Positive Semidefinite Rank-based Correlation Matrix Estimation with Application to Semiparametric Graph Estimation

Tuo Zhao* Kathryn Roeder† Han Liu‡

Abstract

Many statistical methods gain robustness and flexibility by sacrificing convenient computational structures. In this paper, we illustrate this fundamental tradeoff by studying a semiparametric graph estimation problem in high dimensions. We explain how novel computational techniques help to solve this type of problem. In particular, we propose a nonparanormal neighborhood pursuit algorithm to estimate high dimensional semiparametric graphical models with theoretical guarantees. Moreover, we provide an alternative view to analyze the tradeoff between computational efficiency and statistical error under a smoothing optimization framework. Though this paper focuses on the problem of graph estimation, the proposed methodology is widely applicable to other problems with similar structures. We also report thorough experimental results on text, stock, and genomic datasets.

1 Introduction

Undirected graphical models provide a powerful framework for exploring relationships between a large number of variables (Lauritzen, 1996; Wille et al., 2004; Honorio et al., 2009). More specifically, we can represent a d-dimensional random vector $\boldsymbol{X} = (X_1, ..., X_d)^T$ by an undirected graph $\mathcal{G} = (V, E)$, where V contains nodes corresponding to the d variables in $\boldsymbol{X}$, and the edge set E describes the conditional independence relationships between $X_1, ..., X_d$. Let $\boldsymbol{X}_{\backslash\{j,k\}} = \{X_\ell : \ \ell \neq j, k\}$, the distribution of $\boldsymbol{X}$ is Markov to $\mathcal{G}$ if X_j is independent of X_k given $\boldsymbol{X}_{\backslash\{j,k\}}$ for all $(j,k) \notin E$. In this paper, we aim to estimate the graph $\mathcal{G}$ based on n data points of $\boldsymbol{X}$.

A popular distributional assumption for estimating high dimensional undirected graph is multivariate Gaussian, i.e., $\boldsymbol{X} \sim N(\boldsymbol{\mu}, \Sigma)$. Under this assumption, the graph estimation problem

*Department of Computer Science, Johns Hopkins University, Baltimore, MD 21218, USA; e-mail: tour@cs.jhu.edu. Research supported by NSF Grant III-1116730.

†Department of Statistics, Carnegie Mellon University, Pittsburgh, PA 15213, USA; e-mail: roeder@stat.cmu.edu.Research supported by National Institute of Mental Health grant MH057881.

‡Department of Operations Research Research and Financial Engineering, Princeton University, Princeton, NJ 08544, USA; e-mail: hanliu@princeton.edu. Research supported by NSF Grant III-1116730.

can be solved by examining the sparsity pattern of the precision matrix $\mathbf{\Omega} = \Sigma^{-1}$ (Dempster, 1972). There are two major approaches for estimating high dimensional Gaussian graphical models: (i) graphical lasso (Banerjee et al., 2008; Yuan and Lin, 2007; Friedman et al., 2007), which maximizes the ℓ_1-penalized Gaussian likelihood and simultaneously estimates the precision matrix $\mathbf{\Omega}$ and the graph $\mathcal{G}$, (ii) neighborhood pursuit (Meinshausen and Bühlmann, 2006), which maximizes the ℓ_1-penalized pseudo-likelihood and only estimates the graph structure $\mathcal{G}$. Scalable software packages such as `glasso` and `huge` have been developed to implement these algorithms (Zhao et al., 2012a; Friedman et al., 2007). Both methods are consistent in graph recovery under suitable conditions (Meinshausen and Bühlmann, 2006; Ravikumar et al., 2011). However, these two methods have been observed to behave differently in practical applications. In many applications, the neighborhood pursuit method is preferred due to its computational simplicity. Many other Gaussian graph estimation methods have also been proposed and most of them fall in these two categories (Li and Gui, 2006; Lam and Fan, 2009; Peng et al., 2009; Shojaie and Michailidis, 2010; Yuan, 2010; Cai et al., 2011; Yin and Li, 2011; Jalali et al., 2012; Sun and Li, 2012; Sun and Zhang, 2012).

The Gaussian graphical model crucially relies on the normality assumption, which is restrictive in real applications. To relax this assumption, Liu et al. (2009) propose the semiparametric nonparanormal model. They assume that there exists a set of nondecreasing transformations $\boldsymbol{f} = \{f_j\}_{j=1}^d$, which make the transformed random vector $\boldsymbol{f}(\boldsymbol{X}) = (f_1(X_1), \ldots, f_d(X_d))^T$ follow a Gaussian distribution, i.e., $f(\boldsymbol{X}) \sim N(\mathbf{0}, \mathbf{\Omega}^{-1})$. Liu et al. (2009) show that, under the nonparanormal model, the graph $\mathcal{G}$ is still encoded by the sparsity pattern of $\mathbf{\Omega}$. To estimate $\mathbf{\Omega}$, Liu et al. (2012) propose a rank-based estimator named nonparanormal SKEPTIC, which first calculates a transformed Kendall's tau correlation matrix, and then plugs the estimated correlation matrix into the graphical lasso. Such a procedure has been proven to attain the same parametric rates of convergence as the graphical lasso (Liu et al., 2012). However, the transformed Kendall's tau correlation matrix has no positive semidefiniteness guarantee, and directly plugging it into the neighborhood pursuit may lead to a nonconvex formulation. Therefore applying the transformed Kendall's tau matrix to the neighborhood pursuit approach is challenging to both computational and theoretical analyses.

In this paper, we propose a novel projection algorithm to handle this challenge. More specifically, we project the possibly indefinite transformed Kendall's tau matrix into the cone of all positive semidefinite matrices with respect to a smoothed elementwise ℓ_∞ norm, which is induced by the smoothing approach (Nesterov, 2005). We provide both computational and theoretical analyses of the obtained procedure. Computationally, the smoothed elementwise ℓ_∞ norm has nice computational properties that allow us to develop an accelerated proximal gradient algorithm with a convergence rate of $O(\epsilon^{-1/2})$, where ϵ is the desired accuracy of the objective value (Nesterov, 1988). Theoretically, we prove that the proposed projection method preserves the concentration property of the transformed Kendall's tau matrix in high dimensions. This important result enables us to prove graph estimation consistency of the neighborhood pursuit method for

nonparanormal graph estimation (Meinshausen and Bühlmann, 2006).

Besides these new computational and theoretical analyses, we also provide an alternative view to analyze the tradeoff between computational efficiency and statistical error under the smoothing optimization framework. In existing literature (Nesterov, 2005; Chen et al., 2012), the smoothing approach is usually considered as a tradeoff between computational efficiency and approximation error. Thus the smoothness has to be controlled to avoid a large approximation error, which results in a slower convergence rate $O(\epsilon^{-1})$. In this paper, we analyze this tradeoff by directly considering the statistical error. We show that the smoothing approach simultaneously preserves the good statistical properties and enjoys the computational efficiency. Some preliminary results in this paper first appeared in Zhao et al. (2012b) without the technical proofs. Here we provide detailed proofs, more experimental results (including both simulated and real datasets), and a comprehensive comparison between the projection method and other competitors.

Although this paper targets the semiparametric graph estimation problem, the proposed methodology is widely applicable to other statistical methods for nonparanormal models, such as copula discriminant analysis and the semiparametric sparse inverse column operator (Han et al., 2013; Zhao and Liu, 2013). Taking the result in this paper as an initial start, we expect more sophisticated and stronger follow-up work that applies to problems with similar structure.

The rest of this paper is organized as follows: In §2 we briefly review the transformed Kendall's tau estimator; In §3, we describe the projection algorithm and nonparanormal neighborhood pursuit algorithm; In §4, we analyze the statistical properties of the proposed procedures; In §5 and §6, we present experimental results on both simulated and real datasets; In §7, we discuss other related correlation estimators and draw conclusions.

2 Background

We start with some notations. Given a vector $\boldsymbol{v} = (v_1, \ldots, v_d)^T \in \mathbb{R}^d$, we define vector norms $\|\boldsymbol{v}\|_1 = \sum_j |v_j|$, $\|\boldsymbol{v}\|_2^2 = \sum_j v_j^2$, $\|\boldsymbol{v}\|_\infty = \max_j |v_i|$. Given two symmetric matrices $\mathbf{A} = \left[\mathbf{A}_{jk}\right] \in \mathbb{R}^{d\times d}$ and $\mathbf{B} = \left[\mathbf{B}_{jk}\right] \in \mathbb{R}^{d\times d}$, we de note $\Lambda_{\min}(\mathbf{A})$ and $\Lambda_{\max}(\mathbf{A})$ as the smallest and largest eigenvalues of $\mathbf{A}$ respectively; We also denote $\left\langle \mathbf{A}, \mathbf{B} \right\rangle = \mathrm{tr}(\mathbf{A}^T\mathbf{B})$ as the inner product of $\mathbf{A}$ and $\mathbf{B}$. We define matrix norms as $\left\|\left|\mathbf{A}\right|\right\|_1 = \sum_{j,k} |\mathbf{A}_{jk}|$, $\left\|\left|\mathbf{A}\right|\right\|_\infty = \max_{j,k} |\mathbf{A}_{jk}|$, $\|\mathbf{A}\|_{\mathrm{F}}^2 = \sum_{j,k} |\mathbf{A}_{jk}|^2$, and $\|\mathbf{A}\|_\infty = \max_j \sum_k |\mathbf{A}_{jk}|$. We denote $\boldsymbol{v}_{\setminus j} = (v_1, \ldots, v_{j-1}, v_{j+1}, \ldots, v_d)^T \in \mathbb{R}^{d-1}$ as the subvector of $\boldsymbol{v}$ with the j^{th} entry removed. We denote $\mathbf{A}_{\setminus i,\setminus j}$ as the submatrix of $\mathbf{A}$ with the i^{th} row and the j^{th} column removed. We denote $\mathbf{A}_{i,\setminus j}$ as the i^{th} row of $\mathbf{A}$ with its j^{th} entry removed. If I and J are two sets of indices, then $\mathbf{A}_{IJ}$ denotes the submatrix of $\mathbf{A}$ whose rows and columns are indexed by I and J.

2.1 *Nonparanormal* SKEPTIC

The nonparanormal distribution extends the Gaussian distribution by separately modeling the conditional independence structure and marginal distributions. It assumes that there exists a set of transformations such that the transformed random vector follows a Gaussian distribution.

Definition 2.1. *Let $\boldsymbol{f} = \{f_1, ..., f_d\}$ be a collection of nondecreasing univariate functions and $\Sigma^* \in \mathbb{R}^{d\times d}$ be a correlation matrix with* $\mathrm{diag}(\Sigma^*) = \mathbf{1}$. *We say that a d-dimensional random vector $\boldsymbol{X} = (X_1, ..., X_d)^T$ follows a nonparanormal distribution, denoted by*

$$\boldsymbol{X} \sim NPN(\boldsymbol{f}, \Sigma^*),$$

if $\boldsymbol{f}(\boldsymbol{X}) = (f_1(X_1), ..., f_d(X_d))^T \sim N(0, \Sigma^*)$.

For continuous distributions, Liu et al. (2009) prove that the nonparanormal family is equivalent to the Gaussian copula family (Klaassen and Wellner, 1997; Tsukahara, 2005). Moreover, the nonparanormal models have the same nice property as Gaussian models that the conditional independence graph can be characterized by the sparsity pattern of $\boldsymbol{\Omega}^* = (\Sigma^*)^{-1}$. To estimate $\boldsymbol{\Omega}^*$, Liu et al. (2012) propose a rank-based method – namely *nonparanormal* SKEPTIC for estimating the correlation matrix. They exploit the Kendall's tau statistic to directly estimate the unknown correlation matrix. This approach avoids explicitly calculating the marginal transformation functions $\{f_j\}_{j=1}^d$ and achieves the optimal parametric rates of convergence.

More specifically, let $\boldsymbol{x}^1, ..., \boldsymbol{x}^n$ be n independent observations of $\boldsymbol{X}$, where $\boldsymbol{x}^i = (x_1^i, ..., x_d^i)^T$. Then the Kendall's tau statistic $\widehat{\tau}_{jk}$ is defined as follows,

$$\widehat{\tau}_{jk} = \begin{cases} \dfrac{2}{n(n-1)} \displaystyle\sum_{i<i'} \mathrm{sign}\left(x_j^i - x_j^{i'}\right)\left(x_k^i - x_k^{i'}\right) & k \neq j \\ 1 & k = j \end{cases}.$$

Though $\widehat{\tau}_{jk}$ is not consistent for Σ^*_{jk}, the bias can be corrected by resorting to a transformed Kendall's tau estimator $\widehat{\mathbf{S}} = [\widehat{\mathbf{S}}_{jk}] \in \mathbb{R}^{d\times d}$, defined as $\widehat{\mathbf{S}}_{jk} = \sin\left(\pi \cdot \widehat{\tau}_{jk}/2\right)$. Liu et al. (2012) further characterize the following concentration property of the transformed Kendall's tau estimator.

Lemma 2.2 (Liu et al. (2012))**.** *Suppose that $\boldsymbol{X} \sim NPN(\boldsymbol{f}, \Sigma^*)$. There exists a universal constant κ_1, such that*

$$\mathbb{P}\left(\left|\left|\left|\widehat{\mathbf{S}} - \Sigma^*\right|\right|\right|_\infty \leq \kappa_1 \sqrt{\frac{\log d}{n}}\right) \geq 1 - \frac{1}{d^3}. \tag{2.1}$$

Existing literature often uses (2.1) as an important condition to achieve parameter estimation and model selection consistency in high dimensions (Meinshausen and Bühlmann, 2006; Ravikumar et al., 2011; Liu et al., 2012). Therefore in the next section, we propose a projection method to obtain a positive semidefinite replacement of $\widehat{\mathbf{S}}$, which also preserves this concentration property.

2.2 Possible Indefiniteness

Here we present a typical example to illustrate the possible indefiniteness of the transformed Kendall's tau matrix. We sample 100 data from a 200-dimensional normal distribution $N(0, \mathbf{I}_{200})$ and calculate the minimum eigenvalue of the transformed Kendall's tau matrix. We repeat the

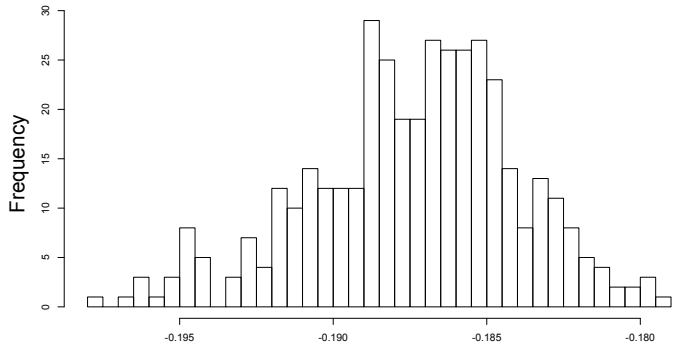


Figure 1: The histogram of minimum eigenvalues of the transformed Kendall's tau matrix. Across all 400 replications, the minimum eigenvalues are always negative.

sampling for 400 times and present the histogram of minimum eigenvalues in Figure 1. We see that these minimum eigenvalues are close to -0.187. For all 400 replications, the minimum eigenvalues are always negative.

Such indefiniteness causes trouble when we apply the neighborhood pursuit algorithm to nonparanormal graph estimation. The main reason is that the neighborhood pursuit is formulated as a quadratic program, when the transformed Kendall's tau matrix in the quadratic term is indefinite, the objective function is no longer convex. Thus existing quadratic programming solvers cannot guarantee a global solution in polynomial time. In the next section, we present a projection method to circumvent this difficulty.

3 Methodology

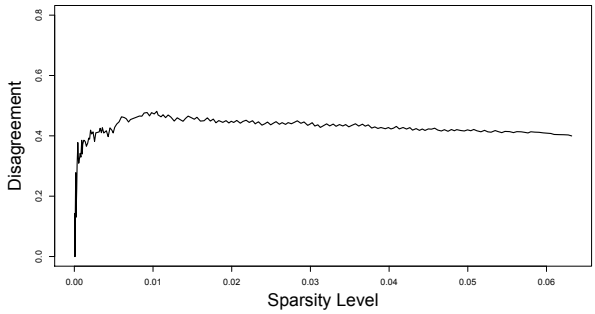
Before introducing the proposed method, we first motivate our interest in the nonparanormal neighborhood pursuit.

3.1 Neighborhood Pursuit v.s. Graphical lasso

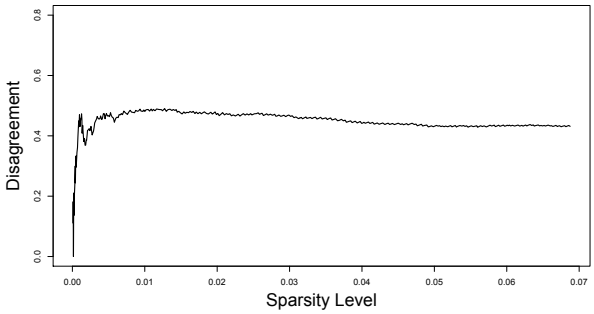
Ravikumar et al. (2011) study the sufficient conditions for perfect graph recovery for the graphical lasso and neighborhood pursuit. They numerically verify that the neighborhood pursuit has better sample complexity than the graphical lasso.

In many real applications, neighborhood pursuit and graphical lasso behave differently. For example, we apply both methods to a stock market dataset (See §6 for more details). We calculate the solution paths over a grid of regularization parameters and align the obtained paths by their sparsity levels. We further calculate the proportion of edges on which these two methods disagree with each other. More specifically, we adjust the regularization parameters of the neighborhood pursuit and graphical lasso such that the obtained graphs have approximately the same number of edges. For example, if one graph has 100 edges and the other graph has 102 edges— with 40

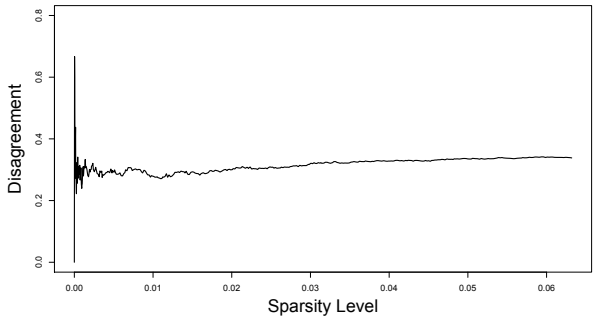
shared edges—the disagreement between these two graphs is $(60/100 + 62/102)/2 = 0.6039$. As shown in Figure 2(a), there exists a large amount of disagreement between the graphical lasso estimate and the neighborhood pursuit estimate. A similar phenomenon is also found in Figure 2(b), where the nonparanormal neighborhood pursuit shows significant difference from the nonparanormal graphical lasso. In addition, we compare the Gaussian graph estimation and nonparanormal graph estimation in the same way. From Figures 2(c) and 2(d), we see that the neighborhood pursuit and graphical lasso also behave differently from their corresponding nonparanormal based methods.



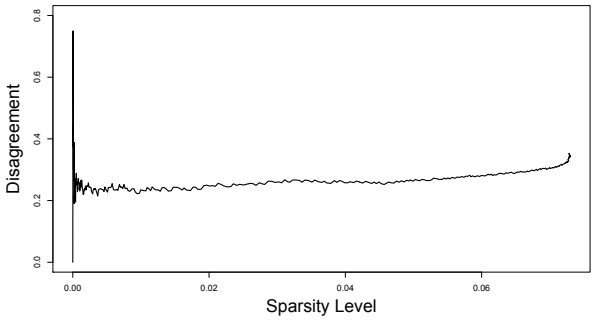
(a) Neighborhood Pursuit v.s. Graphical Lasso



(b) Nonparanormal Neighborhood Pursuit v.s. Nonparanormal Graphical Lasso



(c) Neighborhood Pursuit v.s. Nonparanormal Neighborhood Pursuit



(d) Graphical Lasso v.s. Nonparanormal Graphical Lasso

Figure 2: The quantitative comparison between the solution paths recovered by the graphical lasso, neighborhood pursuit, nonparanormal graphical lasso, and nonparanormal neighborhood pursuit. The neighborhood pursuit shows a large amount of disagreement with the graphical lasso in both Gaussian and nonparanormal graph estimation. Additionally, the neighborhood pursuit and graphical lasso also behave differently from their corresponding nonparanormal based methods.

Given the above merit of the neighborhood pursuit in graph recovery and its different empirical behavior, we are interested in extending this method to the nonparanormal graph estimation.

3.2 Nonparanormal Neighborhood Pursuit

Recall that $\boldsymbol{f}(\boldsymbol{X}) = (f_1(X_1),\ldots,f_d(X_d))^T$, and let $\boldsymbol{f}_{\backslash j}(\boldsymbol{X}_{\backslash j})$ be the $(d-1)$-dimensional subvector of $\boldsymbol{f}(\boldsymbol{X})$ with the j^{th} entry $f_j(X_j)$ removed. Since the conditional distribution of $f_j(X_j)$ given $\boldsymbol{f}_{\backslash j}(X_{\backslash j})$ remains Gaussian, i.e.,

$$f_j(X_j)|\boldsymbol{f}_{\backslash j}(\boldsymbol{X}_{\backslash j}) \sim N\left(\Sigma^*_{j,\backslash j}(\Sigma^*_{\backslash j,\backslash j})^{-1}\boldsymbol{f}_{\backslash j}(\boldsymbol{X}_{\backslash j}), \Sigma^*_{j,j} - \Sigma^*_{j,\backslash j}(\Sigma^*_{\backslash j,\backslash j})^{-1}\Sigma^*_{\backslash j,j}\right), \tag{3.1}$$

then (3.1) can be equivalently represented by a linear regression model:

$$f_j(X_j) = \boldsymbol{f}_{\backslash j}(\boldsymbol{X}_{\backslash j})^T\mathbf{B}^*_{\backslash j,j} + Z_j,$$

where $Z_j \sim N\left(0, \Sigma^*_{j,j} - \Sigma^*_{j,\backslash j}(\Sigma^*_{\backslash j,\backslash j})^{-1}\Sigma^*_{\backslash j,j}\right)$ and $\mathbf{B}^* \in \mathbb{R}^{d\times d}$ is defined as

$$\mathbf{B}^*_{\backslash j,j} = (\Sigma^*_{\backslash j,\backslash j})^{-1}\Sigma^*_{\backslash j,j} \text{ and } \mathbf{B}^*_{j,j} = 0 \text{ for all } 1 \le j \le d. \tag{3.2}$$

Consequently, the nonparanormal neighborhood pursuit solves

$$\widehat{\mathbf{B}}_{\backslash j,j} = \underset{\mathbf{B}_{j,j}=0}{\operatorname{argmin}}\, \mathbf{B}^T_{\backslash j,j}\widetilde{\mathbf{S}}_{\backslash j,\backslash j}\mathbf{B}_{\backslash j,j} - 2\widetilde{\mathbf{S}}^T_{\backslash j,j}\mathbf{B}_{\backslash j,j} + \lambda\|\mathbf{B}_{\backslash j,j}\|_1 \text{ for all } j = 1,\ldots,d, \tag{3.3}$$

where $\widetilde{\mathbf{S}}$ is a positive semidefinite replacement of the transformed Kendall's tau matrix $\widehat{\mathbf{S}}$. The optimization problem in (3.3) can be efficiently solved by the coordinate descent algorithm (Friedman et al., 2007). Since $\widehat{\mathbf{B}}$ may not be symmetric, we need an additional symmetrization procedure to obtain a graph estimator. That is, let $\mathbf{E}^*$ be the adjacency matrix of the underlying graph $\mathcal{G}^*$, we estimate $\mathbf{E}^*$ by

$$\widehat{\mathbf{E}} = [\widehat{\mathbf{E}}_{jk}] = [I(\max\{|\widehat{\mathbf{B}}_{jk}|, |\widehat{\mathbf{B}}_{kj}|\} > 0)].$$

In the next subsection, we explain how to obtain the positive semidefinite replacement $\widetilde{\mathbf{S}}$ with theoretical guarantees in high dimensions.

3.3 Positive Semidefinite Projection

Our proposed projection method is motivated by the following problem:

$$\overline{\mathbf{S}} = \underset{\mathbf{S}}{\operatorname{argmin}} \left|\left|\left|\widehat{\mathbf{S}} - \mathbf{S}\right|\right|\right|_\infty \text{ subject to } \mathbf{S} \succeq 0. \tag{3.4}$$

Since $\overline{\mathbf{S}}$ is the minimizer of (3.4), and Σ^* is a feasible solution to (3.4), by the triangle inequality, we have

$$\left|\left|\left|\Sigma^* - \overline{\mathbf{S}}\right|\right|\right|_\infty \le \left|\left|\left|\widehat{\mathbf{S}} - \overline{\mathbf{S}}\right|\right|\right|_\infty + \left|\left|\left|\widehat{\mathbf{S}} - \Sigma^*\right|\right|\right|_\infty \le 2\left|\left|\left|\widehat{\mathbf{S}} - \Sigma^*\right|\right|\right|_\infty. \tag{3.5}$$

Thus combining (3.5) with Lemma 2.2, it is straightforward to show that

$$\left|\left|\left|\overline{\mathbf{S}} - \Sigma^*\right|\right|\right|_\infty = O_P(\sqrt{\log d/n}). \tag{3.6}$$

However, (3.4) is computationally expensive because of its nonsmoothness. To gain computational efficiency, we apply the smoothing approach in Nesterov (2005) to solve (3.4) at the expense of a controllable accuracy loss. For any matrix $\mathbf{A} \in \mathbb{R}^{d\times d}$, we consider the following smooth surrogate of the elementwise ℓ_∞ norm

$$\left\|\left\|\mathbf{A}\right\|\right\|_\infty^\mu = \max_{\|\mathbf{U}\|_1 \le 1} \left\langle \mathbf{U}, \mathbf{A} \right\rangle - \frac{\mu}{2}\|\mathbf{U}\|_{\mathrm{F}}^2, \tag{3.7}$$

where $\mu > 0$ is a smoothing parameter. The first term in (3.7) is well known as the Fenchel's dual representation, and the second term is the proximity function of $\mathbf{U}$. We call $\left\|\left\|\cdot\right\|\right\|_\infty^\mu$ smoothed elementwise ℓ_∞ norm. The next lemma characterizes the solution to (3.7).

Lemma 3.1. *Let* $\widetilde{\mathbf{U}}^{(\mathbf{A})} = [\widetilde{\mathbf{U}}_{jk}^{(\mathbf{A})}]$ *be the optimal solution to* (3.7)*, we have*

$$\widetilde{\mathbf{U}}_{jk}^{(\mathbf{A})} = \operatorname{sign}\left(\mathbf{A}_{jk}\right) \max\left\{\left|\frac{\mathbf{A}_{jk}}{\mu}\right| - \gamma, 0\right\}, \tag{3.8}$$

where γ *is the minimum nonnegative constant such that* $\left\|\left\|\widetilde{\mathbf{U}}^{(\mathbf{A})}\right\|\right\|_1 \le 1$.

The proof of Lemma 3.1 is provided in Appendix C. The naive algorithm for calculating γ is to sort the matrix, which has an average-case complexity of $O(d^2 \log d)$. See Appendix D for a more efficient algorithm with an average-case complexity of $O(d^2)$.

Figure 3 shows several two-dimensional examples of the elementwise ℓ_∞ norm smoothed by different μ's. We see that increasing μ makes the function smoother, but induces a larger approximation error. The smoothed elementwise ℓ_∞ norm is convex, and has a gradient as follows,

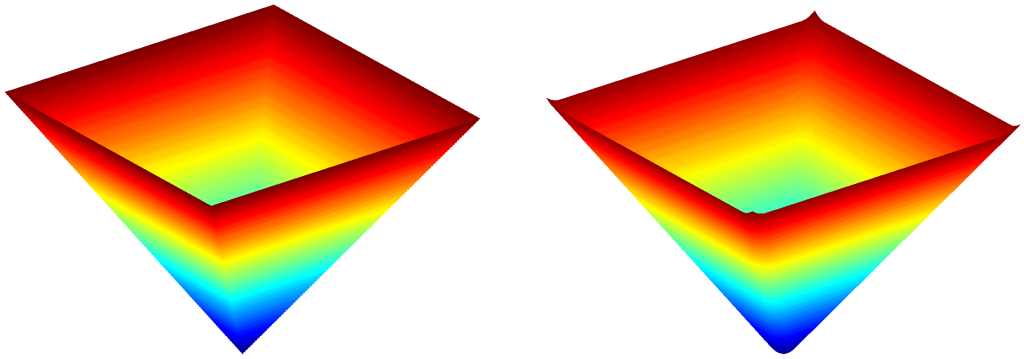
$$\nabla\left\|\left\|\widehat{\mathbf{S}} - \mathbf{S}\right\|\right\|_\infty^\mu = \frac{\partial\left\|\left\|\widehat{\mathbf{S}} - \mathbf{S}\right\|\right\|_\infty^\mu}{\partial(\widehat{\mathbf{S}} - \mathbf{S})} \cdot \frac{\partial(\widehat{\mathbf{S}} - \mathbf{S})}{\partial \mathbf{S}} = -\widetilde{\mathbf{U}}^{(\widehat{\mathbf{S}}-\mathbf{S})}. \tag{3.9}$$

Recall that $\widetilde{\mathbf{U}}^{(\widehat{\mathbf{S}}-\mathbf{S})}$ is essentially a soft thresholding function, therefore it is continuous in $\mathbf{S}$ with a Lipchitz constant $1/\mu$. Since the smoothed elementwise ℓ_∞ norm has the above nice computational properties with a controllable loss in accuracy (See §4 for more details), we focus on the following smooth relaxed optimization problem,

$$\widetilde{\mathbf{S}} = \operatorname*{argmin}_{\mathbf{S}} \left\|\left\|\widehat{\mathbf{S}} - \mathbf{S}\right\|\right\|_\infty^\mu \text{ subject to } \mathbf{S} \succeq 0. \tag{3.10}$$

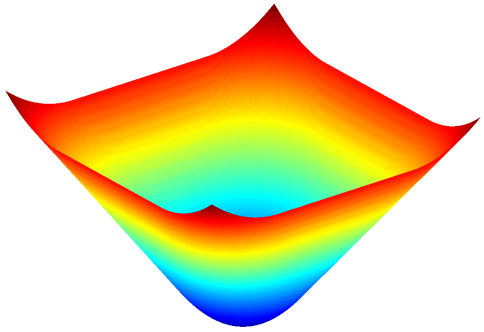
3.4 Accelerated Proximal Gradient Algorithm

We propose an accelerated proximal gradient algorithm as in Nesterov (1988) to solve the optimization problem in (3.10). Unlike most existing accelerated proximal gradient algorithms for unconstrained minimization problems (Ji and Ye, 2009; Chen et al., 2012; Zhao and Liu, 2012; Nesterov, 2005; Beck and Teboulle, 2009), the proposed algorithm can handle the positive semidefinite constraint in (3.10). Before we proceed with the details of the algorithm, we first define the

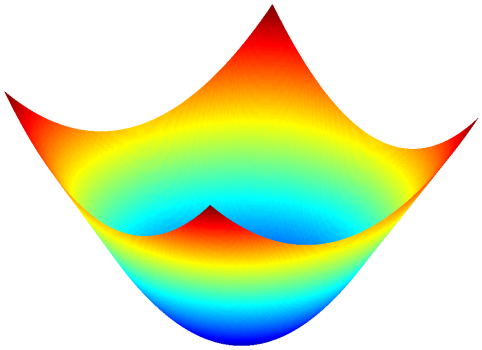


(a) No smoothing ($\mu = 0$)

(b) $\mu = 0.1$



(c) $\mu = 0.5$



(d) $\mu = 1$

Figure 3: The elementwise ℓ_∞ norm (a) and smoothed elementwise ℓ_∞ norms with $\mu = 0.1, 0.5, 1$ (b-d). Increasing μ makes the function smoother, but induces a larger approximation error.

projection operator as

$$\Pi_+(\mathbf{A}) = \operatorname*{argmin}_{\mathbf{B}} \|\mathbf{B} - \mathbf{A}\|_{\mathrm{F}}^2 \text{ subject to } \mathbf{B} \succeq 0, \tag{3.11}$$

where $\mathbf{A} \in \mathbb{R}^{d\times d}$ is a symmetric matrix. $\Pi_+(\mathbf{A})$ is the projection of the symmetric matrix $\mathbf{A}$ to the cone of all positive semidefinite matrices with respect to the Frobenius norm. $\Pi_+(\mathbf{A})$ has a closed form solution as shown in the following lemma.

Lemma 3.2. *Suppose that* $\mathbf{A}$ *has an eigenvalue decomposition* $\mathbf{A} = \sum_{j=1}^{d} \sigma_j \boldsymbol{v}_j \boldsymbol{v}_j^T$, *where* σ_j*'s are the eigenvalues, and* $\boldsymbol{v}_j$*'s are the corresponding eigenvectors. We have*

$$\Pi_+(\mathbf{A}) = \sum_{j=1}^{d} \max\{\sigma_j, 0\} \boldsymbol{v}_j \boldsymbol{v}_j^T. \tag{3.12}$$

The proof of Lemma 3.2 is provided in Appendix E.

Now we start to derive the algorithm. We define two sequences of auxiliary variables $\mathbf{M}^{(t)}$ and $\mathbf{W}^{(t)}$ with $\mathbf{M}^{(0)} = \mathbf{W}^{(0)} = \mathbf{S}^{(0)}$, and a sequence of weights $\theta_t = 2/(1+t)$ for $t = 1, 2, \ldots$. At the t^{th}

iteration, we calculate the auxiliary variable $\mathbf{M}^{(t)}$ as

$$\mathbf{M}^{(t)} = (1-\theta_t)\mathbf{S}^{(t-1)} + \theta_t \mathbf{W}^{(t-1)}. \tag{3.13}$$

We then evaluate the gradient using

$$\mathbf{G}^{(t)} = \frac{\partial \left\|\left\|\widehat{\mathbf{S}} - \mathbf{M}^{(t)}\right\|\right\|_\infty^\mu}{\partial \mathbf{M}^{(t)}} = \left[-\text{sign}\left(\widehat{\mathbf{S}}_{jk} - \mathbf{M}_{jk}^{(t)}\right) \cdot \max\left\{\left|\frac{\widehat{\mathbf{S}}_{jk} - \mathbf{M}_{jk}^{(t)}}{\mu}\right| - \gamma, 0\right\}\right]. \tag{3.14}$$

We consider the following quadratic approximation:

$$Q(\mathbf{W}, \mathbf{W}^{(t-1)}, \mu) = \left\|\left\|\widehat{\mathbf{S}} - \mathbf{W}^{(t-1)}\right\|\right\|_\infty^\mu + \left\langle \mathbf{G}^{(t)}, \mathbf{W} - \mathbf{W}^{(t-1)} \right\rangle + \frac{1}{2\eta_t\theta_t}\|\mathbf{W} - \mathbf{W}^{(t-1)}\|_{\text{F}}^2, \tag{3.15}$$

where η_t is the step-size for the t^{th} iteration. Then we take

$$\mathbf{W}^{(t)} = \underset{\mathbf{W}\geq 0}{\text{argmin}}\, Q(\mathbf{W}, \mathbf{W}^{(t-1)}, \mu) = \Pi_+\left(\mathbf{W}^{(t-1)} - \frac{\eta_t}{\theta_t}\mathbf{G}^{(t)}\right). \tag{3.16}$$

We further calculate $\mathbf{S}^{(t)}$ for the t^{th} iteration as follows,

$$\mathbf{S}^{(t)} = (1-\theta_t)\mathbf{S}^{(t-1)} + \theta_t \mathbf{W}^{(t)}. \tag{3.17}$$

Let ε be the target precision, the algorithm stops when $\left|\left\|\left\|\widehat{\mathbf{S}} - \mathbf{S}^{(t)}\right\|\right\|_\infty^\mu - \left\|\left\|\widehat{\mathbf{S}} - \mathbf{S}^{(t-1)}\right\|\right\|_\infty^\mu\right| \leq \varepsilon\mu$.

Remark 3.3. *A conservative choice of η_t is $\eta_t = \mu$ in all iterations. Since μ is the Lipchitz constant of $\mathbf{G}^{(t)}$, we always have*

$$\begin{aligned}
\left\|\left\|\widehat{\mathbf{S}} - \mathbf{S}^{(t)}\right\|\right\|_\infty^\mu &\leq \left\|\left\|\widehat{\mathbf{S}} - \mathbf{M}^{(t)}\right\|\right\|_\infty^\mu + \left\langle \mathbf{G}^{(t)}, \mathbf{S}^{(t)} - \mathbf{M}^{(t)} \right\rangle + \frac{1}{2\mu}\|\mathbf{S}^{(t)} - \mathbf{M}^{(t)}\|_{\text{F}}^2 \\
&\leq \left\|\left\|\widehat{\mathbf{S}} - \mathbf{M}^{(t)}\right\|\right\|_\infty^\mu + \left\langle \mathbf{G}^{(t)}, \mathbf{S}^{(t)} - \mathbf{M}^{(t)} \right\rangle + \frac{1}{2\mu\theta_t}\|\mathbf{S}^{(t)} - \mathbf{M}^{(t)}\|_{\text{F}}^2 \\
&= Q(\mathbf{S}^{(t)}, \mathbf{M}^{(t)}, \mu),
\end{aligned}$$

where the second inequality comes from $\theta_t \leq 1$. To gain a better empirical performance, we can adopt the backtracking line search with a sequence of non-increasing step-sizes η_t's for $t = 1, 2....$ More specifically, we start with a large enough η_1, and within each iteration we choose the minimum integer z such that

$$\left\|\left\|\widehat{\mathbf{S}} - \mathbf{S}^{(t)}\right\|\right\|_\infty^\mu \leq Q(\mathbf{S}^{(t)}, \mathbf{M}^{(t)}, \eta_t) \;\; \textit{with} \;\; \eta_t = \max\{\mu, m^z\eta_{t-1}\}, \tag{3.18}$$

where $m \in (0,1)$ is the shrinkage parameter.

We summarize the accelerated proximal gradient algorithm in Algorithm 1.

The following theorem establishes the worst-case convergence rate of the proposed accelerated proximal gradient algorithm.

Algorithm 1 The Accelerated Proximal Gradient Algorithm.

Input: $\widehat{\mathbf{S}}$, $\mathbf{S}^{(0)} = \mathbf{M}^{(0)} = \mathbf{W}^{(0)}$, μ, $\theta_t = 2/(1+t)$, ε
Output: $\widetilde{\mathbf{S}} = \mathbf{S}^{(t)}$
Initialize: $t = 1$
repeat
1: Compute the auxiliary variables $\mathbf{M}^{(t)}$ using (3.13)
2: Compute the gradient of (3.10) at $\mathbf{M}^{(t)}$ using (3.14)
3: (Optional) Compute η_t by the backtracking line search using (3.18)
4: Compute the auxiliary variables $\mathbf{W}^{(t)}$ using (3.16)
5: Compute the solution $\mathbf{S}^{(t)}$ using (3.17)
6: $t = t + 1$
until $\left|\left|\left|\widehat{\mathbf{S}} - \mathbf{S}^{(t)}\right|\right|\right|_\infty^\mu - \left|\left|\left|\widehat{\mathbf{S}} - \mathbf{S}^{(t-1)}\right|\right|\right|_\infty^\mu \Big| \leq \varepsilon\mu$.

Theorem 3.4. *To achieve the desired accuracy ϵ such that $\left|\left|\left|\widehat{\mathbf{S}} - \mathbf{S}^{(t)}\right|\right|\right|_\infty^\mu - \left|\left|\left|\widehat{\mathbf{S}} - \widetilde{\mathbf{S}}\right|\right|\right|_\infty^\mu < \epsilon$, the required number of iterations is at most*

$$t = \sqrt{\frac{2\|\mathbf{S}^{(0)} - \widetilde{\mathbf{S}}\|_{\mathrm{F}}^2}{\mu\epsilon}} - 1 = O(\epsilon^{-1/2}\mu^{-1/2}).$$

Theorem 3.4 is a direct result of Nesterov (1988). It guarantees that our derived algorithm achieves the optimal convergence rate for minimizing (3.10) over all first order computational algorithms. Existing literature considers the smoothing approach as a tradeoff between computational efficiency and approximation error (Nesterov, 2005; Chen et al., 2012). Therefore they analyze the convergence rate with respect to the optimal solution to the original problem (3.4). They choose to set the smoothing parameter μ small enough (e.g., $\mu = \epsilon/2$) to avoid a large approximation error, and eventually get a slower convergence rate $O\left(\epsilon^{-1}\right)$.

In contrast, we directly analyze the tradeoff between the computational efficiency and statistical error (Agarwal et al., 2012). Although $\widetilde{\mathbf{S}}$ is not the optimal solution to the original problem (3.4), our analysis in §4 will show that choosing a larger μ (e.g., $\mu \asymp \sqrt{\log d/n}$) can still make $\widetilde{\mathbf{S}}$ concentrate to Σ^* with a rate similar to (2.1) in high dimensions. This boosts the computational performance.

4 Statistical Theory

We first present the rate of convergence of $\widetilde{\mathbf{S}}$ under the elementwise ℓ_∞ norm.

Theorem 4.1. *Suppose that $\boldsymbol{X} \sim NPN(\boldsymbol{f}, \Sigma^*)$, there exist universal constants κ_2 and κ_3 such that by taking $\mu = \kappa_2\sqrt{\log d/n}$, we have the optimum to* (3.10), *$\widetilde{\mathbf{S}}$ satisfying*

$$\mathbb{P}\left(\left|\left|\left|\widetilde{\mathbf{S}} - \Sigma^*\right|\right|\right|_\infty \leq \kappa_3\sqrt{\frac{\log d}{n}}\right) \geq 1 - \frac{1}{d^3}. \tag{4.1}$$

The proof of Theorem 4.1 is provided in Appendix A. Theorem 4.1 implies that we can choose a reasonably large μ to gain the computational efficiency without losing statistical efficiency.

Remark 4.2. *By plugging $\mu = \kappa_2\sqrt{\log d/n}$ into Theorem 3.4, we obtain a more refined convergence rate of the proposed computational algorithm as $O(\epsilon^{-1/2}(\log d/n)^{-1/4})$.*

Now we analyze the graph recovery performance of the nonparanormal neighborhood pursuit under suitable conditions. Let I_j denote the set of the neighbors of node j, and J_j denote the set of the non-neighbors of node j. For all $j = 1,\ldots,d$, we assume that

$$\text{Assumption 1} \qquad \|\Sigma^*_{J_jI_j}(\Sigma^*_{I_jI_j})^{-1}\|_\infty \le \alpha,$$
$$\text{Assumption 2} \qquad \Lambda_{\min}(\Sigma^*_{I_jI_j}) \ge \delta,\ \ \|(\Sigma^*_{I_jI_j})^{-1}\|_\infty \le \psi,$$

where $\alpha \in (0,1)$, $\delta > 0$, and $\psi < \infty$ are all constants. Assumptions 1 and 2 have been extensively studied in existing literature (Zhao and Yu, 2006; Zou, 2006; Wainwright, 2009). Assumption 1 is known as the irrepresentable condition, which requires that the correlation between the non-neighborhood and neighborhood moderate. Assumption 2 is known as the minimum curvature condition, which requires that the correlation within the neighborhood cannot be too large. We then present the results on graph recovery consistency.

Theorem 4.3. *Recall that $\mathbf{E}^*$ denotes the adjacency matrix of $\mathcal{G}^*$, and*

$$\tau = \min_{\mathbf{E}^*_{jk}\neq 0} |\mathbf{B}^*_{jk}|, \tag{4.2}$$

we assume that $\mathbf{\Sigma}^*$ satisfies Assumptions 1 and 2. Let $s = \max_j |I_j|$. If we choose $\lambda \le \min\{\tau/\psi, 2\}$, then for large enough n such that

$$\sqrt{\frac{\log d}{n}} \le \min\left\{\frac{\lambda(1-\alpha)}{26\psi\alpha\kappa_3 s}, \frac{\lambda(1-\alpha)}{26\psi\kappa_3 s}, \frac{\lambda(1-\alpha)}{26(\alpha+1)}, \frac{\tau}{14\psi^2 s\kappa_3}, \frac{\tau}{14\psi\kappa_3}, \frac{\delta}{2s\kappa_3}\right\},$$

we have

$$\mathbb{P}\left(\widehat{\mathbf{E}} = \mathbf{E}^*\right) \ge 1 - \frac{1}{d^3}.$$

Moreover, we have $\mathbb{P}\left(\widehat{\mathbf{E}} = \mathbf{E}^*\right) \to 1$, if the following conditions hold:
Condition 1: α, δ, and ψ are constants, which do not scale with n, d, and s;
Condition 2: τ scales with n, d, and s as

$$\frac{s\log d}{\tau^2 n} \to 0 \ \text{ and } \ \frac{s^2\log d}{n} \to 0;$$

Condition 3: λ scales with τ, n, d, and s as

$$\frac{\lambda}{\tau} \to 0 \text{ and } \ \frac{s^2\log d}{\lambda^2 n} \to 0.$$

The proof of Theorem 4.3 is provided in Appendix B. It guarantees that we can correctly recover the underlying graph structure with high probability.

5 Numerical Simulations

In our numerical simulations, we use six different graphs with 200 nodes ($d = 200$) including neighborhood graph, clique graph, band graph, lattice graph, mixed scale-free graph, and hybrid graph to generate the precision matrix $\mathbf{\Omega}^*$ (shown in Figure 4):

• **Neighborhood**. For each node, we independently sample a random vector from a uniform distribution over $[0,1]^2 \subset \mathbb{R}^2$. Let $\boldsymbol{V}_i \in \mathbb{R}^2$ denote the random vector for node i, then we set an edge between node i and node j with probability $(2\pi)^{-1/2}\exp(-\|\boldsymbol{V}_i - \boldsymbol{V}_j\|_2^2/\phi)$. We set $\phi = 100$ for the simulations in §5.1, §5.2, and §5.3.

• **Clique**. The nodes are evenly partitioned into g disjoint groups and each group contains d/g nodes. The subgraph of each group is fully connected graph. We set $g = 20$ for the simulations in §5.1, §5.2, and §5.3.

• **Band**. Each node is assigned a coordinate j with $j = 1,...,d$. Two nodes are connected by an edge whenever the corresponding points are at distance no more than g. We set $g = 2$ for the simulations in §5.1, §5.2, and §5.3.

• **Lattice**. Each node is assigned a two dimensional coordinate (j,k) with $j = 1,...,g$ and $k = 1,...,d/g$. Two nodes are connected by an edge whenever the corresponding points are at distance 1. We set $g = 10$ for the simulations in §5.1, §5.2, and §5.3.

• **Mixed Scale-free**. The nodes are evenly partitioned into 4 groups. The nodes from different groups are disconnected. The subgraph of each group of nodes is a scale free graph. The degree distribution of the scale-free graph follows a power law. The graph is generated by the preferential attachment mechanism. The graph begins with an initial band graph of 10 nodes with $g = 1$. New nodes are added to the graph one at a time. Each new node is connected to existing node with a probability that is proportional to the number of degrees that the existing nodes already have. Formally, the probability p_i that the new node is connected to node i is, $p_i = \frac{k_i}{\sum_j k_j}$, where k_i is the degree of node i.

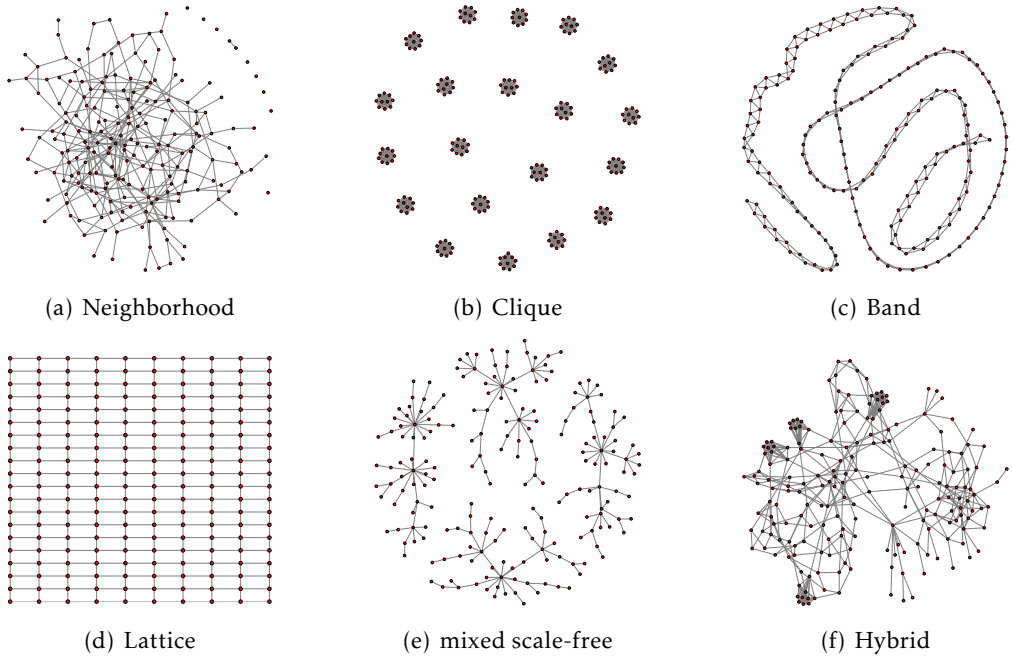
• **Hybrid**. The nodes are evenly partitioned in to 5 groups, named S_1-S_5. The subgraph of S_1 is a neighborhood graph with $\phi = 25$; The subgraph of S_2 is a clique graph with $g = 4$; The subgraph of S_3 is a band graph with $g = 2$; The subgraph of S_4 is a lattice graph with $g = 10$; The subgraph of S_5 is a scale-free graph. In addition, we set an edge between a node in S_k and a node in S_{k+1} with probability 0.01, independently of the other edges for $k = 1,...,4$.

Recall that $\mathbf{E}^*$ denotes the binary adjacency matrix, we calculate

$$\mathbf{\Sigma}^* = \mathcal{C}_2\{(\mathbf{E}^* + (0.5 - \Lambda_{\min}(\mathbf{E}^*)) \cdot \mathbf{I}_{200})^{-1}\},$$

where $\mathcal{C}_2$ is the rescaling operator that converts a covariance matrix to the corresponding correlation matrix. We then generate 100 observations from the Gaussian distribution with covariance

matrix $(\mathbf{\Omega}^*)^{-1}$. We further adopt the power function $g(t) = t^5$ to convert the Gaussian data to the nonparanormal data.



(a) Neighborhood (b) Clique (c) Band

(d) Lattice (e) mixed scale-free (f) Hybrid

Figure 4: Six different graph patterns in the simulation studies.

We use an ROC curve to evaluate the graph recovery performance. Since $d > n$, we cannot obtain the solution paths for the full range of sparsity levels, therefore we restrict the range of false positive rates to be from 0 to 0.1 for computational convenience. For the proposed accelerated proximal gradient algorithm, we set the target precision $\varepsilon = 10^{-3}$ and the shrinkage parameter of the backtracking line search $m = 0.5$.

5.1 Positive Semidefiniteness v.s. Indefiniteness

In this subsection, we first demonstrate the effectiveness of the proposed projection method. The empirical performance of our computational algorithm using different smoothing parameters ($\mu = 0.4603$, 0.1455, 0.0460, 0.0146, and 0.0046) are presented in Tables 1 and 2 on all six graphs (averaged over 100 replications with standard errors in parentheses). We evaluate the computational performance based on the objective value $\left\|\widehat{\mathbf{S}} - \widetilde{\mathbf{S}}\right\|_\infty$ and the estimation error $\left\|\Sigma^* - \widetilde{\mathbf{S}}\right\|_\infty$. We see that smaller μ's attain smaller objective values because of smaller approximation errors.

However, we see that changing μ does not make much difference in the estimation error. In terms of the computational cost, $\mu = 0.0046 \approx 0.002\sqrt{\log d/n}$ is up to 24 times slower than $\mu = 0.4603 \approx 2\sqrt{\log d/n}$. Therefore a reasonably large μ can greatly reduce computational burden with almost no loss of statistical efficiency. Moreover, we also find that our projection method not only guarantees the positive semidefiniteness but also attains smaller estimation error than the

original transformed Kendall's tau matrix.

Table 1: Quantitive comparison between the proposed projection method and transformed Kendall's tau estimator on the neighborhood, clique, and band graphs. "Kendall" denotes the transformed Kendall's tau estimator. Timing results are evaluated in seconds.

Neighborhood	$\mu = 0.4603$	$\mu = 0.1455$	$\mu = 0.0460$	$\mu = 0.0146$	$\mu = 0.0046$	Kendall
Timing	0.7542 (0.0372)	1.3353 (0.0521)	3.3128 (0.1173)	6.4969 (0.3719)	15.395 (2.1716)	N.A. (N.A.)
Obj. Val.	0.0102 (0.0001)	0.0094 (0.0001)	0.0086 (0.0001)	0.0085 (0.0001)	0.0082 (0.0017)	N.A. (N.A.)
Est. Err.	0.4184 (0.0234)	0.4183 (0.0236)	0.4186 (0.0239)	0.4186 (0.0240)	0.4183 (0.0243)	0.4250 (0.0248)
Clique	$\mu = 0.4603$	$\mu = 0.1455$	$\mu = 0.0460$	$\mu = 0.0146$	$\mu = 0.0046$	Kendall
Timing	0.7359 (0.0301)	1.3650 (0.0449)	3.3983 (0.1349)	6.2133 (0.1888)	15.872 (1.4772)	N.A. (N.A.)
Obj. Val.	0.0103 (0.0002)	0.0094 (0.0001)	0.0085 (0.0001)	0.0082 (0.0001)	0.0079 (0.0003)	N.A. (N.A.)
Est. Err.	0.4170 (0.0271)	0.4168 (0.0268)	0.4170 (0.0268)	0.4171 (0.0268)	0.4174 (0.0268)	0.4245 (0.0271)
Band	$\mu = 0.4603$	$\mu = 0.1455$	$\mu = 0.0460$	$\mu = 0.0146$	$\mu = 0.0046$	Kendall
Timing	0.7110 (0.0200)	1.2955 (0.0342)	3.2849 (0.0908)	6.4094 (0.3692)	16.6467 (2.2016)	N.A. (N.A.)
Obj. Val.	0.0104 (0.0002)	0.0094 (0.0001)	0.0086 (0.0001)	0.0082 (0.0001)	0.0080 (0.0001)	N.A. (N.A.)
Est. Err.	0.4311 (0.0319)	0.4309 (0.0318)	0.4309 (0.0317)	0.4310 (0.0317)	0.4308 (0.0312)	0.4832 (0.0316)

We then compare the graph recovery performance of the projection method with the original transformed Kendall's tau estimator. Figure 5 shows the average ROC curves over 100 replications[1]. Since ROC curves corresponding to different μ's are almost identical, we only present the ROC curves corresponding to $\mu = 0.4603$, which is of our main interest. We see that the projection method achieves better performance than the transformed Kendall's tau estimator in graph

[1]The ROC curves from different replications are first aligned by regularization parameters. The averaged ROC curve shows the false positive and true positive rate averaged over all replications w.r.t. each regularization parameter.

Table 2: Quantitive comparison between the proposed projection method and transformed Kendall's tau estimator on the lattice, mixed scale-free, and hybrid graphs. "Kendall" denotes the transformed Kendall's tau estimator. Timing results are evaluated in second.

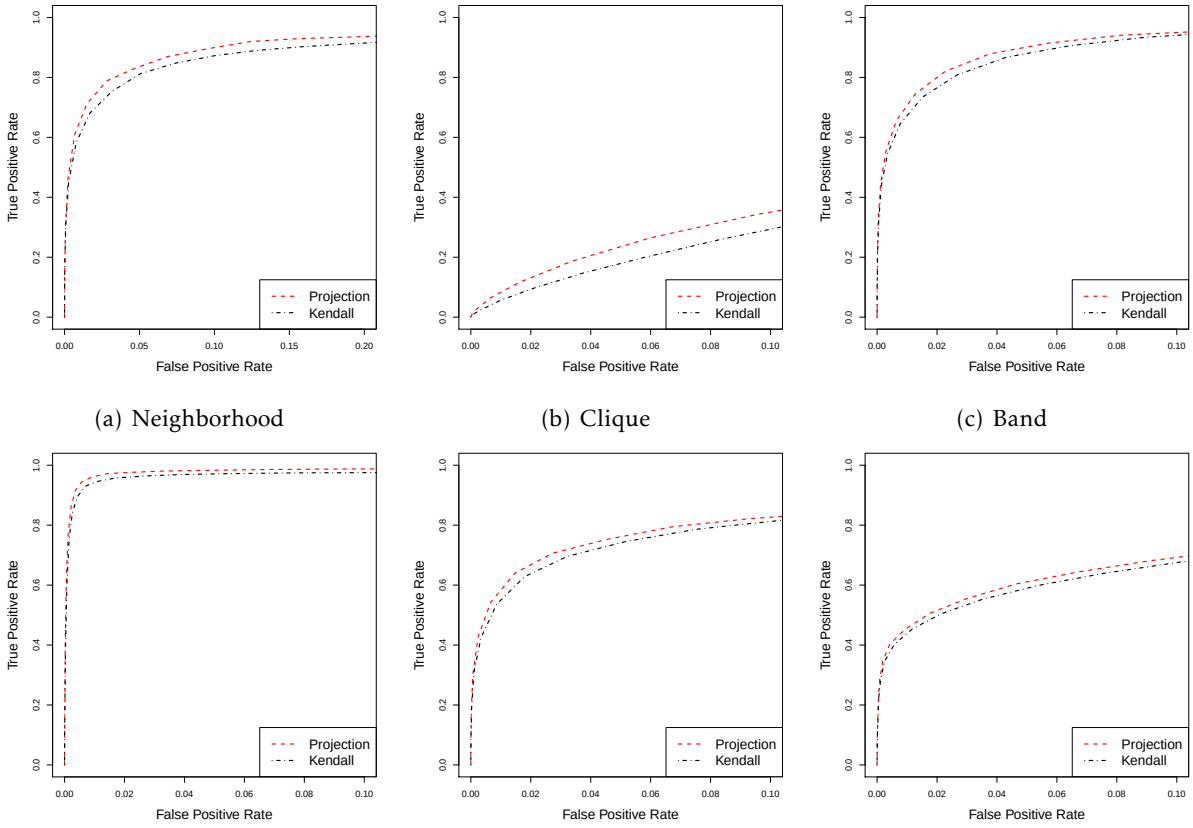
Lattice	$\mu = 0.4603$	$\mu = 0.1455$	$\mu = 0.0460$	$\mu = 0.0146$	$\mu = 0.0046$	Kendall
Timing	0.8290 (0.1110)	1.3592 (0.0527)	3.2477 (0.0679)	6.0479 (0.1312)	14.516 (3.8660)	N.A. (N.A.)
Obj. Val.	0.0104 (0.0002)	0.0094 (0.0001)	0.0086 (0.0001)	0.0084 (0.0001)	0.0082 (0.0011)	N.A. (N.A.)
Est. Err.	0.4098 (0.0185)	0.4098 (0.0185)	0.4099 (0.0181)	0.4098 (0.0180)	0.4100 (0.0181)	0.4176 (0.0180)

Mixed Scale-free	$\mu = 0.4603$	$\mu = 0.1455$	$\mu = 0.0460$	$\mu = 0.0146$	$\mu = 0.0046$	Kendall
Timing	0.7045 (0.0833)	1.2641 (0.0920)	3.3132 (0.1382)	6.3076 (0.1331)	16.7927 (1.8876)	N.A. (N.A.)
Obj. Val.	0.0103 (0.0002)	0.0094 (0.0001)	0.0085 (0.0001)	0.0082 (0.0002)	0.0082 (0.0006)	N.A. (N.A.)
Est. Err.	0.4224 (0.0223)	0.4228 (0.0221)	0.4230 (0.0221)	0.4231 (0.0221)	0.4231 (0.0220)	0.4310 (0.0221)

Hybrid	$\mu = 0.4603$	$\mu = 0.1455$	$\mu = 0.0460$	$\mu = 0.0146$	$\mu = 0.0046$	Kendall
Timing	0.7939 (0.0791)	1.4681 (0.1459)	3.7757 (0.1722)	6.2499 (0.3817)	14.2512 (2.2610)	N.A. (N.A.)
Obj. Val.	0.0104 (0.0002)	0.0094 (0.0001)	0.0085 (0.0001)	0.0085 (0.0006)	0.0081 (0.0006)	N.A. (N.A.)
Est. Err.	0.4113 (0.0223)	0.4108 (0.0221)	0.4112 (0.0232)	0.4112 (0.0232)	0.4111 (0.0232)	0.4187 (0.0234)

recovery for all six graphs.

In summary, these simulation results show that the projection method provides a computational tractable solution and achieves better graph recovery performance than the indefinite transformed Kendall's tau estimator.



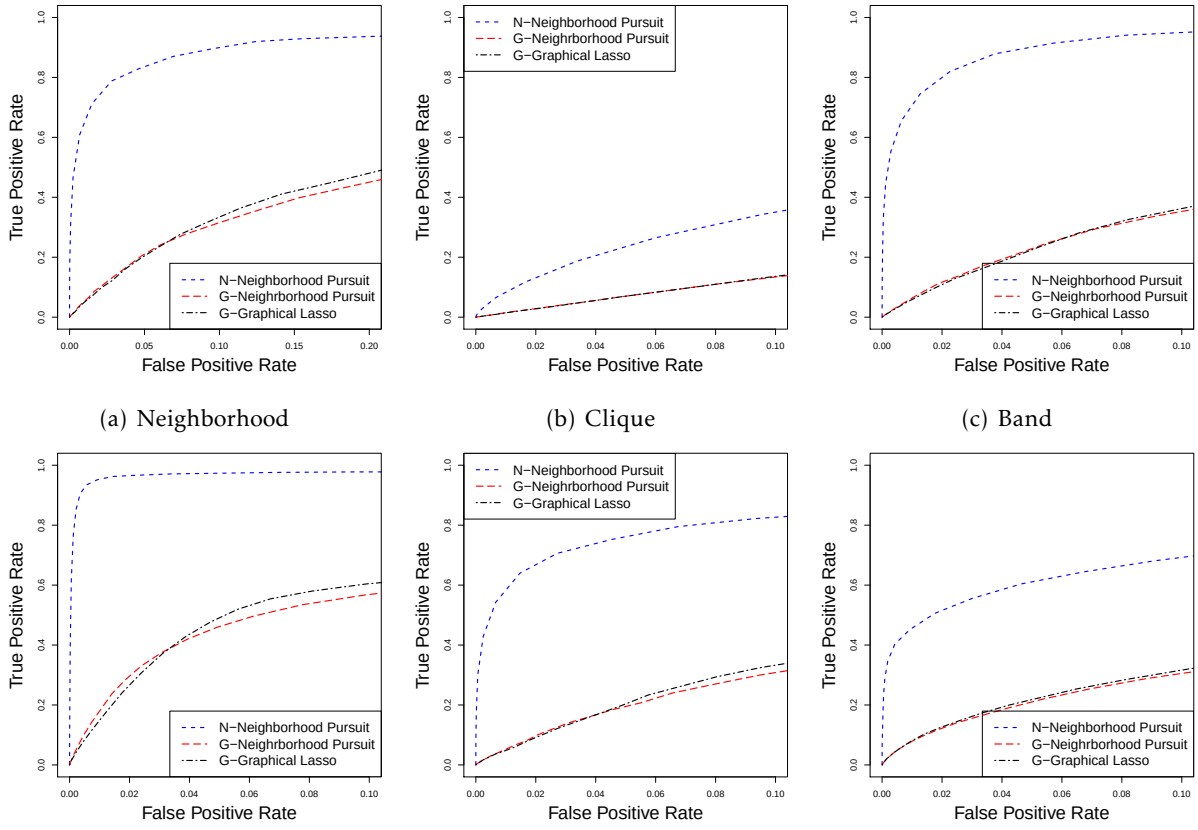
(a) Neighborhood (b) Clique (c) Band

(d) Lattice (e) Mixed Scale-free (f) Hybrid

Figure 5: Average ROC curves of the neighborhood pursuit when combing with different correlation estimators. "Kendall" represents the transformed Kendall's tau estimator, and "Projection" represents the projection method. We see that the proposed projection method achieves better graph recovery performance than the transformed Kendall's tau estimator for all the six graphs.

5.2 Nonparanormal Neighborhood Pursuit v.s. Gaussian Neighborhood Pursuit

This subsection is similar to the numerical studies in Liu et al. (2012), and we compare our proposed method with the Gaussian neighborhood pursuit, which directly combines the Pearson correlation estimator with the neighborhood pursuit and graphical lasso approaches. The main difference is that our experiment is conducted under the setting $d > n$. The smoothing parameter μ is chosen to be 0.4603. The average ROC curves over 100 replications are presented in Figure 6. As can be seen, the nonparanormal neighborhood pursuit outperforms the Gaussian neighborhood pursuit and Gaussian graphical lasso throughout all the 6 graphs.



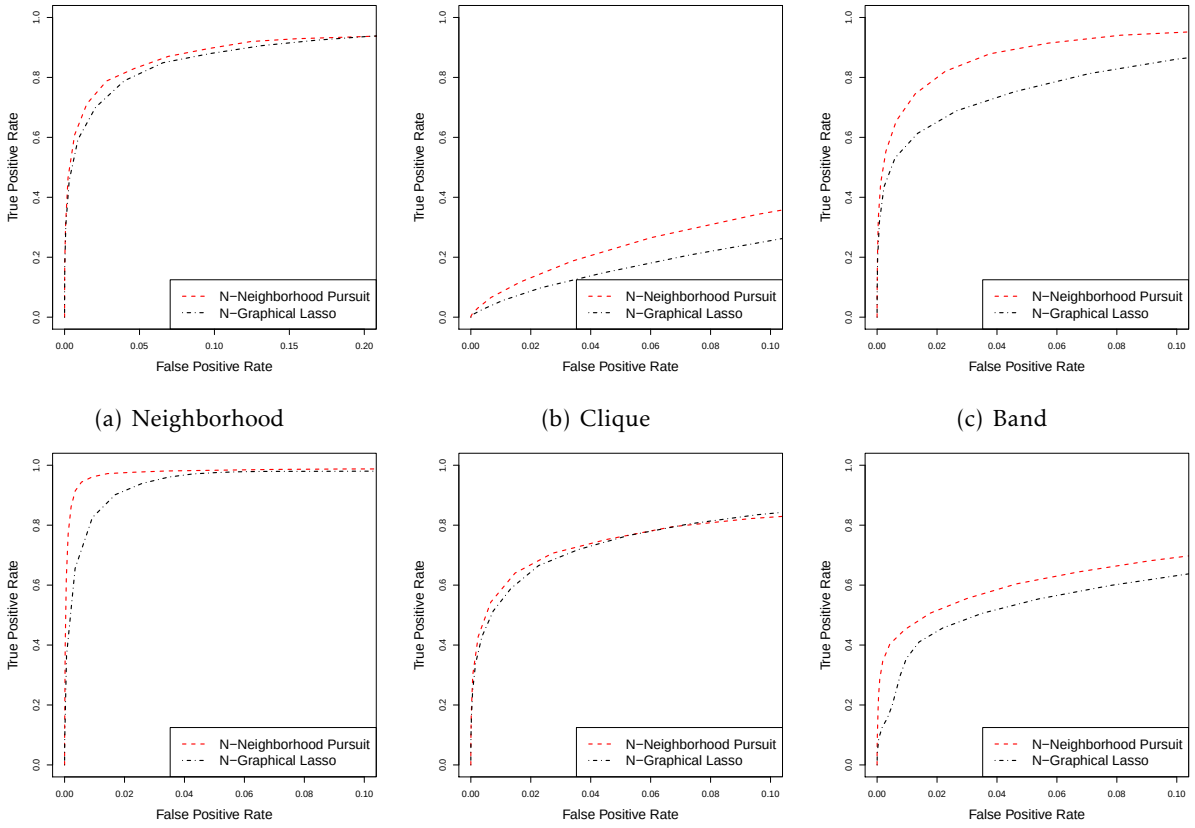
(a) Neighborhood (b) Clique (c) Band

(d) Lattice (e) Mixed Scale-free (f) Hybrid

Figure 6: Average ROC curves of the nonparanormal neighborhood pursuit, Gaussian neighborhood pursuit, and Gaussian graphical lasso. "N-Neighborhood pursuit" represents our proposed nonparanormal neighborhood pursuit. "G-Neighborhood pursuit" represents Gaussian neighborhood pursuit, which combines the neighborhood pursuit and the Pearson correlation estimator. "G-Graphical Lasso" represents Gaussian neighborhood pursuit, which combines the graphical lasso and the Pearson correlation estimator. We see that the nonparanormal neighborhood pursuit outperforms the Gaussian neighborhood pursuit and Gaussian graphical lasso for all the six graphs.

5.3 Nonparanormal Neighborhood Pursuit v.s. Nonparanormal Graphical Lasso

In this subsection, we compare the proposed method with the nonparanormal graphical lasso. The smoothing parameter μ is chosen to be 0.4603. The average ROC curves over 100 replications are presented in Figure 7. We see that the nonparanormal neighborhood pursuit achieves better graph recovery performance than the nonparanormal graphical lasso for all the 6 graphs.



(a) Neighborhood (b) Clique (c) Band

(d) Lattice (e) Mixed Scale-free (f) Hybrid

Figure 7: Average ROC curves of the nonparanormal neighborhood pursuit and nonparanormal graphical lasso. "N-Neighborhood Pursuit" represents the nonparanormal neighborhood pursuit. "N-Graphical Lasso" represents the nonparanormal graphical lasso. We see that the nonparanormal neighborhood pursuit achieves better graph recovery performance than the nonparanormal graphical lasso for all the six graphs.

6 Data Analysis

We present three real data examples. Throughout this subsection, Gaussian graphs are obtained by combining the neighborhood pursuit with the Pearson correlation estimator, while nonparanormal graphs are obtained by the nonparanormal neighborhood pursuit.

We use the following stability graph estimator (Meinshausen and Bühlmann, 2010; Liu et al., 2010) to conduct graph selection:

(1) Calculate the solution path using all samples and choose the regularization parameter at sparsity level θ;

(2) Randomly select $\xi \times 100\%$ of all samples without replacement, and estimate the graph using the selected samples with the regularization parameter chosen in (1);

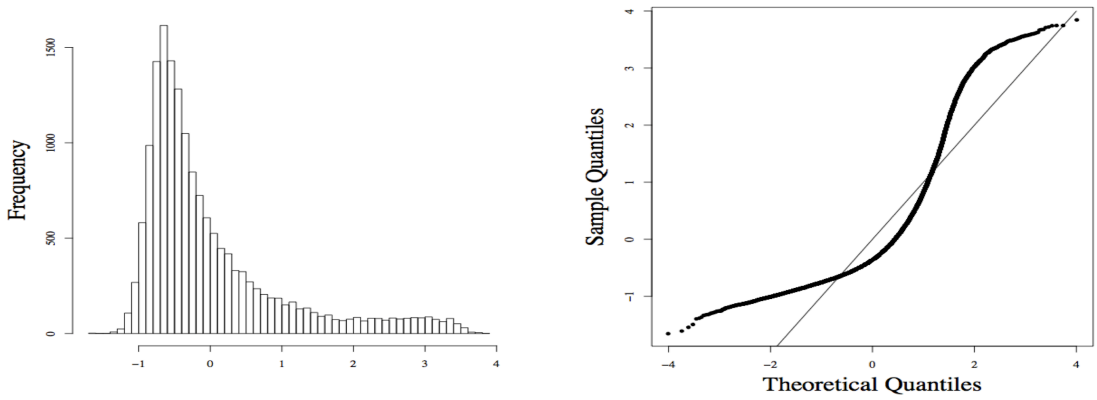
(3) Repeat (2) for 500 times and retain edges that appear with frequency no less than 95%.

We select (θ, ξ) based on two criteria: (1) The obtained graphs should be sparse to ease visualization, interpretation, and computation. (2) The obtained results should be stable. Thus by manually tuning the regularization parameter over a refined grid, we eventually set (θ, ξ) as $(0.04, 0.1)$, $(0.1, 0.5)$, and $(0.10, 0.75)$ respectively for the topic modeling, stock market, and arabidopsis datasets.

6.1 Topic Graph

The topic graph is originally used in Blei and Lafferty (2007) to illustrate the effectiveness of the correlated topic modeling for extracting K "topics" that occur in a collection of documents (corpus). We first estimate the topic proportion for each document. The topic proportion of each document is represented in a K-dimensional simplex. The whole corpus used by Blei and Lafferty (2007) contains 16,351 documents with 19,088 unique terms. Blei and Lafferty (2007) set $K = 100$ and fit a topic model to the articles published in *Science* from 1990 to 1999. Thus each document is represented by a 100-dimensional vector, with each entry corresponding to one of the 100 topics.

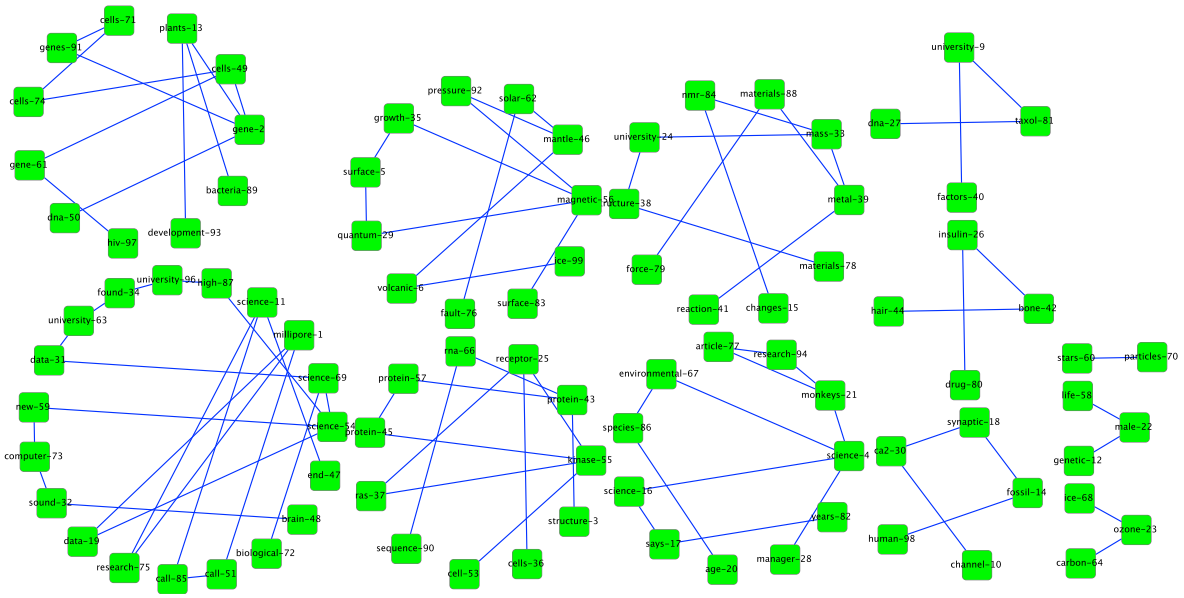
Here we are interested in visualizing the relationship among the topics using the following topic graph: The nodes represent individual topics and neighboring nodes represent highly related topics. In Blei and Lafferty (2007), the topic proportion is assumed to be approximately normal after the log-transformation. To obtain the topic graph, they calculate the Pearson correlation matrix of 100 topics, and plug it into the neighborhood pursuit. When we perform the Kolmogorov-Smirnov test for each topic, however, we find that some of them strongly violate the normality assumption (e.g. Figure 8 shows the histogram and normal qq plot of Topic 4). This motivates our choice of the nonparanormal neighborhood pursuit approach.



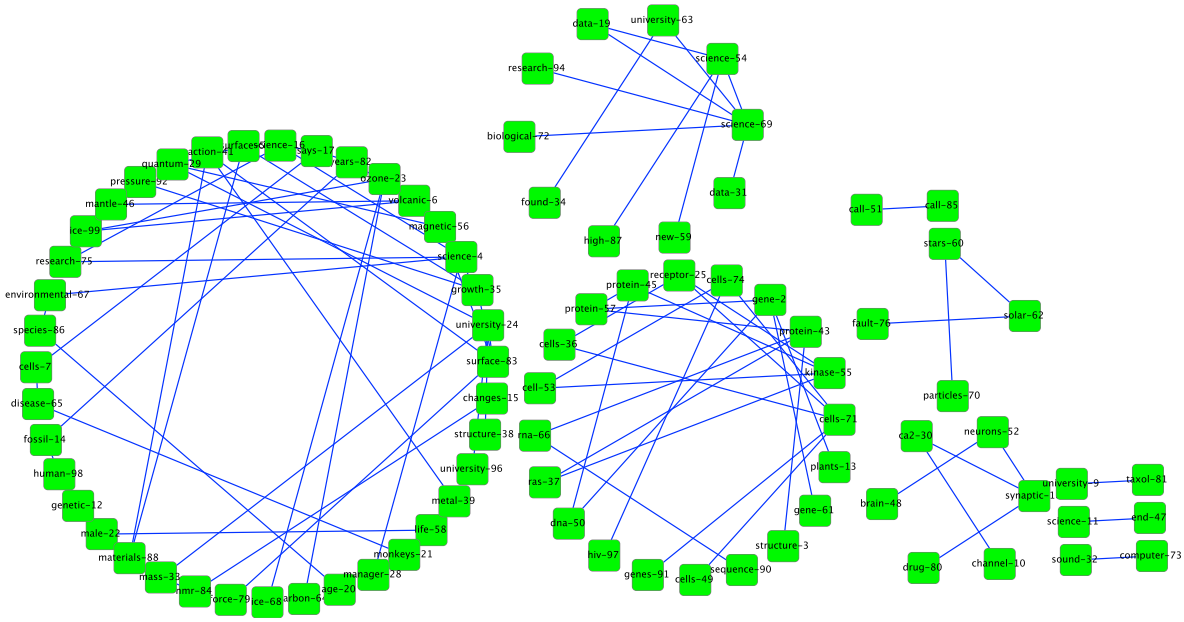
(a) Histogram (b) Normal qq plot

Figure 8: Both the histogram and normal qq plot show significant violation of the normality.

The estimated topic graphs are shown in Figures 9, where the clustering information can be



(a) Nonparanormal Graph



(b) Gaussian Graph

Figure 9: Two topic graphs estimated using the nonparanormal neighborhood pursuit and Gaussian neighborhood pursuit. The nonparanormal graph contains 6 mid-size modules and 6 small modules, while the Gaussian graph contains 1 large module, 2 mid-size modules, and 6 small modules.

read directly from the graphs[2]. The nonparanormal graph contains 6 mid-size modules and 6

[2]Here each topic is labelled with the most frequent word and an index. See

small modules, while the Gaussian graph contains 1 large module, 2 mid-size modules, and 6 small modules. We see that the refined structures discovered by the nonparanormal approach clearly improves the interpretability of the graph. Here we provide a few examples:

(1) Topics closely related to the climate change in Antarctica, such as "ice-68", "ozone-23", and "carbon-64", are clustered in the same module;

(2) Topics closely related to the environmental ecology, such as "monkey-21", "science-4", "species-86", and "environmental-67", are clustered in the same module;

(3) Topics closely related to modern physics, such as "quantum-29", "magnetic-55", "pressure-92", and "solar-62", are clustered in the same module;

(4) Topics closely related to the material mechanics, such as "structure-38", "material-78", "force-79", "metal-39", and "reaction-41", are clustered in the same module.

In contrast, we see that the Gaussian graph mixes all these topics together and clusters them into a large module.

Moreover, with a subsampling ratio of 0.1, the sample size (n = 1,635) is much larger than the dimension (d = 100). We find that all transformed Kendall's tau estimates are positive definite. Thus the proposed projection method is not required.

6.2 S&P 500 Stock Market Graph

We acquire closing prices of all S&P 500 stocks for all the days when the market was open between January 1, 2003 and January 1, 2005. It results in 504 samples of the 452 stocks. The dataset is transformed by calculating the log-ratio of the price at time t to price at time $t-1$, and further standardized by mean zero and variance one. We plot the data points for the first 100 stocks in Figure 10. We highlight a data point in red if its absolute value is greater than 3. We can see that a large number of potential outliers exist. They may affect the quality of the estimated graph.

Since the transformed Kendall's tau estimator is rank-based, it is more robust to outliers than the Pearson correlation estimator.

These 452 stocks belong to 10 different Global Industry Classification Standard (GICS) sectors. We present the obtained graphs in Figure 11. Each stock is represented by a node, which is colored according to its GICS sector. We see that stocks from the same GICS sectors show the tendency to be clustered with each other. We highlight several densely connected modules in the nonparanormal graph, and by color coding we see that the nodes in the same dense module belong to the same sector of the market. In contrast, these modules are shown to be sparse in the Gaussian graph. Especially for the blue nodes, many of them are observed as isolated nodes, which means the stocks they represent are (both marginally and conditionally) independent to the others. This is contrary to common beliefs. Overall, we see that the nonparanormal graph has

http://www.cs.cmu.edu/ lemur/science/topics.html for more details about topic summaries.

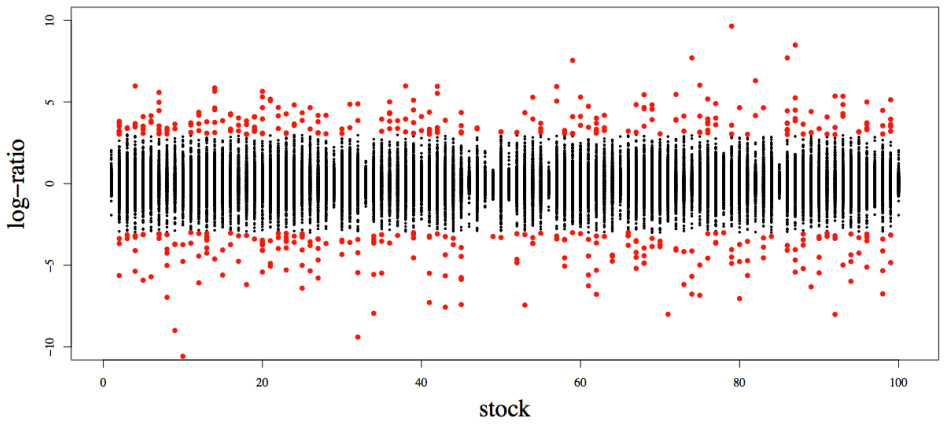
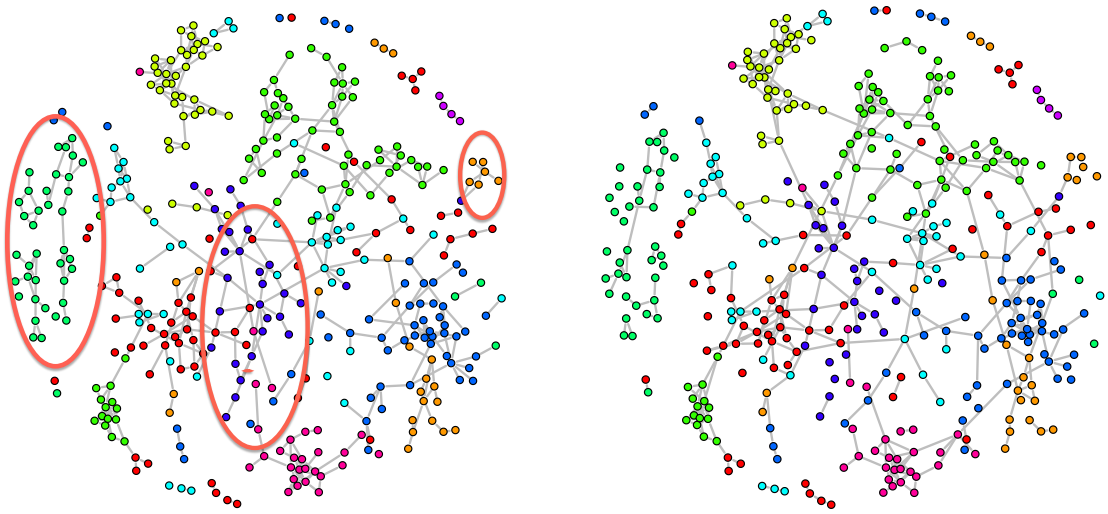


Figure 10: Stock Market Dataset. We can see a large amounted of the outliers (Red dots). Their existence may affect the quality of the estimated graph.



(a) Nonparanormal Graph (b) Gaussian Graph

Figure 11: Stock Graphs. Several densely connected modules are found in the nonparanormal graph, while they are sparser in the corresponding Gaussian graph. The color shows all nodes in this module belong to the same sector of the market.

more refined structures than Gaussian graph such that more meaningful relationships could be revealed.

Moreover, with a subsampling ratio of 0.5, the sample size ($n = 252$) is smaller than the dimension ($d = 452$) and the transformed Kendall's tau estimator is indefinite. By the positive semidefinite projection, we can exploit the convexity of the problem and obtain a high quality graph estimator. The smoothing parameter for the projection method $\mu = 0.3115 \approx 2\sqrt{\log d/n}$ is the same as our previous simulations.

6.3 Gene Graph

This dataset includes 118 gene expression profiles from arabidposis thaliana that originally appeared in Wille et al. (2004). Our analysis focuses on gene expression from 39 genes involved in two isoprenoid pathways: 16 from the mevalonate (MVA) pathway are located in the cytoplasm, 18 from the plastidial (MEP) pathway are located in the chloroplast, and 5 are located in the mitochondria. While the two pathways generally operate independently, crosstalk is known to happen (Wille et al., 2004). Our goal is to recover the gene regulatory graph (network), with special interest in crosstalk.

Though the estimated graphs shown in Figure 12 are similar, there exist subtle differences with potentially interesting implications. Both highlight three tightly connected clusters. With the exception of AACT1 and HMGR1 which are part of the MVA pathway (yellow), most of the genes in the cluster to the left are from the MEP pathway (green). The only gene that changes clusters in the two estimated graphs is GGPPS12; This gene, which is also a member of the MEP pathway, is correctly positioned in the nonparanormal graph. MECPS is clearly a hub gene for this pathway.

Prior investigation suggests that the connections from genes AACT1 and HMGR2 to hub gene MECPS indicate primary sources of the crosstalk between the MEP and MVA pathways and these edges are present in both graphs. We highlight the edge connecting HMGR1 and MECPS, which appears only in the nonparanormal graph. Our analysis suggests that this link constitutes another possible crosstalk between these two pathways. Further investigation of the gene expression levels reveals that the distribution of MECPS is strongly non-Gaussian (Figure 13 shows the histogram and normal qq plot of "MECPS"). This lack of normality might explain why the Gaussian neighborhood pursuit does not detect the link between HMGR1 and MECPS.

Moreover, with a subsampling ratio of 0.75, the sample size ($n = 88$) is much larger than the dimension ($d = 39$). We find that all transformed Kendall's tau estimates are positive definite. Thus the proposed projection method is not required.

7 Discussion and Conclusion

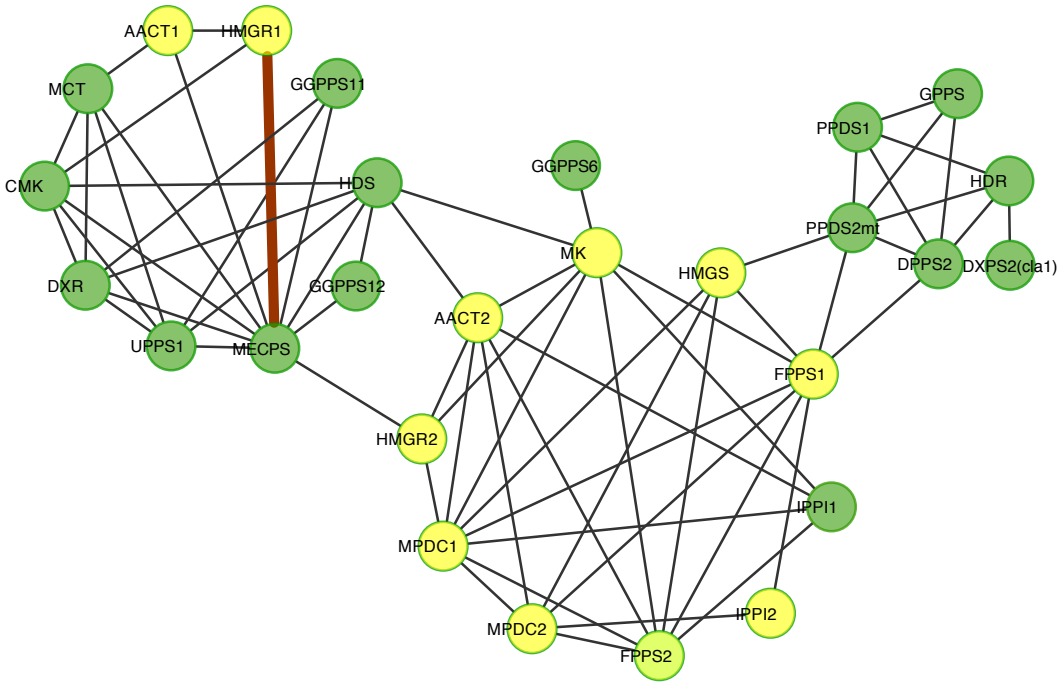
In addition to the projection method, there are two alternative heuristic approaches to obtain a positive semidefinite replacement of $\widehat{\mathbf{S}}$. One approach is based on the following optimization problem (Rousseeuw and Molenberghs, 1993),

$$\widetilde{\mathbf{S}} = \underset{\mathbf{S}}{\operatorname{argmin}} \|\widehat{\mathbf{S}} - \mathbf{S}\|_{\mathrm{F}}^2 \text{ subject to } \mathbf{S} \succeq 0. \tag{7.1}$$

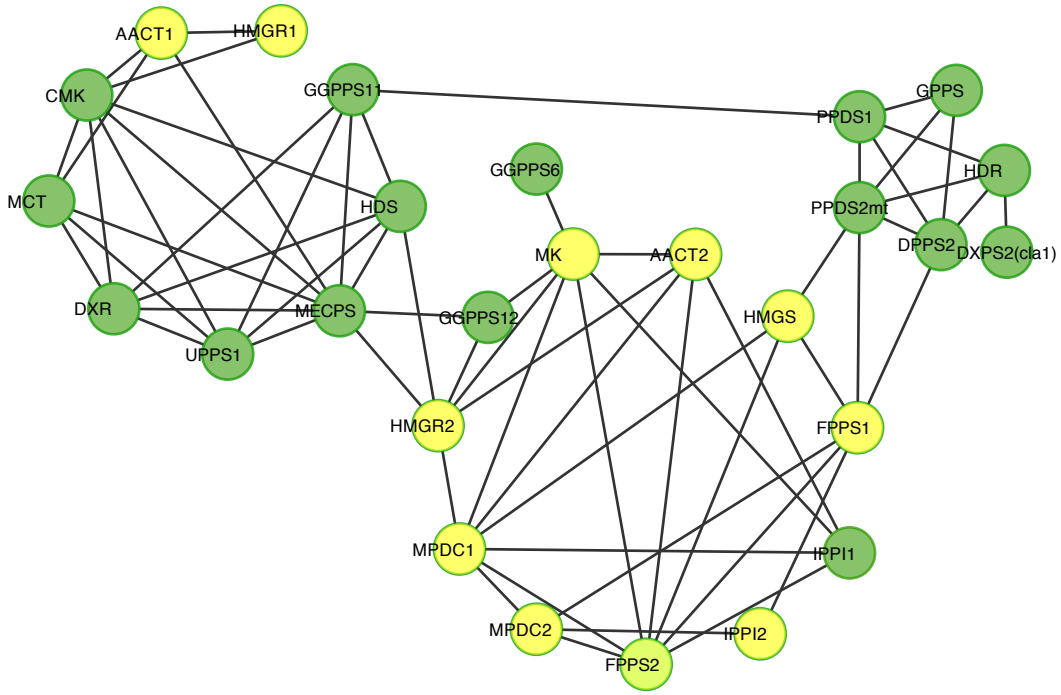
By Lemma 3.2, (7.1) has a closed form solution as

$$\widetilde{\mathbf{S}} = \sum_{j=1}^{d} \max\{\sigma_j, 0\} \cdot \boldsymbol{v}_j \boldsymbol{v}_j^T, \tag{7.2}$$

where σ_j's are eigenvalues of $\widehat{\mathbf{S}}$, and $\boldsymbol{v}_j$'s are the corresponding eigenvectors.

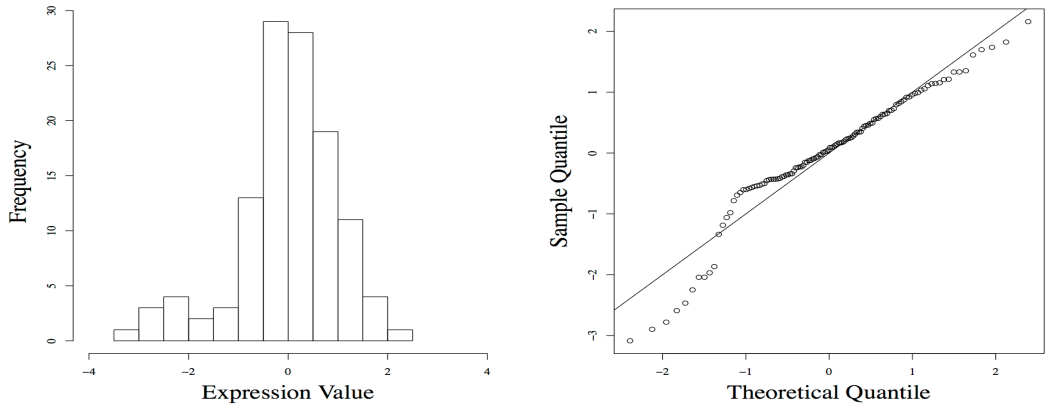


(a) Nonparanormal Graph



(b) Gaussian Graph

Figure 12: Gene regulatory graphs with isolated notes omitted. Genes belonging to the MVA and MEP pathways are colored in yellow and green, respectively. GGPPS12 is correctly positioned by the nonparanormal neighborhood pursuit. Our analysis also suggests that a link between "HMGR1" and "MECPS" constitutes another key point of crosstalk between these two pathways.



(a) Histogram (b) Normal qq plot

Figure 13: Both the histogram and normal qq plot show violation of the normality.

Remark 7.1. *Similar to our projection method,* (7.1) *also projects* $\widehat{\mathbf{S}}$ *onto the cone of all positive semidefinite matrices, but with respect to the Frobenius norm. The theoretical property of such a "truncation" estimator is not clear.*

The other approach is directly adding a positive value to all diagonal entries of $\widehat{\mathbf{S}}$ as follows,

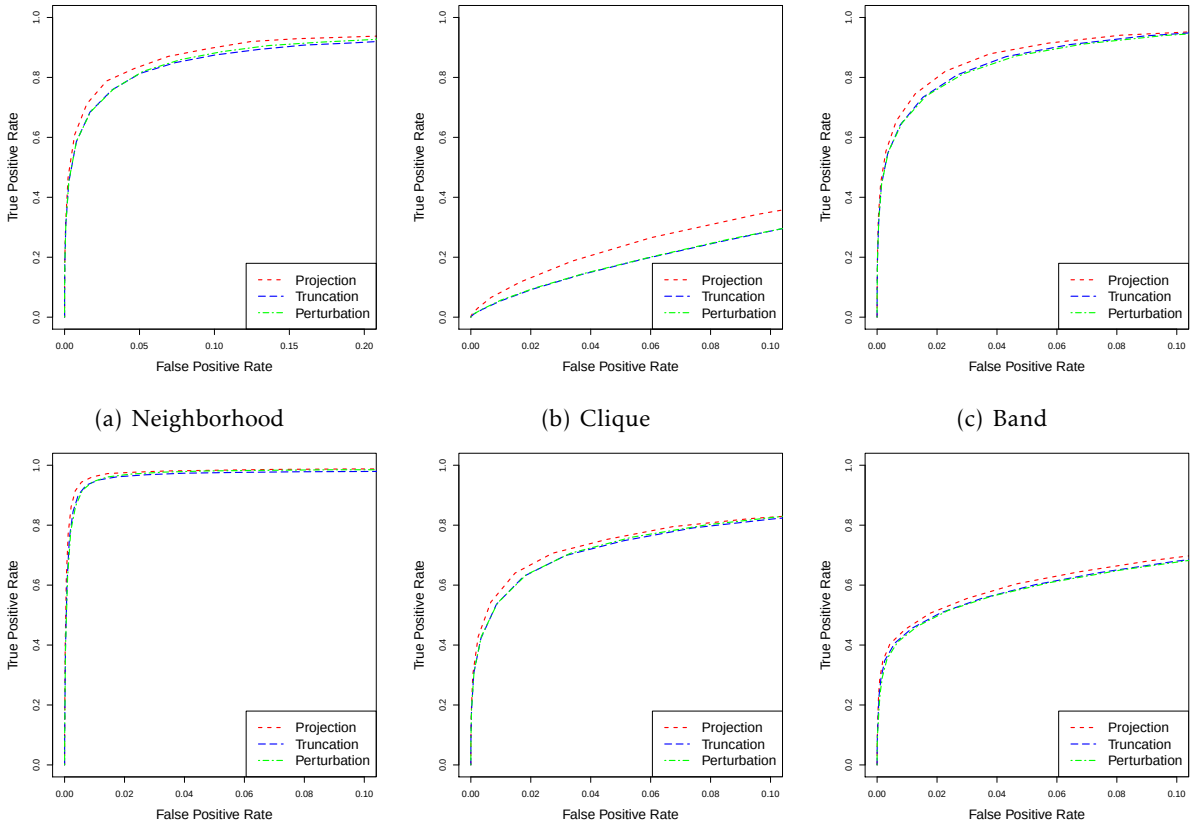
$$\widetilde{\mathbf{S}} = \widehat{\mathbf{S}} + |\min_j \sigma_j| \cdot \mathbf{I}. \tag{7.3}$$

Such a "perturbation" estimator has been used in many classical statistical methods such as the regularized linear discriminant analysis (Guo et al., 2007) and the ridge regression (Hoerl and Kennard, 1970). However, no theoretical analysis has been established for this approach.

We compare the proposed projection method in §3.10 with the above two estimators using the same settings as our simulations in §5. Figure 14 shows the average ROC curves over 100 replications. We see that the projection method achieves better graph recovery performance than the competitors for all the six graphs.

Two closely related methods are copula discriminant analysis (Han et al., 2013), and semiparametric sparse inverse column operator (Zhao and Liu, 2013). Similar to the nonparanormal neighborhood pursuit, both methods are formulated as ℓ_1 regularized quadratic programs. When the transformed Kendall's tau matrix is indefinite, their computational formulations become nonconvex. Thus the proposed projection approach can also benefit these two methods to achieve better empirical performance with theoretical guarantees. More details can be found in Han et al. (2013); Zhao and Liu (2013).

In this paper, we propose a projection method to handle the possible indefiniteness of the transformed Kendall's tau matrix in semiparametric graph estimation. We derive a computationally tractable optimization algorithm to secure the positive semidefiniteness of the estimated correlation matrix in high dimensions. The theoretical study, combined with the proposed projection method, shows that the neighborhood pursuit achieves graph estimation consistency for



(a) Neighborhood (b) Clique (c) Band

(d) Lattice (e) Mixed Scale-free (f) Hybrid

Figure 14: Average ROC curves of the neighborhood pursuit when combing with different correlation estimators. "Projection" represents the projection method. "Truncation" represents the estimator defined in (7.2). "Perturbation" represents the estimator defined in (7.3). We see that the projection method achieves better graph recovery performance than the other two estimators for all the six graphs.

nonparanormal models under suitable conditions. More importantly, this nonparanormal graph estimation problem illustrates a fundamental tradeoff between statistics and computation. Our result shows that it is possible to simultaneously gain robustness of estimation and modeling flexibility without losing good computational structures such as convexity and smoothness. The proposed methodology is theoretically justifiable and applicable to a wide range of problems.

A Proof of Theorem 4.1

Proof. We define

$$\widehat{\mathbf{U}} = \underset{\|\mathbf{U}\|_1 \leq 1}{\operatorname{argmax}} \langle \mathbf{U}, \mathbf{A} \rangle. \tag{A.1}$$

This implies that

$$\left\|\mathbf{A}\right\|_{\infty} - \frac{\mu}{2} \overset{\text{(i)}}{\leq} \langle \widehat{\mathbf{U}}, \mathbf{A} \rangle - \frac{\mu}{2} \|\widehat{\mathbf{U}}\|_{\mathrm{F}}^2 \overset{\text{(ii)}}{\leq} \left\|\mathbf{A}\right\|_{\infty}^{\mu} \leq \langle \widetilde{\mathbf{U}}, \mathbf{A} \rangle \overset{\text{(iii)}}{\leq} \left\|\mathbf{A}\right\|_{\infty}, \tag{A.2}$$

where (i) comes from the fact $\|\widehat{\mathbf{U}}\|_{\mathrm{F}}^2 \leq \left\|\widehat{\mathbf{U}}\right\|_1^2$, (ii) comes from the fact that $\left\|\mathbf{A}\right\|_{\infty}^{\mu}$ is obtained by maximizing (3.7), and (iii) comes from the fact that $\widehat{\mathbf{U}}$ is the maximizer to (A.1). A direct result of (A.2) is

$$\left\|\widetilde{\mathbf{S}} - \Sigma^*\right\|_{\infty} \leq \left\|\widetilde{\mathbf{S}} - \Sigma^*\right\|_{\infty}^{\mu} + \frac{\mu}{2}. \tag{A.3}$$

Since $\overline{\mathbf{S}}$ is a feasible solution to (3.10), we have

$$\left\|\widetilde{\mathbf{S}} - \Sigma^*\right\|_{\infty}^{\mu} \leq \left\|\overline{\mathbf{S}} - \Sigma^*\right\|_{\infty}^{\mu}. \tag{A.4}$$

Recall (3.5), we further have

$$\left\|\widetilde{\mathbf{S}} - \Sigma^*\right\|_{\infty}^{\mu} \leq \left\|\overline{\mathbf{S}} - \Sigma^*\right\|_{\infty} \leq 2\left\|\widehat{\mathbf{S}} - \Sigma^*\right\|_{\infty}. \tag{A.5}$$

Combing (A.3) and (A.5), we further have

$$\left\|\widetilde{\mathbf{S}} - \Sigma^*\right\|_{\infty} \leq 2\left\|\widehat{\mathbf{S}} - \Sigma^*\right\|_{\infty} + \frac{\mu}{2}. \tag{A.6}$$

Since $\mu = \kappa_2\sqrt{\log d/n}$, combining (A.6) and Lemma 2.2, we eventually have

$$\mathbb{P}\left(\left\|\widetilde{\mathbf{S}} - \Sigma^*\right\|_{\infty} \leq \kappa_3\sqrt{\frac{\log d}{n}}\right) \geq \mathbb{P}\left(\left\|\widehat{\mathbf{S}} - \Sigma^*\right\|_{\infty} \leq \kappa_1\sqrt{\frac{\log d}{n}}\right) \geq 1 - \frac{1}{d^3}, \tag{A.7}$$

where $\kappa_3 = 2\kappa_1 + \kappa_2/2$. □

B Proof of Theorem 4.3

Before we proceed with the proof, we need to introduce the following lemmas. Their detailed proofs are provided in Appendices E.1–E.5.

Lemma B.1. *Let $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{d\times d}$ be two symmetric matrices, we have $\|\mathbf{AB}\|_{\infty} \leq \|\mathbf{A}\|_{\infty}\|\mathbf{B}\|_{\infty}$.*

Lemma B.2. *Let $\mathbf{B}, \widehat{\mathbf{B}} \in \mathbb{R}^{d\times d}$ be two invertible symmetric matrices and $\|\mathbf{B} - \widehat{\mathbf{B}}\|_{\infty}\|\mathbf{B}^{-1}\|_{\infty} < \frac{1}{2}$, then we have*

$$\|\widehat{\mathbf{B}}^{-1} - \mathbf{B}^{-1}\|_{\infty} \leq 2\|\mathbf{B}^{-1}\|_{\infty}^2\|\widehat{\mathbf{B}} - \mathbf{B}\|_{\infty}. \tag{B.1}$$

Lemma B.3. *Let $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{d\times d}$ be invertible symmetric matrices with*

$$\|\mathbf{AB}^{-1}\|_{\infty} \leq \alpha, \|\mathbf{B}^{-1}\|_{\infty} \leq \psi \text{ and } \|\mathbf{B} - \widehat{\mathbf{B}}\|_{\infty}\|\mathbf{B}^{-1}\|_{\infty} < \frac{1}{2}. \tag{B.2}$$

We have

$$\|\widehat{\mathbf{A}}\widehat{\mathbf{B}}^{-1} - \mathbf{AB}^{-1}\|_{\infty} \leq 2\psi\|\widehat{\mathbf{A}} - \mathbf{A}\|_{\infty} + 2\alpha\psi\|\widehat{\mathbf{B}} - \mathbf{B}\|_{\infty}. \tag{B.3}$$

Lemma B.4. *Let $\widehat{\mathbf{A}}, \mathbf{A} \in \mathbb{R}^{d\times d}$ be symmetric matrices and $\widehat{\boldsymbol{v}}, \boldsymbol{v} \in \mathbb{R}^d$ be vectors, we have*

$$\|\widehat{\mathbf{A}}\widehat{\boldsymbol{v}} - \mathbf{A}\boldsymbol{v}\|_\infty \leq \|\widehat{\mathbf{A}} - \mathbf{A}\|_\infty \|\widehat{\boldsymbol{v}} - \boldsymbol{v}\|_\infty + \|\widehat{\mathbf{A}} - \mathbf{A}\|_\infty \|\boldsymbol{v}\|_\infty + \|\mathbf{A}\|_\infty \|\widehat{\boldsymbol{v}} - \boldsymbol{v}\|_\infty. \tag{B.4}$$

Lemma B.5. *Assuming that Σ^* satisfies Assumption 2 and $\left\|\left\|\widetilde{\mathbf{S}} - \Sigma^*\right\|\right\|_\infty \leq \kappa_3\sqrt{\log d/n}$ and $s = \max_j \left|I_j\right|$, we have*

$$\|\widetilde{\mathbf{S}}_{I_jI_j} - \Sigma^*_{I_jI_j}\|_\infty \leq \kappa_3 s\sqrt{\frac{\log d}{n}}, \tag{B.5}$$

$$\|\widetilde{\mathbf{S}}_{J_jI_j} - \Sigma^*_{J_jI_j}\|_\infty \leq \kappa_3 s\sqrt{\frac{\log d}{n}}, \tag{B.6}$$

$$\|\widetilde{\mathbf{S}}_{\backslash j,j} - \Sigma^*_{\backslash j,j}\|_\infty \leq \kappa_3\sqrt{\frac{\log d}{n}}. \tag{B.7}$$

Moreover, for large enough n, we have that $\widetilde{\mathbf{S}}_{I_jI_j}$ is invertible and $\Lambda_{\min}(\widetilde{\mathbf{S}}_{I_jI_j}) \geq \delta/2$.

Proof. We adopt a similar strategy in Zhao and Yu (2006); Meinshausen and Bühlmann (2006); Zou (2006); Wainwright (2009); Mai et al. (2012), and all the analysis assumes that the following condition holds,

$$\left\|\left\|\widetilde{\mathbf{S}} - \Sigma^*\right\|\right\|_\infty \leq \kappa_3\sqrt{\frac{\log d}{n}}. \tag{B.8}$$

Thus by Lemma B.5, for large enough n, $\widetilde{\mathbf{S}}_{I_jI_j}$ is invertible and positive definite.

We define the following optimization problem with an auxiliary variable $\widehat{\boldsymbol{\beta}} \in \mathbb{R}^d$,

$$\widehat{\boldsymbol{\beta}}_{I_j} = \underset{\boldsymbol{\beta}_{I_j}}{\operatorname{argmin}}\, \boldsymbol{\beta}_{I_j}^T\widetilde{\mathbf{S}}_{I_jI_j}\boldsymbol{\beta}_{I_j} - 2\widetilde{\mathbf{S}}_{I_j,j}^T\boldsymbol{\beta}_{I_j} + \lambda\|\boldsymbol{\beta}_{I_j}\|_1. \tag{B.9}$$

Since $\widetilde{\mathbf{S}}_{I_jI_j}$ is positive definite, (B.9) is strongly convex and has an unique solution

$$\widehat{\boldsymbol{\beta}}_{I_j} = (\widetilde{\mathbf{S}}_{I_jI_j})^{-1}\left(\widetilde{\mathbf{S}}_{I_j,j} - \frac{\lambda}{2}\zeta_{I_j}\right), \tag{B.10}$$

where ζ_{I_j} is the subgradient of $\|\boldsymbol{\beta}_{I_j}\|_1$ at $\boldsymbol{\beta}_{I_j} = \widehat{\boldsymbol{\beta}}_{I_j}$. We set $\widehat{\boldsymbol{\beta}}_{J_j} = 0$ and $\widehat{\boldsymbol{\beta}}_j = 0$. Thus $\widehat{\boldsymbol{\beta}}$ exactly recovers the neighborhood structure of node j, if $\widehat{\boldsymbol{\beta}}_{I_j}$ does not contain any zero entry.

Now we will show that given (B.8) and large enough n, $\widehat{\boldsymbol{\beta}}$ is the optimal solution to (3.3). We need to verify the optimality conditions of (3.3) as follows:

$$\widetilde{\mathbf{S}}_{I_jI_j}\widehat{\boldsymbol{\beta}}_{I_j} - \widetilde{\mathbf{S}}_{I_j,j} + \frac{\lambda}{2}\zeta_{I_j} = 0, \tag{B.11}$$

$$\widetilde{\mathbf{S}}_{J_jI_j}\widehat{\boldsymbol{\beta}}_{I_j} - \widetilde{\mathbf{S}}_{J_j,j} + \frac{\lambda}{2}\zeta_{J_j} = 0. \tag{B.12}$$

Note that (B.10) already implies (B.11). By plugging (B.10) into (B.12), we have

$$\widetilde{\mathbf{S}}_{J_jI_j}(\widetilde{\mathbf{S}}_{I_jI_j})^{-1}\widetilde{\mathbf{S}}_{I_j,j} - \widetilde{\mathbf{S}}_{J_jI_j}(\widetilde{\mathbf{S}}_{I_jI_j})^{-1}\frac{\lambda}{2}\zeta_{I_j} - \widetilde{\mathbf{S}}_{J_j,j} + \frac{\lambda}{2}\zeta_{J_j} = 0, \tag{B.13}$$

where ζ_{J_j} is the subgradient of $\|\beta_{J_j}\|_1$ at $\beta_{J_j} = 0$. Therefore, we only need to verify

$$\begin{aligned}
&\|\zeta_{J_j}\|_\infty \le 2\lambda^{-1}\|\widetilde{\mathbf{S}}_{J_jI_j}(\widetilde{\mathbf{S}}_{I_jI_j})^{-1}\widetilde{\mathbf{S}}_{I_j,j} - \widetilde{\mathbf{S}}_{J_j,j}\|_\infty + \|\widetilde{\mathbf{S}}_{J_jI_j}(\widetilde{\mathbf{S}}_{I_jI_j})^{-1}\zeta_{I_j}\|_\infty \\
&\le 2\lambda^{-1}\|\widetilde{\mathbf{S}}_{J_jI_j}(\widetilde{\mathbf{S}}_{I_jI_j})^{-1}\widetilde{\mathbf{S}}_{I_j,j} - \widetilde{\mathbf{S}}_{J_j,j}\|_\infty + \|\widetilde{\mathbf{S}}_{J_jI_j}(\widetilde{\mathbf{S}}_{I_jI_j})^{-1} - \Sigma^*_{J_jI_j}(\Sigma^*_{I_jI_j})^{-1}\|_\infty + \alpha, \qquad \text{(B.14)}
\end{aligned}$$

where the last inequality comes from $\|\zeta_{I_j}\|_\infty \le 1$, Assumption 1, and the triangle inequality. It suffices to show that

$$2\lambda^{-1}\|\widetilde{\mathbf{S}}_{J_jI_j}(\widetilde{\mathbf{S}}_{I_jI_j})^{-1}\widetilde{\mathbf{S}}_{I_j,j} - \widetilde{\mathbf{S}}_{J_j,j}\|_\infty + \|\widetilde{\mathbf{S}}_{J_jI_j}(\widetilde{\mathbf{S}}_{I_jI_j})^{-1} - \Sigma^*_{J_jI_j}(\Sigma^*_{I_jI_j})^{-1}\|_\infty < 1-\alpha. \qquad \text{(B.15)}$$

By (3.2) we have $\Sigma^*_{\setminus j,\setminus j}\mathbf{B}^*_{\setminus j,j} = \Sigma^*_{\setminus j,j}$ and $\mathbf{B}^*_{J_j,j} = 0$, we can further rewrite (3.2) as

$$\Sigma^*_{I_j,I_j}\mathbf{B}^*_{I_j,j} = \Sigma^*_{I_j,j} \text{ and } \Sigma^*_{J_j,I_j}\mathbf{B}^*_{I_j,j} = \Sigma^*_{J_j,j}, \qquad \text{(B.16)}$$

which implies that

$$\Sigma^*_{J_jI_j}(\Sigma^*_{I_jI_j})^{-1}\Sigma^*_{I_j,j} = \Sigma^*_{J_j,j}. \qquad \text{(B.17)}$$

Thus we have

$$\begin{aligned}
&\|\widetilde{\mathbf{S}}_{J_jI_j}(\widetilde{\mathbf{S}}_{I_jI_j})^{-1}\widetilde{\mathbf{S}}_{I_j,j} - \widetilde{\mathbf{S}}_{J_j,j}\|_\infty \\
\le&\|\widetilde{\mathbf{S}}_{J_jI_j}(\widetilde{\mathbf{S}}_{I_jI_j})^{-1}\widetilde{\mathbf{S}}_{I_j,j} - \Sigma^*_{J_j,j}\|_\infty + \|\widetilde{\mathbf{S}}_{J_j,j} - \Sigma^*_{J_j,j}\|_\infty \\
\overset{\text{(iv)}}{=}&\|\widetilde{\mathbf{S}}_{J_jI_j}(\widetilde{\mathbf{S}}_{I_jI_j})^{-1}\widetilde{\mathbf{S}}_{I_j,j} - \Sigma^*_{J_jI_j}(\Sigma^*_{I_jI_j})^{-1}\Sigma^*_{I_j,j}\|_\infty + \|\widetilde{\mathbf{S}}_{J_j,j} - \Sigma^*_{J_j,j}\|_\infty \\
\le&\|\widetilde{\mathbf{S}}_{J_jI_j}(\widetilde{\mathbf{S}}_{I_jI_j})^{-1} - \Sigma^*_{J_jI_j}(\Sigma^*_{I_jI_j})^{-1}\|_\infty\|\widetilde{\mathbf{S}}_{I_j,j} - \Sigma^*_{I_j,j}\|_\infty + \|\widetilde{\mathbf{S}}_{J_j,j} - \Sigma^*_{J_j,j}\|_\infty \\
&+ \|\widetilde{\mathbf{S}}_{J_jI_j}(\widetilde{\mathbf{S}}_{I_jI_j})^{-1} - \Sigma^*_{J_jI_j}(\Sigma^*_{I_jI_j})^{-1}\|_\infty\|\Sigma^*_{I_j,j}\|_\infty + \|\Sigma^*_{J_jI_j}(\Sigma^*_{I_jI_j})^{-1}\|_\infty\|\widetilde{\mathbf{S}}_{I_j,j} - \Sigma^*_{I_j,j}\|_\infty \\
\overset{\text{(v)}}{\le}&\|\widetilde{\mathbf{S}}_{J_jI_j}(\widetilde{\mathbf{S}}_{I_jI_j})^{-1} - \Sigma^*_{J_jI_j}(\Sigma^*_{I_jI_j})^{-1}\|_\infty\|\widetilde{\mathbf{S}}_{\setminus j,j} - \Sigma^*_{\setminus j,j}\|_\infty \\
&+ \|\widetilde{\mathbf{S}}_{J_jI_j}(\widetilde{\mathbf{S}}_{I_jI_j})^{-1} - \Sigma^*_{J_jI_j}(\Sigma^*_{I_jI_j})^{-1}\|_\infty + (\alpha+1)\|\widetilde{\mathbf{S}}_{\setminus j,j} - \Sigma^*_{\setminus j,j}\|_\infty, \qquad \text{(B.18)}
\end{aligned}$$

where (iv) comes from (B.17) and (v) comes from Lemma B.4 with $\widehat{\mathbf{A}} = \widetilde{\mathbf{S}}_{J_jI_j}(\widetilde{\mathbf{S}}_{I_jI_j})^{-1}\widetilde{\mathbf{S}}_{I_j,j}$, $\mathbf{A} = \Sigma^*_{J_jI_j}(\Sigma^*_{I_jI_j})^{-1}$, $\widehat{\boldsymbol{v}} = \widetilde{\mathbf{S}}_{I_j,j}$, and $\boldsymbol{v} = \Sigma^*_{I_j,j}$. When $\|\widetilde{\mathbf{S}}_{\setminus j,j} - \Sigma^*_{\setminus j,j}\|_\infty \le 1$, we need to verify

$$\left(1+\frac{4}{\lambda}\right)\|\widetilde{\mathbf{S}}_{J_jI_j}(\widetilde{\mathbf{S}}_{I_jI_j})^{-1} - \Sigma^*_{J_jI_j}(\Sigma^*_{I_jI_j})^{-1}\|_\infty + \frac{2(\alpha+1)}{\lambda}\|\widetilde{\mathbf{S}}_{\setminus j,j} - \Sigma^*_{\setminus j,j}\|_\infty < 1-\alpha. \qquad \text{(B.19)}$$

Then by Lemma B.3 with $\widehat{\mathbf{A}} = \widetilde{\mathbf{S}}_{J_jI_j}$, $\mathbf{A} = \Sigma^*_{J_jI_j}$, $\widehat{\mathbf{B}} = \widetilde{\mathbf{S}}_{I_jI_j}$, and $\mathbf{B} = \Sigma^*_{I_jI_j}$, we only need to verify that, when $\lambda \le 2$,

$$12\psi\alpha\|\widetilde{\mathbf{S}}_{I_jI_j} - \Sigma^*_{I_jI_j}\|_\infty + 12\psi\|\widetilde{\mathbf{S}}_{J_jI_j} - \Sigma^*_{J_jI_j}\|_\infty + 2(\alpha+1)\|\widetilde{\mathbf{S}}_{\setminus j,j} - \Sigma^*_{\setminus j,j}\|_\infty < \lambda(1-\alpha).$$

It is easy to see that given $\lambda \le 2$ and $\alpha < 1$, the above inequality holds when

$$\begin{aligned}
\|\widetilde{\mathbf{S}}_{I_jI_j} - \Sigma^*_{I_jI_j}\|_\infty < \frac{\lambda(1-\alpha)}{26\psi\alpha}, \quad \|\widetilde{\mathbf{S}}_{J_jI_j} - \Sigma^*_{J_jI_j}\|_\infty < \frac{\lambda(1-\alpha)}{26\psi}, \\
\|\widetilde{\mathbf{S}}_{\setminus j,j} - \Sigma^*_{\setminus j,j}\|_\infty < \frac{\lambda(1-\alpha)}{26(\alpha+1)} < 1. \qquad \text{(B.20)}
\end{aligned}$$

Then we need to show that $\widehat{\boldsymbol{\beta}}_{I_j}$ has no zero elements. Here we aim to secure a sufficient condition $\tau > \|\widehat{\boldsymbol{\beta}}_{I_j} - \mathbf{B}^*_{I_j,j}\|_\infty$. Since

$$\begin{aligned}\|\widehat{\boldsymbol{\beta}}_{I_j} - \mathbf{B}^*_{I_j,j}\|_\infty &= \|(\widetilde{\mathbf{S}}_{I_jI_j})^{-1}\widetilde{\mathbf{S}}_{I_j,j} - (\Sigma^*_{I_jI_j})^{-1}\Sigma^*_{I_j,j} - \frac{\lambda}{2}(\widetilde{\mathbf{S}}_{I_jI_j})^{-1}\zeta_{I_j})\|_\infty \\ &\le \|(\widetilde{\mathbf{S}}_{I_jI_j})^{-1}\widetilde{\mathbf{S}}_{I_j,j} - (\Sigma^*_{I_jI_j})^{-1}\Sigma^*_{I_j,j}\|_\infty + \frac{\lambda}{2}\|(\widetilde{\mathbf{S}}_{I_jI_j})^{-1}\|_\infty, \end{aligned} \tag{B.21}$$

by Lemma B.4 with $\widehat{\mathbf{A}} = (\widetilde{\mathbf{S}}_{I_jI_j})^{-1}$, $\mathbf{A} = (\Sigma^*_{I_jI_j})^{-1}$, $\widehat{\boldsymbol{v}} = \widetilde{\mathbf{S}}_{I_j,j}$, and $\boldsymbol{v} = \Sigma^*_{I_j,j}$, we have

$$\begin{aligned}&\|(\widetilde{\mathbf{S}}_{I_jI_j})^{-1}\widetilde{\mathbf{S}}_{I_j,j} - (\Sigma^*_{I_jI_j})^{-1}\Sigma^*_{I_j,j}\|_\infty \\ \le&\|(\widetilde{\mathbf{S}}_{I_jI_j})^{-1} - (\Sigma^*_{I_jI_j})^{-1}\|_\infty\|\widetilde{\mathbf{S}}_{I_j,j} - \Sigma^*_{I_j,j}\|_\infty + \|(\widetilde{\mathbf{S}}_{I_jI_j})^{-1} - (\Sigma^*_{I_jI_j})^{-1}\|_\infty + \psi\|\widetilde{\mathbf{S}}_{I_j,j} - \Sigma^*_{I_j,j}\|_\infty \\ \le&2\|(\widetilde{\mathbf{S}}_{I_jI_j})^{-1} - (\Sigma^*_{I_jI_j})^{-1}\|_\infty + \psi\|\widetilde{\mathbf{S}}_{\setminus,j} - \Sigma^*_{\setminus,j}\|_\infty, \end{aligned} \tag{B.22}$$

where the last inequality holds when $\|\widetilde{\mathbf{S}}_{I_j,j} - \Sigma^*_{I_j,j}\|_\infty \le 1$. Combining

$$\frac{\lambda}{2}\|(\widetilde{\mathbf{S}}_{I_jI_j})^{-1}\|_\infty \le \frac{\lambda}{2}\|(\widetilde{\mathbf{S}}_{I_jI_j})^{-1} - (\Sigma^*_{I_jI_j})^{-1}\|_\infty + \frac{\lambda\psi}{2}$$

with (B.22), we have

$$\begin{aligned}\|\widehat{\boldsymbol{\beta}}_{I_j} - \mathbf{B}^*_{I_j,j}\|_\infty &\le \Big(\frac{\lambda}{2} + 2\Big)\|(\widetilde{\mathbf{S}}_{I_jI_j})^{-1} - (\Sigma^*_{I_jI_j})^{-1}\|_\infty + \psi\|\widetilde{\mathbf{S}}_{\setminus,j} - \Sigma^*_{\setminus,j}\|_\infty + \frac{\lambda\psi}{2} \\ &\le \Big(\lambda + 4\Big)\psi^2\|\widetilde{\mathbf{S}}_{I_jI_j} - \Sigma^*_{I_jI_j}\|_\infty + \psi\|\widetilde{\mathbf{S}}_{\setminus,j} - \Sigma^*_{\setminus,j}\|_\infty + \frac{\lambda\psi}{2}, \end{aligned} \tag{B.23}$$

where the last inequality comes from Lemma B.2. Now we only need to show that

$$6\psi^2\|\widetilde{\mathbf{S}}_{I_jI_j} - \Sigma^*_{I_jI_j}\|_\infty + \psi\|\widetilde{\mathbf{S}}_{\setminus,j} - \Sigma^*_{\setminus,j}\|_\infty < \frac{\tau}{2}, \tag{B.24}$$

when $\lambda\psi \le \tau$ and $\lambda \le 2$. It is easy to verify that (B.24) holds when

$$\|\widetilde{\mathbf{S}}_{I_jI_j} - \Sigma^*_{I_jI_j}\|_\infty < \frac{\tau}{14\psi^2} \text{ and } \|\widetilde{\mathbf{S}}_{\setminus,j} - \Sigma^*_{\setminus,j}\|_\infty < \frac{\tau}{14\psi}. \tag{B.25}$$

By applying Lemma B.5 and Theorem 4.1 to (B.20) and (B.25), for large enough n such that

$$\sqrt{\frac{\log d}{n}} \le \min\left\{\frac{\lambda(1-\alpha)}{26\psi\alpha\kappa_3 s}, \frac{\lambda(1-\alpha)}{26\psi\kappa_3 s}, \frac{\lambda(1-\alpha)}{26(\alpha+1)}, \frac{\tau}{14\psi^2 s\kappa_3}, \frac{\tau}{14\psi\kappa_3}, \frac{\delta}{2s\kappa_3}\right\},$$

we have

$$\mathbb{P}(\widehat{\mathbf{E}} = \mathbf{E}^*) \ge 1 - \frac{1}{d^3}.$$

Under Conditions 1, 2, and 3, we obtain $\mathbb{P}(\widehat{\mathbf{E}} = \mathbf{E}^*) \to 1$. □

C Proof of Lemma 3.1

Proof. Equation (3.7) can be rewritten as

$$\left\|\left\|\mathbf{A}\right\|\right\|_{\infty}^{\mu} = \min_{\left\|\left\|\mathbf{U}\right\|\right\|_1 \leq 1} \frac{\mu}{2}\|\mathbf{U} - \mathbf{A}/\mu\|_{\mathrm{F}}^2. \tag{C.1}$$

By the Lagrangian duality, we know that there exists some constant $\gamma > 0$ such that

$$\left\|\left\|\mathbf{A}\right\|\right\|_{\infty}^{\mu} = \min_{\mathbf{U}} \|\mathbf{U} - \mathbf{A}/\mu\|_{\mathrm{F}}^2 + \gamma\left\|\left\|\mathbf{U}\right\|\right\|_1 \tag{C.2}$$

holds as a Lagrangian form equivalent to (C.1). (C.2) results in the soft thresholding operation as follow

$$\widetilde{\mathbf{U}}_{jk} = \operatorname{sign}\left(\mathbf{A}_{jk}\right) \cdot \max\left\{\left|\frac{\mathbf{A}_{jk}}{\mu}\right| - \gamma, 0\right\}, \tag{C.3}$$

which completes the proof. □

D Fast Projection Algorithm for elementwise ℓ_1-norm ball

By carefully examining (3.8), we find that it is a special case of the quadratic Knapsack problem (Brucker, 1984; Pardalos and Kovoor, 1990). We first define $\mathbf{A}' \in \mathbb{R}^{d\times d}$ with $\mathbf{A}'_{jk} = |\mathbf{A}|_{jk}/\mu$. Given the decreasing order statistics of all elements in $\mathbf{A}'$ as $\mathbf{A}'_{(1)}, \mathbf{A}'_{(2)}, ..., \mathbf{A}'_{(d^2)}$, (3.9) is equivalent to find u such that

$$\sum_{j=1}^{u-1}\left(\mathbf{A}'_{(j)} - \mathbf{A}'_{(u)}\right) < 1 \text{ and } \sum_{j=1}^{u+1}\left(\mathbf{A}'_{(j)} - \mathbf{A}'_{(u)}\right) \geq 1. \tag{D.1}$$

Then with u and $\mathbf{A}'_{(u)}$, we can calculate γ as follows,

$$\gamma = \frac{1}{u}\left(\sum_{j=1}^{u}\mathbf{A}'_{(j)} - 1\right). \tag{D.2}$$

Similar to the fast median algorithm (Corman et al., 2001), Algorithm 1 identifies u and the pivot value $\mathbf{A}'_{(u)}$ using a divide and conquer procedure (without sorting the data). In each iteration we either eliminate elements shown to be strictly smaller than $\mathbf{A}'_{(u)}$ or update the partial sum leading to (D.1). This algorithm has an average-case complexity of $O(d^2)$. Similar algorithms can be found in Liu and Ye (2009); Duchi et al. (2008) for the lasso problem.

E Proof of Lemma 3.2

Proof. The eigenvalue decomposition of $\mathbf{A}$ can be rewritten as $\mathbf{A} = \mathbf{V}\mathbf{Z}\mathbf{V}^T$ with

$$\mathbf{Z} = \operatorname{diag}(\sigma_1, \sigma_2, ..., \sigma_d) \text{ and } \mathbf{V} = (\boldsymbol{v}_1, \boldsymbol{v}_2, ..., \boldsymbol{v}_d). \tag{E.1}$$

Algorithm 2 The elementwise ℓ_1 norm projection algorithm.

Input: $\mathbf{A}' \in \mathbb{R}^{d\times d}$

Initialize: $\mathcal{S}_0 = \{(j,k) \mid j = 1,...,d \text{ and } k = 1,...,d\}$, $w = 0$,$u = 0$

repeat

1: randomly pick $(j',k') \in \mathcal{S}_0$

2: partition $\mathcal{S}_0$:

$$\mathcal{S}_1 = \{(j,k) \in \mathcal{S}_1 \mid \mathbf{A}'_{jk} \geq \mathbf{A}'_{j'k'}\}$$
$$\mathcal{S}_2 = \{(j,k) \in \mathcal{S}_2 \mid \mathbf{A}'_{jk} < \mathbf{A}'_{j'k'}\}$$

3: Calculate $\mathbf{\Delta}_w = |\mathcal{S}_1|$ and $\mathbf{\Delta}_u = \sum_{(j,k)\in\mathcal{S}_1} \mathbf{A}'_{jk}$

4: If $(u + \mathbf{\Delta}_u) - (w + \mathbf{\Delta}_w)\mathbf{A}'_{j'k'} \leq 1$

$u = u + \mathbf{\Delta}_u$; $w = w + \mathbf{\Delta}_w$; $\mathcal{S}_0 \leftarrow \mathcal{S}_2$

else

$\mathcal{S}_0 \leftarrow \mathcal{S}_1 \setminus \{(j',k')\}$

until $\mathcal{S}_0 = \emptyset$

Output: $\gamma = (w-1)/u$

Note is that $\mathbf{V}$ is a unitary matrix. Since the Frobenius norm is invariant to $\mathbf{V}$, we have

$$\min_{\mathbf{B}\geq 0} \|\mathbf{B} - \mathbf{A}\|_{\mathrm{F}}^2 = \min_{\mathbf{B}\geq 0} \|\mathbf{V}^T(\mathbf{B} - \mathbf{A})\mathbf{V}\|_{\mathrm{F}}^2 = \min_{\mathbf{B}\geq 0} \|\mathbf{V}^T\mathbf{B}\mathbf{V}^T - \mathbf{Z}\|_{\mathrm{F}}^2. \tag{E.2}$$

Then it is easy to verify that (E.2) is minimized when

$$\mathbf{V}^T\mathbf{B}\mathbf{V} = \mathbf{R} = \mathrm{diag}(\widetilde{\sigma_1}, \widetilde{\sigma_2}, ..., \widetilde{\sigma_d}), \tag{E.3}$$

where $\widetilde{\sigma_j} = \max\{\sigma_j, 0\}$. Therefore we have

$$\mathbf{B} = \mathbf{V}\mathbf{R}\mathbf{V}^T, \tag{E.4}$$

which completes the proof. □

E.1 Proof of Lemma B.1

Proof.

$$\begin{aligned}\|\mathbf{A}\mathbf{B}\|_\infty = \max_i \sum_j \sum_k |\mathbf{A}_{ik}\mathbf{B}_{kj}| &= \max_i \sum_k |\mathbf{A}_{ik}| \sum_j |\mathbf{B}_{kj}| \\ &\leq \left(\max_i \sum_k |\mathbf{A}_{ik}|\right)\left(\max_\ell \sum_j |\mathbf{B}_{\ell j}|\right) = \|\mathbf{A}\|_\infty \|\mathbf{B}\|_\infty. \end{aligned} \tag{E.5}$$

□

E.2 Proof of Lemma B.2

Proof.

$$\begin{aligned}\|\widehat{\mathbf{B}}^{-1}-\mathbf{B}^{-1}\|_\infty &= \|\widehat{\mathbf{B}}^{-1}(\mathbf{B}-\widehat{\mathbf{B}})\mathbf{B}^{-1})\|_\infty \le \|\widehat{\mathbf{B}}^{-1}\|_\infty\|\mathbf{B}-\widehat{\mathbf{B}}\|_\infty\|\mathbf{B}^{-1}\|_\infty \\ &\le \left(\|\widehat{\mathbf{B}}^{-1}-\mathbf{B}^{-1}\|_\infty + \|\mathbf{B}^{-1}\|_\infty\right)\|\mathbf{B}-\widehat{\mathbf{B}}\|_\infty\|\mathbf{B}^{-1}\|_\infty. \end{aligned} \tag{E.6}$$

Therefore we have

$$\|\widehat{\mathbf{B}}^{-1}-\mathbf{B}^{-1}\|_\infty \le \frac{\|\mathbf{B}^{-1}\|_\infty^2}{1-\|\mathbf{B}-\widehat{\mathbf{B}}\|_\infty\|\mathbf{B}^{-1}\|_\infty}\|\widehat{\mathbf{B}}-\mathbf{B}\|_\infty, \tag{E.7}$$

which completes the proof. □

E.3 Proof of Lemma B.3

Proof. We have the following decomposition,

$$\left\{(\widehat{\mathbf{A}}-\mathbf{A})(\widehat{\mathbf{B}}^{-1}-\mathbf{B}^{-1})\right\} = \widehat{\mathbf{A}}\widehat{\mathbf{B}}^{-1}-\mathbf{A}\mathbf{B}^{-1} + \left\{(\widehat{\mathbf{A}}-\mathbf{A})\mathbf{B}^{-1}\right\} + \left\{\mathbf{A}(\mathbf{B}^{-1}-\widehat{\mathbf{B}}^{-1})\right\}. \tag{E.8}$$

Since

$$\begin{aligned}\left\{\mathbf{A}(\mathbf{B}^{-1}-\widehat{\mathbf{B}}^{-1})\right\} &= \mathbf{A}\mathbf{B}^{-1}(\mathbf{I}-\mathbf{B}\widehat{\mathbf{B}}^{-1}) = \mathbf{A}\mathbf{B}^{-1}(\widehat{\mathbf{B}}-\mathbf{B})\widehat{\mathbf{B}}^{-1} \\ &= \mathbf{A}\mathbf{B}^{-1}(\widehat{\mathbf{B}}-\mathbf{B})(\widehat{\mathbf{B}}^{-1}-\mathbf{B}^{-1}) + \mathbf{A}\mathbf{B}^{-1}(\widehat{\mathbf{B}}-\mathbf{B})\mathbf{B}^{-1}, \end{aligned} \tag{E.9}$$

then we have

$$\begin{aligned}\|\widehat{\mathbf{A}}\widehat{\mathbf{B}}^{-1}-\mathbf{A}\mathbf{B}^{-1}\|_\infty &\le \left\{\|\widehat{\mathbf{A}}-\mathbf{A}\|_\infty\|\widehat{\mathbf{B}}^{-1}-\mathbf{B}^{-1}\|_\infty\right\} + \left\{\|\widehat{\mathbf{A}}-\mathbf{A}\|_\infty\psi\right\} \\ &\quad + \left\{\alpha\|\widehat{\mathbf{B}}-\mathbf{B}\|_\infty\|\widehat{\mathbf{B}}^{-1}-\mathbf{B}^{-1}\|_\infty\right\} + \left\{\alpha\|\widehat{\mathbf{B}}-\mathbf{B}\|_\infty\psi\right\} \\ &= \left\{\|\widehat{\mathbf{A}}-\mathbf{A}\|_\infty(\|\widehat{\mathbf{B}}^{-1}-\mathbf{B}^{-1}\|_\infty+\psi) + (\alpha\|\widehat{\mathbf{B}}-\mathbf{B}\|_\infty(\|\widehat{\mathbf{B}}^{-1}-\mathbf{B}^{-1}\|_\infty+\psi)\right. \\ &= (\|\widehat{\mathbf{A}}-\mathbf{A}\|_\infty+\alpha\|\widehat{\mathbf{B}}-\mathbf{B}\|_\infty)(\psi+\|\widehat{\mathbf{B}}^{-1}-\mathbf{B}^{-1}\|_\infty). \end{aligned} \tag{E.10}$$

By (E.6) in Lemma B.2

$$\begin{aligned}\|\widehat{\mathbf{B}}^{-1}-\mathbf{B}^{-1}\|_\infty &\le \left(\|\widehat{\mathbf{B}}^{-1}-\mathbf{B}^{-1}\|_\infty + \|\mathbf{B}^{-1}\|_\infty\right)\|\mathbf{B}-\widehat{\mathbf{B}}\|_\infty\|\mathbf{B}^{-1}\|_\infty \\ &= \frac{1}{2}\|\widehat{\mathbf{B}}^{-1}-\mathbf{B}^{-1}\|_\infty + \frac{1}{2}\psi, \end{aligned} \tag{E.11}$$

we have

$$\|\widehat{\mathbf{B}}^{-1}-\mathbf{B}^{-1}\|_\infty \le \psi. \tag{E.12}$$

By combing (E.10) and (E.12), we complete the proof. □

E.4 Proof of Lemma B.4

Proof.

$$\begin{aligned}\|\widehat{\mathbf{A}}\widehat{\boldsymbol{v}}-\mathbf{A}\boldsymbol{v}\|_\infty &= \|\widehat{\mathbf{A}}\widehat{\boldsymbol{v}}-\mathbf{A}\widehat{\boldsymbol{v}}+\mathbf{A}\widehat{\boldsymbol{v}}-\mathbf{A}\boldsymbol{v}\|_\infty \\ &\le \|\widehat{\mathbf{A}}\widehat{\boldsymbol{v}}-\mathbf{A}\widehat{\boldsymbol{v}}\|_\infty + \|\mathbf{A}\widehat{\boldsymbol{v}}-\mathbf{A}\boldsymbol{v}\|_\infty \\ &\le \|\widehat{\mathbf{A}}-\mathbf{A}\|_\infty(\|\widehat{\boldsymbol{v}}-\boldsymbol{v}\|_\infty+\|\boldsymbol{v}\|_\infty)+\|\mathbf{A}\|_\infty\|\widehat{\boldsymbol{v}}-\boldsymbol{v}\|_\infty, \end{aligned} \tag{E.13}$$

which completes the proof. □

E.5 Proof of Lemma B.5

Proof. We have

$$\|\widetilde{\mathbf{S}}_{I_jI_j}-\Sigma^*_{I_jI_j}\|_\infty = \max_{k\in I_j}\sum_{\ell\in I_j}|\widetilde{\mathbf{S}}_{k\ell}-\Sigma^*_{k\ell}| \le \kappa_3 s\sqrt{\frac{\log d}{n}},$$

$$\|\widetilde{\mathbf{S}}_{J_jI_j}-\Sigma^*_{J_jI_j}\|_\infty = \max_{k\in J_j}\sum_{\ell\in I_j}|\widetilde{\mathbf{S}}_{k\ell}-\Sigma^*_{k\ell}| \le \kappa_3 s\sqrt{\frac{\log d}{n}},$$

$$\|\widetilde{\mathbf{S}}_{\setminus j,j}-\Sigma^*_{\setminus j,j}\|_\infty = \max_{k\neq j}|\widetilde{\mathbf{S}}_{kj}-\Sigma^*_{kj}| \le \kappa_3\sqrt{\frac{\log d}{n}}.$$

Let $s_j = |I_j| \le s$. For arbitrary $\boldsymbol{v}\in\mathbb{R}^{s_j}$, we have

$$\begin{aligned}\boldsymbol{v}^T(\widetilde{\mathbf{S}}_{I_jI_j})\boldsymbol{v} &= \boldsymbol{v}^T(\widetilde{\mathbf{S}}_{I_jI_j}-\Sigma^*_{I_jI_j}+\Sigma^*_{I_jI_j})\boldsymbol{v} \\ &= \boldsymbol{v}^T\Sigma^*_{I_jI_j}\boldsymbol{v}-\boldsymbol{v}^T(\Sigma^*_{I_jI_j}-\widetilde{\mathbf{S}}_{I_jI_j})\boldsymbol{v} \\ &\ge \Lambda_{\min}(\Sigma^*_{I_jI_j})\|\boldsymbol{v}\|_2^2-\|\boldsymbol{v}\|_1^2\left|\left|\left|\Sigma^*_{I_jI_j}-\widetilde{\mathbf{S}}_{I_jI_j}\right|\right|\right|_\infty \\ &\ge \delta\|\boldsymbol{v}^T\|_2^2-s_j\|\boldsymbol{v}\|_2^2\cdot\kappa_3\sqrt{\frac{\log d}{n}},\end{aligned}$$

where the last inequality comes from the fact $\boldsymbol{v}\in\mathbb{R}^{s_j}$. Thus for large enough n such that

$$\sqrt{\frac{\log d}{n}} \le \frac{\delta}{2s\kappa_3}, \tag{E.14}$$

we have

$$\boldsymbol{v}^T(\widetilde{\mathbf{S}}_{I_jI_j})\boldsymbol{v} \ge \frac{\delta}{2}\|\boldsymbol{v}^T\|_2^2. \tag{E.15}$$

Since $\boldsymbol{v}$ and j are arbitrary, we further have

$$\Lambda_{\min}(\widetilde{\mathbf{S}}_{I_jI_j}) \ge \frac{\delta}{2} \text{ for all } j=1,\ldots,d. \tag{E.16}$$

□

References

Agarwal, A., Negahban, S. and Wainwright, M. J. (2012). Fast global convergence of gradient methods for high-dimensional statistical recovery. *The Annals of Statistics* **40** 2452–2482.

Banerjee, O., Ghaoui, L. E. and d'Aspremont, A. (2008). Model selection through sparse maximum likelihood estimation. *Journal of Machine Learning Research* **9** 485–516.

Beck, A. and Teboulle, M. (2009). A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences* **2** 183–202.

Blei, D. and Lafferty, J. (2007). A correlated topic model of science. *Annals of Applied Statistics* **1** 17–35.

Brucker, P. (1984). An o(n) algorithm for quadratic knapsack problems. *Operations Research Letters* **3** 163–166.

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** 594–607.

Chen, X., Lin, Q., Kim, S., Carbonell, J. and Xing, E. (2012). A smoothing proximal gradient method for general structured sparse regression. *Annals of Applied Statistics* To appear.

Corman, T., Leiserson, C., Rivest, R. and Stein, C. (2001). *Introduction to algorithms*. MIT Press.

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

Duchi, J., Shwartz, S., Singer, Y. and Chandra, T. (2008). Efficient projections onto the l1-ball for learning in high dimensions. *International Conference on Machine Learning* 272–279.

Friedman, J., T. Hastie, H. H. and Tibshirani, R. (2007). Pathwise coordinate optimization. *Annals of Applied Statistics* **1** 302–332.

Guo, Y., Hastie, T. and Tibshirani, R. (2007). Regularized linear discriminant analysis and its application in microarrays. *Biostatistics* **8** 86–100.

Han, F., Zhao, T. and Liu, H. (2013). Coda: High dimensional copula discriminant analysis. *Journal of Machine Learning Research* **14** 629–671.

Hoerl, A. E. and Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics* **12** 55–67.

Honorio, J., Ortiz, L., Samaras, D., Paragios, N., and Goldstein, R. (2009). Sparse and locally constant gaussian graphical models. *Advances in Neural Information Processing Systems* 745–753.

Jalali, A., Johnson, C. and Ravikumar, P. (2012). High-dimensional sparse inverse covariance estimation using greedy methods. *International Conference on Artificial Intelligence and Statistics* To appear.

Ji, S. and Ye, J. (2009). An accelerated gradient method for trace norm minimization. In *Proceedings of the 26th Annual International Conference on Machine Learning*. ACM.

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

Lam, C. and Fan, J. (2009). Sparsistency and rates of convergence in large covariance matrix estimation. *Annals of Statistics* **37** 42–54.

Lauritzen, S. (1996). *Graphical models*, vol. 17. Oxford University Press, USA.

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

Liu, H., Han, F., Yuan, M., Lafferty, J. and Wasserman, L. (2012). High dimensional semiparametric gaussian copula graphical models. *Annals of Statistics* To appear.

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

Liu, H., Roeder, K. and Wasserman, L. (2010). Stability approach to regularization selection for high dimensional graphical models. *Advances in Neural Information Processing Systems* .

Liu, J. and Ye, J. (2009). Efficient euclidean projections in linear time. *International Conference on Machine Learning* .

Mai, Q., Zou, H. and Yuan, M. (2012). A direct approach to sparse discriminant analysis in ultra-high dimensions. *Biometrika* 1–14.

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

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

Nesterov, Y. (1988). On an approach to the construction of optimal methods of smooth convex functions. *Ékonom. i. Mat. Metody* **24** 509–517.

Nesterov, Y. (2005). Smooth minimization of non-smooth functions. *Mathematical Programming* **103** 127–152.

Pardalos, P. M. and Kovoor, N. (1990). An algorithm for a singly constrained class of quadratic programs subject to upper and lower bounds. *Mathematical Programming* **46** 312–328.

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** 735–746.

Ravikumar, P., Wainwright, M., Raskutti, G. and Yu, B. (2011). High-dimensional covariance estimation by minimizing ℓ_1-penalized log-determinant divergence. *Electronic Journal of Statistics* **5** 935–980.

Rousseeuw, P. and Molenberghs, G. (1993). Transformation of non positive semidefinite correlation matrices. *Communications in Statistics-Theory and Methods* **22** 965–984.

Shojaie, A. and Michailidis, G. (2010). Penalized likelihood methods for estimation of sparse high-dimensional directed acyclic graphs. *Biometrika* **97** 519–538.

Sun, H. and Li, H. (2012). Robust gaussian graphical modeling via ℓ_1 penalization. *Biometrics* To appear.

Sun, T. and Zhang, C.-H. (2012). Sparse matrix inversion with scaled lasso. Tech. rep., Department of Statistics, Rutgers University.

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

Wainwright, M. (2009). Sharp thresholds for highdimensional and noisy sparsity recovery using $\ell_1$constrained quadratic programming. *IEEE Transactions on Information Theory* **55** 2183–2201.

Wille, A., Zimmermann, P., Vranova, E., Frholz, A., Laule, O., Bleuler, S., Hennig, L., Prelic, A., von Rohr, P., Thiele, L., Zitzler, E., Gruissem, W. and Bühlmann, P. (2004). Sparse graphical gaussian modeling of the isoprenoid gene network in arabidopsis thaliana. *Genome Biology* **5** R92.

Yin, J. and Li, H. (2011). A sparse conditional gaussian graphical model for analysis of genetical genomics data. *Annals of Applied Statistics* **5** 2630–2650.

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. (2007). Model selection and estimation in the gaussian graphical model. *Biometrika* **94** 19–35.

Zhao, P. and Yu, B. (2006). On model selection consistency of lasso. *Journal of Machine Learning Research* **7** 2541–2563.

Zhao, T. and Liu, H. (2012). Sparse additive machine. In *Proceedings of the 15th International Conference on Artificial Intelligence*.

Zhao, T. and Liu, H. (2013). Semiparametric sparse column inverse operator. Tech. rep., Department of Computer Science, Johns Hopkins University.

Zhao, T., Liu, H., Roeder, K., Lafferty, J. and Wasserman, L. (2012a). The huge package for high-dimensional undirected graph estimation in r. *Journal of Machine Learning Research* To appear.

Zhao, T., Roeder, K. and Liu, H. (2012b). Smooth-projected neighborhood pursuit for high-dimensional nonparanormal graph estimation. *Advances in Neural Information Processing Systems* .

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