stable edges.

Chapter 3

Estimating Sparse Nonparanormal Graphical Models

3.1 Introduction

Estimating covariance or precision matrices is of fundamental importance in multivariate statistical methodologies and applications. In particular, when data follow a joint normal distribution, i.e. $\boldsymbol{X} = (X_1, \cdots, X_p) \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, the precision matrix $\boldsymbol{\Theta} = \boldsymbol{\Sigma}^{-1}$ can be directly translated into a Gaussian graphical model. The Gaussian graphical model serves as a non-causal structured approach to explore the complex systems consisting of Gaussian random variables, and among many interesting applications are gene expression genomics (Friedman, 2004; Wille et al., 2004), image processing (Li, 2009), and macroeconomics determinants study (Dobra et al., 2009). The precision matrix plays a critical role in the Gaussian graphical models because the zero entries in $\boldsymbol{\Theta} = \left(\theta_{ij}\right)_{1 \leq i,j \leq p}$ precisely capture the desired conditional independencies, i.e. $\theta_{ij} = 0$ if and only if $X_i \perp\!\!\!\perp X_j | \ \boldsymbol{X} \setminus \{X_i, X_j\}$. See Lauritzen (1996) and Edwards (2000) for more details.

The sparsity pursuit in precision matrices was initially considered by Dempster (1972) as the covariance selection problem. In the statistical literature, multiple testing methods have been employed for network exploration in the Gaussian

graphical models (Edwards, 2000; Drton and Perlman, 2004). With rapid advances of the high-throughput technology (e.g. microarray, functional magnetic reasoning imaging), estimation of a sparse graphical model has become increasingly important in the high-dimensional setting. Some well-developed penalization techniques have been used for estimating sparse Gaussian graphical models. In a highly-cited paper, Meinshausen and Bühlmann (2006) proposed the neighborhood selection scheme which tries to discover the smallest index set ne_α for each variable X_α satisfying $X_\alpha \perp\!\!\!\perp \boldsymbol{X} \setminus \{X_\alpha, \boldsymbol{X}_{ne_\alpha}\} | \boldsymbol{X}_{ne_\alpha}$. Meinshausen and Bühlmann (2006) further proposed to use the LASSO (Tibshirani, 1996) to fit each neighborhood regression model. Afterwards, one can summarize the zero patterns by aggregation via union or intersection. Yuan (2010) instead considered the Dantzig selector (Candes and Tao, 2007) as an alternative to the LASSO penalized least squares in the neighborhood selection scheme to estimate the precision matrix. Peng et al. (2009) proposed the joint neighborhood LASSO selection. Penalized likelihood methods have been studied for Gaussian graphical modeling. (Yuan and Lin, 2007; Banerjee et al., 2008; Friedman et al., 2008; Rothman et al., 2008). Friedman et al. (2008) developed a fast blockwise coordinate descent algorithm called Graphical LASSO for efficiently solving the LASSO penalized Gaussian graphical model. Witten et al. (2011) further developed a faster version of the Graphical LASSO. Rate of convergence under the Frobenius norm was established by Rothman et al. (2008). Ravikumar et al. (2008) obtained the convergence rate under the elementwise ℓ_∞ norm and the spectral norm. Lam and Fan (2009) studied the non-convex penalized Gaussian graphical model where a non-convex penalty such as SCAD (Fan and Li, 2001) is used to replace the LASSO penalty in order to overcome the bias issue of the LASSO penalization. Zhou et al. (2011) proposed a hybrid method for estimating sparse Gaussian graphical models: they first infer a sparse Gaussian graphical model structure via thresholding neighborhood selection and then estimate the precision matrix of the submodel by maximum

likelihood. Cai et al. (2011) recently proposed a constrained ℓ_1 minimization estimator called CLIME for estimating sparse precision matrices and established its convergence rates under the elementwise ℓ_∞ norm and Frobenius norm.

Although the normality assumption can be relaxed if we only focus on estimating a precision matrix, it plays an essential role in making the neat connection between a sparse precision matrix and a sparse Gaussian graphical model. Without normality, we ought to be very cautious when translating the output of a good sparse precision matrix estimation algorithm into an interpretable sparse Gaussian graphical model. However, the normality assumption often fails in reality. For example, the observed data are often skewed or have heavy tails. To illustrate the issue of non-normality in real applications, let us consider the gene expression data to construct isoprenoid genetic regulatory network in Arabidposis thaliana, which included 16 genes from the mevalonate (MVA) pathway in the cytosolic, 18 genes from the plastidial (MEP) pathway in the chloroplast, and also 5 encode proteins in the mitochondrial. The data, initially used by Wille et al. (2004), contained the gene expression measurements of 39 genes assayed on $n = 118$ Affymetrix GeneChip microarrays. For more details about the data, we refer readers to Wille et al. (2004) and Gilbert et al. (2009). This dataset was later re-analyzed by Li and Gui (2006); Drton and Perlman (2007) and Gilbert et al. (2009) in the context of Gaussian graphical modeling after taking the log-transformation of the expression data. However, the normality assumption may be inappropriate for this dataset even after the log-transformation. To show this, we conduct the normality test at the significance level of 0.05 as in Table Table 3.1, it is clear that at most 9 out of 39 genes would pass any of three normality tests, and even after log-transformation, at least 60% genes reject the null hypothesis of normality. Under Bonferroni correction there are still over 30% genes that fail to pass any normality test. Figure Figure 3.1 plots the histograms of two key isoprenoid genes MECPS in the MEP pathway and MK in the MVA pathway after the log-

Table 3.1: Testing for Normality of the gene expression data in the Arabidposis thaliana data. This table illustrate the number out of 39 genes rejecting the null hypothesis of normality at the significance level of 0.05.

	critical value	Cramer–von Mises	Lilliefors	Shapiro–Francia
raw data	0.05	30	30	35
	0.05/39	24	26	28
log data	0.05	29	24	33
	0.05/39	14	12	16

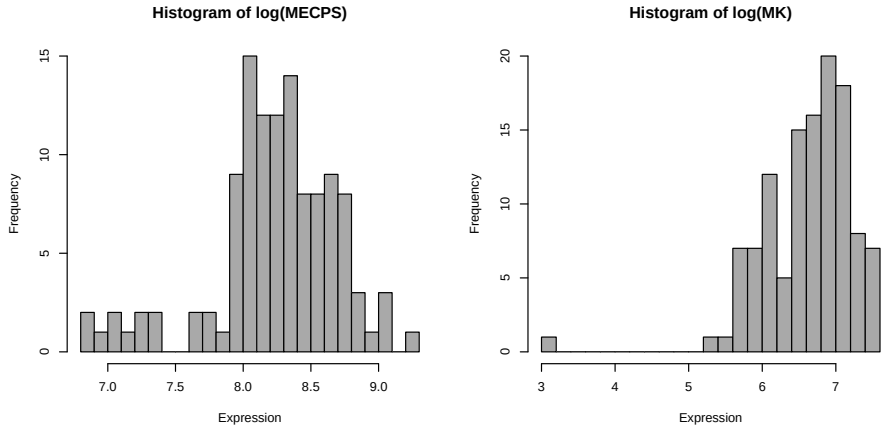


Figure 3.1: Illustration of the non-normality after the log-transformation preprocessing in the isoprenoid gene expression data.

transformation preprocessing, clearly showing the non-normality of the data after the log-transformation.

Using transformation to achieve normality is a classical idea in statistical modeling. The celebrated Box-Cox transformation is widely used in regression analysis. However, any parametric modeling of the transformation suffers from model misspecification which could lead to misleading inference results. In this chapter we take a nonparametric transformation strategy to handle the non-normality issue. Let $F(\cdot)$ be the CDF of a continuous random variable X and $\Phi^{-1}(\cdot)$ be the inverse of the CDF of $N(0,1)$. Consider the transformation from X to Z by $Z = \Phi^{-1}(F(X))$. Then it

is easy to see that Z is standard normal regardless of F. Motivated by this simple probabilistic result, we consider modeling the data by the nonparanormal model:

> **The nonparanormal model:** $\boldsymbol{X} = (X_1, \cdots, X_p)$ *follows a nonparanormal distribution if there is a vector of unknown univariate monotone increasing transformations, denoted by* $\boldsymbol{f} = (f_1, \cdots, f_p)$*, such that the transformed random vector follows a multivariate normal distribution with mean 0 and covariance* $\boldsymbol{\Sigma}$*:*
>
> $$\boldsymbol{f}(\boldsymbol{X}) = (f_1(X_1), \cdots, f_p(X_p)) \sim N_p(0, \boldsymbol{\Sigma}), \tag{3.1}$$
>
> where without loss of generality the diagonals of $\boldsymbol{\Sigma}$ are all equal to 1.

Note that model (3.1) implies that $f_j(X_j)$ is a standard normal random variable. Thus, f_j must be $\Phi^{-1} \circ F_j$ where F_j is the CDF of X_j. Since the marginal normality can be always achieved by transformations, model (3.1) basically assumes that after transformation those marginally normal distributed variables also follow a joint normal distribution. Define $\boldsymbol{Z} = (Z_1, \ldots, Z_p) = (f_1(X_1), \cdots, f_p(X_p))$. By the joint normality assumption of $\boldsymbol{Z}$, we know that $\theta_{ij} = 0$ if and only if $Z_i \perp\!\!\!\perp Z_j | \ \boldsymbol{Z} \setminus \{Z_i, Z_j\}$. Interestingly, we have that

$$Z_i \perp\!\!\!\perp Z_j | \ \boldsymbol{Z} \setminus \{Z_i, Z_j\} \Longleftrightarrow X_i \perp\!\!\!\perp X_j | \ \boldsymbol{X} \setminus \{X_i, X_j\}.$$

Therefore, a sparse precision matrix $\boldsymbol{\Theta}$ can be directly translated into a sparse graphical model for presenting the original variables. In other words, the nonparanormal model nicely retains the good interpretability of the Gaussian model. We follow Liu et al. (2009) to call model (3.1) the nonparanormal model, but model (3.1) is in fact a semiparametric Gaussian copula model. The semiparametric Gaussian copula model is a nice combination of flexibility and interpretability. Semiparametric Gaussian cop-

ulas have generated a lot of interests in statistics, econometrics and finance (Klaassen and Wellner, 1997; Song, 2000; Tsukahara, 2005; Chen and Fan, 2006). Much of the existing theoretical work on the inference of semiparametric Gaussian copulas focuses on the classical asymptotic setting where the dimension is fixed and the sample size goes to infinity.

In this work we primarily focus on estimating $\mathbf{\Theta}$ which is then used to construct a nonparanormal graphical model. As for the nonparametric transformation function, by the expression $f_j = \Phi^{-1} \circ F_j$, we have a natural estimator for the transformation function of the jth variable as $\hat{f}_j = \Phi^{-1} \circ \hat{F}_j^+$ where $\hat{F}_j^+$ is a Winsorized empirical CDF of the jth variables. Note that the Winsorization is used to avoid infinity value and to achieve better bias-variance tradeoff; see Liu et al. (2009) for detailed discussion. In this chapter we show that we can directly estimate $\mathbf{\Theta}$ without estimating these nonparametric transformation functions at all. This statement seems to be a bit surprising because a natural estimation scheme is a two-stage procedure: first estimate f_j and then apply a well-developed sparse Gaussian graphical model estimation method to the transformed data $\hat{\boldsymbol{z}}_i = \hat{\boldsymbol{f}}(\boldsymbol{x}_i), 1 \leq i \leq n$. Liu et al. (2009) have actually studied this "plug-in" estimation approach. They proposed a winsorized estimator of the nonparametric transformation function and used the Graphical LASSO in the second stage. They established convergence rate of the "plug-in" estimator when p is restricted to a polynomial order of n. However, the "plug-in" approach does not yield a satisfactory rate of convergence, for the rate of convergence can be established for the Gaussian graphical model even when p grows with n almost exponentially fast (Ravikumar et al., 2008). As noted in Liu et al. (2009), it is very challenging to push the theory of the "plug-in" approach to handle exponentially large dimensions. The "plug-in" estimator has a much slower rate of convergence, largely due to the fact that it requires uniform convergence over p nonparametric functions in order to ensure a certain rate of convergence for estimating $\mathbf{\Theta}$, which was proved only for polynomial

large dimensions in Liu et al. (2009). One might ask if using a better estimator for the transformation functions could improve the rate of convergence such that p could be allowed to be nearly exponentially large relative to n. This is a legitimate direction for research. We do not pursue this direction in this work. Instead, we show that we could use a rank-based estimation approach to achieve the exact same goal without estimating these transformation functions at all.

Our estimator is constructed in two steps. First, we propose a nonparametric rank-based sample estimate of $\mathbf{\Sigma}$ and prove its rate of convergence under the matrix entry-wise ℓ_∞ norm. As the second step, we compute a sparse estimator $\mathbf{\Theta}$ from the rank-based sample estimate of $\mathbf{\Sigma}$. For that purpose, we consider several regularized rank estimators, including the *rank-based Graphical LASSO*, the *rank-based neighborhood Dantzig selector* and the *rank-based CLIME*. The complete methodological details are presented in Section 3.2. In Section 3.3 we establish theoretical properties of the proposed rank-based estimators regarding both precision matrix estimation and graphical model selection. Section 3.4 contains simulation results and the rank-based analysis of the isoprenoid genetic regulatory network. Section 3.5 contains the concluding remarks, and technical proofs are presented in the Appendix B.

3.2 Proposed Methodology

We first introduce some necessary notation. For a matrix $\boldsymbol{A} = (a_{ij})$, we define its entry-wise ℓ_1 norm as $\|\boldsymbol{A}\|_1 = \sum_{(i,j)} |a_{ij}|$, and its entry-wise ℓ_∞ norm as $\|\boldsymbol{A}\|_{\max} = \max_{(i,j)} |a_{ij}|$. For a vector $\boldsymbol{v} = (v_1, \cdots, v_l)$, we define its ℓ_1 norm as $\|\boldsymbol{v}\|_{\ell_1} = \sum_j |v_j|$ and its ℓ_∞ norm as $\|\boldsymbol{v}\|_{\ell_\infty} = \max_j |v_j|$. To simplify notation, define $\boldsymbol{M}_{A,B}$ as the sub-matrix of $\boldsymbol{M}$ with row indexes A and column indexes B, and define $\boldsymbol{v}_A$ as the sub-vector of $\boldsymbol{v}$ with indexes A. Let (k) be the index set $\{1, \ldots, k-1, k+1, \ldots, p\}$. Denote by $\mathbf{\Sigma}_{(k)} = \mathbf{\Sigma}_{(k),(k)}$ the sub-matrix of $\mathbf{\Sigma}$ with both k-th row and column removed, and

denote by $\boldsymbol{\sigma}_{(k)} = \boldsymbol{\Sigma}_{(k),k}$ the vector including all the covariances associated with the k-th variable. In the same fashion, we can also define $\boldsymbol{\Theta}_{(k)}$, $\boldsymbol{\theta}_{(k)}$, and so on.

3.2.1 The oracle procedures

Suppose an oracle knows the underlying transformation vector, then the oracle could easily recover "oracle data" by applying these true transformations, i.e. $\boldsymbol{z}_i = \boldsymbol{f}(\boldsymbol{x}_i)$ for $1 \leq i \leq n$. Before presenting our rank-based estimators, it is helpful to revisit the "oracle" procedures that are defined based on the "oracle data".

- **The oracle Graphical LASSO.** Let $\hat{\boldsymbol{\Sigma}}^o$ be the sample covariance matrix for the "oracle" data, and then the "oracle" log-profile-likelihood becomes

  $$\log\det(\boldsymbol{\Theta}) - \mathrm{tr}(\hat{\boldsymbol{\Sigma}}^o\boldsymbol{\Theta}).$$

  The "oracle" Graphical LASSO solves the following ℓ_1 penalized likelihood problem:

  $$\min_{\boldsymbol{\Theta}\succ 0} -\log\det(\boldsymbol{\Theta}) + \mathrm{tr}(\hat{\boldsymbol{\Sigma}}^o\boldsymbol{\Theta}) + \lambda\sum_{i\neq j}|\theta_{ij}|. \tag{3.2}$$

- **The oracle neighborhood LASSO selection.** Under the nonparanormal model (3.1), for each $k = 1, \ldots, p$, the "oracle" variable Z_k given $\boldsymbol{Z}_{(k)}$ is normally distributed as

  $$N\left(\boldsymbol{Z}_{(k)}^T\boldsymbol{\Sigma}_{(k)}^{-1}\boldsymbol{\sigma}_{(k)}, 1 - \boldsymbol{\sigma}_{(k)}^T\boldsymbol{\Sigma}_{(k)}^{-1}\sigma_{(k)}\right),$$

  which can be equivalently written as

  $$Z_k = \boldsymbol{Z}_{(k)}^T\boldsymbol{\beta}_k + \varepsilon_k$$

with

$$\boldsymbol{\beta}_k = \boldsymbol{\Sigma}_{(k)}^{-1}\boldsymbol{\sigma}_{(k)} \quad \text{and} \quad \varepsilon_k \sim N(0, 1 - \boldsymbol{\sigma}_{(k)}^T\boldsymbol{\Sigma}_{(k)}^{-1}\boldsymbol{\sigma}_{(k)}).$$

Notice that $\boldsymbol{\beta}_k$ and ε_k are closely related to the precision matrix $\boldsymbol{\Theta}$, i.e.

$$\theta_{kk} = 1/\operatorname{Var}(\varepsilon_k)$$

and

$$\boldsymbol{\theta}_{(k)} = -\boldsymbol{\beta}_k/\operatorname{Var}(\varepsilon_k).$$

Thus for the k-th variable, $\boldsymbol{\theta}_{(k)}$ and $\boldsymbol{\beta}_k$ share exactly the same sparsity pattern. Following Meinshausen and Bühlmann (2006), the oracle neighborhood LASSO selection obtains the solution $\hat{\boldsymbol{\beta}}_k^o$ from the following LASSO penalized least squares problem,

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^{p-1}} \frac{1}{n}\sum_{i=1}^{n}(z_{ik} - \boldsymbol{z}_{i(k)}^T\boldsymbol{\beta})^2 + \lambda\|\boldsymbol{\beta}\|_{\ell_1}, \tag{3.3}$$

and then the sparsity pattern of $\boldsymbol{\Theta}$ can be estimated by integrating the neighborhood support set of $\hat{\boldsymbol{\beta}}_k^o = (\hat{\beta}_{kj}^o)_{j\neq k}$, i.e. $\hat{ne}_k = \{j : \ \hat{\beta}_{kj}^o \neq 0\}$ via intersection or union.

We notice the fact that

$$\frac{1}{n}\sum_{i=1}^{n}(z_{ik} - \boldsymbol{z}_{i(k)}^T\boldsymbol{\beta})^2 = \boldsymbol{\beta}^T\hat{\boldsymbol{\Sigma}}_{(k)}^o\boldsymbol{\beta} - 2\boldsymbol{\beta}^T\hat{\boldsymbol{\sigma}}_{(k)}^o + \hat{\sigma}_{kk}^o$$

Then (3.3) can be written in the following equivalent form

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^{p-1}} \boldsymbol{\beta}^T\hat{\boldsymbol{\Sigma}}^o_{(k)}\boldsymbol{\beta} - 2\boldsymbol{\beta}^T\hat{\boldsymbol{\sigma}}^o_{(k)} + \lambda\|\boldsymbol{\beta}\|_{\ell_1}. \tag{3.4}$$

- **The oracle neighborhood Dantzig selector.** Following Yuan (2010), the LASSO penalized least squares in (3.3) in the neighborhood approach can be replaced with the Dantzig selector as follows:

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^{p-1}} \|\boldsymbol{\beta}\|_{\ell_1} \quad \text{subject to} \quad \|\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{z}_{i(k)}(\boldsymbol{z}^T_{i(k)}\boldsymbol{\beta} - z_{ik})\|_{\ell_\infty} \leq \lambda. \tag{3.5}$$

Then the sparsity pattern of $\boldsymbol{\Theta}$ can be similarly estimated by integration via intersection or union. Furthermore, we notice that

$$\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{z}_{i(k)}(\boldsymbol{z}^T_{i(k)}\boldsymbol{\beta} - z_{ik}) = \hat{\boldsymbol{\Sigma}}^o_{(k)}\boldsymbol{\beta} - \hat{\boldsymbol{\sigma}}^o_{(k)}.$$

Then (3.5) can be written in the following equivalent form

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^{p-1}} \|\boldsymbol{\beta}\|_{\ell_1} \quad \text{subject to} \quad \|\hat{\boldsymbol{\Sigma}}^o_{(k)}\boldsymbol{\beta} - \hat{\boldsymbol{\sigma}}^o_{(k)}\|_{\ell_\infty} \leq \lambda. \tag{3.6}$$

- **The oracle CLIME.** Following Cai et al. (2011) we can estimate sparse precision matrices by solving a constrained ℓ_1 minimization problem:

$$\arg\min_{\boldsymbol{\Theta}} \|\boldsymbol{\Theta}\|_1 \quad \text{subject to} \quad \|\hat{\boldsymbol{\Sigma}}^o\boldsymbol{\Theta} - \boldsymbol{I}\|_{\max} \leq \lambda. \tag{3.7}$$

Cai et al. (2011) compared the CLIME and the Graphical LASSO and show that the CLIME enjoys nice theoretical properties without assuming the irrepresentable condition of Ravikumar et al. (2008) for the Graphical LASSO.

3.2.2 The proposed rank-based estimators

The existing theoretical results in the literature can be directly applied to these oracle estimators. However, the "oracle data" $\boldsymbol{z}_1, \boldsymbol{z}_2, \ldots, \boldsymbol{z}_n$ are unavailable and thus the above-mentioned "oracle" procedures are not genuine estimators. Naturally we wish to construct a genuine estimator that can mimic the oracle estimator. To this end, we can derive an alternative estimator of $\boldsymbol{\Sigma}$ based on the actual data $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n$ and then feed this genuine covariance estimator to the Graphical LASSO, the neighborhood selection or CLIME. To implement this natural idea, we propose a rank-based estimation scheme. Note that $\boldsymbol{\Sigma}$ can be viewed as the correlation matrix as well, i.e. $\sigma_{ij} = \mathrm{corr}(\boldsymbol{z}_i, \boldsymbol{z}_j)$. Let $(x_{1i}, x_{2i}, \ldots, x_{ni})$ be the observed values of variable X_i. We convert them to ranks denoted by $\boldsymbol{r}_i = (r_{1i}, r_{2i}, \ldots, r_{ni})$. Spearman's rank correlation $\hat{r}_{ij}$ is defined as Pearson's correlation between $\boldsymbol{r}_i$ and $\boldsymbol{r}_j$, i.e. $\hat{r}_{ij} = \mathrm{corr}(\boldsymbol{r}_i, \boldsymbol{r}_j)$. Spearman's rank correlation is a nonparametric measure of dependence between two variables. It is important to note that $\boldsymbol{r}_i$ are the ranks of the "oracle" data. Therefore, $\hat{r}_{ij}$ is also identical to the Spearman's rank correlation between the "oracle" variables Z_i, Z_j. In other words, in the framework of rank-based estimation, we can treat the observed data as the "oracle" data and avoid estimating p nonparametric transformation functions.

The nonparanormal model implies that (Z_i, Z_j) follows a bivariate normal distribution with correlation parameter σ_{ij}. Then a classical result due to Kendall (1948) shows the relationship between σ_{ij} and $\hat{r}_{ij}$ is as follows

$$\lim_{n \to +\infty} \mathrm{E}(\hat{r}_{ij}) = \frac{6}{\pi} \arcsin(\frac{1}{2}\sigma_{ij}), \tag{3.8}$$

which indicates that $\hat{r}_{ij}$ is a biased estimator of σ_{ij}. To correct the bias, Kendall

(1948) suggested using the adjusted Spearman's rank correlation

$$\hat{r}_{ij}^s = 2\sin(\frac{\pi}{6}\hat{r}_{ij}). \tag{3.9}$$

Combining (3.8) and (3.9) we see that $\hat{r}_{ij}^s$ is an asymptotically unbiased estimator of σ_{ij}. Naturally we define the rank-based sample estimate of $\mathbf{\Sigma}$ as follows

$$\hat{\boldsymbol{R}}^s = (\hat{r}_{ij}^s)_{1\le i,j\le p}.$$

In Section 3 we show $\hat{\boldsymbol{R}}^s$ is a good estimator of $\mathbf{\Sigma}$. Then we naturally come up with the following rank-based estimators of $\mathbf{\Theta}$ by using the Graphical LASSO, the neighborhood Dantzig selector and CLIME:

- **The rank-based Graphical LASSO:**

$$\hat{\mathbf{\Theta}}_g^s = \arg\min_{\mathbf{\Theta}\succ 0} -\log\det(\mathbf{\Theta}) + \mathrm{tr}(\hat{\boldsymbol{R}}^s\mathbf{\Theta}) + \lambda\sum_{i\neq j}|\theta_{ij}|. \tag{3.10}$$

- **The rank-based neighborhood Dantzig selector:** A rank-based estimate of $\boldsymbol{\beta}_k$ can be solved by

$$\hat{\boldsymbol{\beta}}_k^{s.nd} = \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^{p-1}} \|\boldsymbol{\beta}\|_{\ell_1} \quad \text{subject to} \quad \|\hat{\boldsymbol{R}}_{(k)}^s\boldsymbol{\beta} - \hat{\boldsymbol{r}}_{(k)}^s\|_{\ell_\infty} \le \lambda. \tag{3.11}$$

  The support of $\mathbf{\Theta}$ can be estimated from the support of $\hat{\boldsymbol{\beta}}_1^s, \cdots, \hat{\boldsymbol{\beta}}_p^s$ via integration by union or intersection as in Meinshausen and Bühlmann (2006). We can also construct the rank-based precision matrix estimator

$$\hat{\mathbf{\Theta}}_{nd}^s = (\hat{\theta}_{ij}^{s.nd})_{1\le i,j\le p}$$

with

$$\hat{\theta}_{kk}^{s.nd} = \left((\hat{\boldsymbol{\beta}}_k^{s.nd})^T \hat{\boldsymbol{R}}_{(k)}^s \hat{\boldsymbol{\beta}}_k^{s.nd} - 2(\hat{\boldsymbol{\beta}}_k^{s.nd})^T \hat{\boldsymbol{r}}_{(k)}^{s.nd} + 1 \right)^{-1}$$

and

$$\hat{\boldsymbol{\theta}}_{(k)}^{s.nd} = -\hat{\theta}_{kk}^{s.nd} \hat{\boldsymbol{\beta}}_k^{s.nd}$$

for $k = 1, \ldots, p$. We can symmetrize $\hat{\boldsymbol{\Theta}}_{nd}^s$ by solving the following optimization problem (Yuan, 2010),

$$\check{\boldsymbol{\Theta}}_{nd}^s = \arg \min_{\boldsymbol{\Theta}:\ \boldsymbol{\Theta}=\boldsymbol{\Theta}'} \|\boldsymbol{\Theta} - \hat{\boldsymbol{\Theta}}_{nd}^s\|_{\ell_1}.$$

We also consider using the adaptive Dantzig selector in the rank-based neighborhood estimation in order to achieve better graphical model selection performance. See Section 3.3.2 for more details.

- **The rank-based CLIME:**

$$\hat{\boldsymbol{\Theta}}_c^s = \arg \min_{\boldsymbol{\Theta}} \|\boldsymbol{\Theta}\|_1 \quad \text{subject to} \quad \|\hat{\boldsymbol{R}}^s \boldsymbol{\Theta} - \boldsymbol{I}\|_{\max} \leq \lambda. \tag{3.12}$$

By Lemma 1 in Cai et al. (2011) the above optimization problem can be further decomposed into p subproblems of vector minimization, i.e. for $k = 1, \ldots, p$,

$$\hat{\boldsymbol{\theta}}_k^{s.c} = \arg \min_{\boldsymbol{\theta} \in \mathbb{R}^p} \|\boldsymbol{\theta}\|_{\ell_1} \quad \text{subject to} \quad \|\hat{\boldsymbol{R}}^s \boldsymbol{\theta} - \boldsymbol{e}_k\|_{\ell_\infty} \leq \lambda, \tag{3.13}$$

where $\boldsymbol{e}_k$'s are the natural basis in $\mathbb{R}^p$. Then $\hat{\boldsymbol{\Theta}}_c^s$ is exactly equivalent to $(\hat{\boldsymbol{\theta}}_1^{s.c}, \cdots, \hat{\boldsymbol{\theta}}_p^{s.c})$. Note that $\hat{\boldsymbol{\Theta}}_c^s$ could be asymmetric. Following Cai et al. (2011)

we consider

$$\breve{\mathbf{\Theta}}_c^s = (\breve{\theta}_{ij}^{s.c})_{1\le i,j\le p}$$

with

$$\breve{\theta}_{ij}^{s.c} = \hat{\theta}_{ij}^{s.c} I_{\{|\hat{\theta}_{ij}^{s.c}|\le|\hat{\theta}_{ji}^{s.c}|\}} + \hat{\theta}_{ji}^{s.c} I_{\{|\hat{\theta}_{ij}^{s.c}|>|\hat{\theta}_{ji}^{s.c}|\}}.$$

In the original CLIME paper Cai et al. (2011) proposed to use hard thresholding for graphical model selection. Borrowing the basic idea from the adaptive LASSO (Zou, 2006), we propose an adaptive version of the rank-based CLIME in order to achieve better graphical model selection performance. See Section 3.3.3 for more details.

We would like to point out that using ranks of the raw data is a fundamental idea in the classical nonparametric statistics. The rank-based methods has been widely considered in several fields of statistics including hypothesis testing, point and interval estimates, and various simultaneous inference procedures, for example, the Friedman's test in analysis of variance, and the Wilcoxon signed-rank test. We refer Lehmann (1998) for more discussions about the rank-based methods. In this chapter, we show that this classical idea is still powerful for an interesting high-dimensional statistical problem.

3.2.3 Rank-based neighborhood LASSO?

One might consider the rank-based neighborhood LASSO defined as follows:

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^{p-1}} \boldsymbol{\beta}^T \hat{\boldsymbol{R}}_{(k)}^s \boldsymbol{\beta} - 2\boldsymbol{\beta}^T \hat{\boldsymbol{r}}_{(k)}^s + \lambda\|\boldsymbol{\beta}\|_{\ell_1}. \tag{3.14}$$

However there is a technical problem for the above definition. The Spearman's rank correlation matrix $\hat{\boldsymbol{R}}$ is always positive semidefinite, but $\hat{\boldsymbol{R}}^s$ could become indefinite after the unbiasedness adjustment. To our best knowledge, Devlin et al. (1975) was the first to point out the indefinite issue of the estimated rank correlation matrix. Here we also use a toy example to illustrate this point. Consider the 3×3 correlation matrix

$$\boldsymbol{A} = \begin{pmatrix} 1 & 0.7 & 0 \\ 0.7 & 1 & 0.7 \\ 0 & 0.7 & 1 \end{pmatrix}.$$

Note that $\boldsymbol{A}$ is positive-definite with eigenvalues 1.99, 1.00 and 0.01, but $2\sin(\frac{\pi}{6}\boldsymbol{A})$ becomes indefinite with eigenvalues 2.01, 1.00 and -0.01. The negative eigenvalues will make (3.14) an ill-defined optimization problem. Fortunately, the positive definite issue does not cause any problem for the Graphical LASSO, Dantzig selector and CLIME. Notice that the diagonal elements of $\hat{\boldsymbol{R}}^s$ are obviously strictly positive, and thus Lemma 3 in Ravikumar et al. (2008) suggests that the rank-based graphical lasso always has a unique positive definite solution for any regularization parameter $\lambda > 0$. The rank-based neighborhood Dantzig selector and the rank-based CLIME are still well-defined even when $\hat{\boldsymbol{R}}^s_{(k)}$ becomes indefinite, and the according optimization algorithms also tolerate the indefiniteness of $\hat{\boldsymbol{R}}^s_{(k)}$.

We can construct a hybrid neighborhood estimator by using both the LASSO and Dantzig selector. We can always first check whether $\hat{\boldsymbol{R}}^s$ is positive definite or not. If it is positive definite, we perform the rank-based neighborhood LASSO. If not, we perform the rank-based neighborhood Dantzig selector instead. We have tried this hybrid estimator in numerical studies and found that it works very similarly to the rank-based neighborhood Dantzig selector. This can be understood by the fact that the LASSO penalized least squares and Dantzig selector in generally work very similarly (Bickel et al., 2009; Efron et al., 2007; James et al., 2009). Due to space

consideration we do not present the hybrid method in this chapter.

3.3 Theoretical Properties

For a vector $\boldsymbol{v} = (v_1, \ldots, v_l)$, let $\|\boldsymbol{v}\|_{\min}$ denote the minimum absolute value, i.e. $\|\boldsymbol{v}\|_{\min} = \min_j |v_j|$. For a matrix $\boldsymbol{A} = (a_{ij})$, we define the following matrix norms: the matrix ℓ_1 norm $\|\boldsymbol{A}\|_{\ell_1} = \max_j \sum_i |a_{ij}|$, the matrix ℓ_∞ norm $\|\boldsymbol{A}\|_{\ell_\infty} = \max_i \sum_j |a_{ij}|$, and the Frobenius norm $\|\boldsymbol{A}\|_F = (\sum_{(i,j)} a_{ij}^2)^{1/2}$. For any symmetric matrix, its matrix ℓ_1 norm coincides its matrix ℓ_∞ norm. Denote by $\lambda_{\min}(\boldsymbol{A})$ and $\lambda_{\max}(\boldsymbol{A})$ the smallest and largest eigenvalues of $\boldsymbol{A}$ respectively. Define $\boldsymbol{\Sigma}^*$ as the true covariance matrix, and let $\boldsymbol{\Theta}^*$ be its inverse. Let $\mathcal{A}$ be the support set of $\boldsymbol{\Theta}^*$, and denote $\mathcal{A}^c$ as the complement of $\mathcal{A}$. Let $d = \max_j \sum_{i:\ i \neq j} I_{\{\theta_{ij}^* \neq 0\}}$ be the maximal degree over the underlying graph corresponding to $\boldsymbol{\Theta}^*$, and let $s = \sum_{(i,j):\ i \neq j} I_{\{\theta_{ij}^* \neq 0\}}$ be the total degree over the whole graph.

In this section we establish the rate of convergence and the graphical model selection consistency for the proposed rank-based estimators. The main conclusion drawn from these theoretical results is that the rank-based graphical lasso, neighborhood Dantzig selector and CLIME work as well as their oracle counterparts in terms of the rates of convergence and graphical model selection consistency. To this end, we first provide useful concentration bounds concerning the accuracy of the rank-based sample correlation matrix estimator $\hat{\boldsymbol{R}}^s$.

Lemma 3.1 For $0 < \varepsilon < 1$ and $n \geq \frac{12\pi}{\varepsilon}$, there exits some absolute constant $c_0 > 0$ such that we have the following concentration bounds

$$\begin{aligned} \Pr(|\hat{r}_{ij}^s - \sigma_{ij}| > \varepsilon) &\leq 2\exp(-c_0 n\varepsilon^2); \\ \Pr(\|\hat{\boldsymbol{R}}^s - \boldsymbol{\Sigma}\|_{\max} > \varepsilon) &\leq p^2 \exp(-c_0 n\varepsilon^2). \end{aligned}$$

□

3.3.1 On the rank-based Graphical LASSO

Denote by $\psi_{\min} = \min_{(i,j)\in\mathcal{A}} |\theta^*_{ij}|$ the minimal entry of $\mathbf{\Theta}^*$ in the absolute scale. Define $K_{\mathbf{\Sigma}^*} = \|\mathbf{\Sigma}^*_{\mathcal{A}\mathcal{A}}\|_{\ell_\infty}$ and $K_{\boldsymbol{H}^*} = \|(\boldsymbol{H}^*_{\mathcal{A}\mathcal{A}})^{-1}\|_{\ell_\infty}$. For notation convenience, we define $\boldsymbol{H}^*$ as the Kronecker product $\mathbf{\Sigma}^* \bigotimes \mathbf{\Sigma}^*$.

Theorem 3.1 Assume that $\|\boldsymbol{H}^*_{\mathcal{A}^c\mathcal{A}}(\boldsymbol{H}^*_{\mathcal{A}\mathcal{A}})^{-1}\|_{\ell_\infty} < 1-\kappa$ for $\kappa \in (0,1)$.

(a) Rate of convergence: if the regularization parameter λ is chosen such that

$$\lambda < \frac{1}{6(1+\frac{\kappa}{4})K_{\mathbf{\Sigma}^*}K_{\boldsymbol{H}^*}\max\{1,(1+\frac{4}{\kappa})K^2_{\mathbf{\Sigma}^*}K_{\boldsymbol{H}^*}\}} \cdot \frac{1}{d},$$

with probability at least $1-p^2\exp(-\frac{\kappa^2}{16}c_0 n\lambda^2)$, the rank-based *Graphical LASSO* estimator $\hat{\mathbf{\Theta}}^s_g$ satisfies that $\hat{\theta}^{s.g}_{ij} = 0$ for any $(i,j) \in \mathcal{A}^c$ and moreover

$$\|\hat{\mathbf{\Theta}}^s_g - \mathbf{\Theta}^*\|_{\max} \le 2K_{\boldsymbol{H}^*}(1+\frac{\kappa}{4})\lambda.$$

(b) Graphical model selection consistency: picking a regularization parameter λ to satisfy that

$$\lambda < \min\left\{\frac{\psi_{\min}}{2(1+\frac{\kappa}{4})K_{\boldsymbol{H}^*}}, \frac{1}{6(1+\frac{\kappa}{4})K_{\mathbf{\Sigma}^*}K_{\boldsymbol{H}^*}\max\{1,(1+\frac{4}{\kappa})K^2_{\mathbf{\Sigma}^*}K_{\boldsymbol{H}^*}\}} \cdot \frac{1}{d}\right\},$$

then with probability at least $1-p^2\exp(-\frac{\kappa^2}{16}c_0 n\lambda^2)$, $\hat{\mathbf{\Theta}}^s_g$ is sign consistent satisfying that $\mathrm{sign}(\hat{\theta}^{s.g}_{ij}) = \mathrm{sign}(\theta^*_{ij})$ for any $(i,j) \in \mathcal{A}$ and $\hat{\theta}^{s.g}_{ij} = 0$ for any $(i,j) \in \mathcal{A}^c$.□

In Theorem 3.1, the condition that $\|\boldsymbol{H}^*_{\mathcal{A}^c\mathcal{A}}(\boldsymbol{H}^*_{\mathcal{A}\mathcal{A}})^{-1}\|_{\ell_\infty} < 1-\kappa$ is also referred to as the *irrepresentable condition* for studying the theoretical properties of the Graphical LASSO (Ravikumar et al., 2008). We can obtain a more straightforward understanding of Theorem 3.1 by considering its asymptotic consequences.

Corollary 3.1 Suppose there is a constant $\kappa \in (0,1)$ such that

$$\|\boldsymbol{H}^*_{\mathcal{A}^c\mathcal{A}}(\boldsymbol{H}^*_{\mathcal{A}\mathcal{A}})^{-1}\|_{\ell_\infty} < 1-\kappa.$$

Assume that $\psi_{\min}$, $K_{\boldsymbol{\Sigma}^*}$ and $K_{\boldsymbol{H}^*}$ are all fixed constants.

(a) Rates of convergence: further assume $n \gg d^2 \log p$, and pick λ such that

$$\frac{1}{d} \gg \lambda = O(\sqrt{\frac{\log p}{n}}).$$

Then we have

$$\|\hat{\boldsymbol{\Theta}}^s_g - \boldsymbol{\Theta}^*\|_{\max} = O_P(\sqrt{\frac{\log p}{n}}).$$

Furthermore, the convergence rates in both Frobenius and matrix ℓ_1-norms can also be obtained as follows,

$$\begin{aligned} \|\hat{\boldsymbol{\Theta}}^s_g - \boldsymbol{\Theta}^*\|_F &= O_P(\sqrt{\frac{(s+p)\log p}{n}}); \\ \|\hat{\boldsymbol{\Theta}}^s_g - \boldsymbol{\Theta}^*\|_{\ell_1} &= O_P(\sqrt{\frac{\min\{s+p, d^2\}\log p}{n}}). \end{aligned}$$

(b) Graphical model selection consistency: further assume $n \gg d^2 \log p$, and pick a regularization parameter λ satisfying that

$$\frac{1}{d} \gg \lambda = O(\sqrt{\frac{\log p}{n}}).$$

Then we have the sign consistency that $\text{sign}(\hat{\theta}^{s.g}_{ij}) = \text{sign}(\theta^*_{ij})$ for any $(i,j) \in \mathcal{A}$ and $\text{sign}(\hat{\theta}^{s.g}_{ij}) = 0$ for any $(i,j) \in \mathcal{A}^c$. □

Under the same conditions of Theorem 3.1 and Corollary 3.1, by the results in Ravikumar et al. (2008), we know that the conclusions Theorem 3.1 and Corollary 3.1 hold for the oracle Graphical LASSO. In other words, the rank-based Graphical LASSO estimator is comparable to its oracle counterpart in terms of rates of convergence and graphical model selection consistency.

3.3.2 On the rank-based neighborhood Dantzig selector

Define $b = \min_k \theta^*_{kk}$, $B = \lambda_{max}(\mathbf{\Theta}^*)$ and $M = \|\mathbf{\Theta}^*\|_{\ell_1}$. For each variable X_k, define the corresponding active set $\mathcal{A}_k = \{j \neq k : \theta^*_{kj} \neq 0\}$ with the maximal cardinality $d = \max_k |\mathcal{A}_k|$. Then we can organize $\boldsymbol{\theta}^*_{(k)}$ and $\mathbf{\Theta}^*_{(k)}$ with respect to $\mathcal{A}_k$ as

$$\boldsymbol{\theta}^*_{(k)} = (\boldsymbol{\theta}^*_{\mathcal{A}_k}, \boldsymbol{\theta}^*_{\mathcal{A}_k^c})$$

and

$$\mathbf{\Theta}^*_{(k)} = \begin{pmatrix} \mathbf{\Theta}^*_{\mathcal{A}_k\mathcal{A}_k} & \mathbf{\Theta}^*_{\mathcal{A}_k\mathcal{A}_k^c} \\ \mathbf{\Theta}^*_{\mathcal{A}_k^c\mathcal{A}_k} & \mathbf{\Theta}^*_{\mathcal{A}_k^c\mathcal{A}_k^c} \end{pmatrix}.$$

In the similar way, we can partition $\boldsymbol{\sigma}^*_{(k)}$ and $\mathbf{\Sigma}^*_{(k)}$ with respect to $\mathcal{A}_k$ as well.

Theorem 3.2 Pick the λ such that $d\lambda = o(1)$ and $bn\lambda \geq 12\pi M$. With a probability at least $1 - p^2 \exp(-c_0 \frac{b^2}{M^2} n\lambda^2)$, there exists some quantity $C_{b,B,M} > 0$ depending on b, B and M only such that

$$\|\breve{\mathbf{\Theta}}^s_{nd} - \mathbf{\Theta}^*\|_{\ell_1} \leq \|\hat{\mathbf{\Theta}}^s_{nd} - \mathbf{\Theta}^*\|_{\ell_1} \leq C_{b,B,M} d\lambda. \qquad \square$$

Corollary 3.2 Assume that $n \gg d^2 \log p$, and pick a tuning parameter λ such that

$$\frac{1}{d} \gg \lambda = O(\sqrt{\frac{\log p}{n}}).$$

Suppose both b, B and M are fixed constants. Then we have

$$\|\breve{\mathbf{\Theta}}^s_{nd} - \mathbf{\Theta}^*\|_{\ell_1} = O_P(d\sqrt{\frac{\log p}{n}});$$

$$\|\hat{\mathbf{\Theta}}^s_{nd} - \mathbf{\Theta}^*\|_{\ell_1} = O_P(d\sqrt{\frac{\log p}{n}}). \qquad \square$$

Yuan (2010) established the rates of convergence of the neighborhood Dantzig selector under the ℓ_1 norm, which can be directly applied to the oracle neighborhood Dantzig selector under the nonparanormal model. Comparing Theorem 3.2 and Corollary 3.2 to the results in Yuan (2010), we see that the rank-based neighborhood Dantzig selector and the oracle neighborhood Dantzig selector achieve the same rates of convergence.

The Dantzig selector and the LASSO are closely related (Bickel et al., 2009; Efron et al., 2007; James et al., 2009). Similar to the LASSO, the Dantzig selector tends to over-select. Zou (2006) proposed the adaptive weighting idea to develop the adaptive LASSO which improves the selection performance of the LASSO and corrects its bias too. The very same idea can be used to improve the selection performance of Dantzig selector which leads to the adaptive Dantzig selector (Dicker and Lin, 2009). We can easily extend the rank-based neighborhood Dantzig selector to the rank-based neighborhood adaptive Dantzig selector. Given adaptive weights $\boldsymbol{w}_k$, consider

$$\hat{\boldsymbol{\beta}}^{s.nad}_k = \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^{p-1}} \|\boldsymbol{w}_k \circ \boldsymbol{\beta}\|_{\ell_1} \quad \text{subject to} \quad |\hat{\boldsymbol{R}}^s_{(k)}\boldsymbol{\beta} - \hat{\boldsymbol{r}}^s_{(k)}| \le \lambda \boldsymbol{w}_k, \tag{3.15}$$

where $\circ$ denotes the Hadamard product, and $\boldsymbol{a}_{d\times 1} \leq \boldsymbol{b}_{d\times 1}$ denotes the set of entry-wise inequalities $a_i \leq b_i$ $(1 \leq i \leq d)$ for ease of notation. In both our theoretical analysis and numerical implementation, we utilize the optimal solution $\hat{\boldsymbol{\beta}}_k^{s.nd}s$ from the rank-based Dantzig selector to construct the adaptive weights $\boldsymbol{w}_k$ by

$$\boldsymbol{w}_k^d = (|\hat{\boldsymbol{\beta}}_k^{s.nd}| + \frac{1}{n})^{-1}. \tag{3.16}$$

Define $\boldsymbol{\beta}^*_{\mathcal{A}_k} = (\boldsymbol{\Theta}^*_{\mathcal{A}_k\mathcal{A}_k})^{-1}\boldsymbol{\theta}^*_{\mathcal{A}_k}$, and let $\boldsymbol{\beta}^*_k = (\boldsymbol{\beta}^*_{\mathcal{A}_k},\ \mathbf{0})$. Thus the support of $\boldsymbol{\beta}^*_k$ exactly coincides with that of $\boldsymbol{\theta}^*_{(k)}$, and then it is further equivalent to the active set $\mathcal{A}_k$. Define $\psi_k = \|\boldsymbol{\beta}^*_{\mathcal{A}_k}\|_{\min}$, $G_k = \|(\boldsymbol{\Sigma}^*_{\mathcal{A}_k\mathcal{A}_k})^{-1}\|_{\ell_\infty}$ and $H_k = \|\boldsymbol{\Sigma}^*_{\mathcal{A}_k^c\mathcal{A}_k}(\boldsymbol{\Sigma}^*_{\mathcal{A}_k\mathcal{A}_k})^{-1}\|_{\ell_\infty}$.

Theorem 3.3 For each $k = 1, 2, \cdots, p$, we pick $\lambda = \lambda_{dantzig}$ as in (3.11) such that

$$\lambda_{dantzig} \geq \frac{12\pi M}{bn} \quad \text{and} \quad o(1) = d\lambda_{dantzig} \leq \min\left\{\frac{\psi_k}{2C_0},\ \frac{1}{4C_0 d(\psi_k + 2G_k)} - \frac{1}{C_0 n}\right\},$$

where $C_0 = 4B^2(2 + \frac{b}{M})$. We further pick $\lambda = \lambda_{adantzig}$ as in (3.15) such that

$$\frac{\psi_k^2}{8G_k} \geq \lambda_{adantzig} \geq \max\left\{\frac{12\pi}{n},\ (C_0 d\lambda_{dantzig} + \frac{1}{n}) \cdot \frac{H_k\psi_k}{G_k}\right\}.$$

and

$$o(1) = d\lambda_{adantzig} \leq \min\left\{\lambda_{\min}(\boldsymbol{\Sigma}^*_{\mathcal{A}_k\mathcal{A}_k}),\ \frac{1}{2G_k},\ \frac{\psi_k}{8G_k(\psi_k + G_k)}\right\},$$

Then we choose $\boldsymbol{w}_k = \boldsymbol{w}_k^d$ as in (3.16) for any k. With a probability at least $1 - p^2\exp(-c_0 n \cdot \min(\lambda^2_{adantzig}, \frac{b^2}{M^2}\lambda^2_{dantzig}))$, for each $k = 1, 2, \cdots, p$, the rank-based adaptive Dantzig selector finds the unique optimal solution $\hat{\boldsymbol{\beta}}_k^{s.nad} = (\hat{\boldsymbol{\beta}}_{\mathcal{A}_k}^{s.nad}, \hat{\boldsymbol{\beta}}_{\mathcal{A}_k^c}^{s.nad})$ with $\text{sign}(\hat{\boldsymbol{\beta}}_{\mathcal{A}_k}^{s.nad}) = \text{sign}(\boldsymbol{\beta}^*_{\mathcal{A}_k})$ and $\hat{\boldsymbol{\beta}}_{\mathcal{A}_k^c}^{s.nad} = \mathbf{0}$, and thus the rank-based neighborhood adaptive Dantzig selector is sign consistent for the Graphical model selection. $\square$

Corollary 3.3 Assume that b, B, M, ψ_k, G_k and H_k $(1 \leq k \leq p)$ are all fixed constants, and also assume that $n \gg d^4 \log p$. Suppose that for each k,

$$\lambda_{\min}(\mathbf{\Sigma}^*_{\mathcal{A}_k\mathcal{A}_k}) \gg d^2\sqrt{\frac{\log p}{n}},$$

and then we pick the regularization parameters $\lambda_{dantzig}$ and $\lambda_{adantzig}$ such that

$$\frac{1}{d} \geq \lambda_{dantzig} = O(\sqrt{\frac{\log p}{n}}),$$

and

$$\min\left\{\frac{1}{d}\cdot\lambda_{\min}(\mathbf{\Sigma}^*_{\mathcal{A}_k\mathcal{A}_k}),\ \frac{1}{d}\right\} \gg \lambda_{adantzig} \gg d\lambda_{dantzig} = O(d\sqrt{\frac{\log p}{n}}).$$

Then with probability tending to 1, for each $k = 1, 2, \cdots, p$, the rank-based adaptive Dantzig selector with $\boldsymbol{w}_k = \boldsymbol{w}_k^d$ as in (3.16) finds the unique optimal solution of (3.15) to be $\hat{\boldsymbol{\beta}}_k^{s.nad} = (\hat{\boldsymbol{\beta}}_{\mathcal{A}_k}^{s.nad}, \hat{\boldsymbol{\beta}}_{\mathcal{A}_k^c}^{s.nad})$ with $\text{sign}(\hat{\boldsymbol{\beta}}_{\mathcal{A}_k}^{s.nad}) = \text{sign}(\boldsymbol{\beta}^*_{\mathcal{A}_k})$ and $\hat{\boldsymbol{\beta}}_{\mathcal{A}_k^c}^{s.nad} = \mathbf{0}$, and thus the rank-based neighborhood adaptive Dantzig selector is sign consistent for the Graphical model selection.. □

Our treatment of the adaptive Dantzig selector is fundamentally different from Dicker and Lin (2009). Dicker and Lin (2009) focused on the classical linear regression model and constructed the adaptive weights as inverse of the absolute values of ordinary least square estimator. Their theoretical results only hold in the classical setting with a fixed p as n go to infinity. In our problem p can be much bigger than n. The choice of adaptive weights in (3.16) plays a critical role to establish the graphical model selection consistency for the adaptive Dantzig selector under the high-dimensional setting, where p is at a nearly exponential rate to n. Our technical analysis uses some key ideas such as the strong duality and the complementary

slackness from the linear optimization theory (Bertsimas and Tsitsiklis, 1997; Boyd and Vandenberghe, 2004).

3.3.3 On the rank-based CLIME

Compared to the graphical lasso, the CLIME can enjoy nice theoretical properties without assuming the irrepresentable condition (Cai et al., 2011). This continues to hold when comparing the rank-based graphical lasso and the rank-based CLIME.

Theorem 3.4 Recall that $M = \|\mathbf{\Theta}^*\|_{\ell_1}$. Pick a λ such that $n\lambda \geq 12\pi M$. With a probability at least $1 - p^2 \exp(-\frac{c_0}{M^2} n\lambda^2)$, we have

$$\|\hat{\mathbf{\Theta}}_c^s - \mathbf{\Theta}^*\|_{\max} \leq 2M\lambda.$$

Moreover, assume that $n \gg d^2 \log p$ and suppose M is a fixed constant. Pick a regularization parameter $\lambda = O(\sqrt{\frac{\log p}{n}})$. Then we have

$$\|\hat{\mathbf{\Theta}}_c^s - \mathbf{\Theta}^*\|_{\max} = O_P(\sqrt{\frac{\log p}{n}}). \qquad \square$$

Theorem 3.4 is parallel to Theorem 6 in Cai et al. (2011) which can be used to establish the rate of convergence of the oracle CLIME.

To improve graphical model selection performance, Cai et al. (2011) suggested an additional thresholding step by applying the entrywise hard-thresholding rule to $\hat{\mathbf{\Theta}}_c^s$:

$$HT(\hat{\mathbf{\Theta}}_c^s) = (\hat{\theta}_{ij}^{s.c} \cdot I_{\{|\hat{\theta}_{ij}^{s.c}| \geq \tau_n\}})_{1 \leq i,j \leq p}. \tag{3.17}$$

where $\tau_n \geq 2M\lambda$ is the threshold, and λ is given in Theorem 3.4. To establish the graphical model selection consistency, Theorem 7 in Cai et al. (2011) requires

$$\tau_n < \min_{(i,j)\in\mathcal{A}} \frac{1}{2}|\theta^*_{ij}|.$$

We could apply the thresholding idea to the rank-based CLIME. However, we prefer to develop an adaptive version of the rank-based CLIME by using the adaptive penalization idea in the adaptive LASSO and the adaptive Dantzig selector.

Given an adaptive weight matrix $\boldsymbol{W}$ we define the rank-based adaptive CLIME as follows:

$$\hat{\boldsymbol{\Theta}}^{s}_{ac} = \arg\min_{\boldsymbol{\Theta}} \|\boldsymbol{W} \circ \boldsymbol{\Theta}\|_1 \quad \text{subject to} \quad |\hat{\boldsymbol{R}}^{s}\boldsymbol{\Theta} - \boldsymbol{I}| \leq \lambda \boldsymbol{W}, \tag{3.18}$$

where $\boldsymbol{A}_{p\times p} \leq \boldsymbol{B}_{p\times p}$ is a simplified expression for the set of inequalities $a_{ij} \leq b_{ij}$ for all $1 \leq i, j \leq p$. Write $\boldsymbol{W} = (\boldsymbol{w}_1, \cdots, \boldsymbol{w}_p)$. By Lemma 1 in Cai et al. (2011) the above linear programming problem in (3.18) is exactly equivalent to p vector minimization subproblems:

$$\hat{\boldsymbol{\theta}}^{s.ac}_{k} = \arg\min_{\boldsymbol{\theta}\in\mathbb{R}^p} \|\boldsymbol{w}_k \circ \boldsymbol{\theta}\|_{\ell_1} \quad \text{subject to} \quad |\hat{\boldsymbol{R}}^{s}\boldsymbol{\theta} - \boldsymbol{e}_k| \leq \lambda \boldsymbol{w}_k.$$

for $k = 1, \ldots, p$, and $\hat{\boldsymbol{\Theta}}^{s}_{ac}$ is equal to $(\hat{\boldsymbol{\theta}}^{s.ac}_1, \cdots, \hat{\boldsymbol{\theta}}^{s.ac}_p)$. In both our theory and implementation, we utilize the rank-based CLIME's optimal solution $\hat{\boldsymbol{\Theta}}^{s}_{c}$ to construct an adaptive weight matrix $\boldsymbol{W}$ by

$$\boldsymbol{W}^c = (|\hat{\boldsymbol{\Theta}}^{s}_{c}| + \frac{1}{n})^{-1}. \tag{3.19}$$

We now prove the graphical model selection consistency of the rank-based adaptive CLIME. Denote $\boldsymbol{\Theta}^*$ as $(\boldsymbol{\theta}^*_1, \cdots, \boldsymbol{\theta}^*_p)$ and define $\tilde{\mathcal{A}}_k = \mathcal{A}_k \cup \{k\}$. Then we

can organize $\boldsymbol{\theta}_k^*$ and $\boldsymbol{\Sigma}^*$ with respect to $\tilde{\mathcal{A}}_k$ and $\tilde{\mathcal{A}}_k^c$. For $k = 1, \ldots, p$, we define $\tilde{G}_k = \|(\boldsymbol{\Sigma}^*_{\tilde{\mathcal{A}}_k \tilde{\mathcal{A}}_k})^{-1}\|_{\ell_\infty}$ and $\tilde{H}_k = \|\boldsymbol{\Sigma}^*_{\tilde{\mathcal{A}}_k^c \tilde{\mathcal{A}}_k}(\boldsymbol{\Sigma}^*_{\tilde{\mathcal{A}}_k \tilde{\mathcal{A}}_k})^{-1}\|_{\ell_\infty}$.

Theorem 3.5 Recall $\psi_{\min} = \min_{(i,j)\in\mathcal{A}} |\theta^*_{ij}|$. Pick $\lambda = \lambda_{clime}$ as in (3.12) such that

$$\min\left\{\frac{\psi_{\min}}{4M},\ \frac{1}{4M(\psi_{\min} + 2\tilde{G}_k)d} - \frac{2}{Mn}\right\} \geq \lambda_{clime} \geq \frac{12\pi M}{n} \quad \text{and} \quad d\lambda_{clime} = o(1),$$

and further pick $\lambda = \lambda_{aclime}$ as in (3.18) such that

$$\frac{\psi^2_{\min}}{8\tilde{G}_k} \geq \lambda_{aclime} \geq \max\left\{\frac{12\pi}{n},\ (2M\lambda_{clime} + \frac{1}{n}) \cdot \frac{\tilde{H}_k \psi_{\min}}{\tilde{G}_k}\right\},$$

and

$$o(1) = d\lambda_{aclime} \leq \min\left\{\lambda_{\min}(\boldsymbol{\Sigma}^*_{\mathcal{A}_k \mathcal{A}_k}),\ \frac{1}{2\tilde{G}_k},\ \frac{\psi_{\min}}{4\tilde{G}_k(\psi_{\min} + \tilde{G}_k)}\right\}$$

Besides, we choose the adaptive weights $\boldsymbol{W} = \boldsymbol{W}^c$ as in (3.19). Then the rank-based adaptive CLIME's optimal solution $\hat{\boldsymbol{\Theta}}^s_{ac}$ is sign consistent for the graphical model selection, i.e. $\text{sign}(\hat{\theta}^{s.ac}_{ij}) = \text{sign}(\theta^*_{ij})$ for any $(i,j) \in \mathcal{A}$ and $\text{sign}(\hat{\theta}^{s.ac}_{ij}) = 0$ for any $(i,j) \in \mathcal{A}^c$, with a probability at least $1 - p^2 \exp(-c_0 n \min(\lambda^2_{aclime}, \frac{1}{M^2}\lambda^2_{clime}))$. □

Corollary 3.4 Assume that M, $\psi_{\min}$, $\tilde{G}_k$ and $\tilde{H}_k$ $(1 \leq k \leq p)$ are all fixed constants. Let $n \gg d^2 \log p$. Suppose that for each k, we have

$$\lambda_{\min}(\boldsymbol{\Sigma}^*_{\mathcal{A}_k \mathcal{A}_k}) \gg d\sqrt{\frac{\log p}{n}},$$

and then we pick the regularization parameters λ_{clime} and λ_{aclime} such that

$$\frac{1}{d} \geq \lambda_{clime} = O(\sqrt{\frac{\log p}{n}}),$$

and

$$\min\left\{\frac{1}{d}\cdot\lambda_{\min}(\mathbf{\Sigma}^*_{\mathcal{A}_k\mathcal{A}_k}),\ \frac{1}{d}\right\}\gg\lambda_{aclime}\gg\lambda_{clime}.$$

Then we choose the adaptive weights $\boldsymbol{W}=\boldsymbol{W}^c$ as in (3.19). With probability tending to 1, the rank-based adaptive CLIME's solution $\hat{\mathbf{\Theta}}^s_{ac}$ is sign consistent, i.e. $\text{sign}(\hat{\theta}^{s.ac}_{ij})=\text{sign}(\theta^*_{ij})$ for $(i,j)\in\mathcal{A}$ and $\text{sign}(\hat{\theta}^{s.ac}_{ij})=0$ for $(i,j)\in\mathcal{A}^c$. □

The choice of adaptive weights in (3.19) plays a critical role to establish the graphical model selection consistency for the rank-based adaptive CLIME estimator under the nonparanormal model in the high-dimensional setting. Similarly as the rank-based adaptive Dantzig selector, the rank-based adaptive CLIME does not require the strong ir-representable condition to establish the sparsity recovery property.

3.4 Numerical Properties

In this section we present both simulation studies and real examples to demonstrate the finite sample performance of the proposed rank-based estimators.

3.4.1 Monte-Carlo simulations

In the simulation study, we consider both Gaussian data and nonparanormal data. In models 1–4 we draw n independent samples from $N_p(0,\mathbf{\Sigma})$ with four different $\mathbf{\Theta}$:

Model 1: $\theta_{ii}=1$ and $\theta_{i,i+1}=0.5$;

Model 2: $\theta_{ii}=1$, $\theta_{i,i+1}=0.4$ and $\theta_{i,i+2}=\theta_{i,i+3}=0.2$;

Model 3: Randomly choose 16 nodes to be the hub nodes in $\mathbf{\Theta}$, and each of them connects with 5 distinct nodes with $\mathbf{\Theta}_{ij}=0.2$. Elements, not associated with

hub nodes, are set as 0 in $\mathbf{\Theta}$. The diagonal element σ is chosen similarly as that in the previous model.

Model 4: $\mathbf{\Theta} = \mathbf{\Theta}_0 + \sigma I$, where $\mathbf{\Theta}_0$ is a zero-diagonal symmetric matrix. Each off-diagonal element $\mathbf{\Theta}_{0ij}$ independently follows a point mass $0.99\delta_0 + 0.01\delta_{0.2}$, and the diagonal element σ is set to be the absolute value of the minimal negative eigenvalue of $\mathbf{\Theta}_0$ to ensure the semi-positive-definiteness of $\mathbf{\Theta}$.

Model 1–2 are used in Yuan and Lin (2007), and Model 4 is similarly considered as a random network in Rothman et al. (2008) and Cai et al. (2011). Model 3 is to mimic the hub network. In Model 1b–4b we first generate n independent data from $N_p(0, \mathbf{\Sigma})$ and then transfer the normal data using transformation functions

$$\boldsymbol{g} = [f_1^{-1}, f_2^{-1}, f_3^{-1}, f_4^{-1}, f_5^{-1}, f_1^{-1}, f_2^{-1}, f_3^{-1}, f_4^{-1}, f_5^{-1}, \ldots],$$

where $f_1(x) = x$, $f_2(x) = \log(x)$, $f_3(x) = x^{\frac{1}{3}}$, $f_4(x) = \log(\frac{x}{1-x})$ and

$$f_5(x) = f_2(x)I_{\{x<-1\}} + f_1(x)I_{\{-1\le x\le 1\}} + (f_4(x-1)+1)I_{\{x>1\}}.$$

In all cases we let $n = 300$ and $p = 100$. We summarize the total number of zero entries out of the total $p = k(k-1)/2 = 4950$ off-diagonal entries in the precision matrix for Model 1–4 in Table Table 3.2.

Table 3.2: Summary of the total number of zero entries in the precision matrix for Model 1–4.

Simulation Model	1	2	3	4
# Nonzero Entries	99	390	80	≈50

Table 3.3 summarizes all the estimators investigated in our study. For each estimator, the tuning parameter is chosen by cross-validation. For all neighborhood

approaches (i.e. MB, R-NDS and R-NADS), we summarize the zero patterns by both aggregation strategies via union or intersection, denoted as the suffixes *.au* and *.ai* respectively. Estimation accuracy is measured by the average matrix ℓ_1-norm over 100 replications, and selection accuracy is evaluated by the average false positive/negative.

Table 3.3: List of all estimators in the Monte Carlo simulation study.

Notation	Details
GLASSO	Penalized likelihood estimation via Graphical LASSO
MB	Neighborhood selection via LASSO (Meinshausen and Bühlmann, 2006)
NDS	Neighborhood selection via Dantzig selector
CLIME	Constrained ℓ_1 minimization estimator
LLW	The "plug-in" extension of GLASSO (Liu et al., 2009)
R-GLASSO	Proposed rank-based extension of GLASSO
R-NDS	Proposed rank-based neighborhood selection by Dantzig selector
R-NADS	Proposed rank-based neighborhood selection by adaptive Dantzig selector
R-CLIME	Proposed rank-based extension of CLIME
R-ACLIME	Proposed rank-based adaptive extension of CLIME

The simulation results are summarized in Table 3.4–Table 3.7. First of all, we can see that the Graphical LASSO, neighborhood selection and CLIME do not have satisfactory performance under Models 1b–4b due to the lack of ability to handle non-normality. Second, the three rank-based estimators perform similarly to their oracle counterparts. Note that in Models 1b–4b the oracle Graphical LASSO, the oracle neighborhood Dantzig and the oracle CLIME are actually the Graphical LASSO, the neighborhood Dantzig and the CLIME in Models 1–4. In terms of precision matrix estimation the rank-based CLIME seems to be the best, while the rank-based adaptive CLIME has the best graphical model selection performance.

Table 3.4: Estimation performance in the Gaussian graphical model.

	Gaussian			
	Model 1	Model 2	Model 3	Model 4
GLASSO	1.06	2.18	0.98	0.89
	(0.02)	(0.03)	(0.01)	(0.01)
LLW	1.28	2.23	0.99	0.93
	(0.01)	(0.03)	(0.01)	(0.01)
R-GLASSO	1.23	2.26	0.93	0.95
	(0.01)	(0.02)	(0.01)	(0.01)
NDS	1.11	2.15	0.86	0.73
	(0.02)	(0.02)	(0.01)	(0.01)
R-NDS	1.21	2.24	0.91	0.82
	(0.01)	(0.03)	(0.01)	(0.01)
CLIME	1.03	2.04	0.83	0.75
	(0.01)	(0.02)	(0.01)	(0.01)
R-CLIME	1.17	2.21	0.89	0.86
	(0.01)	(0.03)	(0.01)	(0.01)

Table 3.5: Estimation performance in the Nonparanormal graphical model.

	Nonparanormal			
	Model 1b	Model 2b	Model 3b	Model 4b
GLASSO	2.36	4.48	2.12	1.91
	(0.01)	(0.08)	(0.04)	(0.03)
LLW	1.28	2.23	0.99	0.93
	(0.01)	(0.03)	(0.01)	(0.01)
R-GLASSO	1.23	2.26	0.93	0.95
	(0.01)	(0.02)	(0.01)	(0.01)
NDS	2.32	4.03	1.78	1.58
	(0.01)	(0.05)	(0.05)	(0.03)
R-NDS	1.21	2.24	0.91	0.82
	(0.01)	(0.03)	(0.01)	(0.01)
CLIME	2.11	4.23	2.05	1.37
	(0.02)	(0.04)	(0.03)	(0.02)
R-CLIME	1.17	2.21	0.89	0.86
	(0.01)	(0.03)	(0.01)	(0.01)

Table 3.6: Selection performance in the Gaussian graphical model. Selection accuracy is measured by counts of false negative (#FN) or false positive (#FP). The standard errors are shown in the parenthesis.

	Model 1		Model 2		Model 3		Model 4	
	#FN	#FP	#FN	#FP	#FN	#FP	#FN	#FP
GLASSO	0.00	521.21	263.16	45.21	0.00	114.48	0.03	35.33
	(0.00)	(1.91)	(0.58)	(1.26)	(0.00)	(1.94)	(0.02)	(1.29)
LLW	0.00	518.84	264.18	43.45	0.00	116.02	0.04	35.08
	(0.00)	(1.91)	(0.56)	(1.34)	(0.00)	(2.01)	(0.02)	(1.19)
R-GLASSO	0.00	505.77	264.86	48.01	0.00	114.89	0.03	37.13
	(0.00)	(1.67)	(0.57)	(1.57)	(0.00)	(2.17)	(0.02)	(1.07)
MB.au	0.00	154.81	232.99	89.61	0.00	44.03	0.02	41.22
	(0.00)	(1.29)	(0.74)	(1.37)	(0.00)	(0.81)	(0.01)	(0.77)
R-NDS.au	0.00	163.78	230.77	118.46	0.00	69.16	0.03	49.31
	(0.00)	(1.27)	(0.79)	(2.12)	(0.00)	(0.92)	(0.02)	(0.88)
R-NADS.au	0.00	80.90	218.69	83.62	0.00	60.75	0.03	48.59
	(0.00)	(2.52)	(1.02)	(2.90)	(0.00)	(1.04)	(0.02)	(0.92)
MB.ai	0.00	30.62	260.76	21.79	0.00	9.42	0.04	9.58
	(0.00)	(0.53)	(0.55)	(0.60)	(0.00)	(0.31)	(0.02)	(0.34)
R-NDS.ai	0.00	38.62	259.66	29.34	0.00	11.52	0.07	11.87
	(0.00)	(0.52)	(0.61)	(0.68)	(0.00)	(0.40)	(0.04)	(0.40)
R-NADS.ai	0.06	14.92	256.16	24.62	0.00	10.54	0.08	10.98
	(0.02)	(0.11)	(0.68)	(0.79)	(0.00)	(0.36)	(0.04)	(0.38)
CLIME	0.00	143.88	263.77	34.71	0.00	32.53	0.02	32.59
	(0.00)	(0.10)	(0.57)	(1.42)	(0.00)	(0.78)	(0.01)	(1.17)
R-CLIME	0.00	148.24	265.81	38.23	0.00	37.44	0.04	36.56
	(0.01)	(3.11)	(1.22)	(2.55)	(0.05)	(2.45)	(0.33)	(1.18)
R-ACLIME	0.00	82.53	264.74	34.52	0.00	29.83	0.07	31.09
	(0.00)	(0.13)	(0.63)	(2.60)	(0.00)	(0.61)	(0.03)	(1.02)

Table 3.7: Selection performance in the Nonparanormal graphical model. Selection accuracy is measured by counts of false negative (#FN) or false positive (#FP). The standard errors are shown in the parenthesis.

	Model 1b		Model 2b		Model 3b		Model 4b	
	#FN	#FP	#FN	#FP	#FN	#FP	#FN	#FP
GLASSO	58.81	470.05	286.40	44.70	9.82	134.70	8.06	44.20
	(0.35)	(5.30)	(0.74)	(1.48)	(0.41)	(2.08)	(0.36)	(1.33)
LLW	0.00	518.84	264.18	43.45	0.00	116.02	0.04	35.08
	(0.00)	(1.91)	(0.56)	(1.34)	(0.00)	(2.01)	(0.02)	(1.19)
R-GLASSO	0.00	505.77	264.86	48.01	0.00	114.89	0.03	37.13
	(0.00)	(1.67)	(0.57)	(1.57)	(0.00)	(2.17)	(0.02)	(1.07)
MB.au	56.28	472.86	283.15	61.69	12.99	99.10	8.28	57.65
	(0.26)	(4.11)	(0.64)	(1.04)	(0.46)	(1.31)	(0.36)	(0.90)
R-NDS.au	0.00	163.78	230.77	118.46	0.00	69.16	0.03	49.31
	(0.00)	(1.27)	(0.79)	(2.12)	(0.00)	(0.92)	(0.02)	(0.88)
R-NADS.au	0.00	80.90	218.69	83.62	0.00	60.75	0.03	48.59
	(0.00)	(2.52)	(1.02)	(2.90)	(0.00)	(1.04)	(0.02)	(0.92)
MB.ai	68.68	197.44	304.71	22.72	16.88	50.25	11.67	23.88
	(0.16)	(1.12)	(0.61)	(0.56)	(0.52)	(0.92)	(0.42)	(0.50)
R-NDS.ai	0.00	38.62	259.66	29.34	0.00	11.52	0.08	11.87
	(0.00)	(0.52)	(0.61)	(0.68)	(0.00)	(0.40)	(0.04)	(0.40)
R-NADS.ai	0.06	14.92	256.16	24.62	0.00	10.54	0.08	10.98
	(0.02)	(0.11)	(0.68)	(0.79)	(0.00)	(0.36)	(0.04)	(0.38)
CLIME	47.14	385.95	286.16	45.25	10.02	123.31	7.87	46.38
	(0.39)	(1.99)	(0.74)	(1.45)	(0.41)	(2.11)	(0.36)	(1.34)
R-CLIME	0.00	148.24	265.81	38.23	0.00	37.44	0.04	36.56
	(0.01)	(3.11)	(1.22)	(2.55)	(0.05)	(2.45)	(0.33)	(1.18)
R-ACLIME	0.00	82.53	264.74	34.52	0.00	29.83	0.07	31.09
	(0.00)	(0.13)	(0.63)	(2.60)	(0.00)	(0.61)	(0.03)	(1.02)

3.4.2 Applications to gene expression genomics

We illustrate our proposed rank-based estimators on a real data set to recover the isoprenoid genetic regulatory network in Arabidposis thaliana including 16 genes from the mevalonate (MVA) pathway in the cytosolic, 19 genes from the plastidial (MEP) pathway in the chloroplast, and also 5 encode proteins in the mitochondrial. The data, initially used by Wille et al. (2004), contained the gene expression measurements of 39 genes (excluding protein GGPPS7 in the MEP pathway) assayed on $n = 118$ Affymetrix GeneChip microarrays.

Table 3.8: Bootstrap selection in the isoprenoid gene pathway, and counts of stable edges selected by each estimator are listed.

	GLASSO	MB	CLIME	LLW	R-GLASSO	R-NADS	R-ACLIME
Stable Edge	100	101	67	87	88	50	52

We use seven estimators (GLASSO, MB, CLIME, LLW, R-GLASSO, R-NADS and R-ACLIME) to reconstruct the regulatory network. The first three estimators are performed after taking the log-transformation of the original data, and the other four estimators are directly applied to the original data. To be more conservative, we only consider the integration by union for the neighborhood selection procedures. We compare the final model decided using the Bootstrap method with the cutoff value as 0.8. For each estimator, we draw 100 independent Bootstrap samples, and perform the estimator for each Bootstrap sample. Then, the final model only includes genes selected by at least 80 times over 100 Bootstrap samples. We report the counts of the selected edges for each estimator in Table 3.8. We also compare pairwise intersection of the selected edges among different estimators. More than 70% of the selected edges by GLASSO, MB or CLIME turn out to be validated by both LLW and R-GLASSO, and more than 40% of the selected edges by GLASSO, MB or CLIME are justified by

R-NADS and R-ACLIME. Moreover, though in different means, the selected models all support the biological argument that the interaction between the pathways do exist although both pathways operate independently under normal conditions (Laule et al., 2003; Rodriguez-Concepcion et al., 2004). Thus, the normality assumption after log-transformation appears to be effective in this means.

Chapter 4

Conclusion

4.1 Summary of Contributions

In Chapter 2, we studied the estimation of a sparse high-dimensional Ising model for the binary data. In methodology, we propose efficient procedures to learn a sparse Ising model via the penalized composite conditional likelihood with non-concave penalties. In computation, we design a novel coordinate-minorization-ascent algorithm to efficiently solve the underlying optimization problem, which combines both strengths of coordinate-ascent and minorization-maximization principles. We extend the theory of non-concave penalized likelihood to penalized composite conditional likelihood estimation under the NP-dimensionality, and provide a rigorous discussion of the optimality of the computed local solution. Our proposed procedure is applied to study the Human Immunodeficiency Virus type 1 protease structure based on data from the Stanford HIV Drug Resistance Database. Our statistical learning results match the known biological findings very well, although no prior biological information is used in the data analysis procedure.

In Chapter 3, we studied the estimation of a sparse nonparanormal graphical model for the high-dimensional non-normal data. In methodology, we propose a unified regularized rank estimation scheme under the robust nonparanormal graphical model.

In particular, we study the rank-based Graphical LASSO, the rank-based Dantzig selector and the rank-based CLIME. In theory, we establish their theoretical properties in the setting where the dimension is nearly exponentially large relative to the sample size, and we show that the proposed rank-based estimators work as well as the oracle procedures with the oracle information of transformations.

4.2 Future Directions

Estimation of the High-dimensional Log-linear Graphical Model.
In this thesis, we have proposed efficient procedures to effectively estimate sparse ising models and sparse nonparanomal networks for binary data and continuous data respectively. Then, one interesting research direction is to estimate the log-liner graphical model for discovering complex interactions for a large-scale system of discrete data, which is motivated by inferring genetic networks from the high-throughput sequencing data. Recently, in the same spirit of the neighborhood selection in the Gaussian graphical model Meinshausen and Bühlmann (2006) and the binary Markov random field Ravikumar et al. (2010), Allen and Liu (2012) developed the neighborhood ℓ_1-penalized Poisson regression called the Poisson Graphical Lasso to estimate the high-dimensional log-linear graphical model. Therefore, the rigorous theoretical analysis of the graphical model selection property must be developed for the Poisson Graphical Lasso method under the high-dimensional setting. Meanwhile, the penalized composite likelihood approach is worth pursuing as an alternative procedure.

Local Solution Issue of the General Non-convex Penalized Methodology.
Chapter 2 has discussed the local solution issue of the non-concave penalized composite likelihood methodology, and established theory to show that the computed local solution would exactly match the oracle solution with an overwhelming proba-

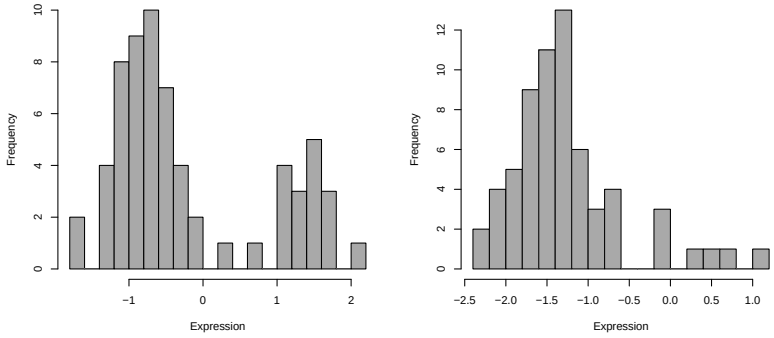
bility. The similar idea of convex relaxation by iterative local linear approximation has been used in the statistical literature such as Candes et al. (2008), Zhang (2010b) and Bradic et al. (2011), while most research works focus on the non-convex penalized least square problem, and none of these papers provides a rigorous theoretical justification for a general non-convex penalization problem. Thus it would be very interesting to generalize the theoretical result as in Theorem 2.3 and Corollary 2.2 of Chapter 2 for the non-concave penalized composite likelihood to more general convex loss functions with a folded concave penalty function, i.e.

$$\min_{\boldsymbol{\beta}} \ell(\boldsymbol{\beta}) + \sum_{j} P_\lambda(|\beta_j|), \tag{4.1}$$

where $\ell(\cdot)$ is a convex loss and $P_\lambda(\cdot)$ satisfies (i). $P'_\lambda(|t|) \geq \lambda$ for $|t| \leq \lambda$; (ii). $P'_\lambda(|t|) = 0$ for $|t| \geq a\lambda$ and $a > 1$. The folded concave penalty includes several popular concave penalty functions proposed in the statistical literature, for example the SCAD (Fan and Li, 2001) and the MCP (Zhang, 2010a).

Large Covariance Matrix Estimation under the Nonparanormal Model. The precision matrix plays an important role in the linear discriminant analysis and the quadratic discriminant analysis, whereas the covariance matrix is indispensable to the principal component analysis and the clustering. In Chapter 3, we have proposed the unified regularized rank procedure to estimate the sparse precision matrix under the nonparanormal model. Then it would be very interesting to extend the regularized rank approach to the large covariance matrix estimation problem under the nonparanormal model, for example Xue and Zou (012a) and Xue and Zou (012b).

One motivating example is to consider the small round blue-cell tumors microarray data (Khan et al., 2001), which have 64 training tissue samples with four types of tumors and 6567 gene expression values for each sample. Rothman et al. (2009)



(A) Gene id: 784224 (B) Gene id: 740604

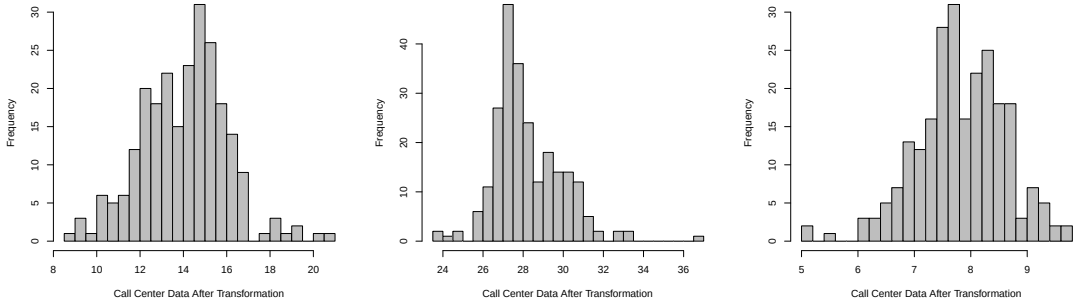
Figure 4.1: Illustration of non-normality in the small round blue-cell data: Panel (A) has two modes, while Panel (B) is highly skewed.

constructed the thresholding covariance matrix estimator on the top 40 and bottom 160 genes selected by the F statistic. Here we conducted various normality tests on these 200 genes and report the results in Table 4.1. More than 60% genes are unable to pass any of four normality tests, and under Bonferroni correction there are still over 30% genes that fail to pass the normality tests. At least 30 of the top 40 genes fail to pass any normality test. Figure 4.1 plots the histograms of two genes to visually illustrate the non-normality.

Table 4.1: Normality test results for the small round blue-cell gene expression data. The counts of genes that fail to pass each normality test are shown in the table.

critical value		Cramer–von Mises	Lilliefors	Pearson's Chi-square
0.05	all 200	153	143	127
	top 40	40	40	39
0.05/200	all 200	84	66	65
	top 40	35	31	30

Another motivating example is to consider the telephone call center data consisting of daily call records from 7:00 am until midnight of a major U.S. northeastern



(A) 7:00am-7:10am (B) 3:30pm-3:40pm (C) 11:50pm-12:00pm

Figure 4.2: Illustration of the non-normality for the call center data

.

financial organization in the year of 2002 (Shen and Huang, 2005). This dataset has observations for 239 weekdays, and each observation recorded the number of telephone calls received for each of the 102 ten-minute intervals of the 17-hour period. In particular, Huang et al. (2006) and Bickel and Levina (2008) preprocessed the data by applying the square root transformation (Brown et al., 2005) to approximate the normality, and then constructed the regularized covariance matrix estimator by banding the Cholesky factors. However, Table 4.2 shows that the suggested square root transformation does help to approximate the normality, but there is still about half of 10-minute recording periods to be non-normal at the significance level of 0.05 even after transformation. Figure 4.2 plots histograms of three ten-minute intervals to visually illustrate the non-normality.

Table 4.2: Testing for Normality of the call center data. The counts of time intervals that fail to pass each normality test are shown in the table.

	Cramer-von Mises	Lilliefors	Pearson Chi-square
raw data	77	76	52
sqrt data	76	70	49

References

Allen, G. and Liu, Z. (2012). A log-linear graphical model for inferring genetic networks from high-throughput sequencing data. *Manuscript*.

Atchley, W., Wollenberg, K., Fitch, W., Terhalle, W., and Dress, A. (2000). Correlations among amino acid sites in bhlh protein domains: an information theoretic analysis. *Molecular Biology and Evolution*, 17(1):164–178.

Banerjee, O., El Ghaoui, L., and d'Aspremont, A. (2008). Model selection through sparse maximum likelihood estimation for multivariate Gaussian or binary data. *Journal of Machine Learning Research*, 9:485–516.

Bertsimas, D. and Tsitsiklis, J. (1997). *Introduction to linear optimization*. Athena Scientific Belmont, MA.

Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society: Series B*, pages 192–236.

Bickel, P. and Levina, E. (2008). Regularized estimation of large covariance matrices. *The Annals of Statistics*, 36(1):199–227.

Bickel, P., Ritov, Y., and Tsybakov, A. (2009). Simultaneous analysis of Lasso and Dantzig selector. *The Annals of Statistics*, 37(4):1705–1732.

Boyd, S. and Vandenberghe, L. (2004). *Convex optimization*. Cambridge University Press.

Bradic, J., Fan, J., and Jiang, J. (2012). Regularization for coxs proportional hazards model with np-dimensionality. *The Annals of Statistics*, 39(6):3092–3120.

Bradic, J., Fan, J., and Wang, W. (2011). Penalized composite quasi-likelihood for ultrahigh dimensional variable selection. *Journal of the Royal Statistical Society: Series B*, 73(3):325–349.

Brown, L., Gans, N., Mandelbaum, A., Sakov, A., Shen, H., Zeltyn, S., and Zhao, L. (2005). Statistical analysis of a telephone call center. *Journal of the American Statistical Association*, 100(469):36–50.

Bühlmann, P. and Meier, L. (2008). Invited discussion on "one-step sparse estimates in nonconcave penalized likelihood models". *The Annals of Statistics*, 36(4):1534–1541.

Cai, T., Liu, W., and Luo, X. (2011). A Constrained ℓ_1 Minimization Approach to Sparse Precision Matrix Estimation. *Journal of the American Statistical Association*, 106(494):594–607.

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

Candes, E., Wakin, M., and Boyd, S. (2008). Enhancing sparsity by reweighted ℓ_1 minimization. *Journal of Fourier Analysis and Applications*, 14(5):877–905.

Chen, X. and Fan, Y. (2006). Estimation of copula-based semiparametric time series models. *Journal of Econometrics*, 130(2):307–335.

Daubechies, I., Defrise, M., and De Mol, C. (2004). An iterative thresholding algorithm for linear inverse problems with a sparsity constraint. *Communications on Pure and Applied Mathematics*, 57(11):1413–1457.

Dempster, A. (1972). Covariance selection. *Biometrics*, 28(1):157–175.

Dempster, A., Laird, N., and Rubin, D. (1977). Maximum likelihood from incomplete data via the em algorithm (with discussion). *Journal of the Royal Statistical Society: Series B*, 39(1):1–38.

Devlin, S., Gnanadesikan, R., and Kettenring, J. (1975). Robust estimation and outlier detection with correlation coefficients. *Biometrika*, 62(3):531–545.

Dicker, L. and Lin, X. (2009). Variable Selection using the Dantzig Selector: asymptotic theory and extensions. *Manuscript*.

Dobra, A., Eicher, T., and Lenkoski, A. (2009). Modeling uncertainty in macroeconomic growth determinants using Gaussian graphical models. *Statistical Methodology*, 7:292–306.

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

Drton, M. and Perlman, M. (2007). Multiple Testing and Error Control in Gaussian Graphical Model Selection. *Statistical Science*, 22(3):430–449.

Edwards, D. (2000). *Introduction to graphical modelling*. Springer Verlag.

Efron, B., Hastie, T., and Tibshirani, R. (2007). Discussion: The Dantzig selector: Statistical estimation when p is much larger than n. *The Annals of Statistics*, 35(6):2358–2364.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96(456):1348–1360.

Fan, J. and Lv, J. (2010). A selective overview of variable selection in high dimensional feature space. *Statistica Sinica*, 20(1):101–148.

Fan, J. and Lv, J. (2011). Nonconcave penalized likelihood with np-dimensionality. *IEEE Transactions on Information Theory*, 57(8):5467–5484.

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

Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1–22.

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

Fu, W. (1998). Penalized regressions: the bridge versus the lasso. *Journal of Computational and Graphical Statistics*, 7(3):397–416.

Genkin, A., Lewis, D., and Madigan, D. (2007). Large-scale bayesian logistic regression for text categorization. *Technometrics*, 49(3):291–304.

Gilbert, H., van der Laan, M., and Dudoit, S. (2009). Joint Multiple Testing Procedures for Graphical Model Selection with Applications to Biological Networks. *U.C. Berkeley Division of Biostatistics Working Paper Series.*, pages 245–290.

Hoeffding, W. (1948). A class of statistics with asymptotically normal distribution. *The Annals of Mathematical Statistics*, 19(3):293–325.

Höfling, H. and Tibshirani, R. (2009). Estimation of sparse binary pairwise markov networks using pseudo-likelihoods. *Journal of Machine Learning Research*, 10:883–906.

Huang, J., Liu, N., Pourahmadi, M., and Liu, L. (2006). Covariance matrix selection and estimation via penalised normal likelihood. *Biometrika*, 93(1):85–98.

Hunter, D. and Lange, K. (2004). A tutorial on mm algorithms. *The American Statistician*, 58(1):30–37.

Irbäck, A., Peterson, C., and Potthast, F. (1996). Evidence for nonrandom hydrophobicity structures in protein chains. *Proceedings of the National Academy of Sciences*, 93(18):9533–9538.

Ising, E. (1925). Beitrag zur theorie des ferromagnetismus. *Zeitschrift für Physik A Hadrons and Nuclei*, 31(1):253–258.

James, G., Radchenko, P., and Lv, J. (2009). Dasso: connections between the dantzig selector and lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(1):127–142.

Kendall, M. (1948). Rank Correlation Methods. *Charles Griffin and Co. Ltd., London*.

Khan, J., Wei, J., Ringnér, M., Saal, L., Ladanyi, M., Westermann, F., Berthold, F., Schwab, M., Antonescu, C., Peterson, C., et al. (2001). Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. *Nature Medicine*, 7(6):673–679.

Klaassen, C. and Wellner, J. (1997). Efficient estimation in the bivariate normal copula model: normal margins are least favourable. *Bernoulli*, 3(1):55–77.

Lam, C. and Fan, J. (2009). Sparsistency and rates of convergence in large covariance matrix estimation. *The Annals of Statistics*, 37(6):4254–4278.

Lange, K., Hunter, D., and Yang, I. (2000). Optimization transfer using surrogate objective functions. *Journal of Computational and Graphical Statistics*, 9(1):1–20.

Laule, O., Fürholz, A., Chang, H., Zhu, T., Wang, X., Heifetz, P., Gruissem, W., and Lange, M. (2003). Crosstalk between cytosolic and plastidial pathways of isoprenoid biosynthesis in Arabidopsis thaliana. *Proceedings of the National Academy of Sciences*, 100(11):6866–6871.

Lauritzen, S. (1996). *Graphical models*. Oxford University Press, USA.

Lehmann, E. (1998). *Nonparametrics: statistical methods based on ranks*. Prentice Hall Upper Saddle River, New Jersey.

Li, H. and Gui, J. (2006). Gradient directed regularization for sparse Gaussian concentration graphs, with applications to inference of genetic networks. *Biostatistics*, 7(2):302–317.

Li, S. (2009). *Markov random field modeling in image analysis*. Springer-Verlag New York Inc.

Lindsay, B. (1988). Composite likelihood methods. *Contemporary Mathematics*, 80(1):221–239.

Liu, H., Lafferty, J., and Wasserman, L. (2009). The nonparanormal: semiparametric estimation of high dimensional undirected graphs. *Journal of Machine Learning Research*, 10:1–37.

Liu, Y., Eyal, E., and Bahar, I. (2008). Analysis of correlated mutations in hiv-1 protease using spectral clustering. *Bioinformatics*, 24(10):1243–1250.

Lv, J. and Fan, Y. (2009). A unified approach to model selection and sparse recovery using regularized least squares. *The Annals of Statistics*, 37(6A):3498–3528.

Majewski, J., Li, H., and Ott, J. (2001). The ising model in physics and statistical genetics. *The American Journal of Human Genetics*, 69(4):853–862.

Markowitz, M., Mo, H., Kempf, D., Norbeck, D., Bhat, T., Erickson, J., and Ho, D. (1995). Selection and analysis of human immunodeficiency virus type 1 variants with increased resistance to abt-538, a novel protease inhibitor. *Journal of Virology*, 69(2):701–706.

McDiarmid, C. (1989). On the method of bounded differences. *Surveys in Combinatorics*, 141:148–188.

Meier, L., Van De Geer, S., and Bühlmann, P. (2008). The group lasso for logistic regression. *Journal of the Royal Statistical Society: Series B*, 70(1):53–71.

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

Meinshausen, N. and Bühlmann, P. (2010). Stability selection (with discussion). *Journal of the Royal Statistical Society: Series B*, 72(4):417–473.

Muzammil, S., Ross, P., and Freire, E. (2003). A major role for a set of non-active site mutations in the development of hiv-1 protease drug resistance. *Biochemistry*, 42(3):631–638.

Ohtaka, H., Schön, A., and Freire, E. (2003). Multidrug resistance to hiv-1 protease inhibition requires cooperative coupling between distal mutations. *Biochemistry*, 42(46):13659–13666.

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

Ravikumar, P., Wainwright, M., and Lafferty, J. (2010). High-dimensional ising model selection using l1-regularized logistic regression. *The Annals of Statistics*, 38(3):1287–1319.

Ravikumar, P., Wainwright, M., Raskutti, G., and Yu, B. (2008). High-dimensional covariance estimation by minimizing ℓ_1-penalized log-determinant divergence. *Advances in Neural Information Processing Systems*.

Rhee, S., Liu, T., Ravela, J., Gonzales, M., and Shafer, R. (2004). Distribution of human immunodeficiency virus type 1 protease and reverse transcriptase mutation patterns in 4,183 persons undergoing genotypic resistance testing. *Antimicrobial Agents and Chemotherapy*, 48(8):3122–3126.

Rhee, S., Taylor, J., Wadhera, G., Ben-Hur, A., Brutlag, D., and Shafer, R. (2006). Genotypic predictors of human immunodeficiency virus type 1 drug resistance. *Proceedings of the National Academy of Sciences*, 103(46):17355–17360.

Rodriguez-Concepcion, M., Fores, O., Martinez-Garcia, J., Gonzalez, V., Phillips, M., Ferrer, A., and Boronat, A. (2004). Distinct light-mediated pathways regulate the biosynthesis and exchange of isoprenoid precursors during Arabidopsis seedling development. *The Plant Cell*, 16(1):144.

Rothman, A., Bickel, P., Levina, E., and Zhu, J. (2008). Sparse permutation invariant covariance estimation. *Electronic Journal of Statistics*, 2:494–515.

Rothman, A., Levina, E., and Zhu, J. (2009). Generalized thresholding of large covariance matrices. *Journal of the American Statistical Association*, 104(485):177–186.

Schelldorfer, J., Bühlmann, P., and DE GEER, S. (2011). Estimation for high-dimensional linear mixed-effects models using ℓ_l-penalization. *Scandinavian Journal of Statistics*, 38(2):197–214.

Shen, H. and Huang, J. (2005). Analysis of call centre arrival data using singular value decomposition. *Applied Stochastic Models in Business and Industry*, 21(3):251–263.

Song, P. (2000). Multivariate dispersion models generated from gaussian copula. *Scandinavian Journal of Statistics*, 27(2):305–320.

Städler, N., Bühlmann, P., and Van De Geer, S. (2010). ℓ_1-penalization for mixture regression models (with discussion). *Test*, 19(2):209–256.

Stauffer, D. (2008). Social applications of two-dimensional ising models. *American Journal of Physics*, 76:470.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B*, 58(1):267–288.

Tisdale, M., Myers, R., Maschera, B., Parry, N., Oliver, N., and Blair, E. (1995). Cross-resistance analysis of human immunodeficiency virus type 1 variants individually selected for resistance to five different protease inhibitors. *Antimicrobial Agents and Chemotherapy*, 39(8):1704–1710.

Tseng, P. (1988). Coordinate ascent for maximizing nondifferentiable concave functions. *Technical Report LIDS-P, 1840*, Massachusetts Institute of Technology, Laboratory for Information and Decision Systems.

Tsukahara, H. (2005). Semiparametric estimation in copula models. *Canadian Journal of Statistics*, 33(3):357–375.

Varin, C. (2008). On composite marginal likelihoods. *Advances in Statistical Analysis*, 92(1):1–28.

Varin, C., Reid, N., and Firth, D. (2011). An overview of composite likelihood methods. *Statistica Sinica*, 21(1):5–42.

Wang, H., Li, R., and Tsai, C. (2007). Tuning parameter selectors for the smoothly clipped absolute deviation method. *Biometrika*, 94(3):553–568.

Wille, A., Zimmermann, P., Vranová, E., Fürholz, A., Laule, O., Bleuler, S., Hennig, L., Prelic, A., Von Rohr, P., Thiele, L., et al. (2004). Sparse graphical Gaussian modeling of the isoprenoid gene network in Arabidopsis thaliana. *Genome Biology*, 5(11):R92.1–13.

Witten, D., Friedman, J., and Simon, N. (2011). New insights and faster computations for the graphical lasso. *Journal of Computational and Graphical Statistics*, 20(4):892–900.

Wu, M., Cai, T., and Lin, X. (2010). A parametric permutation test for regression coefficients in lasso regularized regression for high dimensional data. *Technical Report*, Harvard University.

Wu, T. and Lange, K. (2008). Coordinate descent algorithms for lasso penalized regression. *The Annals of Applied Statistics*, 2(1):224–244.

Wu, T., Schiffer, C., Gonzales, M., Taylor, J., Kantor, R., Chou, S., Israelski, D., Zolopa, A., Fessel, W., and Shafer, R. (2003). Mutation patterns and structural correlates in human immunodeficiency virus type 1 protease following different protease inhibitor treatments. *Journal of Virology*, 77(8):4836–4847.

Xue, L. and Zou, H. (2011). Regularized rank-based estimation of high-dimensional Nonparanormal graphical models. *Technical Report, University of Minnesota*.

Xue, L. and Zou, H. (2012a). Rank-based tapering estimation of bandable covariance matrices. *Technical Report, University of Minnesota*.

Xue, L. and Zou, H. (2012b). On estimating sparse correlation matrices of semiparametric Gaussian copulas. *Technical Report, University of Minnesota*.

Xue, L., Zou, H., and Cai, T. (2012). Non-concave penalized composite conditional likelihood estimation of sparse Ising models. *The Annals of Statistics, to appear*.

Yuan, M. (2010). High-dimensional inverse covariance matrix estimation via linear programming. *Journal of Machine Learning Research*, 11:2261–2286.

Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B*, 68(1):49–67.

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

Zhang, C. (2010a). Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics*, 38(2):894–942.

Zhang, T. (2010b). Analysis of multi-stage convex relaxation for sparse regularization. *Journal of Machine Learning Research*, 11:1081–1107.

Zhao, P. and Yu, B. (2007). On model selection consistency of lasso. *Journal of Machine Learning Research*, 7(2):2541–2563.

Zhou, S., Rutimann, P., Xu, M., and Buhlmann, P. (2011). High-dimensional covariance estimation based on Gaussian graphical models. *Journal of Machine Learning Research*, 12:2975–3026.

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

Zou, H. and Li, R. (2008). One-step sparse estimates in nonconcave penalized likelihood models (with discussion). *The Annals of Statistics*, 36(4):1509–1566.

Appendix A

Proof of Chapter 2

Before presenting the proof, we first define some useful quantities. The score functions of the negative composite likelihood ($-\ell^{(j)}$) and the Hessian matrices are defined as follows:

$$\psi_k^{(j)} = -\frac{\partial \ell^{(j)}(\boldsymbol{\beta}^{(j)})}{\partial \beta_{jk}} = \frac{1}{N}\sum_{n=1}^{N} x_{jn}x_{kn}(\theta_{jn} - 1), \quad k \neq j$$

$$H_{k_1,k_2}^{(j)} = -\frac{\partial^2 \ell^{(j)}(\boldsymbol{\beta}^{(j)})}{\partial \beta_{jk_1}\partial \beta_{jk_2}} = \frac{1}{N}\sum_{n=1}^{N} x_{k_1n}x_{k_2n}(1 - \theta_{jn})\theta_{jn}, \quad k_1, k_2 \neq j$$

Similarly, let ψ be the score function of $-\ell_c$ such that

$$\psi_{(jk)} = \frac{\partial - \ell_c(\boldsymbol{\beta})}{\partial \beta_{jk}}$$

for $1 \leq j < k \leq K$. By definition we have the following identities

$$\psi_{(jk)} = \psi_k^{(j)} + \psi_j^{(k)}.$$

In what follows we write $\psi^* = \psi(\boldsymbol{\beta}^*)$.

A.1 Proof of Theorem 2.1

Proof A.1 We first prove part (1).

Consider $V(\boldsymbol{\alpha}_{\mathcal{A}}) = -\ell_c(\boldsymbol{\beta}^*_{\mathcal{A}} + d_N\boldsymbol{\alpha}_{\mathcal{A}}) + \ell_c(\boldsymbol{\beta}^*_{\mathcal{A}})$ and its minimizer is

$$\widetilde{\boldsymbol{\alpha}}^{hmle}_{\mathcal{A}} = \frac{1}{d_N}(\widetilde{\boldsymbol{\beta}}^{hmle}_{\mathcal{A}} - \boldsymbol{\beta}^*_{\mathcal{A}}).$$

By definition, $V(\widetilde{\boldsymbol{\alpha}}^{hmle}_{\mathcal{A}}) \leq V(\mathbf{0}) = 0$. Fix any $R > 0$ and consider any $\boldsymbol{\alpha}_{\mathcal{A}}$ satisfying $\|\boldsymbol{\alpha}_{\mathcal{A}}\|_2 = R$. Using Taylor's expansion, we know that

$$\begin{aligned} V(\boldsymbol{\alpha}_{\mathcal{A}}) &= d_N\boldsymbol{\alpha}^T_{\mathcal{A}}\psi^*_{\mathcal{A}} + \frac{1}{2}d^2_N\boldsymbol{\alpha}^T_{\mathcal{A}}H^*_{\mathcal{A}\mathcal{A}}\boldsymbol{\alpha}_{\mathcal{A}} + \frac{1}{2}d^2_N\boldsymbol{\alpha}^T_{\mathcal{A}}\left[H_{\mathcal{A}\mathcal{A}}(\boldsymbol{\beta}(t)) - H^*_{\mathcal{A}\mathcal{A}}\right]\boldsymbol{\alpha}_{\mathcal{A}} \\ &\equiv T_1 + T_2 + T_3, \end{aligned} \tag{A.1}$$

for some $t \in [0,1]$ and $\boldsymbol{\beta}(t) = \boldsymbol{\beta}^*_{\mathcal{A}} + td_N\boldsymbol{\alpha}_{\mathcal{A}}$. Note that $E[\psi^*_{\mathcal{A}}] = 0$ and $\|\psi^*_{\mathcal{A}}\|_\infty \leq 2$. By Cauchy-Schwartz inequality $|\boldsymbol{\alpha}^T_{\mathcal{A}}\psi^*_{\mathcal{A}}| \leq 2\sqrt{s}R$. Using Hoeffding's inequality we have

$$\Pr(T_1 \geq -d_N\epsilon) \leq \exp(-\frac{N\epsilon^2}{8sR^2}). \tag{A.2}$$

For the second term we first have $T_2 \geq \frac{d^2_N}{2}\lambda_{\min}(H^*_{\mathcal{A}\mathcal{A}})R^2$. Each entry of H^* is between $-\frac{1}{2}$ and $\frac{1}{2}$. Thus Hoeffding's inequality and the union bound yield

$$\Pr(\|H^{(N)}_j - H_j\|^2_F \geq \frac{b^2}{4}) \leq 2s^2\exp(-N\frac{b^2}{2s^2}).$$

So by the inequality that $\lambda_{\min}(H^*_{\mathcal{A}\mathcal{A}}) \geq b - \|H^*_{\mathcal{A}\mathcal{A}} - E[H^*_{\mathcal{A}\mathcal{A}}]\|_F$ we have

$$\Pr(T_2 \geq d^2_N bR^2/4) \geq 1 - 2s^2\exp(-\frac{Nb^2}{2s^2}). \tag{A.3}$$

For $|T_3|$, let $\lambda_{\max}(\frac{1}{N}\sum_{n=1}^N \boldsymbol{x}_{\mathcal{A}n}\boldsymbol{x}^T_{\mathcal{A}n}) = B_N$. Define $\bar{\eta}_{jn}(\boldsymbol{\beta}) = \theta_{jn}\big(1-\theta_{jn}\big)\big(2\theta_{jn}-1\big)$.

Using the mean value theorem, for some $t' \in [0,t]$ and $\boldsymbol{\beta}(t') = \boldsymbol{\beta}^*_{\mathcal{A}} + t' d_N \boldsymbol{\alpha}_{\mathcal{A}}$, we have

$$\begin{aligned} |T_3| &= \frac{d_N^3}{2} |\frac{1}{N} \sum_n \sum_{j=1}^K \sum_{\substack{k_1 \neq j \\ k_2 \neq j}} \alpha_{jk_1} \alpha_{jk_2} x_{k_1 n} x_{k_2 n} t' \bar{\eta}_{jn}(\boldsymbol{\beta}(t'))(\sum_{k' \neq j} \alpha_{jk'} x_{jn} x_{k'n})| \\ &\leq \frac{d_N^3}{2} (\frac{\sqrt{sR^2}}{4}) \cdot (2B_N \sum_{(j,k) \in \mathcal{A}} \alpha_{jk}^2) = \frac{d_N^3 B_N}{4} \sqrt{s} R^3. \end{aligned} \tag{A.4}$$

In the last step we have used $|\bar{\eta}_{jn}(\boldsymbol{\beta}(t'))| \leq \frac{1}{4}$ for any j and $\boldsymbol{\alpha}_{\mathcal{A}^c} = 0$. Moreover, $B_N \leq B + \|\frac{1}{N} \sum_{n=1}^N \boldsymbol{x}_{\mathcal{A}n} \boldsymbol{x}_{\mathcal{A}n}^T - E[\boldsymbol{x}_{\mathcal{A}} \boldsymbol{x}_{\mathcal{A}}^T]\|_F$. Since $x_{jn} = \pm 1$, we apply Hoeffding's inequality and the union bound to obtain the following probability bound

$$\Pr(\|\frac{1}{N} \sum_{n=1}^N \boldsymbol{x}_{\mathcal{A}n} \boldsymbol{x}_{\mathcal{A}n}^T - E[\boldsymbol{x}_{\mathcal{A}} \boldsymbol{x}_{\mathcal{A}}^T]\|_F \geq B/2) \leq 2s^2 \exp(-\frac{NB^2}{8s^2}),$$

which leads to

$$\Pr(|T_3| \leq \frac{3 d_N^3 B}{8} \sqrt{s} R^3) \geq 1 - 2s^2 \exp(-\frac{NB^2}{8s^2}). \tag{A.5}$$

Taking $R < \frac{b}{3B} \frac{\sqrt{N}}{s}$ and combining (A.2), (A.3) and (A.5), we have

$$T_1 + T_2 + T_3 \geq \frac{bR^2}{8} d_N^2 - \frac{3B}{8} R^3 d_N^3 \sqrt{s} > 0$$

with probability at least $1 - \tau_1$. Thus, the convexity of V implies that

$$\Pr\left(\|\widetilde{\boldsymbol{\beta}}^{hmle}_{\mathcal{A}} - \boldsymbol{\beta}^*_{\mathcal{A}}\|_2 \leq \sqrt{\frac{s}{N}} R \right) \geq 1 - \tau_1.$$

We now prove part (2). First, we show that if $\min_{(j,k) \in \mathcal{A}} |\widetilde{\beta}^{hmle}_{jk}| > a\lambda$ and $\|\psi_{\mathcal{A}^c}(\widehat{\boldsymbol{\beta}}^{oracle})\|_\infty \leq \lambda$, then $\widehat{\boldsymbol{\beta}}^{oralce}$ is a local maximizer of $\ell_c(\boldsymbol{\beta}) - \sum_{(j,k)} P_\lambda(|\beta_{jk}|)$. To see that, consider a small ball of radius t with $\widehat{\boldsymbol{\beta}}^{oralce}$ being the center. Let $\boldsymbol{\beta}$ be

any point in the ball. So $\|\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}^{oralce}\|_2 \le t$. Clearly, for a sufficiently small t we have $\min_{(j,k)\in\mathcal{A}}|\beta_{jk}| > a\lambda$ and $\max_{(j,k)\in\mathcal{A}^c}|\beta_{jk}| < \lambda$. By Taylor's expansion we have

$$\begin{aligned}
& \{-\ell_c(\boldsymbol{\beta}) + \sum_{(j,k)} P_\lambda(|\beta_{jk}|)\} - \{-\ell_c(\widehat{\boldsymbol{\beta}}^{oracle}) + \sum_{(j,k)} P_\lambda(|\hat{\beta}_{jk}^{oracle}|)\} \\
= \;& (\boldsymbol{\beta}_{\mathcal{A}} - \widetilde{\boldsymbol{\beta}}^{hmle})^T \psi_{\mathcal{A}^c}(\widehat{\boldsymbol{\beta}}^{oracle}) + \frac{1}{2}(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}^{oracle})^T H(\boldsymbol{\beta}')(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}^{oracle}) \\
& + \sum_{(j,k)\in\mathcal{A}^c} \lambda|\beta_{jk}| \\
\ge \;& \sum_{(j,k)\in\mathcal{A}^c} (\lambda - |\psi_{(jk)}(\widehat{\boldsymbol{\beta}}^{oracle})|)|\beta_{jk}| \\
\ge \;& 0.
\end{aligned}$$

A probability bound for the event of $\min_{(j,k)\in\mathcal{A}}|\tilde{\beta}_{jk}^{hmle}| > a\lambda$ is given by

$$\begin{aligned}
& \Pr(\min_{(j,k)\in\mathcal{A}} |\tilde{\beta}_{jk}^{hmle}| > a\lambda) \\
\ge \;& \Pr\left(\|\widetilde{\boldsymbol{\beta}}_{\mathcal{A}}^{hmle} - \boldsymbol{\beta}_{\mathcal{A}}^*\|_2 \le \sqrt{\frac{s}{N}} R_*\right) \\
\ge \;& 1 - \exp(-R_*^2 \frac{b^2}{8^3}) - 2s^2 \exp(-\frac{N}{s^2}\frac{b^2}{2}) - 2s^2 \exp(-\frac{N}{s^2}\frac{B^2}{8}).
\end{aligned} \tag{A.6}$$

Now consider $\Pr(\|\psi_{\mathcal{A}^c}(\widehat{\boldsymbol{\beta}}^{oracle})\|_\infty < \lambda)$. There exists some $t \in [0,1]$ such that

$$\psi(\widehat{\boldsymbol{\beta}}^{oracle}) = \psi(\boldsymbol{\beta}^*) + H^*(\widehat{\boldsymbol{\beta}}^{oracle} - \boldsymbol{\beta}^*) + r \tag{A.7}$$

where $r = (H(\boldsymbol{\beta}^* + t(\widehat{\boldsymbol{\beta}}^{oracle} - \boldsymbol{\beta}^*)) - H^*)(\widehat{\boldsymbol{\beta}}^{oracle} - \boldsymbol{\beta}^*)$. Note $\psi_{\mathcal{A}}(\widehat{\boldsymbol{\beta}}^{oracle}) = 0$, so

$$\widetilde{\boldsymbol{\beta}}_{\mathcal{A}} - \boldsymbol{\beta}_{\mathcal{A}}^* = (H_{\mathcal{A}\mathcal{A}}^*)^{-1}(-\psi_{\mathcal{A}} - r_{\mathcal{A}}).$$

Then $\|\psi_{\mathcal{A}^c}(\widehat{\boldsymbol{\beta}}^{oracle})\|_\infty \leq \lambda$ becomes

$$\|H^*_{\mathcal{A}^c\mathcal{A}}(H^*_{\mathcal{A}\mathcal{A}})^{-1}(-\psi_{\mathcal{A}} - r_{\mathcal{A}}) + \psi_{\mathcal{A}^c} + r_{\mathcal{A}^c}\|_\infty \leq \lambda$$

which is guaranteed if

$$(\|H^*_{\mathcal{A}^c\mathcal{A}}(H^*_{\mathcal{A}\mathcal{A}})^{-1}\|_\infty + 1)(\|\psi\|_\infty + \|r\|_\infty) \leq \lambda.$$

Therefore we have a simple lower bound for $\Pr(\|\psi_{\mathcal{A}^c}(\widehat{\boldsymbol{\beta}}^{oracle})\|_\infty \leq \lambda)$:

$$\begin{aligned} & \Pr(\|\psi_{\mathcal{A}^c}(\widehat{\boldsymbol{\beta}}^{oracle})\|_\infty \leq \lambda) \\ > \; & 1 - \Pr(\|H^*_{\mathcal{A}^c\mathcal{A}}(H^*_{\mathcal{A}\mathcal{A}})^{-1}\|_\infty > 2\phi) - \Pr(\|\psi\|_\infty > \frac{\lambda}{4\phi+2}) - \Pr(\|r\|_\infty > \frac{\lambda}{4\phi+2}). \end{aligned}$$

Using Hoeffding's inequality and the union bound we have

$$\Pr(\|\psi\|_\infty \leq \frac{\lambda}{4\phi+2}) \geq 1 - K^2 \exp(-\frac{N\lambda^2}{128(\phi+\frac{1}{2})^2}). \tag{A.8}$$

Write $\boldsymbol{\alpha} = \widetilde{\boldsymbol{\beta}}^{hmle} - \boldsymbol{\beta}^*$, and thus $\boldsymbol{\alpha}_{\mathcal{A}^c} = 0$. By the mean value theorem we have a bound for $r_{(jk)}$

$$\begin{aligned} |r_{(jk)}| &= |\frac{1}{N}\sum_{n=1}^{N}\sum_{k_2\neq j}\sum_{k'\neq j} x_{kn}x_{jn}x_{k_2n}x_{k'n}\alpha_{jk_2}\alpha_{jk'}t'\bar{\eta}_{jn}(\boldsymbol{\beta}(t')) \\ &\quad + \frac{1}{N}\sum_{n=1}^{N}\sum_{j_2\neq k}\sum_{j'\neq k} x_{jn}x_{kn}x_{j_2n}x_{j'n}\alpha_{kj_2}\alpha_{kj'}t'\bar{\eta}_{kn}(\boldsymbol{\beta}(t'))| \\ &\leq B_N \cdot \|\widetilde{\boldsymbol{\beta}}_{\mathcal{A}} - \boldsymbol{\beta}^*_{\mathcal{A}}\|_2^2. \end{aligned}$$

In the last step we have used $|\bar{\eta}_{jn}(\boldsymbol{\beta}(t'))| \leq \frac{1}{4}$ for any j and $\boldsymbol{\alpha}_{\mathcal{A}^c} = 0$. Moreover, recall

that

$$B_N \leq B + \|\frac{1}{N}\sum_{n=1}^{N} \boldsymbol{x}_{\mathcal{A}n}\boldsymbol{x}_{\mathcal{A}n}^T - E[\boldsymbol{x}_{\mathcal{A}}\boldsymbol{x}_{\mathcal{A}}^T]\|_F.$$

Thus,

$$\Pr\left(\|r\|_\infty < \frac{\lambda}{4\phi+2}\right) \geq 1-\exp(\frac{-N\lambda}{3B(2\phi+1)s}\frac{b^2}{8^3})-2s^2\exp(\frac{-Nb^2}{2s^2})-2s^2\exp(\frac{-NB^2}{8s^2}). \tag{A.9}$$

For notation convenience define

$$c = \|(E[H^*_{\mathcal{A}\mathcal{A}}])^{-1}\|_\infty \leq \sqrt{s}\|(E[H^*_{\mathcal{A}\mathcal{A}}])^{-1}\|_2$$

and

$$\delta = \|H^*_{\mathcal{A}^c\mathcal{A}}(H^*_{\mathcal{A}\mathcal{A}})^{-1} - E[H^*_{\mathcal{A}^c\mathcal{A}}](E[H^*_{\mathcal{A}\mathcal{A}}])^{-1}\|_\infty$$

$$\delta_1 = \|(H^*_{\mathcal{A}\mathcal{A}})^{-1} - (E[H^*_{\mathcal{A}\mathcal{A}}])^{-1}\|_\infty$$

$$\delta_2 = \|H^*_{\mathcal{A}\mathcal{A}} - E[H^*_{\mathcal{A}\mathcal{A}}]\|_\infty$$

$$\delta_3 = \|H^*_{\mathcal{A}^c\mathcal{A}} - E[H^*_{\mathcal{A}^c\mathcal{A}}]\|_\infty$$

Then by definition

$$\begin{aligned}
\delta &= \|(H^*_{\mathcal{A}^c\mathcal{A}} - E[H^*_{\mathcal{A}^c\mathcal{A}}])((H^*_{\mathcal{A}\mathcal{A}})^{-1} - (E[H^*_{\mathcal{A}\mathcal{A}}])^{-1}) \\
&\quad + E[H^*_{\mathcal{A}^c\mathcal{A}}](E[H^*_{\mathcal{A}\mathcal{A}}])^{-1}(-H^*_{\mathcal{A}\mathcal{A}} + E[H^*_{\mathcal{A}\mathcal{A}}])(H^*_{\mathcal{A}\mathcal{A}})^{-1} \\
&\quad + (H^*_{\mathcal{A}^c\mathcal{A}} - E[H^*_{\mathcal{A}^c\mathcal{A}}])(E[H^*_{\mathcal{A}\mathcal{A}}])^{-1}\|_\infty \\
&\leq \delta_3\delta_1 + \phi\delta_2\|(H^*_{\mathcal{A}\mathcal{A}})^{-1}\|_\infty + \delta_3 c \\
&\leq \delta_3\delta_1 + \phi(c+\delta_1)\delta_2 + \delta_3 c.
\end{aligned}$$

Note that

$$\begin{aligned}
\delta_1 &= \|(H^*_{\mathcal{A}\mathcal{A}})^{-1}(E[H^*_{\mathcal{A}\mathcal{A}}] - H^*_{\mathcal{A}\mathcal{A}})(E[H^*_{\mathcal{A}\mathcal{A}}])^{-1}\|_\infty \\
&\leq \|(H^*_{\mathcal{A}\mathcal{A}})^{-1}\|_\infty \cdot \|E[H^*_{\mathcal{A}\mathcal{A}}] - H^*_{\mathcal{A}\mathcal{A}}\|_\infty \cdot \|(E[H^*_{\mathcal{A}\mathcal{A}}])^{-1}\|_\infty \\
&\leq (\delta_1 + c)\delta_2 c.
\end{aligned}$$

Hence as long as $\delta_2 c < 1$, both $\delta_1 \leq \frac{\delta_2 c^2}{1-\delta_2 c}$ and $\delta \leq (\delta_3 + \phi\delta_2)\frac{c}{1-\delta_2 c}$ hold. Then we have

$$\begin{aligned}
\Pr(\delta_2 < \frac{1}{4c}) &\geq 1 - \Pr(\|H^*_{\mathcal{A}^c\mathcal{A}} - E[H^*_{\mathcal{A}^c\mathcal{A}}]\|_{\max} > \frac{1}{4cs}) \\
&\geq 1 - 2s^2\exp(-\frac{N}{8c^2s^2}). \qquad \text{(A.10)}
\end{aligned}$$

$$\begin{aligned}
\Pr(\delta_3 < \frac{\phi}{2c}) &\geq 1 - \Pr(\|H^*_{\mathcal{A}^c\mathcal{A}} - E[H^*_{\mathcal{A}^c\mathcal{A}}]\|_{\max} > \frac{\phi}{4cs}) \\
&\geq 1 - K^2 s\exp(-\frac{N\phi^2}{2c^2s^2}). \qquad \text{(A.11)}
\end{aligned}$$

Finally we have $c \leq \sqrt{s}/b$. Therefore, part (2) is proven by combining (A.6), (A.8), (A.9) and (A.10), (A.11). This completes the proof. □

A.2 Proof of Theorem 2.2

Proof A.2 We first prove part (1).

Consider

$$V(\boldsymbol{\alpha}_{\mathcal{A}}) = (\ell_c(\boldsymbol{\beta}^*_{\mathcal{A}} + d_N\boldsymbol{\alpha}_{\mathcal{A}}) - \ell_c(\boldsymbol{\beta}^*_{\mathcal{A}})) + \lambda \sum_{(j,k)\in\mathcal{A}} (|\beta^*_{jk} + d_N\alpha_{jk}| - |\beta^*_{jk}|) \quad \text{(A.12)}$$

and its minimizer is

$$\widetilde{\boldsymbol{\alpha}}_{\mathcal{A}} = \frac{1}{d_N}(\widetilde{\boldsymbol{\beta}}_{\mathcal{A}} - \boldsymbol{\beta}^*_{\mathcal{A}}).$$

By definition, $V(\widetilde{\boldsymbol{\alpha}}_{\mathcal{A}}) \leq V(\mathbf{0}) = 0$. Fix a $R > 0$ and consider any $\boldsymbol{\alpha}_{\mathcal{A}}$ satisfying $\|\boldsymbol{\alpha}_{\mathcal{A}}\|_2 = R$. Using Taylor expansion, for some $t \in [0,1]$ and $\boldsymbol{\beta}(t) = \boldsymbol{\beta}^*_{\mathcal{A}} + td_N\boldsymbol{\alpha}_{\mathcal{A}}$,

$$\begin{aligned} V(\boldsymbol{\alpha}_{\mathcal{A}}) &= d_N\boldsymbol{\alpha}^T_{\mathcal{A}}\psi^*_{\mathcal{A}} + \frac{1}{2}d^2_N\boldsymbol{\alpha}^T_{\mathcal{A}}H^*_{\mathcal{A}\mathcal{A}}\boldsymbol{\alpha}_{\mathcal{A}} + \frac{1}{2}d^2_N\boldsymbol{\alpha}^T_{\mathcal{A}}\left[H_{\mathcal{A}\mathcal{A}}(\boldsymbol{\beta}(t)) - H^*_{\mathcal{A}\mathcal{A}}\right]\boldsymbol{\alpha}_{\mathcal{A}} \\ &\quad + \lambda \sum_{(j,k)\in\mathcal{A}} (|\beta^*_{jk} + d_N\alpha_{jk}| - |\beta^*_{jk}|) \\ &\equiv T_1 + T_2 + T_3 + T_4. \end{aligned} \quad \text{(A.13)}$$

We derive an upper bound for $|T_1|$. Note that $E[\psi^*_{\mathcal{A}}] = 0$ and $\|\psi^*_{\mathcal{A}}\|_\infty \leq 2$. By Cauchy-Schwartz inequality $|\boldsymbol{\alpha}^T_{\mathcal{A}}\psi^*_{\mathcal{A}}| \leq 2\sqrt{s}R$. Using Hoeffding's inequality we have

$$\Pr(T_1 \geq -d_N\epsilon) \leq \exp(-\frac{N\epsilon^2}{8sR^2}). \quad \text{(A.14)}$$

We derive a lower bound for T_2. First note that $T_2 \geq \frac{d^2_N}{2}\lambda_{\min}(H^*_{\mathcal{A}\mathcal{A}})R^2$.

$$\begin{aligned} \lambda_{\min}(H^*_{\mathcal{A}\mathcal{A}}) &\geq \lambda_{\min}(E[H^*_{\mathcal{A}\mathcal{A}}]) + \lambda_{\min}(H^*_{\mathcal{A}\mathcal{A}} - E[H^*_{\mathcal{A}\mathcal{A}}]) \\ &\geq b - \|H^*_{\mathcal{A}\mathcal{A}} - E[H^*_{\mathcal{A}\mathcal{A}}]\|_F \end{aligned}$$

Each entry of H^* is between $-\frac{1}{2}$ and $\frac{1}{2}$. Thus Hoeffding's inequality and the union bound yield

$$\Pr(\|H_j^{(N)} - H_j\|_F^2 \geq \frac{b^2}{4}) \leq 2s^2 \exp(-N\frac{b^2}{2s^2}). \tag{A.15}$$

So we have

$$\Pr(T_2 \geq d_N^2 bR^2/4) \geq 1 - 2s^2 \exp(-\frac{Nb^2}{2s^2}). \tag{A.16}$$

We derive an upper bound for $|T_3|$. Let $\lambda_{\max}(\frac{1}{N}\sum_{n=1}^N \boldsymbol{x}_{\mathcal{A}n}\boldsymbol{x}_{\mathcal{A}n}^T) = B_N$. Define $\bar{\eta}_{jn}(\boldsymbol{\beta}) = \theta_{jn}\big(1-\theta_{jn}\big)\big(2\theta_{jn}-1\big)$. Using the mean value theorem, we have that, for some $t' \in [0,t]$ and $\boldsymbol{\beta}(t') = \boldsymbol{\beta}_{\mathcal{A}}^* + t' d_N \boldsymbol{\alpha}_{\mathcal{A}}$,

$$\begin{aligned} |T_3| &= \frac{d_N^3}{2}|\frac{1}{N}\sum_n \sum_{j=1}^K \sum_{\substack{k_1\neq j\\ k_2\neq j}} \alpha_{jk_1}\alpha_{jk_2}x_{k_1n}x_{k_2n}t'\bar{\eta}_{jn}(\boldsymbol{\beta}(t'))(\sum_{k'\neq j}\alpha_{jk'}x_{jn}x_{k'n})| \\ &\leq \frac{d_N^3}{2}(\frac{\sqrt{sR^2}}{4})\cdot(2B_N\sum_{(j,k)\in\mathcal{A}}\alpha_{jk}^2) = \frac{d_N^3 B_N}{4}\sqrt{s}R^3. \end{aligned} \tag{A.17}$$

In the last step we have used $|\bar{\eta}_{jn}(\boldsymbol{\beta}(t'))| \leq \frac{1}{4}$ for any j. Moreover,

$$\begin{aligned} B_N &\leq B + \lambda_{\max}(\tfrac{1}{N}\textstyle\sum_{n=1}^N \boldsymbol{x}_{\mathcal{A}n}\boldsymbol{x}_{\mathcal{A}n}^T - E[\boldsymbol{x}_{\mathcal{A}}\boldsymbol{x}_{\mathcal{A}}^T]) \\ &\leq B + \|\tfrac{1}{N}\textstyle\sum_{n=1}^N \boldsymbol{x}_{\mathcal{A}n}\boldsymbol{x}_{\mathcal{A}n}^T - E[\boldsymbol{x}_{\mathcal{A}}\boldsymbol{x}_{\mathcal{A}}^T]\|_F \end{aligned}$$

Since $x_{jn} = \pm 1$, we apply Hoeffding's inequality and the union bound to obtain the following probability bound

$$\Pr(\|\frac{1}{N}\sum_{n=1}^N \boldsymbol{x}_{\mathcal{A}n}\boldsymbol{x}_{\mathcal{A}n}^T - E[\boldsymbol{x}_{\mathcal{A}}\boldsymbol{x}_{\mathcal{A}}^T]\|_F \geq B/2) \leq 2s^2\exp(-\frac{NB^2}{8s^2}),$$

which leads to

$$\Pr(|T_3| \leq \frac{3d_N^3 B}{8}\sqrt{s}R^3) \geq 1 - 2s^2\exp(-\frac{NB^2}{8s^2}). \tag{A.18}$$

For T_4 it is easy to see

$$|T_4| \leq \lambda d_N \sum_{(j,k)\in\mathcal{A}} |\alpha_{jk}| \leq \lambda d_N\sqrt{s}R. \tag{A.19}$$

Let $\epsilon = d_N R^2\frac{b}{8}$ then

$$T_1 + T_2 + T_3 + T_4 \geq \frac{bR^2}{8}d_N^2 - \frac{3B}{8}R^3 d_N^3\sqrt{s} - \lambda d_N\sqrt{s}R \equiv L(d_N, \lambda) \tag{A.20}$$

with probability at least

$$1 - \exp(-R^2\frac{b^2}{8^3}\frac{Nd_N^2}{s}) - 2s^2\exp(-\frac{N}{s^2}\frac{b^2}{2}) - 2s^2\exp(-\frac{N}{s^2}\frac{B^2}{8}).$$

Furthermore, let $d_N = \lambda\sqrt{s}\frac{16}{bR}$ and $\lambda s < \frac{8b^2}{3B}$. Then $L(d_N, \lambda) > 0$ and hence part (1) is proved.

We now prove part (2). Let $\widetilde{\boldsymbol{\beta}} = (\widetilde{\boldsymbol{\beta}}_{\mathcal{A}}, 0)$. There exists some $t \in [0, 1]$ such that

$$\psi(\widetilde{\boldsymbol{\beta}}) = \psi(\boldsymbol{\beta}^*) + H^*(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^*) + (H(\boldsymbol{\beta}^* + t(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)) - H^*)(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^*) \tag{A.21}$$

We call the third term r (the reminder). To prove $\widetilde{\boldsymbol{\beta}} = (\widetilde{\boldsymbol{\beta}}_{\mathcal{A}}, 0) = \widehat{\boldsymbol{\beta}}^{lasso}$, we only need to show

$$\|\psi_{\mathcal{A}^c}(\tilde{\boldsymbol{\beta}})\|_\infty \leq \lambda. \tag{A.22}$$

On the other hand, we have

$$\psi_{\mathcal{A}}(\tilde{\boldsymbol{\beta}}) = -\lambda z \tag{A.23}$$

for some vector z and $\|z\|_\infty \leq 1$. Using (A.21) we can write (A.23) as

$$\psi_{\mathcal{A}} + H^*_{\mathcal{A}\mathcal{A}}(\widetilde{\boldsymbol{\beta}}_{\mathcal{A}} - \boldsymbol{\beta}^*_{\mathcal{A}}) + r_{\mathcal{A}} = -\lambda z$$

which yields

$$\widetilde{\boldsymbol{\beta}}_{\mathcal{A}} - \boldsymbol{\beta}^*_{\mathcal{A}} = (H^*_{\mathcal{A}\mathcal{A}})^{-1}(-\lambda z - \psi_{\mathcal{A}} - r_{\mathcal{A}}). \tag{A.24}$$

Combining (A.24) and (A.22), we need to show

$$\|H^*_{\mathcal{A}^c\mathcal{A}}(H^*_{\mathcal{A}\mathcal{A}})^{-1}(-\lambda z - \psi_{\mathcal{A}} - r_{\mathcal{A}}) + \psi_{\mathcal{A}^c} + r_{\mathcal{A}^c}\|_\infty \leq \lambda \tag{A.25}$$

which is guaranteed if

$$\|H^*_{\mathcal{A}^c\mathcal{A}}(H^*_{\mathcal{A}\mathcal{A}})^{-1}\|_\infty(\|\psi\|_\infty + \|r\|_\infty + \lambda) + \|\psi\|_\infty + \|r\|_\infty \leq \lambda. \tag{A.26}$$

By the condition $\|E[H^*_{\mathcal{A}^c\mathcal{A}}](E[H^*_{\mathcal{A}\mathcal{A}}])^{-1}\|_\infty \leq 1 - \eta$, (A.26) holds if

$$\|H^*_{\mathcal{A}^c\mathcal{A}}(H^*_{\mathcal{A}\mathcal{A}})^{-1} - E[H^*_{\mathcal{A}^c\mathcal{A}}](E[H^*_{\mathcal{A}\mathcal{A}}])^{-1}\|_\infty < \eta/2, \tag{A.27}$$

$$\|\psi\|_\infty \leq \lambda\frac{\eta/4}{2-\eta/2}, \quad \text{and} \quad \|r\|_\infty \leq \lambda\frac{\eta/4}{2-\eta/2}. \tag{A.28}$$

We first provide probability bounds for the two events in (A.28). Using Hoeffding's

inequality and the union bound, we have

$$\Pr(\|\psi\|_\infty < \lambda\frac{\eta/4}{2-\eta/2}) \geq 1 - K^2\exp(-N\lambda^2\frac{(\eta/4)^2}{8(2-\eta/2)^2}). \tag{A.29}$$

Write $\boldsymbol{\alpha} = \widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^*$, and thus $\boldsymbol{\alpha}_{\mathcal{A}^c} = 0$. Similar to (A.17), by the mean value theorem we have a bound for $r_{(jk)}$

$$\begin{aligned} |r_{(jk)}| &= |\frac{1}{N}\sum_{n=1}^{N}\sum_{k_2\neq j}\sum_{k'\neq j} x_{kn}x_{jn}x_{k_2n}x_{k'n}\alpha_{jk_2}\alpha_{jk'}t'\bar{\eta}_{jn}(\boldsymbol{\beta}(t')) \\ &\quad + \frac{1}{N}\sum_{n=1}^{N}\sum_{j_2\neq k}\sum_{j'\neq k} x_{jn}x_{kn}x_{j_2n}x_{j'n}\alpha_{kj_2}\alpha_{kj'}t'\bar{\eta}_{kn}(\boldsymbol{\beta}(t'))| \\ &\leq B_N\cdot\|\widetilde{\boldsymbol{\beta}}_{\mathcal{A}} - \boldsymbol{\beta}^*_{\mathcal{A}}\|_2^2. \end{aligned} \tag{A.30}$$

In the last step we have used $|\bar{\eta}_{jn}(\boldsymbol{\beta}(t'))| \leq \frac{1}{4}$ for any j and $\boldsymbol{\alpha}_{\mathcal{A}^c} = 0$. Moreover, recall that

$$B_N \leq B + \|\frac{1}{N}\sum_{n=1}^{N}\boldsymbol{x}_{\mathcal{A}n}\boldsymbol{x}_{\mathcal{A}n}^T - E[\boldsymbol{x}_{\mathcal{A}}\boldsymbol{x}_{\mathcal{A}}^T]\|_F.$$

Letting

$$\lambda s < \min(\frac{b^2}{16^2B}\frac{\eta/3}{4-\eta}, \frac{8b^2}{3B})$$

and using part (1), it follows that

$$\Pr\left(\|r\|_\infty < \lambda\frac{\eta/4}{2-\eta/2}\right) \geq 1 - e^{-N\lambda^2/2} - 2s^2[\exp(\frac{-Nb^2}{2s^2}) + \exp(\frac{-NB^2}{8s^2})]. \tag{A.31}$$

We now provide a probability bound for the event in (A.27). First we have

$$c = \|(E[H^*_{\mathcal{A}\mathcal{A}}])^{-1}\|_\infty \leq \sqrt{s}\|(E[H^*_{\mathcal{A}\mathcal{A}}])^{-1}\|_2 \leq \sqrt{s}/b.$$

For notation convenience define

$$\delta = \|H^*_{\mathcal{A}^c\mathcal{A}}(H^*_{\mathcal{A}\mathcal{A}})^{-1} - E[H^*_{\mathcal{A}^c\mathcal{A}}](E[H^*_{\mathcal{A}\mathcal{A}}])^{-1}\|_\infty$$

$$\delta_1 = \|(H^*_{\mathcal{A}\mathcal{A}})^{-1} - (E[H^*_{\mathcal{A}\mathcal{A}}])^{-1}\|_\infty$$

$$\delta_2 = \|H^*_{\mathcal{A}\mathcal{A}} - E[H^*_{\mathcal{A}\mathcal{A}}]\|_\infty$$

$$\delta_3 = \|H^*_{\mathcal{A}^c\mathcal{A}} - E[H^*_{\mathcal{A}^c\mathcal{A}}]\|_\infty$$

Then by definition

$$\begin{aligned}
\delta &= \|(H^*_{\mathcal{A}^c\mathcal{A}} - E[H^*_{\mathcal{A}^c\mathcal{A}}])((H^*_{\mathcal{A}\mathcal{A}})^{-1} - (E[H^*_{\mathcal{A}\mathcal{A}}])^{-1}) \\
&\quad + E[H^*_{\mathcal{A}^c\mathcal{A}}](E[H^*_{\mathcal{A}\mathcal{A}}])^{-1}(-H^*_{\mathcal{A}\mathcal{A}} + E[H^*_{\mathcal{A}\mathcal{A}}])(H^*_{\mathcal{A}\mathcal{A}})^{-1} \\
&\quad + (H^*_{\mathcal{A}^c\mathcal{A}} - E[H^*_{\mathcal{A}^c\mathcal{A}}])(E[H^*_{\mathcal{A}\mathcal{A}}])^{-1}\|_\infty \\
&\leq \delta_3\delta_1 + (1-\eta)\delta_2\|(H^*_{\mathcal{A}\mathcal{A}})^{-1}\|_\infty + \delta_3 c \\
&\leq \delta_3\delta_1 + (1-\eta)(c+\delta_1)\delta_2 + \delta_3 c. \qquad \text{(A.32)}
\end{aligned}$$

Note that

$$\begin{aligned}
\delta_1 &= \|(H^*_{\mathcal{A}\mathcal{A}})^{-1}(E[H^*_{\mathcal{A}\mathcal{A}}] - H^*_{\mathcal{A}\mathcal{A}})(E[H^*_{\mathcal{A}\mathcal{A}}])^{-1}\|_\infty \\
&\leq \|(H^*_{\mathcal{A}\mathcal{A}})^{-1}\|_\infty \cdot \|E[H^*_{\mathcal{A}\mathcal{A}}] - H^*_{\mathcal{A}\mathcal{A}}\|_\infty \cdot \|(E[H^*_{\mathcal{A}\mathcal{A}}])^{-1}\|_\infty \\
&\leq (\delta_1 + c)\delta_2 c.
\end{aligned}$$

Hence as long as $\delta_2 c < 1$ we have $\delta_1 \leq \frac{\delta_2 c^2}{1-\delta_2 c}$ and (A.32) yields

$$\delta \leq (\delta_3 + (1-\eta)\delta_2)\frac{c}{1-\delta_2 c}. \qquad \text{(A.33)}$$

$$\begin{aligned}\Pr(\delta_2 < \frac{\eta}{4c(1-\eta/2)}) &\geq 1 - \Pr(\|H^*_{\mathcal{A}\mathcal{A}} - E[H^*_{\mathcal{A}\mathcal{A}}]\|_{\max} > \frac{\eta}{4cs(1-\eta/2)}) \\ &\geq 1 - 2s^2 \exp(-\frac{N\eta^2}{2c^2s^2(2-\eta)^2}). \end{aligned} \tag{A.34}$$

$$\begin{aligned}\Pr(\delta_3 < \frac{\eta}{4c}) &\geq 1 - \Pr(\|H^*_{\mathcal{A}^c\mathcal{A}} - E[H^*_{\mathcal{A}^c\mathcal{A}}]\|_{\max} > \frac{\eta}{4cs}) \\ &\geq 1 - K^2 s \exp(-\frac{N\eta^2}{8c^2s^2}). \end{aligned} \tag{A.35}$$

Therefore,

$$\Pr(\delta < \eta/2) \geq 1 - 2s^2 \exp(-\frac{N\eta^2}{2c^2s^2(2-\eta)^2}) - K^2 s \exp(-\frac{N\eta^2}{8c^2s^2}) \tag{A.36}$$

$$\geq 1 - 2s^2 \exp(-\frac{Nb^2\eta^2}{2s^3(2-\eta)^2}) - K^2 s \exp(-\frac{Nb^2\eta^2}{8s^3}) \tag{A.37}$$

Combining (A.29), (A.31) and (A.37), we obtain the desired result. This completes the proof. □

A.3 Proof of Corollary 2.1

Proof A.3 Corollary 2.1 follows directly from Theorems 2.1 and 2.2, and thus we omit its proof here. □

A.4 Proof of Theorem 2.3

Proof A.4 Under the event $\{\|\widetilde{\boldsymbol{\beta}}^{(0)} - \boldsymbol{\beta}^*\|_\infty \leq \lambda\}$, we have $|\widetilde{\boldsymbol{\beta}}^{(0)}_{jk}| \leq \lambda$ for $(j,k) \in \mathcal{A}^c$ and $|\widetilde{\boldsymbol{\beta}}^{(0)}_{jk}| \geq a\lambda$ for $(j,k) \in \mathcal{A}$. Therefore, $\widetilde{\boldsymbol{\beta}}^{(1)}$ is the solution of the following penalized

composite likelihood

$$\widehat{\boldsymbol{\beta}}^{(1)} = \arg\max_{\boldsymbol{\beta}}\{\ell_c(\boldsymbol{\beta}) - \lambda \sum_{(j,k)\in\mathcal{A}^c} |\beta_{jk}|\}. \tag{A.38}$$

It turns out that $\widehat{\boldsymbol{\beta}}^{oracle}$ is the global solution of (A.38) under the additional probability event that $\{\|\psi_{\mathcal{A}^c}(\widehat{\boldsymbol{\beta}}^{oracle})\|_\infty \le \lambda\}$. To see this, we observe that for any $\boldsymbol{\beta}$

$$\begin{aligned}
& (-\ell_c(\boldsymbol{\beta}) + \lambda \sum_{(j,k)\in\mathcal{A}^c} |\beta_{jk}|) - (-\ell_c(\widehat{\boldsymbol{\beta}}^{oracle}) + \lambda \sum_{(j,k)\in\mathcal{A}^c} |\hat{\beta}^{oracle}_{jk}|) \\
\ge & \sum_{(j,k)\in\mathcal{A}^c} (\lambda - |\psi_{(jk)}(\widehat{\boldsymbol{\beta}}^{oracle})|) \cdot |\beta_{jk}| \\
\ge & \ 0,
\end{aligned}$$

where we used the convexity of $-\ell_c$. In the proof of Theorem 2.1 we have shown that

$$\begin{aligned}
& \Pr(\|\psi_{\mathcal{A}^c}(\widehat{\boldsymbol{\beta}}^{oracle})\|_\infty > \lambda) \\
< & \ K^2 \exp(-\frac{N\lambda^2}{32(2\phi+1)^2}) + \exp(-\frac{N\lambda}{3B(2\phi+1)s}\frac{b^2}{8^3}) + K^2 s \exp(-\frac{Nb^2}{2s^3}) \\
& +2s^2[\exp(-\frac{b^2 N}{8s^3}) + \exp(-\frac{N}{s^2}\frac{b^2}{2}) + \exp(-\frac{N}{s^2}\frac{B^2}{8})] \\
\equiv & \ \tau_3.
\end{aligned}$$

Therefore, the LLA-CMA algorithm finds the oracle estimator with probability at least $1 - \tau_3 - \Pr(\|\widetilde{\boldsymbol{\beta}}^{(0)} - \boldsymbol{\beta}^*\|_\infty > \lambda)$. This proves part (1).

If we further consider the event $\{\min_{(j,k)\in\mathcal{A}} |\hat{\beta}^{oracle}_{jk}| > a\lambda\}$. Then $\widetilde{\boldsymbol{\beta}}^{(2)}$ is the solution of the following penalized composite likelihood

$$\max_{\boldsymbol{\beta}}\{\ell_c(\boldsymbol{\beta}) - \lambda \sum_{(j,k)\in\mathcal{A}^c} |\beta_{jk}|\},$$

which implies that $\widetilde{\boldsymbol{\beta}}^{(2)} = \widetilde{\boldsymbol{\beta}}^{(1)}$ and hence the LLA loop will stop. From (A.6) we have obtained a probability bound for the event of $\{\min_{(j,k)\in\mathcal{A}} |\hat{\beta}_{jk}^{oracle}| \leq a\lambda\}$ as follows

$$\begin{aligned} & \Pr(\min_{(j,k)\in\mathcal{A}} |\tilde{\beta}_{jk}^{hmle}| \leq a\lambda) \\ \leq\ & \exp(-R_*^2 \frac{b^2}{8^3}) + 2s^2 \exp(-\frac{N}{s^2}\frac{b^2}{2}) + 2s^2 \exp(-\frac{N}{s^2}\frac{B^2}{8}) \\ \equiv\ & \tau_4. \end{aligned}$$

Then we have $\widetilde{\boldsymbol{\beta}}^{(m)} = \widetilde{\boldsymbol{\beta}}^{(1)} = \widehat{\boldsymbol{\beta}}^{oracle}$ for $m = 2, 3, \ldots$ which means the LLA-CMA algorithm converges after two LLA iteration and finds the oracle estimator with probability at least $1 - \tau_3 - \Pr(\|\widetilde{\boldsymbol{\beta}}^{(0)} - \boldsymbol{\beta}^*\|_\infty > \lambda) - \tau_4$. Note that $\tau_3 + \tau_4 = \tau_2$. This proves part (2). □

A.5 Proof of Corollary 2.2

Proof A.5 Part (1) follows directly from Theorem 2.3. We only prove part (2). With the chosen λ^{lasso}, Theorem 2.2 shows that with probability tending to one, $\widehat{\boldsymbol{\beta}}^{lasso} = (\widehat{\boldsymbol{\beta}}_{\mathcal{A}}^{lasso}, \widehat{\boldsymbol{\beta}}_{\mathcal{A}^c}^{lasso})$ satisfies that $\widehat{\boldsymbol{\beta}}_{\mathcal{A}}^{lasso} = \widetilde{\boldsymbol{\beta}}_{\mathcal{A}}$, $\widehat{\boldsymbol{\beta}}_{\mathcal{A}^c}^{lasso} = 0$ and

$$\Pr(\|\widetilde{\boldsymbol{\beta}}_{\mathcal{A}} - \boldsymbol{\beta}_{\mathcal{A}}^*\|_2 \leq 16\lambda^{lasso}\sqrt{s}/b) \to 0.$$

Note that $16\lambda^{lasso}\sqrt{s}/b < \lambda^{scad}$ and $\|\widetilde{\boldsymbol{\beta}}_{\mathcal{A}} - \boldsymbol{\beta}_{\mathcal{A}}^*\|_\infty \leq \|\widetilde{\boldsymbol{\beta}}_{\mathcal{A}} - \boldsymbol{\beta}_{\mathcal{A}}^*\|_2$, we then conclude

$$\tau_0 = \Pr(\|\widehat{\boldsymbol{\beta}}^{lasso} - \boldsymbol{\beta}^*\|_\infty \leq \lambda^{scad}) \to 0. \qquad \square$$

Appendix B

Proof of Chapter 3

B.1 Proof of Lemma 3.1

Proof B.1 First, Spearman's rank correlation $\hat{r}_{ij}$ can be written in terms of the Hoeffding decomposition (Hoeffding, 1948)

$$\hat{r}_{ij} = \frac{n-2}{n+1} u_{ij} + \frac{3}{n+1} d_{ij} \tag{B.1}$$

where

$$d_{ij} = \frac{1}{n(n-1)} \sum_{k \neq l} \text{sign}(x_{ki} - x_{li}) \cdot \text{sign}(x_{kj} - x_{lj}),$$

and

$$u_{ij} = \frac{3}{n(n-1)(n-2)} \sum_{k \neq l, k \neq m, l \neq m} \text{sign}(x_{ki} - x_{li}) \cdot \text{sign}(x_{kj} - x_{mj}). \tag{B.2}$$

Direct calculation yields that

$$\begin{aligned} \mathrm{E}(u_{ij}) &= 3 \cdot \mathrm{E}\big[\mathrm{sign}(x_{ki} - x_{li}) \cdot \mathrm{sign}(x_{kj} - x_{mj})\big] \\ &= 12 \cdot \mathrm{E}\big[(\Phi(Z_i) - \frac{1}{2})(\Phi(Z_j) - \frac{1}{2})\big] \\ &= \frac{6}{\pi} \sin^{-1}(\frac{\sigma_{ij}}{2}), \end{aligned}$$

where the last step follows from Kendall (1948). Then we can write

$$\sigma_{ij} = 2\sin(\frac{\pi}{6}E(u_{ij})).$$

By definition $\hat{r}^s_{ij} = 2\sin(\frac{\pi}{6}\hat{r}_{ij})$. Note that $2\sin(\frac{\pi}{6}\cdot)$ is a Lipschitz function with the Lipschitz constant $\pi/3$. Then we have

$$\Pr(|\hat{r}^s_{ij} - \sigma_{ij}| > \varepsilon) \leq \Pr(|\hat{r}_{ij} - E(u_{ij})| > \frac{3\varepsilon}{\pi}).$$

Applying (B.1) and (B.2) yields that

$$\hat{r}_{ij} - E(u_{ij}) = u_{ij} - E(u_{ij}) + \frac{3}{n+1}d_{ij} - \frac{3}{n+1}u_{ij}.$$

Note that $|u_{ij}| \leq 3$ and $|d_{ij}| \leq 1$. Hence, both

$$|u_{ij}| \leq \frac{\varepsilon}{4\pi}(n+1)$$

and

$$|d_{ij}| \leq \frac{\varepsilon}{4\pi}(n+1)$$

always hold provided that $n > 12\pi/\varepsilon$, which are satisfied by the assumption in Lemma 1. For such chosen n, we have

$$\Pr(|r_{ij} - E(u_{ij})| > \frac{3\varepsilon}{\pi}) \leq \Pr(|u_{ij} - E(u_{ij})| > \frac{3\varepsilon}{2\pi}).$$

Lastly, we observe that u_{ij} is a function of independent samples $(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$. If we replace the t-th sample by some $\tilde{\boldsymbol{x}}_t$, we make a claim that the change in u_{ij} will be bounded as

$$\sup_{\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n,\tilde{\boldsymbol{x}}_t} |u_{ij}(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n) - u_{ij}(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{t-1}, \tilde{\boldsymbol{x}}_t, \boldsymbol{x}_{t+1}, \ldots, \boldsymbol{x}_n)| \leq \frac{15}{n}. \tag{B.3}$$

Then applying the McDiarmid's inequality (McDiarmid, 1989) concludes that for some absolute constant $c_0 > 0$,

$$\Pr(|\hat{r}_{ij}^s - \sigma_{ij}| > \varepsilon) \leq \Pr(|u_{ij} - E(u_{ij})| \geq \frac{3\varepsilon}{2\pi}) \leq 2\exp(-c_0 n\varepsilon^2).$$

Now it remains to verify (B.3) to complete the proof of Lemma 1. We provide a brief proof for this claim. To this end, we assume that $\boldsymbol{x}_t = (x_{1t}, \cdots, x_{pt})'$ is replaced by $\tilde{\boldsymbol{x}}_t = (\tilde{x}_{1t}, \cdots, \tilde{x}_{pt})'$, and we want to prove that the change of u_{ij} is at most $\frac{15}{n}$. Without loss of generality we may assume that

$$n_i = \#\{s : \text{sign}(\tilde{x}_{ti} - x_{si}) = -\text{sign}(x_{ti} - x_{si}),\ s \neq t\}$$

and

$$n_j = \#\{s : \text{sign}(\tilde{x}_{tj} - x_{sj}) = -\text{sign}(x_{tj} - x_{sj}),\ s \neq t\}.$$

Recall that

$$u_{ij} = \frac{3}{n(n-1)(n-2)} \sum_{k\neq l, k\neq m, l\neq m} \text{sign}(x_{ki} - x_{li}) \cdot \text{sign}(x_{kj} - x_{mj}).$$

Then we have

$$\begin{aligned}
& \big|u_{ij}(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n) - u_{ij}(\boldsymbol{x}_1..\boldsymbol{x}_{t-1}, \tilde{\boldsymbol{x}}_t, \boldsymbol{x}_{t+1}..\boldsymbol{x}_n)\big| \\
\leq\ & \frac{3}{n(n-1)(n-2)} \Big| \sum_{k\neq t, k\neq m, m\neq t} \big(\text{sign}(x_{ki} - x_{ti}) - \text{sign}(x_{ki} - \tilde{x}_{ti})\big) \cdot \text{sign}(x_{kj} - x_{mj}) \\
& + \sum_{k\neq t, k\neq l, l\neq t} \big(\text{sign}(x_{kj} - x_{tj}) - \text{sign}(x_{kj} - \tilde{x}_{tj})\big) \cdot \text{sign}(x_{ki} - x_{li}) \\
& + \sum_{l\neq t, m\neq t, l\neq m} \big(\text{sign}(x_{ti} - x_{li}) \cdot \text{sign}(x_{tj} - x_{mj}) - \text{sign}(\tilde{x}_{ti} - x_{li}) \cdot \text{sign}(\tilde{x}_{tj} - x_{mj})\big)\Big| \\
\leq\ & \frac{3}{n(n-1)(n-2)} \cdot 2[n_i(n-2) + n_j(n-2) + n_j(n-1-n_i) + n_i(n-1-n_j)] \\
=\ & \frac{6[(n_i + n_j)(2n-3) - 2n_i n_j]}{n(n-1)(n-2)} \\
\leq\ & \frac{12}{n}(1 + \frac{1}{4} \cdot \frac{1}{(n-1)(n-2)}) \\
\leq\ & \frac{15}{n}
\end{aligned}$$

where the third inequality holds if and only if $n_i = n_j = n - \frac{3}{2}$. This completes the proof of Lemma 3.1. □

B.2 Proof of Theorem 3.1

Proof B.2 By Lemma 3 of Ravikumar et al. (2008), $\hat{\boldsymbol{\Theta}}_g^s \succ 0$ is uniquely characterized by the sub-differential optimality condition,

$$\hat{\boldsymbol{R}}^s - (\hat{\boldsymbol{\Theta}}_g^s)^{-1} + \lambda \hat{\boldsymbol{Z}} = \boldsymbol{0},$$

where $\hat{\boldsymbol{Z}}$ is the sub-differential with respect to $\hat{\boldsymbol{\Theta}}_g^s$ satisfying

$$\hat{z}_{ij} = \begin{cases} 0 & \text{if } i = j; \\ \text{sign}(\hat{\theta}_{ij}^s) & \text{if } i \neq j \ \& \ \hat{\theta}_{ij}^s \neq 0; \\ \in [-1, +1] & \text{if } i \neq j \ \& \ \hat{\theta}_{ij}^s = 0. \end{cases}$$

Define the oracle estimator $\tilde{\boldsymbol{\Theta}}_g^s$ exactly supported in the true support set $\mathcal{A}$ by

$$\tilde{\boldsymbol{\Theta}}_g^s = \arg \min_{\boldsymbol{\Theta} \succ 0, \ \boldsymbol{\Theta}_{\mathcal{A}^c} = \boldsymbol{0}} - \log\det(\boldsymbol{\Theta}) + \text{tr}(\hat{\boldsymbol{\Sigma}}^o \boldsymbol{\Theta}) + \lambda \sum_{i \neq j} |\theta_{ij}|. \tag{B.4}$$

Then we can construct $\tilde{\boldsymbol{Z}}^s$ to satisfy that

$$\hat{\boldsymbol{R}}^s - (\tilde{\boldsymbol{\Theta}}_g^s)^{-1} + \lambda \tilde{\boldsymbol{Z}}^s = \boldsymbol{0}. \tag{B.5}$$

Follow the same line of the proof in Ravikumar et al. (2008), we can show that $\tilde{\boldsymbol{Z}}^s$ is the sub-differential of $\tilde{\boldsymbol{\Theta}}_g^s$. Define $\boldsymbol{\Delta} = \hat{\boldsymbol{\Theta}}_g^s - \boldsymbol{\Theta}^*$, $\boldsymbol{W} = \hat{\boldsymbol{R}}^s - \boldsymbol{\Sigma}^*$ and $\boldsymbol{U} = (\tilde{\boldsymbol{\Theta}}_g^s)^{-1} - \boldsymbol{\Sigma}^* + \boldsymbol{\Sigma}^* \boldsymbol{\Delta} \boldsymbol{\Sigma}^*$. We use the notation $\vec{\boldsymbol{A}} = \text{vec}(\boldsymbol{A})$ to denote the vectorization of the matrix $\boldsymbol{A}$. Notice that $\boldsymbol{H}^* \vec{\boldsymbol{\Delta}} = (\boldsymbol{\Sigma}^* \bigotimes \boldsymbol{\Sigma}^*) \vec{\boldsymbol{\Delta}} = \boldsymbol{\Sigma}^* \boldsymbol{\Delta} \boldsymbol{\Sigma}^*$ by the properties of Kronecker product. Now the equation (B.5) can be written as

$$\boldsymbol{\Sigma}^* \boldsymbol{\Delta} \boldsymbol{\Sigma}^* + \boldsymbol{W} - \boldsymbol{U} + \lambda \hat{\boldsymbol{Z}}^s = \boldsymbol{0},$$

or equivalently as

$$\boldsymbol{H}^* \vec{\boldsymbol{\Delta}} + \vec{\boldsymbol{W}} - \vec{\boldsymbol{U}} + \lambda \vec{\hat{\boldsymbol{Z}}}^{\boldsymbol{s}} = \vec{\boldsymbol{0}}.$$

Now partition the matrices $\boldsymbol{H}^*$ and $\vec{\tilde{\boldsymbol{Z}}}^{\boldsymbol{s}}$ according to $\mathcal{A}$ as

$$\begin{pmatrix} \boldsymbol{H}^*_{\mathcal{A}\mathcal{A}} & \boldsymbol{H}^*_{\mathcal{A}\mathcal{A}^c} \\ \boldsymbol{H}^*_{\mathcal{A}^c\mathcal{A}} & \boldsymbol{H}^*_{\mathcal{A}^c\mathcal{A}^c} \end{pmatrix} \quad \text{and} \quad \begin{pmatrix} \vec{\tilde{\boldsymbol{Z}}}^{\boldsymbol{s}}_{\boldsymbol{\mathcal{A}}} \\ \vec{\tilde{\boldsymbol{Z}}}^{\boldsymbol{s}}_{\boldsymbol{\mathcal{A}^c}} \end{pmatrix}$$

respectively. Similarly, we can partition $\vec{\boldsymbol{\Delta}}$, $\vec{\boldsymbol{W}}$ and $\vec{\boldsymbol{U}}$ according to $\mathcal{A}$ as well. By definition, we have $\vec{\boldsymbol{\Delta}}_{\boldsymbol{\mathcal{A}^c}} = \mathbf{0}$, and then the equation (B.5) is further equivalent to the following equations

$$\begin{aligned} \boldsymbol{H}^*_{\mathcal{A}\mathcal{A}}\vec{\boldsymbol{\Delta}}_{\boldsymbol{\mathcal{A}}} + \vec{\boldsymbol{W}}_{\boldsymbol{\mathcal{A}}} - \vec{\boldsymbol{U}}_{\boldsymbol{\mathcal{A}}} + \lambda\vec{\tilde{\boldsymbol{Z}}}^{\boldsymbol{s}}_{\boldsymbol{\mathcal{A}}} &= \mathbf{0} \\ \boldsymbol{H}^*_{\mathcal{A}^c\mathcal{A}}\vec{\boldsymbol{\Delta}}_{\boldsymbol{\mathcal{A}}} + \vec{\boldsymbol{W}}_{\boldsymbol{\mathcal{A}^c}} - \vec{\boldsymbol{U}}_{\boldsymbol{\mathcal{A}^c}} + \lambda\vec{\tilde{\boldsymbol{Z}}}^{\boldsymbol{s}}_{\boldsymbol{\mathcal{A}^c}} &= \mathbf{0} \end{aligned}$$

To show that $\tilde{\boldsymbol{Z}}^s$ is the sub-differential of $\tilde{\boldsymbol{\Theta}}^s_g$, it remains to prove that $\|\vec{\tilde{\boldsymbol{Z}}}^{\boldsymbol{s}}_{\boldsymbol{\mathcal{A}^c}}\|_{\ell_\infty} < 1$. Solving the above linear equations yields that

$$\lambda\vec{\tilde{\boldsymbol{Z}}}^{\boldsymbol{*}}_{\boldsymbol{\mathcal{A}^c}} = -\vec{\boldsymbol{W}}_{\boldsymbol{\mathcal{A}^c}} + \vec{\boldsymbol{U}}_{\boldsymbol{\mathcal{A}^c}} - \boldsymbol{H}^*_{\mathcal{A}^c\mathcal{A}}(\boldsymbol{H}^*_{\mathcal{A}\mathcal{A}})^{-1}(\vec{\boldsymbol{W}}_{\boldsymbol{\mathcal{A}}} - \vec{\boldsymbol{U}}_{\boldsymbol{\mathcal{A}}} + \lambda\vec{\tilde{\boldsymbol{Z}}}^{\boldsymbol{*}}_{\boldsymbol{\mathcal{A}}}).$$

We claim that $\|\boldsymbol{U}\|_{\max} \leq \frac{\kappa}{4}\lambda$ holds under the event $\mathcal{S}_1 = \{\|\boldsymbol{W}\|_{\max} \leq \frac{\kappa}{4}\lambda\}$, which will be justified later. Then it is easy to see that under the event $\mathcal{S}_1$, the oracle estimator $\tilde{\boldsymbol{\Theta}}^s_g$ is exactly the same as the ℓ_1-penalized semiparametric likelihood estimator $\hat{\boldsymbol{\Theta}}^s_g$ since $\tilde{\boldsymbol{\Theta}}^s_g$ satisfies that $\tilde{\boldsymbol{Z}}^s_{\mathcal{A}} = \text{sign}(\hat{\boldsymbol{\Theta}}^*_{\mathcal{A}})$ by construction and also that

$$\|\vec{\tilde{\boldsymbol{Z}}}^{\boldsymbol{s}}_{\boldsymbol{\mathcal{A}^c}}\|_{\ell_\infty} \leq \frac{\kappa}{2} + (1-\kappa)(\frac{\kappa}{2}+1) = 1 - \frac{\kappa^2}{2} < 1,$$

where we use the assumption $\|\boldsymbol{H}^*_{\mathcal{A}^c\mathcal{A}}(\boldsymbol{H}^*_{\mathcal{A}\mathcal{A}})^{-1}\|_{\ell_\infty} < 1-\kappa$ and the fact $\|\vec{\tilde{\boldsymbol{Z}}}^{\boldsymbol{s}}_{\boldsymbol{\mathcal{A}}}\|_{\ell_\infty} \leq 1$.

Now it remains to prove that $\|\boldsymbol{U}\|_{\max} \leq \frac{\kappa}{4}\lambda$ holds under $\mathcal{S}_1$. Picking λ such that

$$(1+\frac{\kappa}{4})\lambda \leq \frac{1}{6d}\cdot\min\left\{\frac{1}{K_{\boldsymbol{\Sigma}^*}K_{\boldsymbol{H}^*}}, \frac{1}{(1+\frac{4}{\kappa})K^3_{\boldsymbol{\Sigma}^*}K^2_{\boldsymbol{H}^*}}\right\}, \tag{B.6}$$

Lemma 6 in Ravikumar et al. (2008) suggests that under the event $\mathcal{S}_1$,

$$\|\boldsymbol{\Delta}\|_{\max} \leq 2K_{\boldsymbol{H}^*}(\|\boldsymbol{W}\|_{\max} + \lambda) \leq 2K_{\boldsymbol{H}^*}(1 + \frac{\kappa}{4})\lambda.$$

Obviously $\|\boldsymbol{\Delta}\|_{\max} \leq (3dK_{\boldsymbol{\Sigma}^*})^{-1}$ holds. Lemma 5 in Ravikumar et al. (2008) yields

$$\|\boldsymbol{U}\|_{\max} \leq \frac{3}{2}d\|\boldsymbol{\Delta}\|^2_{\max}K^3_{\boldsymbol{\Sigma}^*}. \tag{B.7}$$

Now combining both (B.6) and (B.7) implies that

$$\|\boldsymbol{U}\|_{\max} \leq 6dK^3_{\boldsymbol{\Sigma}^*}K^2_{\boldsymbol{H}^*}(1 + \frac{\kappa}{4})^2\lambda^2 \leq \frac{\kappa}{4}\lambda.$$

Then the rate of convergence under matrix ℓ_1 norm can be derived from the entry-wise ℓ_∞ bound. Besides, as long as λ is chosen to satisfy (B.6) and also that $\lambda \leq \frac{\kappa}{48\pi}n$, we can obtain the following probability lower bound,

$$\Pr(\hat{\boldsymbol{\Theta}}^s_g = \tilde{\boldsymbol{\Theta}}^s_g) \geq \Pr(\mathcal{S}_1) \geq 1 - p^2\exp(-\frac{\kappa^2}{16}c_0 n\lambda^2).$$

The selection consistency can be similarly proved. Under the same event $\mathcal{S}_1$, $\hat{\boldsymbol{\Theta}}^s_g$ satisfies $\|\hat{\boldsymbol{\Theta}}^s_g - \boldsymbol{\Theta}^*\|_{\max} \leq 2K_{\boldsymbol{H}^*}(1 + \frac{\kappa}{4})\lambda$. The sign consistency can be easily obtained as claimed by noting the fact that $\psi_{\min} > 2K_{\boldsymbol{H}^*}(1 + \frac{\kappa}{4})\lambda$. This completes the proof of Theorem 3.1. □

B.3 Proof of Theorem 3.2

Proof B.3 Note we only need to prove the risk bound for $\hat{\boldsymbol{\Theta}}^s_{nd}$ because its risk uniformly dominates that of $\breve{\boldsymbol{\Theta}}^s_{nd}$ by a factor of 2,

$$\|\breve{\boldsymbol{\Theta}}^s_{nd} - \boldsymbol{\Theta}^*\|_{\ell_1} \leq \|\breve{\boldsymbol{\Theta}}^s_{nd} - \hat{\boldsymbol{\Theta}}^s_{nd}\|_{\ell_1} + \|\hat{\boldsymbol{\Theta}}^s_{nd} - \boldsymbol{\Theta}^*\|_{\ell_1} \leq 2\|\hat{\boldsymbol{\Theta}}^s_{nd} - \boldsymbol{\Theta}^*\|_{\ell_1}.$$

To bound the difference between $\hat{\mathbf{\Theta}}^{s}_{nd}$ and $\mathbf{\Theta}^*$ under the ℓ_1-norm, we should bound $|\hat{\theta}^{s.nd}_{kk} - \theta^*_{kk}|$ and $\|\hat{\boldsymbol{\theta}}^{s.nd}_{(k)} - \boldsymbol{\theta}^*_{(k)}\|_{\ell_1}$ for $k = 1, \ldots, p$, respectively. In the sequel we consider the probability event $\{\|\hat{\boldsymbol{R}}^{s} - \mathbf{\Sigma}^*\|_{\max} \leq \frac{b}{M}\lambda\}$. Under this event, we make a claim that for any $k = 1, \ldots, p$, we have

$$\|\hat{\boldsymbol{R}}^{s}_{(k)}\boldsymbol{\beta}^*_k - \hat{\boldsymbol{r}}^{s}_{(k)}\|_{\ell_\infty} \leq \lambda \quad \text{and} \quad \|\hat{\boldsymbol{\beta}}^{s.nd}_k - \boldsymbol{\beta}^*_k\|_{\ell_1} \leq C_0 d\lambda, \tag{B.8}$$

where C_0 is some quantity depending on b, B and M only. The first inequality in the claim (B.8) implies that $\boldsymbol{\beta}^*_k$ is a feasible solution to the optimization problem (3.11) of the rank-based Dantzig selector and thus $\|\hat{\boldsymbol{\beta}}^{s.nd}_k\|_{\ell_1} \leq \|\boldsymbol{\beta}^*_k\|_{\ell_1}$ naturally holds.

Now we derive a probability bound $|\hat{\theta}^{s.nd}_{kk} - \theta^*_{kk}|$ under the same probability event. To this end, we derive a probability upper bound for $|(\hat{\theta}^{s.nd}_{kk})^{-1} - (\theta^*_{kk})^{-1}|$ first. Recall that $(\hat{\theta}^{s.nd}_{kk})^{-1} = (\hat{\boldsymbol{\beta}}^{s.nd}_k)^T \hat{\boldsymbol{R}}^{s}_{(k)} \hat{\boldsymbol{\beta}}^{s.nd}_k - 2(\hat{\boldsymbol{\beta}}^{s.nd}_k)^T \hat{\boldsymbol{r}}^{s}_{(k)} + 1$ and $(\theta^*_{kk})^{-1} = 1 - (\boldsymbol{\sigma}^*_{(k)})^T \boldsymbol{\beta}^*_k$. Note that $\|\hat{\boldsymbol{R}}^{s}_{(k)}\hat{\boldsymbol{\beta}}^{s.nd}_k - \hat{\boldsymbol{r}}^{s}_{(k)}\|_{\ell_\infty} \leq \lambda$ obviously holds since $\hat{\boldsymbol{\beta}}^{s.nd}_k$ is a feasible solution, and thus by the triangle inequality we have

$$\begin{aligned}
& |(\hat{\theta}^{s.nd}_{kk})^{-1} - (\theta^*_{kk})^{-1}| \qquad (B.9)\\
= \;& |(\hat{\boldsymbol{\beta}}^{s.nd}_k)^T \hat{\boldsymbol{R}}^{s}_{(k)} \hat{\boldsymbol{\beta}}^{s.nd}_k - 2(\hat{\boldsymbol{\beta}}^{s.nd}_k)^T \hat{\boldsymbol{r}}^{s}_{(k)} + (\boldsymbol{\sigma}^*_{(k)})^T \boldsymbol{\beta}^*_k| \\
\leq \;& |(\hat{\boldsymbol{\beta}}^{s.nd}_k)^T \hat{\boldsymbol{R}}^{s}_{(k)} \hat{\boldsymbol{\beta}}^{s.nd}_k - (\hat{\boldsymbol{\beta}}^{s.nd}_k)^T \hat{\boldsymbol{r}}^{s}_{(k)}| + |(\hat{\boldsymbol{\beta}}^{s.nd}_k)^T \hat{\boldsymbol{r}}^{s}_{(k)} - (\boldsymbol{\sigma}^*_{(k)})^T \boldsymbol{\beta}^*_k| \\
\leq \;& \|\hat{\boldsymbol{R}}^{s}_{(k)}\hat{\boldsymbol{\beta}}^{s.nd}_k - \hat{\boldsymbol{r}}^{s}_{(k)}\|_{\ell_\infty} \cdot \|\hat{\boldsymbol{\beta}}^{s.nd}_k\|_{\ell_1} + |(\hat{\boldsymbol{r}}^{s}_{(k)})^T \hat{\boldsymbol{\beta}}^{s.nd}_k - (\boldsymbol{\sigma}^*_{(k)})^T \boldsymbol{\beta}^*_k| \\
\leq \;& \lambda\|\hat{\boldsymbol{\beta}}^{s.nd}_k\|_{\ell_1} + |(\hat{\boldsymbol{r}}^{s}_{(k)})^T \hat{\boldsymbol{\beta}}^{s.nd}_k - (\boldsymbol{\sigma}^*_{(k)})^T \hat{\boldsymbol{\beta}}^{s.nd}_k| + |(\boldsymbol{\sigma}^*_{(k)})^T \hat{\boldsymbol{\beta}}^{s.nd}_k - (\boldsymbol{\sigma}^*_{(k)})^T \boldsymbol{\beta}^*_k| \\
\leq \;& \lambda\|\boldsymbol{\beta}^*_k\|_{\ell_1} + \|\hat{\boldsymbol{R}}^{s} - \mathbf{\Sigma}^*\|_{\max} \cdot \|\hat{\boldsymbol{\beta}}^{s.nd}_k\|_{\ell_1} + \|\boldsymbol{\sigma}^*_{(k)}\|_{\ell_\infty} \cdot \|\hat{\boldsymbol{\beta}}^{s.nd}_k - \boldsymbol{\beta}^*_k\|_{\ell_1} \\
\leq \;& (1 + \frac{b}{M}) \cdot \lambda\|\boldsymbol{\beta}^*_k\|_{\ell_1} + \|\hat{\boldsymbol{\beta}}^{s.nd}_k - \boldsymbol{\beta}^*_k\|_{\ell_1}. \qquad (B.10)
\end{aligned}$$

Notice that

$$|\hat{\theta}^{s.nd}_{kk} - \theta^*_{kk}| = |(\hat{\theta}^{s.nd}_{kk})^{-1} - (\theta^*_{kk})^{-1}| \cdot |\hat{\theta}^{s.nd}_{kk}| \cdot |\theta^*_{kk}|,$$

and

$$|\hat{\theta}_{kk}^{s.nd}| \leq |\hat{\theta}_{kk}^{s.nd} - \theta_{kk}^*| + |\theta_{kk}^*|.$$

Then $|\hat{\theta}_{kk}^{s.nd} - \theta_{kk}^*|$ can be upper bounded as follows,

$$|\hat{\theta}_{kk}^{s.nd} - \theta_{kk}^*| \leq \frac{|(\hat{\theta}_{kk}^{s.nd})^{-1} - (\theta_{kk}^*)^{-1}| \cdot |\theta_{kk}^*|^2}{1 - |(\hat{\theta}_{kk}^{s.nd})^{-1} - (\theta_{kk}^*)^{-1}| \cdot |\theta_{kk}^*|}. \tag{B.11}$$

Recall that $\boldsymbol{\beta}_k^* = -(\theta_{kk}^*)^{-1}\boldsymbol{\theta}_{(k)}^*$, and then it is easy to show that

$$\|\boldsymbol{\beta}_k^*\|_{\ell_1} = \frac{\|\boldsymbol{\theta}_{(k)}^*\|_{\ell_1} + |\theta_{kk}^*|}{|\theta_{kk}^*|} - 1 \leq \frac{M}{b} - 1.$$

Thus we combine (B.8) and (B.10)-(B.11) to obtain the following upper bound

$$|\hat{\theta}_{kk}^{s.nd} - \theta_{kk}^*| \leq \frac{B^2[(1+\frac{b}{M})\frac{M}{b}\lambda + C_0 d\lambda]}{1 - B[(1+\frac{b}{M})\frac{M}{b}\lambda + C_0 d\lambda]}, \tag{B.12}$$

where the fact that $|\theta_{kk}^*| \leq \lambda_{\max}(\boldsymbol{\Theta}^*) = B$ is used. Since $d\lambda = o(1)$, for ease of notation, we denote the right hand side of (B.12) as $C_1 d\lambda$ for some quantity $C_1 > 0$.

Next, we want to bound $\|\hat{\boldsymbol{\theta}}_{(k)}^{s.nd} - \boldsymbol{\theta}_{(k)}^*\|_{\ell_1}$. Observe that

$$\hat{\boldsymbol{\theta}}_{(k)}^{s.nd} - \boldsymbol{\theta}_{(k)}^* = -\hat{\theta}_{kk}^{s.nd}\hat{\boldsymbol{\beta}}_k^{s.nd} + \theta_{kk}^*\boldsymbol{\beta}_k^*,$$

and then we have

$$\begin{aligned}
\|\hat{\boldsymbol{\theta}}_{(k)}^{s.nd} - \boldsymbol{\theta}_{(k)}^*\|_{\ell_1} &\leq \|(\hat{\theta}_{kk}^{s.nd} - \theta_{kk}^*)\hat{\boldsymbol{\beta}}_k^{s.nd}\|_{\ell_1} + \|\theta_{kk}^*(\hat{\boldsymbol{\beta}}_k^{s.nd} - \boldsymbol{\beta}_k^*)\|_{\ell_1} \\
&\leq |\hat{\theta}_{kk}^{s.nd} - \theta_{kk}^*| \cdot \|\hat{\boldsymbol{\beta}}_k^{s.nd}\|_{\ell_1} + |\theta_{kk}^*| \cdot \|\hat{\boldsymbol{\beta}}_k^{s.nd} - \boldsymbol{\beta}_k^*\|_{\ell_1} \\
&\leq |\hat{\theta}_{kk}^{s.nd} - \theta_{kk}^*| \cdot \|\boldsymbol{\beta}_k^*\|_{\ell_1} + B \cdot C_0 d\lambda \\
&\leq C_1 d\lambda \cdot b^{-1}M + B \cdot C_0 d\lambda.
\end{aligned} \tag{B.13}$$

Therefore, under the same event, we can combine (B.12) and (B.13) to conclude that

$$\|\hat{\boldsymbol{\Theta}}^{s}_{nd} - \boldsymbol{\Theta}^*\|_{\ell_1} = \max_{1\le k\le p}\left(|\hat{\theta}^{s.nd}_{kk} - \theta^*_{kk}| + \|\hat{\boldsymbol{\theta}}^{s.nd}_{(k)} - \boldsymbol{\theta}^*_{(k)}\|_{\ell_1}\right) \le C_{b,B,M} d\lambda.$$

To complete the proof, we shall prove the claim (B.8) under the probability event $\{\|\hat{\boldsymbol{R}}^{s} - \boldsymbol{\Sigma}^*\|_{\max} \le \frac{b}{M}\lambda\}$. The first part of the claim can be justified by noting that $\boldsymbol{\Sigma}^*_{(k)}\boldsymbol{\beta}^*_k = \boldsymbol{\sigma}^*_{(k)}$ and then we have

$$\begin{aligned}
\|\hat{\boldsymbol{R}}^{s}_{(k)}\boldsymbol{\beta}^* - \hat{\boldsymbol{r}}^{s}_{(k)}\|_{\ell_\infty} &\le \|(\hat{\boldsymbol{R}}^{s}_{(k)} - \boldsymbol{\Sigma}^*_{(k)})\boldsymbol{\beta}^*\|_{\ell_\infty} + \|\boldsymbol{\sigma}^*_{(k)} - \hat{\boldsymbol{r}}^{s}_{(k)}\|_{\ell_\infty} \\
&\le \|\hat{\boldsymbol{R}}^{s} - \boldsymbol{\Sigma}^*\|_{\max}\cdot(\|\boldsymbol{\beta}^*_k\|_{\ell_1} + 1) \\
&\le \lambda.
\end{aligned}$$

For the second part, we first partition $\boldsymbol{\beta}^*_k$ and $\hat{\boldsymbol{\beta}}^{s.nd}_k$ according to the active set $\mathcal{A}_k$ such that $\boldsymbol{\beta}^*_k = (\boldsymbol{\beta}^*_{\mathcal{A}_k}, \mathbf{0})$ and $\hat{\boldsymbol{\beta}}^{s.nd}_k = (\hat{\boldsymbol{\beta}}^{s.nd}_{\mathcal{A}_k}, \hat{\boldsymbol{\beta}}^{s.nd}_{\mathcal{A}^c_k})$. Due to the fact that

$$\|\boldsymbol{\beta}^*_{\mathcal{A}_k} - \hat{\boldsymbol{\beta}}^{s.nd}_{\mathcal{A}_k}\|_{\ell_1} - \|\hat{\boldsymbol{\beta}}^{s.nd}_{\mathcal{A}^c_k}\|_{\ell_1} \ge \|\boldsymbol{\beta}^*_{\mathcal{A}_k}\|_{\ell_1} - \|\hat{\boldsymbol{\beta}}^{s.nd}_{k}\|_{\ell_1} = \|\boldsymbol{\beta}^*_{k}\|_{\ell_1} - \|\hat{\boldsymbol{\beta}}^{s.nd}_{k}\|_{\ell_1} \ge 0,$$

it immediately follows that

$$\|\hat{\boldsymbol{\beta}}^{s.nd}_{\mathcal{A}^c_k}\|_{\ell_1} \le \|\boldsymbol{\beta}^*_{\mathcal{A}_k} - \hat{\boldsymbol{\beta}}^{s.nd}_{\mathcal{A}_k}\|_{\ell_1}.$$

Then, we can apply the Cauchy inequality to obtain that

$$\begin{aligned}
\|\hat{\boldsymbol{\beta}}^{s}_k - \boldsymbol{\beta}^*_k\|_{\ell_1} &\le 2\|\hat{\boldsymbol{\beta}}^{s}_{\mathcal{A}_k} - \boldsymbol{\beta}^*_{\mathcal{A}_k}\|_{\ell_1} \\
&\le 2d^{1/2}\|\hat{\boldsymbol{\beta}}^{s}_{\mathcal{A}_k} - \boldsymbol{\beta}^*_{\mathcal{A}_k}\|_{\ell_2} \\
&\le 2d^{1/2}\lambda^{-1}_{\min}(\boldsymbol{\Sigma}^*_{(k)})\left[(\hat{\boldsymbol{\beta}}^{s.nd}_k - \boldsymbol{\beta}^*_k)^T\boldsymbol{\Sigma}^*_{(k)}(\hat{\boldsymbol{\beta}}^{s.nd}_k - \boldsymbol{\beta}^*_k)\right]^{1/2} \\
&\le 2Bd^{1/2}\cdot\|\hat{\boldsymbol{\beta}}^{s.nd}_k - \boldsymbol{\beta}^*_k\|^{1/2}_{\ell_1}\cdot\|\boldsymbol{\Sigma}^*_{(k)}(\hat{\boldsymbol{\beta}}^{s.nd}_k - \boldsymbol{\beta}^*_k)\|^{1/2}_{\ell_\infty},
\end{aligned}$$

where $\lambda_{\min}^{-1}(\boldsymbol{\Sigma}_{(k)}^*)$ is bounded by $\lambda_{\min}^{-1}(\boldsymbol{\Sigma}^*) = \lambda_{\max}(\boldsymbol{\Theta}^*) \leq B$ in the last inequality. Note that $\boldsymbol{\Sigma}_{(k)}^*\boldsymbol{\beta}_k^* = \boldsymbol{\sigma}_{(k)}^*$ by definition, and thus we can obtain the upper bound

$$\begin{aligned}
& \|\hat{\boldsymbol{\beta}}_k^{s.nd} - \boldsymbol{\beta}_k^*\|_{\ell_1} \\
\leq\ & 4B^2 d \cdot \|\boldsymbol{\Sigma}_{(k)}^*(\hat{\boldsymbol{\beta}}_k^{s.nd} - \boldsymbol{\beta}_k^*)\|_{\ell_\infty} \\
\leq\ & 4B^2 d \cdot (\|(\boldsymbol{\Sigma}_{(k)}^* - \hat{\boldsymbol{R}}_{(k)}^s)\hat{\boldsymbol{\beta}}_k^{s.nd}\|_{\ell_\infty} + \|\hat{\boldsymbol{R}}_{(k)}^s\hat{\boldsymbol{\beta}}_k^{s.nd} - \boldsymbol{\sigma}_{(k)}^*\|_{\ell_\infty}) \\
\leq\ & 4B^2 d \cdot (\|\boldsymbol{\Sigma}_{(k)}^* - \hat{\boldsymbol{R}}_{(k)}^s\|_{\max} \cdot \|\hat{\boldsymbol{\beta}}_k^{s.nd}\|_{\ell_1} + \|\hat{\boldsymbol{R}}_{(k)}^s\hat{\boldsymbol{\beta}}_k^{s.nd} - \hat{\boldsymbol{r}}_{(k)}^s\|_{\ell_\infty} + \|\hat{\boldsymbol{r}}_{(k)}^s - \boldsymbol{\sigma}_{(k)}^*\|_{\ell_\infty}) \\
\leq\ & 4B^2(2 + \frac{b}{M})d\lambda \\
\equiv\ & C_0 d\lambda. \qquad \square
\end{aligned}$$

where $\|\hat{\boldsymbol{R}}^s - \boldsymbol{\Sigma}^*\|_{\max} \leq \frac{b}{M}\lambda$, $\|\hat{\boldsymbol{\beta}}_k^{s.nd}\|_{\ell_1} \leq \|\boldsymbol{\beta}_k^*\|_{\ell_1} \leq \frac{M}{b}$ and $\|\hat{\boldsymbol{R}}_{(k)}^s\hat{\boldsymbol{\beta}}_k^{s.nd} - \hat{\boldsymbol{r}}_{(k)}^s\|_{\ell_\infty} \leq \lambda$ are used in the last inequality. Now we complete the proof of Theorem 3.2.

B.4 Proof of Theorem 3.3

Proof B.4 Throughout the following proof, we consider the probability event

$$\{\|\hat{\boldsymbol{R}}^s - \boldsymbol{\Sigma}^*\|_{\max} \leq \min(\lambda_{adantzig}, \frac{b}{M}\lambda_{dantzig})\}. \tag{B.14}$$

For ease of notation, we define $\lambda_{dantzig} = \lambda_0$ and $\lambda_{adantzig} = \lambda_1$. Recall that $\boldsymbol{\beta}_k^* = -(\theta_{kk}^*)^{-1}\boldsymbol{\theta}_{(k)}^*$, and thus the graphical model selection consistency can be easily obtained as long as we can prove the sign consistency of $\hat{\boldsymbol{\beta}}_k^{s.nad}$ for $k = 1, \cdots, p$. Now we focus on the proof of the sign consistency of $\hat{\boldsymbol{\beta}}_k^{s.nad}$ in the sequel.

Under the probability event (B.14), $\hat{\boldsymbol{R}}_{\mathcal{A}_k\mathcal{A}_k}^s$ is always positive-definite for any k. To see this, we use the Weyl's inequality to obtain that

$$\lambda_{\min}(\hat{\boldsymbol{R}}_{\mathcal{A}_k\mathcal{A}_k}^s) + \lambda_{\max}(\hat{\boldsymbol{R}}_{\mathcal{A}_k\mathcal{A}_k}^s - \boldsymbol{\Sigma}_{\mathcal{A}_k\mathcal{A}_k}^*) \geq \lambda_{\min}(\boldsymbol{\Sigma}_{\mathcal{A}_k\mathcal{A}_k}^*),$$

and then we can derive the positive lower bound for the minimal eigenvalue of $\hat{\boldsymbol{R}}^s_{\mathcal{A}_k\mathcal{A}_k}$.

$$\begin{aligned}
\lambda_{\min}(\hat{\boldsymbol{R}}^s_{\mathcal{A}_k\mathcal{A}_k}) &\geq \lambda_{\min}(\boldsymbol{\Sigma}^*_{\mathcal{A}_k\mathcal{A}_k}) - \|\hat{\boldsymbol{R}}^s_{\mathcal{A}_k\mathcal{A}_k} - \boldsymbol{\Sigma}^*_{\mathcal{A}_k\mathcal{A}_k}\|_{\ell_2} \\
&\geq \lambda_{\min}(\boldsymbol{\Sigma}^*_{\mathcal{A}_k\mathcal{A}_k}) - \|\hat{\boldsymbol{R}}^s_{\mathcal{A}_k\mathcal{A}_k} - \boldsymbol{\Sigma}^*_{\mathcal{A}_k\mathcal{A}_k}\|_F \\
&\geq \lambda_1(d - \sqrt{d(d-1)}) \\
&> \frac{\lambda_1}{2} \\
&> 0.
\end{aligned}$$

To establish the sign consistency of $\hat{\boldsymbol{\beta}}^{s.nad}_k$, we introduce the associated dual variables $\boldsymbol{\alpha}^+_{\boldsymbol{k}} = (\alpha^+_j)_{j\neq k} \in \mathbb{R}^{p-1}_+$ and $\boldsymbol{\alpha}^-_{\boldsymbol{k}} = (\alpha^-_j)_{j\neq k} \in \mathbb{R}^{p-1}_+$ for any k. Then the Lagrange dual function of (3.15) is defined as

$$L(\boldsymbol{\beta}; \boldsymbol{\alpha}^+_{\boldsymbol{k}}, \boldsymbol{\alpha}^-_{\boldsymbol{k}}) = \|\boldsymbol{w}^d_k \circ \boldsymbol{\beta}\|_{\ell_1} + (\hat{\boldsymbol{R}}^s_{(k)}\boldsymbol{\beta} - \hat{\boldsymbol{r}}^s_{(k)} - \lambda_1 \boldsymbol{w}^d_k)^T \boldsymbol{\alpha}^+_{\boldsymbol{k}} + (-\hat{\boldsymbol{R}}^s_{(k)}\boldsymbol{\beta} + \hat{\boldsymbol{r}}^s_{(k)} - \lambda_1 \boldsymbol{w}^d_k)^T \boldsymbol{\alpha}^-_{\boldsymbol{k}}.$$

Let $(\hat{\boldsymbol{R}}^s_{(k)}\boldsymbol{\beta} - \hat{\boldsymbol{r}}^s_{(k)})_j$ denote its j-th coordinate. Due to the strong duality of linear programming (Boyd and Vandenberghe, 2004), the complementary slackness condition holds for the primal problem with respect to any primal and dual solution pair $(\boldsymbol{\beta}, \boldsymbol{\alpha}^+_{\boldsymbol{k}}, \boldsymbol{\alpha}^-_{\boldsymbol{k}})$, which implies that

$$\alpha^+_j \cdot [(\hat{\boldsymbol{R}}^s_{(k)}\boldsymbol{\beta} - \hat{\boldsymbol{r}}^s_{(k)})_j - \lambda_1 w^d_j] = 0$$

and

$$\alpha^-_j \cdot [-(\hat{\boldsymbol{R}}^s_{(k)}\boldsymbol{\beta} - \hat{\boldsymbol{r}}^s_{(k)})_j - \lambda_1 w^d_j] = 0$$

for any $j \neq k$. Observe that only one of α^+_j and α^-_j can be zero because only one of the following two equations $(\hat{\boldsymbol{R}}^s_{(k)}\boldsymbol{\beta} - \hat{\boldsymbol{r}}^s_{(k)})_j = \lambda_1 w^d_j$ and $(\hat{\boldsymbol{R}}^s_{(k)}\boldsymbol{\beta} - \hat{\boldsymbol{r}}^s_{(k)})_j = -\lambda_1 w^d_j$ can hold. Thus we can uniquely define $\boldsymbol{\alpha}_k = \boldsymbol{\alpha}^+_{\boldsymbol{k}} - \boldsymbol{\alpha}^-_{\boldsymbol{k}}$. Then we can rewrite the

Lagrange dual function as

$$\begin{aligned} L(\boldsymbol{\beta};\boldsymbol{\alpha}_k) &= (\boldsymbol{w}_k^d \circ \mathrm{sign}(\boldsymbol{\beta}))^T\boldsymbol{\beta} + (\hat{\boldsymbol{R}}^s_{(k)}\boldsymbol{\beta} - \hat{\boldsymbol{r}}^s_{(k)})^T\boldsymbol{\alpha}_k - \lambda_1(\boldsymbol{\alpha}_{\boldsymbol{k}}^+ + \boldsymbol{\alpha}_{\boldsymbol{k}}^-)^T\boldsymbol{w}_k^d \\ &= (\boldsymbol{w}_k^d \circ \mathrm{sign}(\boldsymbol{\beta}) - \hat{\boldsymbol{R}}^s_{(k)}\boldsymbol{\alpha}_k)^T\boldsymbol{\beta} - \lambda_1\|\boldsymbol{w}_k^d \circ \boldsymbol{\alpha}_k\|_{\ell_1} - \boldsymbol{\alpha}_k^T\hat{\boldsymbol{r}}^s_{(k)}. \end{aligned}$$

By the Lagrange duality theory, the corresponding dual problem of (3.15) is

$$\max_{\boldsymbol{\alpha}\in\mathbb{R}^{p-1}} -\lambda_1\|\boldsymbol{w}_k^d \circ \boldsymbol{\alpha}_k\|_{\ell_1} - \langle\boldsymbol{\alpha}_k, \hat{\boldsymbol{r}}^s_{(k)}\rangle \quad \text{subject to} \quad |\hat{\boldsymbol{R}}^s_{(k)}\boldsymbol{\alpha}_k| \le \boldsymbol{w}_k^d$$

Now we shall construct an optimal primal and dual solution pair $(\tilde{\boldsymbol{\beta}}_k, \tilde{\boldsymbol{\alpha}}_k)$ to the rank-based adaptive Dantzig selector. In addition, we will prove that $(\tilde{\boldsymbol{\beta}}_k, \tilde{\boldsymbol{\alpha}}_k)$ is indeed the unique solution pair to the rank-based adaptive Dantzig selector, and $\tilde{\boldsymbol{\beta}}_k$ is exactly supported in the true active set $\mathcal{A}_k$. To this end, we construct $(\tilde{\boldsymbol{\beta}}_k, \tilde{\boldsymbol{\alpha}}_k)$ as

$$\tilde{\boldsymbol{\alpha}}_k = (\tilde{\boldsymbol{\alpha}}_{\mathcal{A}_k}, \tilde{\boldsymbol{\alpha}}_{\mathcal{A}_k^c}) = (\tilde{\boldsymbol{\alpha}}_{\mathcal{A}_k}, \boldsymbol{0})$$

and

$$\tilde{\boldsymbol{\beta}}_k = (\tilde{\boldsymbol{\beta}}_{\mathcal{A}_k}, \tilde{\boldsymbol{\beta}}_{\mathcal{A}_k^c}) = (\tilde{\boldsymbol{\beta}}_{\mathcal{A}_k}, \boldsymbol{0}),$$

where

$$\tilde{\boldsymbol{\alpha}}_{\mathcal{A}_k} = -(\hat{\boldsymbol{R}}^s_{\mathcal{A}_k\mathcal{A}_k})^{-1}\boldsymbol{w}^d_{\mathcal{A}_k} \circ \mathrm{sign}(\boldsymbol{\beta}^*_{\mathcal{A}_k}),$$

and

$$\tilde{\boldsymbol{\beta}}_{\mathcal{A}_k} = (\hat{\boldsymbol{R}}^s_{\mathcal{A}_k\mathcal{A}_k})^{-1}(\hat{\boldsymbol{r}}^s_{\mathcal{A}_k} + \lambda_1\boldsymbol{w}^d_{\mathcal{A}_k} \circ \mathrm{sign}(\tilde{\boldsymbol{\alpha}}_{\mathcal{A}_k})).$$

In what follows, we first show that $(\tilde{\boldsymbol{\beta}}_k, \tilde{\boldsymbol{\alpha}}_k)$ satisfies four optimality conditions,

which play the central role to prove that $(\tilde{\boldsymbol{\beta}}_k, \tilde{\boldsymbol{\alpha}}_k)$ is a unique optimal primal and dual solution pair. Now, we introduce these four optimality conditions.

$$\begin{aligned}
\hat{\boldsymbol{R}}^s_{\mathcal{A}_k\mathcal{A}_k}\tilde{\boldsymbol{\beta}}_{\mathcal{A}_k} - \hat{\boldsymbol{r}}^s_{\mathcal{A}_k} &= \lambda_1 \boldsymbol{w}^d_{\mathcal{A}_k} \circ \text{sign}(\tilde{\boldsymbol{\alpha}}_{\mathcal{A}_k}) && \text{(B.15)}\\
\hat{\boldsymbol{R}}^s_{\mathcal{A}_k\mathcal{A}_k}\tilde{\boldsymbol{\alpha}}_{\mathcal{A}_k} &= -\boldsymbol{w}^d_{\mathcal{A}_k} \circ \text{sign}(\tilde{\boldsymbol{\beta}}_{\mathcal{A}_k}) && \text{(B.16)}\\
|\hat{\boldsymbol{R}}^s_{\mathcal{A}_k^c\mathcal{A}_k}\tilde{\boldsymbol{\beta}}_{\mathcal{A}_k} - \hat{\boldsymbol{r}}^s_{\mathcal{A}_k^c}| &< \lambda_1 \boldsymbol{w}^d_{\mathcal{A}_k^c} && \text{(B.17)}\\
|\hat{\boldsymbol{R}}^s_{\mathcal{A}_k^c\mathcal{A}_k}\tilde{\boldsymbol{\alpha}}_{\mathcal{A}_k}| &< \boldsymbol{w}^d_{\mathcal{A}_k^c} && \text{(B.18)}
\end{aligned}$$

where (B.15) and (B.17) are primal constraints, and (B.16) and (B.18) are dual constraints. Note that (B.15) can be easily verified by substituting $\tilde{\boldsymbol{\alpha}}_{\mathcal{A}_k}$ and $\tilde{\boldsymbol{\beta}}_{\mathcal{A}_k}$. Under the same probability event (B.14), to show (B.16), (B.17) & (B.18), we also need to derive the upper bounds for the following two useful quantities

$$K_1 = \|(\hat{\boldsymbol{R}}^s_{\mathcal{A}_k\mathcal{A}_k})^{-1} - (\boldsymbol{\Sigma}^*_{\mathcal{A}_k\mathcal{A}_k})^{-1}\|_{\ell_\infty},$$

and

$$K_2 = \|\hat{\boldsymbol{R}}^s_{\mathcal{A}_k^c\mathcal{A}_k}(\hat{\boldsymbol{R}}^s_{\mathcal{A}_k\mathcal{A}_k})^{-1} - \boldsymbol{\Sigma}^*_{\mathcal{A}_k^c\mathcal{A}_k}(\boldsymbol{\Sigma}^*_{\mathcal{A}_k\mathcal{A}_k})^{-1}\|_{\ell_\infty}.$$

Note that

$$K_1 = (\hat{\boldsymbol{R}}^s_{\mathcal{A}_k\mathcal{A}_k})^{-1} \cdot (\hat{\boldsymbol{R}}^s_{\mathcal{A}_k\mathcal{A}_k} - \boldsymbol{\Sigma}^*_{\mathcal{A}_k\mathcal{A}_k}) \cdot (\boldsymbol{\Sigma}^*_{\mathcal{A}_k\mathcal{A}_k})^{-1},$$

and then we have

$$\begin{aligned}
K_1 &\leq \|(\hat{\boldsymbol{R}}^s_{\mathcal{A}_k\mathcal{A}_k})^{-1}\|_{\ell_\infty} \cdot \|\hat{\boldsymbol{R}}^s_{\mathcal{A}_k\mathcal{A}_k} - \boldsymbol{\Sigma}^*_{\mathcal{A}_k\mathcal{A}_k}\|_{\ell_\infty} \cdot \|(\boldsymbol{\Sigma}^*_{\mathcal{A}_k\mathcal{A}_k})^{-1}\|_{\ell_\infty}\\
&\leq d\lambda_1 G_k(G_k + K_1).
\end{aligned}$$

Some simple calculation leads to the following upper bound for K_1.

$$K_1 \leq \frac{d\lambda_1 G_k^2}{1 - d\lambda_1 G_k}$$

On the other hand, K_2 can be upper bounded using triangle inequalities.

$$\begin{aligned} K_2 &\leq \|(\hat{\boldsymbol{R}}^s_{\mathcal{A}_k^c\mathcal{A}_k} - \boldsymbol{\Sigma}^*_{\mathcal{A}_k^c\mathcal{A}_k}) \cdot (\hat{\boldsymbol{R}}^s_{\mathcal{A}_k\mathcal{A}_k})^{-1}\|_{\ell_\infty} + \|\boldsymbol{\Sigma}^*_{\mathcal{A}_k^c\mathcal{A}_k} \cdot ((\boldsymbol{\Sigma}^*_{\mathcal{A}_k\mathcal{A}_k})^{-1} - (\hat{\boldsymbol{R}}^s_{\mathcal{A}_k\mathcal{A}_k})^{-1})\|_{\ell_\infty} \\ &\leq \left(\|\hat{\boldsymbol{R}}^s_{\mathcal{A}_k^c\mathcal{A}_k} - \boldsymbol{\Sigma}^*_{\mathcal{A}_k^c\mathcal{A}_k}\|_{\ell_\infty} + H_k \cdot \|\hat{\boldsymbol{R}}^s_{\mathcal{A}_k\mathcal{A}_k} - \boldsymbol{\Sigma}^*_{\mathcal{A}_k\mathcal{A}_k}\|_{\ell_\infty}\right) \cdot \|(\hat{\boldsymbol{R}}^s_{\mathcal{A}_k\mathcal{A}_k})^{-1}\|_{\ell_\infty} \\ &\leq (d\lambda_1 + H_k \cdot d\lambda_1) \cdot (G_k + K_1) \\ &\leq \frac{d\lambda_1 G_k(1 + H_k)}{1 - d\lambda_1 G_k} \end{aligned}$$

Under the same event (B.14), we make a claim about the adaptive weights $\boldsymbol{w}^d_k$ that

$$\|\boldsymbol{w}^d_{\mathcal{A}_k^c}\|_{\min} \geq \frac{d\lambda_1 G_k + H_k}{2\lambda_1 G_k}\psi_k + \frac{1 + dG_k}{1 - d\lambda_1 G_k}; \tag{B.19}$$

$$\|\boldsymbol{w}^d_{\mathcal{A}_k}\|_\infty \leq \frac{1 - d\lambda_1 G_k}{2\lambda_1 G_k}\psi_k - dG_k - 1. \tag{B.20}$$

These claims are very useful to prove the other three optimality conditions (B.16), (B.17) & (B.18), and their proofs will be provided later.

Now we are ready to prove (B.16), (B.17) and (B.18) for the solution pair $(\tilde{\boldsymbol{\beta}}_k, \tilde{\boldsymbol{\alpha}}_k)$. To prove (B.16), it suffices to show the sign consistency that $\text{sign}(\boldsymbol{\beta}^*_{\mathcal{A}_k}) = \text{sign}(\tilde{\boldsymbol{\beta}}_{\mathcal{A}_k})$ since its left hand side becomes $\hat{\boldsymbol{R}}^s_{\mathcal{A}_k\mathcal{A}_k}\tilde{\boldsymbol{\alpha}}_{\mathcal{A}_k} = -\boldsymbol{w}^d_{\mathcal{A}_k} \circ \text{sign}(\boldsymbol{\beta}^*_{\mathcal{A}_k})$ if we plug in $\tilde{\boldsymbol{\alpha}}_{\mathcal{A}_k}$. Recall that $\boldsymbol{\beta}^*_{\mathcal{A}_k} = (\boldsymbol{\Sigma}^*_{\mathcal{A}_k\mathcal{A}_k})^{-1}\boldsymbol{\sigma}^*_{\mathcal{A}_k}$. Now consider the difference between $\tilde{\boldsymbol{\beta}}_{\mathcal{A}_k}$ and $\boldsymbol{\beta}^*_{\mathcal{A}_k}$.

$$\tilde{\boldsymbol{\beta}}_{\mathcal{A}_k} - \boldsymbol{\beta}^*_{\mathcal{A}_k} = (\hat{\boldsymbol{R}}^s_{\mathcal{A}_k\mathcal{A}_k})^{-1}(\hat{\boldsymbol{r}}^s_{\mathcal{A}_k} - \boldsymbol{\sigma}^*_{\mathcal{A}_k} + \lambda_1 \boldsymbol{w}^d_{\mathcal{A}_k} \circ \text{sign}(\tilde{\boldsymbol{\alpha}}_{\mathcal{A}_k})) - ((\hat{\boldsymbol{R}}^s_{\mathcal{A}_k\mathcal{A}_k})^{-1} - (\boldsymbol{\Sigma}^*_{\mathcal{A}_k\mathcal{A}_k})^{-1})\boldsymbol{\sigma}^*_{\mathcal{A}_k}$$

Then we apply the triangle inequality to obtain an upper bound.

$$\begin{aligned}
\|\tilde{\boldsymbol{\beta}}_{\mathcal{A}_k} - \boldsymbol{\beta}^*_{\mathcal{A}_k}\|_{\ell_\infty} &\leq (G_k + K_1)(\lambda_1 + \lambda_1 \|\boldsymbol{w}^d_{\mathcal{A}_k}\|_{\ell_\infty}) + K_1 \|\boldsymbol{\sigma}^*_{\mathcal{A}_k}\|_{\ell_\infty} \\
&\leq \frac{\lambda_1 G_k}{1 - d\lambda_1 G_k}(1 + \|\boldsymbol{w}^d_{\mathcal{A}_k}\|_{\ell_\infty}) + \frac{d\lambda_1 G_k^2}{1 - d\lambda_1 G_k} \\
&\leq \frac{1}{2}\|\boldsymbol{\beta}^*_{\mathcal{A}_k}\|_{\min} \\
&< \|\boldsymbol{\beta}^*_{\mathcal{A}_k}\|_{\min}
\end{aligned}$$

where the third inequality obviously holds due to the claim (B.20). Then by the above upper bound, the sign consistency of $\text{sign}(\boldsymbol{\beta}^*_{\mathcal{A}_k}) = \text{sign}(\tilde{\boldsymbol{\beta}}_{\mathcal{A}_k})$ is immediately satisfied.

Next, we can obtain the condition (B.18) via the triangular inequality as follows,

$$\begin{aligned}
\|\hat{\boldsymbol{R}}^s_{\mathcal{A}_k^c \mathcal{A}_k} \tilde{\boldsymbol{\alpha}}_{\mathcal{A}_k}\|_{\ell_\infty} &\leq \|\hat{\boldsymbol{R}}^s_{\mathcal{A}_k^c \mathcal{A}_k} (\hat{\boldsymbol{R}}^s_{\mathcal{A}_k \mathcal{A}_k})^{-1} \cdot \boldsymbol{w}^d_{\mathcal{A}_k}\|_{\ell_\infty} \\
&\leq (H_k + K_2) \cdot \|\boldsymbol{w}^d_{\mathcal{A}_k}\|_{\ell_\infty} \\
&\leq \frac{d\lambda_1 G_k + H_k}{1 - d\lambda_1 G_k} \cdot \|\boldsymbol{w}^d_{\mathcal{A}_k}\|_{\ell_\infty} \\
&< \|\boldsymbol{w}^d_{\mathcal{A}_k^c}\|_{\min},
\end{aligned}$$

where the last inequality can be shown by using both claim (B.19) and (B.20).

Now it remains to prove (B.17). Using the facts that $\boldsymbol{\theta}^*_{\mathcal{A}_k^c} = \boldsymbol{0}$ and $\boldsymbol{\Sigma}^* \boldsymbol{\Theta}^* = \boldsymbol{I}$, some simple calculation yields a useful equality that

$$\boldsymbol{\Sigma}^*_{\mathcal{A}_k^c \mathcal{A}_k} (\boldsymbol{\Sigma}^*_{\mathcal{A}_k \mathcal{A}_k})^{-1} \boldsymbol{\sigma}^*_{\mathcal{A}_k} = \boldsymbol{\sigma}^*_{\mathcal{A}_k^c}.$$

Then we can equivalently rewrite the left hand side of (B.17) as follows.

$$\begin{aligned}
\hat{\boldsymbol{R}}^s_{\mathcal{A}_k^c \mathcal{A}_k} \tilde{\boldsymbol{\beta}}_{\mathcal{A}_k} - \hat{\boldsymbol{r}}^s_{\mathcal{A}_k^c} &= \hat{\boldsymbol{R}}^s_{\mathcal{A}_k^c \mathcal{A}_k} (\hat{\boldsymbol{R}}^s_{\mathcal{A}_k \mathcal{A}_k})^{-1} (\hat{\boldsymbol{r}}^s_{\mathcal{A}_k} + \lambda_1 \boldsymbol{w}^d_{\mathcal{A}_k} \circ \text{sign}(\tilde{\boldsymbol{\alpha}}_{\mathcal{A}_k})) - \hat{\boldsymbol{r}}^s_{\mathcal{A}_k^c} \\
&= \hat{\boldsymbol{R}}^s_{\mathcal{A}_k^c \mathcal{A}_k} (\hat{\boldsymbol{R}}^s_{\mathcal{A}_k \mathcal{A}_k})^{-1} (\hat{\boldsymbol{r}}^s_{\mathcal{A}_k} - \boldsymbol{\sigma}^*_{\mathcal{A}_k} + \lambda_1 \boldsymbol{w}^d_{\mathcal{A}_k} \circ \text{sign}(\tilde{\boldsymbol{\alpha}}_{\mathcal{A}_k})) + \\
&\quad (\hat{\boldsymbol{R}}^s_{\mathcal{A}_k^c \mathcal{A}_k} (\hat{\boldsymbol{R}}^s_{\mathcal{A}_k \mathcal{A}_k})^{-1} - \boldsymbol{\Sigma}^*_{\mathcal{A}_k^c \mathcal{A}_k} (\boldsymbol{\Sigma}^*_{\mathcal{A}_k \mathcal{A}_k})^{-1}) \boldsymbol{\sigma}^*_{\mathcal{A}_k} + (\boldsymbol{\sigma}^*_{\mathcal{A}_k^c} - \hat{\boldsymbol{r}}^s_{\mathcal{A}_k^c})
\end{aligned}$$

Again we apply the triangle inequality to obtain an upper bound.

$$\begin{aligned}
\|\hat{\boldsymbol{R}}^s_{\mathcal{A}_k^c\mathcal{A}_k}\tilde{\boldsymbol{\beta}}_{\mathcal{A}_k} - \hat{\boldsymbol{r}}^s_{\mathcal{A}_k^c}\|_\infty &\leq (H_k + K_2)(\lambda_1 + \lambda_1\|\boldsymbol{w}^d_{\mathcal{A}_k}\|_\infty) + K_2\|\boldsymbol{\sigma}^*_{\mathcal{A}_k}\|_\infty + \lambda_1 \\
&\leq \frac{d\lambda_1^2 G_k + \lambda_1 H_k}{1 - d\lambda_1 G_k}(1 + \|\boldsymbol{w}^d_{\mathcal{A}_k}\|_\infty) + \frac{d\lambda_1 G_k(1 + H_k)}{1 - d\lambda_1 G_k} + \lambda_1 \\
&< \lambda_1\|\boldsymbol{w}^d_{\mathcal{A}_k^c}\|_{\min},
\end{aligned}$$

where the last inequality has used the following fact that

$$\begin{aligned}
(1 - d\lambda_1 G_k)\cdot\|\boldsymbol{w}^d_{\mathcal{A}_k^c}\|_{\min} &> (d\lambda_1 G_k + H_k)(\frac{1 - d\lambda_1 G_k}{2\lambda_1 G_k}\psi_k - dG_k) + dG_k(1 + H_k) + 1 \\
&\geq (d\lambda_1 G_k + H_k)\cdot(1 + \|\boldsymbol{w}^d_{\mathcal{A}_k}\|_\infty) + dG_k(1 + H_k) + 1,
\end{aligned}$$

which is due to the claims (B.19) & (B.20).

So far, four optimality conditions has been verified for $(\tilde{\boldsymbol{\beta}}_k, \tilde{\boldsymbol{\alpha}}_k)$. In the sequel, we shall show that $\tilde{\boldsymbol{\beta}}_k$ is indeed a unique optimal solution for the primal problem (3.15). To this end, we first note that due to (B.15)–(B.18), $(\tilde{\boldsymbol{\beta}}_k, \tilde{\boldsymbol{\alpha}}_k)$ are feasible solutions to the primal and dual problems respectively. Next, (B.15) and (B.16) further show that $(\tilde{\boldsymbol{\beta}}_k, \tilde{\boldsymbol{\alpha}}_k)$ satisfy the complementary-slackness conditions for both the primal and the dual problems. Thus, $(\tilde{\boldsymbol{\beta}}_k, \tilde{\boldsymbol{\alpha}}_k)$ are optimal solutions to these problems by Theorem 4.5 in Bertsimas and Tsitsiklis (1997). Now it remains to show the uniqueness. Suppose there exists another optimal solution $\breve{\boldsymbol{\beta}}_k$, and we have $\|\boldsymbol{w}^d_k \circ \breve{\boldsymbol{\beta}}_k\|_{\ell_1} = \|\boldsymbol{w}^d_k \circ \tilde{\boldsymbol{\beta}}_k\|_{\ell_1}$. Let Γ_k denote the support of $\breve{\boldsymbol{\beta}}_k$, and then $\breve{\boldsymbol{\beta}}_k = (\breve{\boldsymbol{\beta}}_{\Gamma_k}, \mathbf{0})$. By the strong duality we have

$$\begin{aligned}
\|\boldsymbol{w}^d_k \circ \breve{\boldsymbol{\beta}}_k\|_{\ell_1} &= \|\boldsymbol{w}^d_k \circ \tilde{\boldsymbol{\beta}}_k\|_{\ell_1} \\
&= -\lambda_1\|\boldsymbol{w}^d_k \circ \tilde{\boldsymbol{\alpha}}_k\|_{\ell_1} - \langle\tilde{\boldsymbol{\alpha}}_k, \hat{\boldsymbol{r}}^s_{(k)}\rangle \\
&= \inf_{\boldsymbol{\beta}} L(\boldsymbol{\beta}; \tilde{\boldsymbol{\alpha}}^+_{\boldsymbol{k}}, \tilde{\boldsymbol{\alpha}}^-_{\boldsymbol{k}}) \\
&\leq L(\breve{\boldsymbol{\beta}}_k; \tilde{\boldsymbol{\alpha}}^+_{\boldsymbol{k}}, \tilde{\boldsymbol{\alpha}}^-_{\boldsymbol{k}}) \\
&\leq \|\boldsymbol{w}^d_k \circ \breve{\boldsymbol{\beta}}_k\|_{\ell_1}.
\end{aligned}$$

Thus $L(\breve{\boldsymbol{\beta}}_k; \tilde{\boldsymbol{\alpha}}_{\boldsymbol{k}}^+, \tilde{\boldsymbol{\alpha}}_{\boldsymbol{k}}^-) = \|\boldsymbol{w}_k^d \circ \breve{\boldsymbol{\beta}}_k\|_{\ell_1}$, which immediately implies that the complementary slackness condition holds for the primal problem, i.e.

$$(\hat{\boldsymbol{R}}_{(k)}^s \breve{\boldsymbol{\beta}}_k - \hat{\boldsymbol{r}}_{(k)}^s - \lambda_1 \boldsymbol{w}_k^d)^T \tilde{\boldsymbol{\alpha}}_{\boldsymbol{k}}^+ = 0 \tag{B.21}$$

and

$$(-\hat{\boldsymbol{R}}_{(k)}^s \breve{\boldsymbol{\beta}}_k + \hat{\boldsymbol{r}}_{(k)}^s - \lambda_1 \boldsymbol{w}_k^d)^T \tilde{\boldsymbol{\alpha}}_{\boldsymbol{k}}^- = 0. \tag{B.22}$$

Now let $\breve{\boldsymbol{\beta}}_{\boldsymbol{k}}^+ = \max(\breve{\boldsymbol{\beta}}_k, \mathbf{0})$ and $\breve{\boldsymbol{\beta}}_{\boldsymbol{k}}^- = \min(\breve{\boldsymbol{\beta}}_k, \mathbf{0})$. Similarly, we can show that the complementary slackness condition also holds for the dual problem, i.e.

$$(\hat{\boldsymbol{R}}_{(k)}^s \tilde{\boldsymbol{\alpha}}_k - \boldsymbol{w}_k^d)^T \breve{\boldsymbol{\beta}}_{\boldsymbol{k}}^+ = 0 \tag{B.23}$$

and

$$(-\hat{\boldsymbol{R}}_{(k)}^s \tilde{\boldsymbol{\alpha}}_k - \boldsymbol{w}_k^d)^T \breve{\boldsymbol{\beta}}_{\boldsymbol{k}}^- = 0. \tag{B.24}$$

Notice that $\tilde{\boldsymbol{\alpha}}_{\mathcal{A}_k} \neq \mathbf{0}$ and $\tilde{\boldsymbol{\alpha}}_{\mathcal{A}_k^c} = \mathbf{0}$ by definition. Then we can equivalently rewrite (B.21)–(B.24) in terms of $\mathcal{A}_k$ and Γ_k as

$$\hat{\boldsymbol{R}}_{\mathcal{A}_k\Gamma_k}^s \breve{\boldsymbol{\beta}}_{\Gamma_k} - \hat{\boldsymbol{r}}_{\mathcal{A}_k}^s = \lambda_1 \boldsymbol{w}_{\mathcal{A}_k}^d \circ \operatorname{sign}(\tilde{\boldsymbol{\alpha}}_{\mathcal{A}_k}) \tag{B.25}$$

and

$$\hat{\boldsymbol{R}}_{\Gamma_k\mathcal{A}_k}^s \tilde{\boldsymbol{\alpha}}_{\mathcal{A}_k} = -\boldsymbol{w}_{\Gamma_k}^d \circ \operatorname{sign}(\breve{\boldsymbol{\beta}}_{\Gamma_k}). \tag{B.26}$$

Observe that in the equation (B.26), for any index j such that $j \in \Gamma_k$ but $j \notin \mathcal{A}_k$, $\hat{\boldsymbol{R}}_{j\mathcal{A}_k}^s \tilde{\boldsymbol{\alpha}}_{\mathcal{A}_k} = -w_j^d \cdot \operatorname{sign}(\breve{\beta}_j)$ cannot hold since it contradicts with (B.18). Then it is

easy to see that $\Gamma_k \subset \mathcal{A}_k$ obviously holds for $\hat{\boldsymbol{\beta}}_{\mathcal{A}_k}$ and $\breve{\boldsymbol{\beta}}_{\Gamma_k}$. We further notice that combining $\Gamma_k \subset \mathcal{A}_k$ and (B.25) immediately imply that both $\hat{\boldsymbol{\beta}}_{\mathcal{A}_k}$ and $\breve{\boldsymbol{\beta}}_{\Gamma_k}$ satisfy the same optimality condition (B.15). Thus the uniqueness follows from (B.15), (B.25) and the non-singularity of $\hat{\boldsymbol{R}}^s_{\mathcal{A}_k\mathcal{A}_k}$.

Now it remains to verify the claims (B.19) and (B.20) under the same event (B.14). In particular, under the event $\{\|\hat{\boldsymbol{R}}^s - \boldsymbol{\Sigma}^*\|_{\max} \leq b\lambda_0/M\}$, Theorem 3.2 has shown that for some quantity $C_0 = 4B^2(2 + \frac{b}{M}) > 0$, we have $\|\hat{\boldsymbol{\beta}}_k^{s.nd} - \boldsymbol{\beta}_k^*\|_{\ell_1} \leq C_0 d\lambda_0$. Then we can derive a lower bound for $\|\boldsymbol{w}^d_{\mathcal{A}_k^c}\|_{\min}$.

$$\|\boldsymbol{w}^d_{\mathcal{A}_k^c}\|_{\min} = \frac{1}{\max\limits_{j\in\mathcal{A}_k^c} |\hat{\beta}_j^s| + \frac{1}{n}} \geq \frac{1}{C_0 d\lambda_0 + \frac{1}{n}},$$

which immediately yields the desired lower bound as in (B.19) by noting that

$$\frac{G_k \cdot d\lambda_1 + H_k}{2G_k \cdot \lambda_1} \cdot \psi_k + \frac{1 + G_k \cdot d}{1 - G_k \cdot d\lambda_1} \leq \frac{H_k\psi_k}{2G_k \cdot \lambda_1} + (\psi_k + 2G_k) \cdot d + 2 \leq \frac{1}{C_0 d\lambda_0 + \frac{1}{n}}$$

where both inequalities follow from the proper choices of tuning parameters λ_0 & λ_1 as stated in the assumptions of Theorem 3.3. On the other hand, we notice that

$$\frac{1 - G_k \cdot d\lambda_1}{2G_k \cdot \lambda_1}\psi_k - dG_k - 1 \geq \frac{\psi_k}{2G_k \cdot \lambda_1} - (\psi_k + G_k) \cdot d - 1 \geq \frac{\psi_k}{4G_k \cdot \lambda_1},$$

where the last inequality follows from the proper choice of λ_1 as in Theorem 3.3, and then we can similarly show the second claim (B.20) by noting that

$$\|\boldsymbol{w}^d_{\mathcal{A}_k}\|_\infty \leq \frac{1}{\min\limits_{j\in\mathcal{A}_k} |\hat{\beta}_j^s|} \leq \frac{1}{\psi_k - C_0 d\lambda_0} \leq \frac{2}{\psi_k} \leq \frac{\psi_k}{4G_k \cdot \lambda_1}$$

where last two inequalities are due to $\psi_k \geq 2C_0 d\lambda_0$ and $\psi_k^2 \geq 8G_k\lambda_1$. Now two aforementioned claims are proved, which completes the proof of Theorem 3.3. $\square$

B.5 Proof of Theorem 3.4

Proof B.5 To bound the difference between $\hat{\boldsymbol{\Theta}}_c^s$ and $\boldsymbol{\Theta}^*$ under the entry-wise ℓ_∞-norm, we consider the probability event $\{\|\hat{\boldsymbol{R}}^s - \boldsymbol{\Sigma}\|_{\max} \leq \lambda/M\}$. First, we show that $\boldsymbol{\Theta}^*$ is always a feasible solution under the above event, i.e.

$$\|\hat{\boldsymbol{R}}^s\boldsymbol{\Theta}^* - \boldsymbol{I}\|_{\max} \leq \|(\hat{\boldsymbol{R}}^s - \boldsymbol{\Sigma}^*)\boldsymbol{\Theta}^*\|_{\max} \leq \|\hat{\boldsymbol{R}}^s - \boldsymbol{\Sigma}\|_{\max} \cdot \|\boldsymbol{\Theta}^*\|_{\ell_1} \leq \lambda.$$

Note that $\hat{\boldsymbol{\Theta}}_c^s$ is the optimal solution, and then $\hat{\boldsymbol{\Theta}}_c^s$ is naturally a feasible solution. Hence by definition, it is easy to see that

$$\|\hat{\boldsymbol{R}}^s\hat{\boldsymbol{\Theta}}_c^s - \boldsymbol{I}\|_{\max} \leq \lambda$$

and

$$\|\hat{\boldsymbol{\Theta}}_c^s\|_{\ell_1} \leq \|\boldsymbol{\Theta}^*\|_{\ell_1}$$

obviously hold. Now we can obtain the desired upper bound under the entrywise matrix ℓ_∞-norm as follows.

$$\begin{aligned}
\|\hat{\boldsymbol{\Theta}}_c^s - \boldsymbol{\Theta}^*\|_{\max} &\leq \|\boldsymbol{\Theta}^*\|_{\ell_1} \cdot \|\boldsymbol{\Sigma}^*\hat{\boldsymbol{\Theta}}_c^s - \boldsymbol{I}\|_{\max} \\
&= M \cdot \|(\boldsymbol{\Sigma}^* - \hat{\boldsymbol{R}}^s)\hat{\boldsymbol{\Theta}}^s + \hat{\boldsymbol{R}}^s\hat{\boldsymbol{\Theta}}_c^s - \boldsymbol{I}\|_{\max} \\
&\leq M \cdot \|\boldsymbol{\Sigma}^* - \hat{\boldsymbol{R}}^s\|_{\max} \cdot \|\hat{\boldsymbol{\Theta}}_c^s\|_{\ell_1} + M \cdot \|\hat{\boldsymbol{R}}^s\hat{\boldsymbol{\Theta}}_c^s - \boldsymbol{I}\|_{\max} \\
&\leq \lambda\|\boldsymbol{\Theta}^*\|_{\ell_1} + M\lambda \\
&= 2M\lambda.
\end{aligned}$$

This completes the proof of Theorem 3.4. □

B.6 Proof of Theorem 3.5

Proof B.6 Throughout the proof, we consider the probability event

$$\{\|\hat{\boldsymbol{R}}^s - \boldsymbol{\Sigma}^*\|_{\max} \leq \min(\lambda_{aclime}, \frac{1}{M}\lambda_{clime})\}. \tag{B.27}$$

For ease of notation, we define $\lambda_{clime} = \lambda_0$ and $\lambda_{aclime} = \lambda_1$. To prove the graphical model selection consistency, it suffices to prove the selection consistency for the p subproblems of vector minimization under the same probability event (B.27).

Under this event, $\hat{\boldsymbol{R}}^s_{\tilde{\mathcal{A}}_k\tilde{\mathcal{A}}_k}$ is always positive-definite for any k due to the Weyl's inequality that

$$\begin{aligned}
\lambda_{\min}(\hat{\boldsymbol{R}}^s_{\tilde{\mathcal{A}}_k\tilde{\mathcal{A}}_k}) &\geq \lambda_{\min}(\boldsymbol{\Sigma}^*_{\tilde{\mathcal{A}}_k\tilde{\mathcal{A}}_k}) - \|\hat{\boldsymbol{R}}^s_{\tilde{\mathcal{A}}_k\tilde{\mathcal{A}}_k} - \boldsymbol{\Sigma}^*_{\tilde{\mathcal{A}}_k\tilde{\mathcal{A}}_k}\|_F \\
&\geq \lambda_1(d - \sqrt{d(d-1)}) \\
&> \frac{\lambda_1}{2} \\
&> 0.
\end{aligned}$$

To establish selection consistency for the k-th subproblem ($1 \leq k \leq p$), we introduce the dual variables $\boldsymbol{\alpha}^+_{\boldsymbol{k}} = (\alpha^+_j)_{1\leq j\leq p} \in \mathbb{R}^p_+$ and $\boldsymbol{\alpha}^-_{\boldsymbol{k}} = (\alpha^-_j)_{1\leq j\leq p} \in \mathbb{R}^p_+$. Let $\boldsymbol{W}^c = (\boldsymbol{w}^c_1, \cdots, \boldsymbol{w}^c_p)$. Then by Boyd and Vandenberghe (2004), the Lagrange dual function is defined as

$$L(\boldsymbol{\theta}; \boldsymbol{\alpha}^+_{\boldsymbol{k}}, \boldsymbol{\alpha}^-_{\boldsymbol{k}}) = \|\boldsymbol{w}^c_k \circ \boldsymbol{\theta}\|_{\ell_1} + (\hat{\boldsymbol{R}}^s\boldsymbol{\theta} - \boldsymbol{e}_k - \lambda_1\boldsymbol{w}^c_k)^T\boldsymbol{\alpha}^+_{\boldsymbol{k}} + (-\hat{\boldsymbol{R}}^s\boldsymbol{\theta} + \boldsymbol{e}_k - \lambda_1\boldsymbol{w}^c_k)^T\boldsymbol{\alpha}^-_{\boldsymbol{k}}.$$

Note that only one of α^+_j and α^-_j can be zero for each $j = 1, \cdots, p$, and then we can uniquely define $\boldsymbol{\alpha} = \boldsymbol{\alpha}^+_{\boldsymbol{k}} - \boldsymbol{\alpha}^-_{\boldsymbol{k}}$. Then we can rewrite the Lagrange dual function in

terms of $\boldsymbol{\alpha}$ as

$$L(\boldsymbol{\theta};\boldsymbol{\alpha}) = (\boldsymbol{w}_k^c \circ \operatorname{sign}(\boldsymbol{\theta}) - \hat{\boldsymbol{R}}^s\boldsymbol{\alpha})^T\boldsymbol{\theta} - \lambda_1\|\boldsymbol{w}_k^c \circ \boldsymbol{\alpha}\|_{\ell_1} - \boldsymbol{\alpha}^T\boldsymbol{e}_k.$$

By the Lagrange duality theory, the dual problem is

$$\max_{\boldsymbol{\alpha}\in\mathbb{R}^p} -\lambda_1\|\boldsymbol{w}_k^c \circ \boldsymbol{\alpha}\|_{\ell_1} - \langle\boldsymbol{\alpha}, \boldsymbol{e}_k\rangle \quad \text{subject to} \quad |\hat{\boldsymbol{R}}^s\boldsymbol{\alpha}| \le \boldsymbol{w}_k^c$$

Now we shall construct an optimal primal and dual solution pair $(\tilde{\boldsymbol{\theta}}_k, \tilde{\boldsymbol{\alpha}}_k)$ to the subproblem. In addition, we further show that $(\tilde{\boldsymbol{\theta}}_k, \tilde{\boldsymbol{\alpha}}_k)$ is actually the unique solution pair, and $\tilde{\boldsymbol{\theta}}_k$ is exactly supported in the true active set $\tilde{\mathcal{A}}_k$. To this end, denote $\boldsymbol{e}_k = (\boldsymbol{e}_{\tilde{\mathcal{A}}_k}, \boldsymbol{e}_{\tilde{\mathcal{A}}_k^c})$, and then we consider

$$\tilde{\boldsymbol{\alpha}}_k = (\tilde{\boldsymbol{\alpha}}_{\tilde{\mathcal{A}}_k}, \tilde{\boldsymbol{\alpha}}_{\tilde{\mathcal{A}}_k^c}) = (\tilde{\boldsymbol{\alpha}}_{\tilde{\mathcal{A}}_k}, \boldsymbol{0})$$

and

$$\tilde{\boldsymbol{\theta}}_k = (\tilde{\boldsymbol{\theta}}_{\tilde{\mathcal{A}}_k}, \tilde{\boldsymbol{\theta}}_{\tilde{\mathcal{A}}_k^c}) = (\tilde{\boldsymbol{\theta}}_{\tilde{\mathcal{A}}_k}, \boldsymbol{0})$$

with

$$\tilde{\boldsymbol{\alpha}}_{\tilde{\mathcal{A}}_k} = -(\hat{\boldsymbol{R}}^s_{\tilde{\mathcal{A}}_k\tilde{\mathcal{A}}_k})^{-1}\boldsymbol{w}^c_{\tilde{\mathcal{A}}_k} \circ \operatorname{sign}(\boldsymbol{\theta}^*_{\tilde{\mathcal{A}}_k}),$$

and

$$\tilde{\boldsymbol{\theta}}_{\tilde{\mathcal{A}}_k} = (\hat{\boldsymbol{R}}^s_{\tilde{\mathcal{A}}_k\tilde{\mathcal{A}}_k})^{-1}(\boldsymbol{e}_{\tilde{\mathcal{A}}_k} + \lambda_1\boldsymbol{w}^c_{\tilde{\mathcal{A}}_k} \circ \operatorname{sign}(\tilde{\boldsymbol{\alpha}}_{\tilde{\mathcal{A}}_k})).$$

First we want to show that $(\tilde{\boldsymbol{\theta}}_k, \tilde{\boldsymbol{\alpha}}_k)$ satisfies the following four optimality conditions:

$$\begin{aligned}
\hat{\boldsymbol{R}}^s_{\tilde{\mathcal{A}}_k\tilde{\mathcal{A}}_k}\tilde{\boldsymbol{\theta}}_{\tilde{\mathcal{A}}_k} - \boldsymbol{e}_{\tilde{\mathcal{A}}_k} &= \lambda_1 \boldsymbol{w}^c_{\tilde{\mathcal{A}}_k} \circ \operatorname{sign}(\tilde{\boldsymbol{\alpha}}_{\mathcal{A}_k}) && \text{(B.28)}\\
\hat{\boldsymbol{R}}^s_{\tilde{\mathcal{A}}_k\tilde{\mathcal{A}}_k}\tilde{\boldsymbol{\alpha}}_{\mathcal{A}_k} &= -\boldsymbol{w}^c_{\tilde{\mathcal{A}}_k} \circ \operatorname{sign}(\tilde{\boldsymbol{\theta}}_{\tilde{\mathcal{A}}_k}) && \text{(B.29)}\\
|\hat{\boldsymbol{R}}^s_{\tilde{\mathcal{A}}_k^c\mathcal{A}_k}\tilde{\boldsymbol{\theta}}_{\tilde{\mathcal{A}}_k}| &< \lambda_1 \boldsymbol{w}^c_{\tilde{\mathcal{A}}_k^c} && \text{(B.30)}\\
|\hat{\boldsymbol{R}}^s_{\tilde{\mathcal{A}}_k^c\mathcal{A}_k}\tilde{\boldsymbol{\alpha}}_{\tilde{\mathcal{A}}_k}| &< \boldsymbol{w}^c_{\tilde{\mathcal{A}}_k^c} && \text{(B.31)}
\end{aligned}$$

where (B.28) and (B.30) are primal constraints, and (B.29) and (B.31) are dual constraints. Note that (B.28) can be easily verified by substituting $\tilde{\boldsymbol{\alpha}}_{\tilde{\mathcal{A}}_k}$ and $\tilde{\boldsymbol{\theta}}_{\tilde{\mathcal{A}}_k}$. To prove the other optimality conditions, we need to consider two related quantities

$$\tilde{K}_1 = \|(\hat{\boldsymbol{R}}^s_{\tilde{\mathcal{A}}_k\tilde{\mathcal{A}}_k})^{-1} - (\boldsymbol{\Sigma}^*_{\tilde{\mathcal{A}}_k\tilde{\mathcal{A}}_k})^{-1}\|_{\ell_\infty}$$

and

$$\tilde{K}_2 = \|\hat{\boldsymbol{R}}^s_{\tilde{\mathcal{A}}_k^c\tilde{\mathcal{A}}_k}(\hat{\boldsymbol{R}}^s_{\tilde{\mathcal{A}}_k\tilde{\mathcal{A}}_k})^{-1} - \boldsymbol{\Sigma}^*_{\tilde{\mathcal{A}}_k^c\tilde{\mathcal{A}}_k}(\boldsymbol{\Sigma}^*_{\tilde{\mathcal{A}}_k\tilde{\mathcal{A}}_k})^{-1}\|_{\ell_\infty}.$$

Under the same probability event (B.27), similarly as in Theorem 3.3, we can obtain the following upper bounds for $\tilde{K}_1$ and $\tilde{K}_2$.

$$\tilde{K}_1 \le \frac{\tilde{G}_k^2}{1 - d\lambda_1\tilde{G}_k} d\lambda_1 \quad \text{and} \quad \tilde{K}_2 \le \frac{\tilde{G}_k(1+\tilde{H}_k)}{1 - d\lambda_1\tilde{G}_k} d\lambda_1.$$

Further, we make a claim about $\boldsymbol{w}^c_k = (\boldsymbol{w}^c_{\tilde{\mathcal{A}}_k}, \boldsymbol{w}^c_{\tilde{\mathcal{A}}_k^c})$, which will be proved later.

$$\begin{aligned}
\|\boldsymbol{w}^c_{\tilde{\mathcal{A}}_k^c}\|_{\min} &\ge \frac{d\lambda_1\tilde{G}_k + \tilde{H}_k}{2\lambda_1\tilde{G}_k}\psi_{\min} + \frac{d\tilde{G}_k}{1 - d\lambda_1\tilde{G}_k}; && \text{(B.32)}\\
\|\boldsymbol{w}^c_{\tilde{\mathcal{A}}_k}\|_\infty &\le \frac{1 - d\lambda_1\tilde{G}_k}{2\lambda_1\tilde{G}_k}\psi_{\min} - d\tilde{G}_k. && \text{(B.33)}
\end{aligned}$$

Now we are ready to prove (B.29), (B.30) and (B.31) for the solution pair $(\tilde{\boldsymbol{\theta}}_k, \tilde{\boldsymbol{\alpha}}_k)$. To prove (B.29), it is enough for us to show that $\text{sign}(\boldsymbol{\theta}^*_{\tilde{\mathcal{A}}_k}) = \text{sign}(\tilde{\boldsymbol{\theta}}_{\tilde{\mathcal{A}}_k})$. Note that $\boldsymbol{\Sigma}^*_{\tilde{\mathcal{A}}_k\tilde{\mathcal{A}}_k}\boldsymbol{\theta}^*_{\tilde{\mathcal{A}}_k} = \boldsymbol{e}_{\tilde{\mathcal{A}}_k}$, and thus $\boldsymbol{\theta}^*_{\tilde{\mathcal{A}}_k} = (\boldsymbol{\Sigma}^*_{\mathcal{A}_k\mathcal{A}_k})^{-1}\boldsymbol{e}_{\tilde{\mathcal{A}}_k}$. Then we can write the difference between $\tilde{\boldsymbol{\theta}}_{\tilde{\mathcal{A}}_k}$ and $\boldsymbol{\theta}^*_{\tilde{\mathcal{A}}_k}$ as

$$\tilde{\boldsymbol{\theta}}_{\tilde{\mathcal{A}}_k} - \boldsymbol{\theta}^*_{\tilde{\mathcal{A}}_k} = (\hat{\boldsymbol{R}}^s_{\tilde{\mathcal{A}}_k\tilde{\mathcal{A}}_k})^{-1}(\lambda_1 \boldsymbol{w}^c_{\tilde{\mathcal{A}}_k} \circ \text{sign}(\tilde{\boldsymbol{\alpha}}_{\tilde{\mathcal{A}}_k})) - ((\hat{\boldsymbol{R}}^s_{\tilde{\mathcal{A}}_k\tilde{\mathcal{A}}_k})^{-1} - (\boldsymbol{\Sigma}^*_{\tilde{\mathcal{A}}_k\tilde{\mathcal{A}}_k})^{-1})\boldsymbol{e}_{\tilde{\mathcal{A}}_k}.$$

Next, we apply the triangle inequality to obtain the upper bound

$$\begin{aligned}
\|\tilde{\boldsymbol{\theta}}_{\tilde{\mathcal{A}}_k} - \boldsymbol{\theta}^*_{\tilde{\mathcal{A}}_k}\|_{\ell_\infty} &\leq (\tilde{G}_k + \tilde{K}_1)\cdot\lambda_1\|\boldsymbol{w}^c_{\tilde{\mathcal{A}}_k}\|_\infty + \tilde{K}_1 \\
&\leq \frac{\lambda_1\tilde{G}_k\cdot\|\boldsymbol{w}^c_{\tilde{\mathcal{A}}_k}\|_\infty}{1-d\lambda_1\tilde{G}_k} + \frac{d\lambda_1\tilde{G}_k^2}{1-d\lambda_1\tilde{G}_k} \\
&\leq \frac{1}{2}\psi_{\min} \\
&< \|\boldsymbol{\theta}^*_{\tilde{\mathcal{A}}_k}\|_{\min}
\end{aligned}$$

where the third inequality holds by the claim (B.33). Then the desired sign consistency of $\text{sign}(\boldsymbol{\theta}^*_{\tilde{\mathcal{A}}_k}) = \text{sign}(\tilde{\boldsymbol{\theta}}_{\tilde{\mathcal{A}}_k})$ obviously holds.

Moreover, we can easily obtain (B.31) via the following triangular inequality,

$$\begin{aligned}
\|\hat{\boldsymbol{R}}^s_{\tilde{\mathcal{A}}_k^c\tilde{\mathcal{A}}_k}\tilde{\boldsymbol{\alpha}}_{\tilde{\mathcal{A}}_k}\|_{\ell_\infty} &\leq \|\hat{\boldsymbol{R}}^s_{\tilde{\mathcal{A}}_k^c\tilde{\mathcal{A}}_k}(\hat{\boldsymbol{R}}^s_{\tilde{\mathcal{A}}_k\tilde{\mathcal{A}}_k})^{-1}\boldsymbol{w}^c_{\tilde{\mathcal{A}}_k}\|_{\ell_\infty} \\
&\leq (\tilde{H}_k + \tilde{K}_2)\cdot\|\boldsymbol{w}^c_{\tilde{\mathcal{A}}_k}\|_\infty \\
&\leq \frac{d\lambda_1\tilde{G}_k + \tilde{H}_k}{1-d\lambda_1\tilde{G}_k}\cdot\|\boldsymbol{w}^c_{\tilde{\mathcal{A}}_k}\|_\infty \\
&\leq \frac{d\lambda_1\tilde{G}_k + \tilde{H}_k}{2\lambda_1\tilde{G}_k}\psi_{\min} \\
&< \|\boldsymbol{w}^c_{\tilde{\mathcal{A}}_k^c}\|_{\min},
\end{aligned}$$

where the fourth inequality obviously holds due to the claims (B.32) and (B.33).

Now it remains to prove (B.30). Partition $\boldsymbol{\Sigma}^*\boldsymbol{\theta}_k = \boldsymbol{e}_k$ with respect to $\tilde{\mathcal{A}}_k$, and

then we will have $\boldsymbol{\Sigma}^*_{\tilde{\mathcal{A}}_k\tilde{\mathcal{A}}_k}\boldsymbol{\theta}^*_{\tilde{\mathcal{A}}_k} = \boldsymbol{e}_{\tilde{\mathcal{A}}_k}$ and $\boldsymbol{\Sigma}^*_{\tilde{\mathcal{A}}^c_k\tilde{\mathcal{A}}_k}\boldsymbol{\theta}^*_{\tilde{\mathcal{A}}_k} = \boldsymbol{e}_{\tilde{\mathcal{A}}^c_k}$. Thus we have

$$\boldsymbol{\Sigma}^*_{\tilde{\mathcal{A}}^c_k\tilde{\mathcal{A}}_k}(\boldsymbol{\Sigma}^*_{\tilde{\mathcal{A}}_k\tilde{\mathcal{A}}_k})^{-1}\boldsymbol{e}_{\tilde{\mathcal{A}}_k} = \boldsymbol{e}_{\tilde{\mathcal{A}}^c_k} = \mathbf{0}.$$

Then we can rewrite the left hand side of (B.30) as follows.

$$\begin{aligned}\hat{\boldsymbol{R}}^s_{\tilde{\mathcal{A}}^c_k\tilde{\mathcal{A}}_k}\tilde{\boldsymbol{\theta}}_{\tilde{\mathcal{A}}_k} &= \hat{\boldsymbol{R}}^s_{\tilde{\mathcal{A}}^c_k\tilde{\mathcal{A}}_k}(\hat{\boldsymbol{R}}^s_{\tilde{\mathcal{A}}_k\tilde{\mathcal{A}}_k})^{-1}\cdot(\lambda_1\boldsymbol{w}^c_{\tilde{\mathcal{A}}_k}\circ\operatorname{sign}(\tilde{\boldsymbol{\alpha}}_{\tilde{\mathcal{A}}_k}))\\ &\quad+(\hat{\boldsymbol{R}}^s_{\tilde{\mathcal{A}}^c_k\tilde{\mathcal{A}}_k}(\hat{\boldsymbol{R}}^s_{\tilde{\mathcal{A}}_k\tilde{\mathcal{A}}_k})^{-1}-\boldsymbol{\Sigma}^*_{\tilde{\mathcal{A}}^c_k\tilde{\mathcal{A}}_k}(\boldsymbol{\Sigma}^*_{\tilde{\mathcal{A}}_k\tilde{\mathcal{A}}_k})^{-1})\boldsymbol{e}_{\tilde{\mathcal{A}}_k}\end{aligned}$$

Here we apply the triangle inequality again to obtain the upper bound.

$$\|\hat{\boldsymbol{R}}^s_{\tilde{\mathcal{A}}^c_k\tilde{\mathcal{A}}_k}\tilde{\boldsymbol{\theta}}_{\tilde{\mathcal{A}}_k}\|_\infty \le (\tilde{H}_k+\tilde{K}_2)\cdot\lambda_1\|\boldsymbol{w}^c_{\tilde{\mathcal{A}}_k}\|_\infty+\tilde{K}_2 < \lambda_1\|\boldsymbol{w}^c_{\tilde{\mathcal{A}}^c_k}\|_{\min},$$

where the last inequality has used the fact that

$$\begin{aligned}(1-d\lambda_1\tilde{G}_k)\cdot\|\boldsymbol{w}^c_{\tilde{\mathcal{A}}^c_k}\|_{\min} &> (d\lambda_1\tilde{G}_k+\tilde{H}_k)\cdot(\frac{1-d\lambda_1\tilde{G}_k}{2\lambda_1\tilde{G}_k}\psi_{\min}-d\tilde{G}_k)+d\tilde{G}_k(1+\tilde{H}_k)\\ &\ge (d\lambda_1\tilde{G}_k+\tilde{H}_k)\cdot\|\boldsymbol{w}^c_{\tilde{\mathcal{A}}_k}\|_\infty+d\tilde{G}_k(1+\tilde{H}_k),\end{aligned}$$

which is due to the claims (B.32) and (B.33).

Now four optimality conditions are proved. Moreover, the feasibility and the uniqueness of $(\tilde{\boldsymbol{\theta}},\tilde{\boldsymbol{\alpha}})$ can be similarly proved by following the same line of proof as in Theorem 3.3. Thus, it remains to verify both claims (B.32) and (B.33) to complete the proof of Theorem 3.5. Under the event $\{\|\hat{\boldsymbol{R}}^s-\boldsymbol{\Sigma}^*\|_{\max}\le\lambda_0/M\}$, Theorem 3.4 shows that $\|\hat{\boldsymbol{\Theta}}^s_c-\boldsymbol{\Theta}^*\|_{\max}\le 2M\lambda_0$. Then we can derive a lower bound for $\|\boldsymbol{w}^c_{\mathcal{A}^c_k}\|_{\min}$.

$$\|\boldsymbol{w}^c_{\mathcal{A}^c_k}\|_{\min} = \frac{1}{\max\limits_{j\in\mathcal{A}^c_k}|\hat{\theta}^{s.c}_j|+\frac{1}{n}} \ge \frac{1}{2M\lambda_0+\frac{1}{n}},$$

which immediately yields the desired lower bound by noting that

$$\frac{\tilde{G}_k \cdot d\lambda_1 + \tilde{H}_k}{2\tilde{G}_k \cdot \lambda_1}\psi_{\min} + \frac{\tilde{G}_k \cdot d}{1 - \tilde{G}_k \cdot d\lambda_1} \leq \frac{\tilde{H}_k \psi_{\min}}{2\tilde{G}_k \cdot \lambda_1} + (\psi_{\min} + 2\tilde{G}_k) \cdot d \leq \frac{1}{2M\lambda_0 + \frac{1}{n}},$$

where both inequalities follow from the proper choices of tuning parameters λ_0 & λ_1 as stated in the assumptions of Theorem 3.5.

On the other hand, similarly as in Theorem 3.4, we can show that

$$\|\boldsymbol{w}^c_{\mathcal{A}_k}\|_\infty \leq \frac{1}{\min\limits_{j \in \mathcal{A}_k} |\hat{\theta}^{s.c}_j|} \leq \frac{1}{\psi_{\min} - 2M\lambda_0} \leq \frac{2}{\psi_{\min}} \leq \frac{\psi_{\min}}{4\tilde{G}_k \cdot \lambda_1}$$

and

$$\frac{1 - \tilde{G}_k \cdot d\lambda_1}{2\tilde{G}_k \cdot \lambda_1}\psi_{\min} - \tilde{G}_k \geq \frac{\psi_{\min}}{2\tilde{G}_k \cdot \lambda_1} - (\psi_{\min} + \tilde{G}_k) \cdot d \geq \frac{\psi_{\min}}{4\tilde{G}_k \cdot \lambda_1}$$

where both inequalities follow from the proper choices of tuning parameters λ_0 & λ_1 as stated in the assumptions of Theorem 3.5. Now two aforementioned claims, i.e. (B.32) and (B.33), are proved, which completes the proof of Theorem 3.5. □