The Annals of Statistics
2012, Vol. 40, No. 5, 2541–2571
DOI: 10.1214/12-AOS1041

REGULARIZED RANK-BASED ESTIMATION OF HIGH-DIMENSIONAL NONPARANORMAL GRAPHICAL MODELS[1]

BY LINGZHOU XUE AND HUI ZOU

University of Minnesota

A sparse precision matrix can be directly translated into a sparse Gaussian graphical model under the assumption that the data follow a joint normal distribution. This neat property makes high-dimensional precision matrix estimation very appealing in many applications. However, in practice we often face nonnormal data, and variable transformation is often used to achieve normality. In this paper we consider the nonparanormal model that assumes that the variables follow a joint normal distribution after a set of unknown monotone transformations. The nonparanormal model is much more flexible than the normal model while retaining the good interpretability of the latter in that each zero entry in the sparse precision matrix of the nonparanormal model corresponds to a pair of conditionally independent variables. In this paper we show that the nonparanormal graphical model can be efficiently estimated by using a rank-based estimation scheme which does not require estimating these unknown transformation functions. In particular, we study the rank-based graphical lasso, the rank-based neighborhood Dantzig selector and the rank-based CLIME. We establish their theoretical properties in the setting where the dimension is nearly exponentially large relative to the sample size. It is shown that the proposed rank-based estimators work as well as their oracle counterparts defined with the oracle data. Furthermore, the theory motivates us to consider the adaptive version of the rank-based neighborhood Dantzig selector and the rank-based CLIME that are shown to enjoy graphical model selection consistency without assuming the irrepresentable condition for the oracle and rank-based graphical lasso. Simulated and real data are used to demonstrate the finite performance of the rank-based estimators.

1. Introduction. Estimating covariance or precision matrices is of fundamental importance in multivariate statistical methodologies and applications. In particular, when data follow a joint normal distribution, $\mathbf{X} = (X_1, \ldots, X_p) \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, the precision matrix $\boldsymbol{\Theta} = \boldsymbol{\Sigma}^{-1}$ can be directly translated into a Gaussian graphical model. The Gaussian graphical model serves as a noncausal structured approach to explore the complex systems consisting of Gaussian random variables, and it finds many interesting applications in areas such as gene expression genomics

Received February 2012; revised August 2012.

[1]Some results in this paper were reported in a research proposal funded by the Office of Naval Research in November 2010. Supported in part by NSF Grant DMS-08-46068 and a grant from ONR.

MSC2010 subject classifications. Primary 62G05, 62G20; secondary 62F12, 62J07.

Key words and phrases. CLIME, Dantzig selector, graphical lasso, nonparanormal graphical model, rate of convergence, variable transformation.

and macroeconomics determinants study [Dobra, Eicher and Lenkoski (2010), Friedman (2004), Wille et al. (2004)]. The precision matrix plays a critical role in the Gaussian graphical models because the zero entries in $\mathbf{\Theta} = (\theta_{ij})_{p\times p}$ precisely capture the desired conditional independencies, that is, $\theta_{ij} = 0$ if and only if $X_i \perp\!\!\!\perp X_j | \mathbf{X} \setminus \{X_i, X_j\}$ [Edwards (2000), Lauritzen (1996)].

The sparsity pursuit in precision matrices was initially considered by Dempster (1972) as the covariance selection problem. Multiple testing methods have been employed for network exploration in the Gaussian graphical models [Drton and Perlman (2004)]. With rapid advances of the high-throughput technology (e.g., microarray, functional MRI), estimation of a sparse graphical model has become increasingly important in the high-dimensional setting. Some well-developed penalization techniques have been used for estimating sparse Gaussian graphical models. In a highly-cited paper, Meinshausen and Bühlmann (2006) proposed the neighborhood selection scheme which tries to discover the smallest index set ne_α for each variable X_α satisfying $X_\alpha \perp\!\!\!\perp \mathbf{X} \setminus \{X_\alpha, \mathbf{X}_{ne_\alpha}\} | \mathbf{X}_{ne_\alpha}$. Meinshausen and Bühlmann (2006) further proposed to use the lasso [Tibshirani (1996)] to fit each neighborhood regression model. Afterwards, one can summarize the zero patterns by aggregation via union or intersection. Yuan (2010) considered the Dantzig selector [Candes and Tao (2007)] as an alternative to the lasso penalized least squares in the neighborhood selection scheme. Peng et al. (2009) proposed the joint neighborhood lasso selection. Penalized likelihood methods have been studied for Gaussian graphical modeling [Yuan and Lin (2007)]. Friedman, Hastie and Tibshirani (2008) developed a fast blockwise coordinate descent algorithm [Banerjee, El Ghaoui and d'Aspremont (2008)] called graphical lasso for efficiently solving the lasso penalized Gaussian graphical model. Rate of convergence under the Frobenius norm was established by Rothman et al. (2008). Ravikumar et al. (2011) obtained the convergence rate under the elementwise ℓ_∞ norm and the spectral norm. Lam and Fan (2009) studied the nonconvex penalized Gaussian graphical model where a nonconvex penalty such as SCAD [Fan and Li (2001)] is used to replace the lasso penalty in order to overcome the bias issue of the lasso penalization. Zhou et al. (2011) proposed a hybrid method for estimating sparse Gaussian graphical models: they first infer a sparse Gaussian graphical model structure via thresholding neighborhood selection and then estimate the precision matrix of the submodel by maximum likelihood. Cai, Liu and Luo (2011) recently proposed a constrained ℓ_1 minimization estimator called CLIME for estimating sparse precision matrices and established its convergence rates under the elementwise ℓ_∞ norm and Frobenius norm.

Although the normality assumption can be relaxed if we only focus on estimating a precision matrix, it plays an essential role in making the neat connection between a sparse precision matrix and a sparse Gaussian graphical model. Without normality, we ought to be very cautious when translating a good sparse precision matrix estimator into an interpretable sparse Gaussian graphical model. However, the normality assumption often fails in reality. For example, the observed data are

TABLE 1

Testing for normality of the gene expression measurements in the Arabidposis thaliana data. This table illustrates the number out of 39 *genes rejecting the null hypothesis of normality at the significance level of* 0.05

	Critical value	**Cramer–von Mises**	**Lilliefors**	**Shapiro–Francia**
Raw data	0.05	30	30	35
	0.05/39	24	26	28
Log data	0.05	29	24	33
	0.05/39	14	12	16

often skewed or have heavy tails. To illustrate the issue of nonnormality in real applications, let us consider the gene expression data to construct isoprenoid genetic regulatory network in Arabidposis thaliana [Wille et al. (2004)], including 16 genes from the mevalonate (MVA) pathway in the cytosolic, 18 genes from the plastidial (MEP) pathway in the chloroplast and 5 encode proteins in the mitochondrial. This dataset contains gene expression measurements of 39 genes assayed on $n = 118$ Affymetrix GeneChip microarrays. This dataset was analyzed by Wille et al. (2004), Li and Gui (2006) and Drton and Perlman (2007) in the context of Gaussian graphical modeling after taking the log-transformation of the data. However, the normality assumption is still inappropriate even after the log-transformation. To show this, we conduct the normality test at the significance level of 0.05 as in Table 1. It is clear that at most 9 out of 39 genes would pass any of three normality tests. Even after log-transformation, at least 60% genes reject the null hypothesis of normality. With Bonferroni correction there are still over 30% genes that fail to pass any normality test. Figure 1 plots the histograms of two

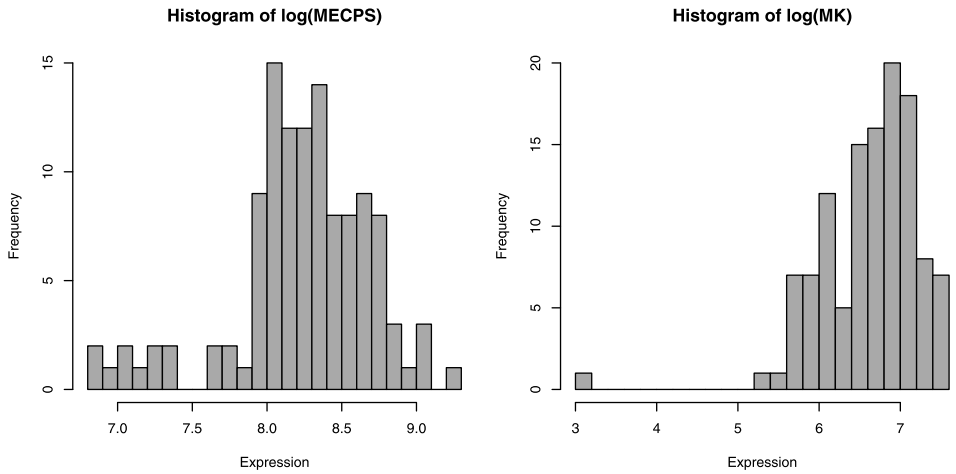

FIG. 1. *Illustration of the nonnormality after the log-transformation preprocessing.*

key isoprenoid genes MECPS in the MEP pathway and MK in the MVA pathway after the log-transformation, clearly showing the nonnormality of the data after the log-transformation.

Using transformation to achieve normality is a classical idea in statistical modeling. The celebrated Box–Cox transformation is widely used in regression analysis. However, any parametric modeling of the transformation suffers from model mis-specification which could lead to misleading inference. In this paper we take a nonparametric transformation strategy to handle the nonnormality issue. Let $F(\cdot)$ be the CDF of a continuous random variable X and $\Phi^{-1}(\cdot)$ be the inverse of the CDF of $N(0,1)$. Consider the transformation from X to Z by $Z = \Phi^{-1}(F(X))$. Then it is easy to see that Z is standard normal regardless of F. Motivated by this simple fact, we consider modeling the data by the following nonparanormal model:

The nonparanormal model: $\mathbf{X} = (X_1, \ldots, X_p)$ follows a p-dimensional nonparanormal distribution if there exists a vector of unknown univariate monotone increasing transformations, denoted by $\mathbf{f} = (f_1, \ldots, f_p)$, such that the transformed random vector follows a multivariate normal distribution with mean 0 and covariance $\mathbf{\Sigma}$,

$$\mathbf{f}(\mathbf{X}) = \big(f_1(X_1), \ldots, f_p(X_p)\big) \sim N_p(0, \mathbf{\Sigma}), \tag{1}$$

where without loss of generality the diagonals of $\mathbf{\Sigma}$ are equal to 1.

Note that model (1) implies that $f_j(X_j)$ is a standard normal random variable. Thus, f_j must be $\Phi^{-1} \circ F_j$ where F_j is the CDF of X_j. The marginal normality is always achieved by transformations, so model (1) basically assumes that those marginally normal-transformed variables are jointly normal as well. We follow Liu, Lafferty and Wasserman (2009) to call model (1) the nonparanormal model, but model (1) is in fact a semiparametric Gaussian copula model. The semiparametric Gaussian copula model is a nice combination of flexibility and interpretability, and it has generated a lot of interests in statistics, econometrics and finance; see Song (2000), Chen and Fan (2006), Chen, Fan and Tsyrennikov (2006) and references therein. Let $\mathbf{Z} = (Z_1, \ldots, Z_p) = (f_1(X_1), \ldots, f_p(X_p))$. By the joint normality assumption of $\mathbf{Z}$, we know that $\theta_{ij} = 0$ if and only if $Z_i \perp\!\!\!\perp Z_j | \mathbf{Z} \setminus \{Z_i, Z_j\}$. Interestingly, we have that

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

Therefore, a sparse $\mathbf{\Theta}$ can be directly translated into a sparse graphical model for presenting the original variables.

In this work we primarily focus on estimating $\mathbf{\Theta}$ which is then used to construct a nonparanormal graphical model. As for the nonparametric transformation function, by the expression $f_j = \Phi^{-1} \circ F_j$, we have a natural estimator for the transformation function of the jth variable as $\hat{f}_j = \Phi^{-1} \circ \hat{F}_j^+$ where $\hat{F}_j^+$ is a Winsorized empirical CDF of the jth variables. Note that the Winsorization is used

to avoid infinity value and to achieve better bias-variance tradeoff; see Liu, Lafferty and Wasserman (2009) for detailed discussion. In this paper we show that we can directly estimate $\mathbf{\Theta}$ without estimating these nonparametric transformation functions at all. This statement seems to be a bit surprising because a natural estimation scheme is a two-stage procedure: first estimate f_j and then apply a well-developed sparse Gaussian graphical model estimation method to the transformed data $\hat{\mathbf{z}}_i = \hat{\mathbf{f}}(\mathbf{x}_i)$, $1 \le i \le n$. Liu, Lafferty and Wasserman (2009) have actually studied this "plug-in" estimation approach. They proposed a Winsorized estimator of the nonparametric transformation function and used the graphical lasso in the second stage. They established convergence rate of the "plug-in" estimator when p is restricted to a polynomial order of n. However, Liu, Lafferty and Wasserman (2009) did not get a satisfactory rate of convergence for the "plug-in" approach, because the rate of convergence can be established for the Gaussian graphical model even when p grows with n almost exponentially fast [Ravikumar et al. (2011)]. As noted in Liu, Lafferty and Wasserman (2009), it is very challenging, if not impossible, to push the theory of the "plug-in" approach to handle exponentially large dimensions. One might ask if using a better estimator for the transformation functions could improve the rate of convergence such that p could be allowed to be nearly exponentially large relative to n. This is a legitimate direction for research. We do not pursue this direction in this work. Instead, we show that we could use a rank-based estimation approach to achieve the exact same goal without estimating these transformation functions at all.

Our estimator is constructed in two steps. First, we propose using the adjusted Spearman's rank correlation to get a nonparametric sample estimate of $\mathbf{\Sigma}$. As the second step, we compute a sparse estimator $\mathbf{\Theta}$ from the rank-based sample estimate of $\mathbf{\Sigma}$. For that purpose, we consider several regularized rank estimators, including the *rank-based graphical lasso*, the *rank-based neighborhood Dantzig selector* and the *rank-based CLIME*. The complete methodological details are presented in Section 2. In Section 3 we establish theoretical properties of the proposed rank-based estimators, regarding both precision matrix estimation and graphical model selection. In particular, we are motivated by the theory to consider the adaptive version of the rank-based neighborhood Dantzig selector and the rank-based CLIME, which can select the true support set with an overwhelming probability without assuming a stringent irrepresentable condition required for the oracle and rank-based graphical lasso. Section 4 contains numerical results and Section 5 has some concluding remarks. Technical proofs are presented in an Appendix.

A referee informed us in his/her review report that Liu et al. (2012) also independently used the rank-based correlation in the context of nonparametric Gaussian graphical model estimation. A major focus in Liu et al. (2012) is the numerical demonstration of the robustness property of the rank-based methods using both Spearman's rho and Kendall's tau when data are contaminated. In the present paper we provide a systematic analysis of the rank-based estimators, and our theoretical analysis further leads to the rank-based adaptive Dantizg selector and the

rank-based adaptive CLIME in order to achieve improved sparsity recovery properties. Our theoretical analysis of the rank-based adaptive Dantizg selector is of independent interest. Although the theory is established for the rank-based estimators using Spearman's rho, the same analysis can be easily adopted to prove the theoretical properties of the rank-based estimators using Kendall's tau rank correlation.

2. Methodology. We first introduce some necessary notation. For a matrix $\mathbf{A} = (a_{ij})$, we define its entry-wise ℓ_1 norm as $\|\mathbf{A}\|_1 = \sum_{(i,j)} |a_{ij}|$, and its entry-wise ℓ_∞ norm as $\|\mathbf{A}\|_{\max} = \max_{(i,j)} |a_{ij}|$. For a vector $\mathbf{v} = (v_1, \ldots, v_l)$, we define its ℓ_1 norm as $\|\mathbf{v}\|_{\ell_1} = \sum_j |v_j|$ and its ℓ_∞ norm as $\|\mathbf{v}\|_{\ell_\infty} = \max_j |v_j|$. To simplify notation, define $\mathbf{M}_{A,B}$ as the sub-matrix of $\mathbf{M}$ with row indexes A and column indexes B, and define $\mathbf{v}_A$ as the sub-vector of $\mathbf{v}$ with indexes A. Let (k) be the index set $\{1, \ldots, k-1, k+1, \ldots, p\}$. Denote by $\mathbf{\Sigma}_{(k)} = \mathbf{\Sigma}_{(k),(k)}$ the sub-matrix of $\mathbf{\Sigma}$ with both kth row and column removed, and denote by $\boldsymbol{\sigma}_{(k)} = \mathbf{\Sigma}_{(k),k}$ the vector including all the covariances associated with the kth variable. In the same fashion, we can also define $\mathbf{\Theta}_{(k)}$, $\boldsymbol{\theta}_{(k)}$, and so on.

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

- *The oracle graphical lasso.* Let $\hat{\mathbf{\Sigma}}^o$ be the sample covariance matrix for the "oracle" data, and then the "oracle" log-profile-likelihood becomes $\log\det(\mathbf{\Theta}) - \operatorname{tr}(\hat{\mathbf{\Sigma}}^o \mathbf{\Theta})$. The "oracle" graphical lasso solves the following ℓ_1 penalized likelihood problem:

$$\min_{\mathbf{\Theta} \succ 0} -\log\det(\mathbf{\Theta}) + \operatorname{tr}\big(\hat{\mathbf{\Sigma}}^o \mathbf{\Theta}\big) + \lambda \sum_{i \neq j} |\theta_{ij}|. \tag{2}$$

- *The oracle neighborhood lasso selection.* Under the nonparanormal model, for each $k = 1, \ldots, p$, the "oracle" variable Z_k given $\mathbf{Z}_{(k)}$ is normally distributed as $N(\mathbf{Z}_{(k)}^T \mathbf{\Sigma}_{(k)}^{-1} \boldsymbol{\sigma}_{(k)}, 1 - \boldsymbol{\sigma}_{(k)}^T \mathbf{\Sigma}_{(k)}^{-1} \sigma_{(k)})$, which can be written as $Z_k = \mathbf{Z}_{(k)}^T \boldsymbol{\beta}_k + \varepsilon_k$ with $\boldsymbol{\beta}_k = \mathbf{\Sigma}_{(k)}^{-1} \boldsymbol{\sigma}_{(k)}$ and $\varepsilon_k \sim N(0, 1 - \boldsymbol{\sigma}_{(k)}^T \mathbf{\Sigma}_{(k)}^{-1} \boldsymbol{\sigma}_{(k)})$. Notice that $\boldsymbol{\beta}_k$ and ε_k are closely related to the precision matrix $\mathbf{\Theta}$, that is, $\theta_{kk} = 1/\operatorname{Var}(\varepsilon_k)$ and $\boldsymbol{\theta}_{(k)} = -\boldsymbol{\beta}_k / \operatorname{Var}(\varepsilon_k)$. Thus for the kth variable, $\boldsymbol{\theta}_{(k)}$ and $\boldsymbol{\beta}_k$ share the same sparsity pattern. Following Meinshausen and Bühlmann (2006), the oracle neighborhood lasso selection obtains the solution $\hat{\boldsymbol{\beta}}_k^o$ from the following lasso penalized least squares problem:

$$\min_{\boldsymbol{\beta} \in \mathbb{R}^{p-1}} \frac{1}{n} \sum_{i=1}^{n} \big(z_{ik} - \mathbf{z}_{i(k)}^T \boldsymbol{\beta}\big)^2 + \lambda \|\boldsymbol{\beta}\|_{\ell_1}, \tag{3}$$

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

We notice the fact that

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

Then (3) can be written in the following equivalent form:

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^{p-1}} \boldsymbol{\beta}^T\hat{\mathbf{\Sigma}}_{(k)}^o\boldsymbol{\beta} - 2\boldsymbol{\beta}^T\hat{\boldsymbol{\sigma}}_{(k)}^o + \lambda\|\boldsymbol{\beta}\|_{\ell_1}. \tag{4}$$

- *The oracle neighborhood Dantzig selector.* Following Yuan (2010) the lasso least squares in (3) can be replaced with the Dantzig selector

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^{p-1}} \|\boldsymbol{\beta}\|_{\ell_1} \qquad \text{subject to } \left\|\frac{1}{n}\sum_{i=1}^{n}\mathbf{z}_{i(k)}\big(\mathbf{z}_{i(k)}^T\boldsymbol{\beta} - z_{ik}\big)\right\|_{\ell_\infty} \leq \lambda. \tag{5}$$

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

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

Then (5) can be written in the following equivalent form:

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^{p-1}} \|\boldsymbol{\beta}\|_{\ell_1} \qquad \text{subject to } \|\hat{\mathbf{\Sigma}}_{(k)}^o\boldsymbol{\beta} - \hat{\boldsymbol{\sigma}}_{(k)}^o\|_{\ell_\infty} \leq \lambda. \tag{6}$$

- *The oracle CLIME.* Following Cai, Liu and Luo (2011) we can estimate precision matrices by solving a constrained ℓ_1 minimization problem,

$$\arg\min_{\mathbf{\Theta}} \|\mathbf{\Theta}\|_1 \qquad \text{subject to } \|\hat{\mathbf{\Sigma}}^o\mathbf{\Theta} - \mathbf{I}\|_{\max} \leq \lambda. \tag{7}$$

Cai, Liu and Luo (2011) compared the CLIME with the graphical lasso, and showed that the CLIME enjoys nice theoretical properties without assuming the irrepresentable condition of Ravikumar et al. (2011) for the graphical lasso.

2.2. *The proposed rank-based estimators.* The existing theoretical results in the literature can be directly applied to these oracle estimators. However, the "oracle data" $\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_n$ are unavailable and thus the above-mentioned "oracle" procedures are not genuine estimators. Naturally we wish to construct a genuine estimator that can mimic the oracle estimator. To this end, we can derive an alternative estimator of $\mathbf{\Sigma}$ based on the actual data $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ and then feed this genuine covariance estimator to the graphical lasso, the neighborhood selection or CLIME. To implement this natural idea, we propose a rank-based estimation scheme. Note

that $\mathbf{\Sigma}$ can be viewed as the correlation matrix as well, that is, $\sigma_{ij} = \mathrm{corr}(\mathbf{z}_i, \mathbf{z}_j)$. Let $(x_{1i}, x_{2i}, \ldots, x_{ni})$ be the observed values of variable X_i. We convert them to ranks denoted by $\mathbf{r}_i = (r_{1i}, r_{2i}, \ldots, r_{ni})$. Spearman's rank correlation $\hat{r}_{ij}$ is defined as Pearson's correlation between $\mathbf{r}_i$ and $\mathbf{r}_j$. Spearman's rank correlation is a nonparametric measure of dependence between two variables. It is important to note that $\mathbf{r}_i$ are the ranks of the "oracle" data. Therefore, $\hat{r}_{ij}$ is also identical to the Spearman's rank correlation between the "oracle" variables Z_i, Z_j. In other words, in the framework of rank-based estimation, we can treat the observed data as the "oracle" data and avoid estimating p nonparametric transformation functions. We make a note here that one may consider other rank correlation measures such as Kendall's tau correlation. To fix the idea we use Spearman's rank correlation throughout this paper.

The nonparanormal model implies that (Z_i, Z_j) follows a bivariate normal distribution with correlation parameter σ_{ij}. Then a classical result due to Kendall (1948) [see also Kruskal (1958)] shows that

$$\lim_{n\to+\infty} \mathrm{E}(\hat{r}_{ij}) = \frac{6}{\pi} \arcsin\left(\frac{1}{2}\sigma_{ij}\right), \tag{8}$$

which indicates that $\hat{r}_{ij}$ is a biased estimator of σ_{ij}. To correct the bias, Kendall (1948) suggested using the adjusted Spearman's rank correlation

$$\hat{r}^s_{ij} = 2\sin\left(\frac{\pi}{6}\hat{r}_{ij}\right). \tag{9}$$

Combining (8) and (9) we see that $\hat{r}^s_{ij}$ is an asymptotically unbiased estimator of σ_{ij}. Naturally we define the rank-based sample estimate of $\mathbf{\Sigma}$ as follows:

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

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

- *The rank-based graphical lasso*:

$$\hat{\mathbf{\Theta}}^s_g = \arg\min_{\mathbf{\Theta}\succ 0} -\log\det(\mathbf{\Theta}) + \mathrm{tr}(\hat{\mathbf{R}}^s\mathbf{\Theta}) + \lambda\sum_{i\neq j}|\theta_{ij}|. \tag{10}$$

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

$$\hat{\boldsymbol{\beta}}^{s.nd}_k = \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^{p-1}} \|\boldsymbol{\beta}\|_{\ell_1} \qquad \text{subject to } \|\hat{\mathbf{R}}^s_{(k)}\boldsymbol{\beta} - \hat{\mathbf{r}}^s_{(k)}\|_{\ell_\infty} \le \lambda. \tag{11}$$

  The support of $\mathbf{\Theta}$ can be estimated from the support of $\hat{\boldsymbol{\beta}}^{s.nd}_1, \ldots, \hat{\boldsymbol{\beta}}^{s.nd}_p$ via aggregation by union or intersection. We can also construct the rank-based precision matrix estimator $\hat{\mathbf{\Theta}}^s_{nd} = (\hat{\theta}^{s.nd}_{ij})_{1\le i,j\le p}$ with

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

and

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

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

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

Theoretical analysis of the rank-based neighborhood Dantzig selector in Section 3.2 motivated us to consider using the adaptive Dantzig selector in the rank-based neighborhood estimation in order to achieve better graphical model selection performance. See Section 3.2 for more details.

- *The rank-based CLIME*:

$$\hat{\boldsymbol{\Theta}}_{c}^{s} = \arg\min_{\boldsymbol{\Theta}} \|\boldsymbol{\Theta}\|_1 \qquad \text{subject to } \|\hat{\mathbf{R}}^{s}\boldsymbol{\Theta} - \mathbf{I}\|_{\max} \le \lambda. \tag{12}$$

Let $\mathbf{e}_k$'s be the natural basis in $\mathbb{R}^p$. By Lemma 1 in Cai, Liu and Luo (2011) the above optimization problem can be further decomposed into p subproblems of vector minimization,

$$\hat{\boldsymbol{\theta}}_{k}^{s.c} = \arg\min_{\boldsymbol{\theta}\in\mathbb{R}^p} \|\boldsymbol{\theta}\|_{\ell_1} \qquad \text{subject to } \|\hat{\mathbf{R}}^{s}\boldsymbol{\theta} - \mathbf{e}_k\|_{\ell_\infty} \le \lambda, \tag{13}$$

for $k = 1, \ldots, p$. Then $\hat{\boldsymbol{\Theta}}_{c}^{s}$ is exactly equivalent to $(\hat{\boldsymbol{\theta}}_{1}^{s.c}, \ldots, \hat{\boldsymbol{\theta}}_{p}^{s.c})$. Note that $\hat{\boldsymbol{\Theta}}_{c}^{s}$ could be asymmetric. Following Cai, Liu and Luo (2011) we consider

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

with $\breve{\theta}_{ij}^{s.c} = \hat{\theta}_{ij}^{s.c} I_{\{|\hat{\theta}_{ij}^{s.c}| \le |\hat{\theta}_{ji}^{s.c}|\}} + \hat{\theta}_{ji}^{s.c} I_{\{|\hat{\theta}_{ij}^{s.c}| > |\hat{\theta}_{ji}^{s.c}|\}}$. In the original paper Cai, Liu and Luo (2011) proposed to use hard thresholding for graphical model selection. Borrowing the basic idea from Zou (2006), we propose an adaptive version of the rank-based CLIME in order to achieve better graphical model selection. See Section 3.3 for more details.

2.3. *Rank-based neighborhood lasso?* One might consider the rank-based neighborhood lasso defined as follows:

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^{p-1}} \boldsymbol{\beta}^T\hat{\mathbf{R}}_{(k)}^{s}\boldsymbol{\beta} - 2\boldsymbol{\beta}^T\hat{\mathbf{r}}_{(k)}^{s} + \lambda\|\boldsymbol{\beta}\|_{\ell_1}. \tag{14}$$

However, there is a technical problem for the above definition. The Spearman's rank correlation matrix $\hat{\mathbf{R}}$ is always positive semidefinite, but the adjusted correlation matrix $\hat{\mathbf{R}}^{s}$ could become indefinite. To our best knowledge, Devlin, Gnanadesikan and Kettenring (1975) were the first to point out the indefinite issue of the

estimated rank correlation matrix. Here we also use a toy example to illustrate this point. Consider the 3×3 correlation matrix

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

Note that $\mathbf{A}$ is positive-definite with eigenvalues 1.99, 1.00 and 0.01, but $2\sin(\frac{\pi}{6}\mathbf{A})$ becomes indefinite with eigenvalues 2.01, 1.00 and -0.01. The negative eigenvalues will make (14) an ill-defined optimization problem. Fortunately, the positive definite issue does not cause any problem for the graphical lasso, Dantzig selector and CLIME. Notice that the diagonal elements of $\hat{\mathbf{R}}^s$ are obviously strictly positive, and thus Lemma 3 in Ravikumar et al. (2011) suggests that the rank-based graphical lasso always has a unique positive definite solution for any regularization parameter $\lambda > 0$. The rank-based neighborhood Dantzig selector and the rank-based CLIME are still well defined, even when $\hat{\mathbf{R}}^s_{(k)}$ becomes indefinite, and the according optimization algorithms also tolerate the indefiniteness of $\hat{\mathbf{R}}^s_{(k)}$. One might consider a positive definite correction of $\hat{\mathbf{R}}^s$ for implementing the neighborhood lasso estimator. However, the resulting estimator shall behave similarly to the rank-based neighborhood Dantzig selector because the lasso penalized least squares and Dantzig selector, in general, work very similarly [Bickel, Ritov and Tsybakov (2009), James, Radchenko and Lv (2009)].

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

In this section we establish theoretical properties for the proposed rank-based estimators. The main conclusion drawn from these theoretical results is that the rank-based graphical lasso, neighborhood Dantzig selector and CLIME work as well as their oracle counterparts in terms of the rates of convergence. We first provide useful concentration bounds concerning the accuracy of the rank-based sample correlation matrix.

LEMMA 1. *Fix any* $0 < \varepsilon < 1$, *and let* $n \geq \frac{12\pi}{\varepsilon}$. *Then there exists some absolute constant* $c_0 > 0$, *and we have the following concentration bounds*:

$$\Pr\bigl(|\hat{r}_{ij}^s - \sigma_{ij}| > \varepsilon\bigr) \leq 2\exp\bigl(-c_0 n\varepsilon^2\bigr),$$

$$\Pr\bigl(\bigl\|\hat{\mathbf{R}}^s - \boldsymbol{\Sigma}\bigr\|_{\max} > \varepsilon\bigr) \leq p^2\exp\bigl(-c_0 n\varepsilon^2\bigr).$$

Lemma 1 is a key ingredient of our theoretical analysis. It basically shows that the rank-based sample estimator of $\boldsymbol{\Sigma}$ works as well as the usual sample covariance estimator of $\boldsymbol{\Sigma}$ based on the "oracle data."

3.1. *Rank-based graphical lasso.* Denote by $\psi_{\min} = \min_{(i,j)\in\mathcal{A}} |\theta_{ij}^*|$ the minimal entry of $\boldsymbol{\Theta}^*$ in the absolute scale. Define $K_{\boldsymbol{\Sigma}^*} = \|\boldsymbol{\Sigma}_{\mathcal{A}\mathcal{A}}^*\|_{\ell_\infty}$ and $K_{\mathbf{H}^*} = \|(\mathbf{H}_{\mathcal{A}\mathcal{A}}^*)^{-1}\|_{\ell_\infty}$. Define $\mathbf{H}^*$ as the Kronecker product $\boldsymbol{\Sigma}^* \otimes \boldsymbol{\Sigma}^*$.

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

(a) *Element-wise maximal bound*: *if* λ *is chosen such that*

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

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

$$\bigl\|\hat{\boldsymbol{\Theta}}_g^s - \boldsymbol{\Theta}^*\bigr\|_{\max} \leq 2K_{\mathbf{H}^*}\biggl(1 + \frac{\kappa}{4}\biggr)\lambda.$$

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

$$\lambda < \min\biggl(\frac{\psi_{\min}}{2(1+\kappa/4)K_{\mathbf{H}^*}}, \frac{d^{-1}}{6(1+\kappa/4)K_{\boldsymbol{\Sigma}^*}K_{\mathbf{H}^*} \cdot \max\{1, (1+4/\kappa)K_{\boldsymbol{\Sigma}^*}^2 K_{\mathbf{H}^*}\}}\biggr),$$

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

In Theorem 1, the condition $\|\mathbf{H}_{\mathcal{A}^c\mathcal{A}}^*(\mathbf{H}_{\mathcal{A}\mathcal{A}}^*)^{-1}\|_{\ell_\infty} < 1 - \kappa$ is also referred as the *irrepresentable condition* for studying the theoretical properties of the graphical lasso [Ravikumar et al. (2011)]. We can obtain a straightforward understanding of Theorem 1 by considering its asymptotic consequences.

COROLLARY 1. *Assume that there is a constant* $\kappa \in (0, 1)$ *such that* $\|\mathbf{H}_{\mathcal{A}^c\mathcal{A}}^*(\mathbf{H}_{\mathcal{A}\mathcal{A}}^*)^{-1}\|_{\ell_\infty} < 1 - \kappa$. *Suppose that* $K_{\boldsymbol{\Sigma}^*}$ *and* $K_{\mathbf{H}^*}$ *are both fixed constants*.

(a) *Rates of convergence*: *assume* $n \gg d^2 \log p$, *and pick a regularization parameter* λ *such that* $d^{-1} \gg \lambda = O((\log p/n)^{1/2})$. *Then we have*

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

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

$$\|\hat{\boldsymbol{\Theta}}_g^s - \boldsymbol{\Theta}^*\|_F = O_P\left(\sqrt{\frac{(s+p)\log p}{n}}\right),$$

$$\|\hat{\boldsymbol{\Theta}}_g^s - \boldsymbol{\Theta}^*\|_{\ell_1} = O_P\left(\sqrt{\frac{\min\{s+p, d^2\}\log p}{n}}\right).$$

(b) *Graphical model selection consistency*: *assume* $\psi_{\min}$ *is also fixed and* $n \gg d^2 \log p$. *Pick a* λ *such that* $d^{-1} \gg \lambda = O((\log p/n)^{1/2})$. *Then we have* $\operatorname{sign}(\hat{\theta}_{ij}^{s,g}) = \operatorname{sign}(\theta_{ij}^*)$, $\forall (i,j) \in \mathcal{A}$ *and* $\operatorname{sign}(\hat{\theta}_{ij}^{s,g}) = 0$, $\forall (i,j) \in \mathcal{A}^c$.

Under the same conditions of Theorem 1 and Corollary 1, by the results in Ravikumar et al. (2011), we know that the conclusions of Theorem 1 and Corollary 1 hold for the oracle graphical lasso. In other words, the rank-based graphical lasso estimator is comparable to its oracle counterpart in terms of rates of convergence.

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

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

Likewise we can partition $\boldsymbol{\sigma}_{(k)}^*$ and $\boldsymbol{\Sigma}_{(k)}^*$ with respect to $\mathcal{A}_k$.

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

$$\|\check{\boldsymbol{\Theta}}_{nd}^s - \boldsymbol{\Theta}^*\|_{\ell_1} \leq 2\|\hat{\boldsymbol{\Theta}}_{nd}^s - \boldsymbol{\Theta}^*\|_{\ell_1} \leq C_{b,B,M} d\lambda.$$

COROLLARY 2. *Suppose that* b, B *and* M *are all fixed. Let* $n \gg d^2 \log p$, *and pick* λ *such that* $d^{-1} \gg \lambda = O((\log p/n)^{1/2})$. *Then we have*

$$\|\check{\boldsymbol{\Theta}}_{nd}^s - \boldsymbol{\Theta}^*\|_{\ell_1} = O_P\left(d\sqrt{\frac{\log p}{n}}\right) \quad \text{and} \quad \|\hat{\boldsymbol{\Theta}}_{nd}^s - \boldsymbol{\Theta}^*\|_{\ell_1} = O_P\left(d\sqrt{\frac{\log p}{n}}\right).$$

Yuan (2010) established the rates of convergence of the neighborhood Dantzig selector under the ℓ_1 norm, which can be directly applied to the oracle neighborhood Dantzig selector under the nonparanormal model. Comparing Theorem 2 and Corollary 2 to the results in Yuan (2010), we see that the rank-based neighborhood Dantzig selector and the oracle neighborhood Dantzig selector achieve the same rates of convergence.

Dantzig selector and the lasso are closely related [Bickel, Ritov and Tsybakov (2009), James, Radchenko and Lv (2009)]. Similarly to the lasso, the Dantzig selector tends to over-select. Zou (2006) proposed the adaptive weighting idea to develop the adaptive lasso which improves the selection performance of the lasso and corrects its bias too. The very same idea can be used to improve the selection performance of Dantzig selector which leads to the adaptive Dantzig selector [Dicker and Lin (2009)]. We can extend the rank-based Dantzig selector to the rank-based adaptive Dantzig selector. Given adaptive weights $\mathbf{w}_k$, consider

$$\hat{\boldsymbol{\beta}}_k^{s.nad} = \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^{p-1}} \|\mathbf{w}_k \circ \boldsymbol{\beta}\|_{\ell_1} \qquad \text{subject to } \left|\hat{\mathbf{R}}^s_{(k)}\boldsymbol{\beta} - \hat{\mathbf{r}}^s_{(k)}\right| \le \lambda \mathbf{w}_k, \tag{15}$$

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

$$\mathbf{w}_k^d = \left(|\hat{\boldsymbol{\beta}}_k^{s.nd}| + \frac{1}{n}\right)^{-1}. \tag{16}$$

Define $\boldsymbol{\beta}^*_{\mathcal{A}_k} = (\boldsymbol{\Theta}^*_{\mathcal{A}_k\mathcal{A}_k})^{-1}\boldsymbol{\theta}^*_{\mathcal{A}_k}$, and let $\boldsymbol{\beta}_k^* = (\boldsymbol{\beta}^*_{\mathcal{A}_k}, \mathbf{0})$. Thus the support of $\boldsymbol{\beta}_k^*$ exactly coincides with that of $\boldsymbol{\theta}^*_{(k)}$, and then it is further equivalent to the active set $\mathcal{A}_k$. Define $\psi_k = \|\boldsymbol{\beta}^*_{\mathcal{A}_k}\|_{\min}$, $G_k = \|(\boldsymbol{\Sigma}^*_{\mathcal{A}_k\mathcal{A}_k})^{-1}\|_{\ell_\infty}$ and $H_k = \|\boldsymbol{\Sigma}^*_{\mathcal{A}_k^c\mathcal{A}_k}(\boldsymbol{\Sigma}^*_{\mathcal{A}_k\mathcal{A}_k})^{-1}\|_{\ell_\infty}$ for $k = 1, 2, \ldots, p$. Let $C_0 = 4B^2(2 + \frac{b}{M})$.

THEOREM 3. *For each k, we pick $\lambda = \lambda_d$ as in (11) satisfying that $\lambda_d \ge \frac{12\pi M}{bn}$ and $o(1) = d\lambda_d \le \min\{\frac{\psi_k}{2C_0}, \frac{1}{4C_0 d(\psi_k+2G_k)} - \frac{1}{C_0 n}\}$, and pick $\lambda = \lambda_{ad}$ as in (15) such that $\frac{\psi_k^2}{8G_k} \ge \lambda_{ad} \ge \max\{\frac{12\pi}{n}, (C_0 d\lambda_d + \frac{1}{n})\frac{H_k\psi_k}{G_k}\}$, and $o(1) = d\lambda_{ad} \le \min\{\lambda_{\min}(\boldsymbol{\Sigma}^*_{\mathcal{A}_k\mathcal{A}_k}), \frac{1}{2G_k}, \frac{\psi_k}{8G_k(\psi_k+G_k)}\}$. In addition, we also choose $\mathbf{w}_k = \mathbf{w}_k^d$ as in (16) for each k. Then with a probability at least $1 - p^2\exp(-c_0 n \cdot \min\{\lambda_{ad}^2, \frac{b^2}{M^2}\lambda_d^2\})$, for each k, the rank-based adaptive Dantzig selector finds the unique solution $\hat{\boldsymbol{\beta}}_k^{s.nad} = (\hat{\boldsymbol{\beta}}^{s.nad}_{\mathcal{A}_k}, \hat{\boldsymbol{\beta}}^{s.nad}_{\mathcal{A}_k^c})$ with $\operatorname{sign}(\hat{\boldsymbol{\beta}}^{s.nad}_{\mathcal{A}_k}) = \operatorname{sign}(\boldsymbol{\beta}^*_{\mathcal{A}_k})$ and $\hat{\boldsymbol{\beta}}^{s.nad}_{\mathcal{A}_k^c} = \mathbf{0}$, and thus the rank-based neighborhood adaptive Dantzig selector is consistent for the graphical model selection.*

COROLLARY 3. *Suppose* b, B, M, ψ_k, G_k *and* H_k $(1 \le k \le p)$ *are all constants. Assume that* $n \gg d^4 \log p$ *and* $\lambda_{\min}(\mathbf{\Sigma}^*_{\mathcal{A}_k\mathcal{A}_k}) \gg d^2(\log p/n)^{1/2}$. *Pick the tuning parameters* λ_d *and* λ_{ad} *such that* $\frac{1}{d} \gg \lambda_d = O((\log p/n)^{1/2})$ *and* $\min\{\frac{1}{d} \cdot \lambda_{\min}(\mathbf{\Sigma}^*_{\mathcal{A}_k\mathcal{A}_k}), \frac{1}{d}\} \gg \lambda_{ad} \gg d\lambda_d$. *Then with probability tending to* 1, *for each* k, *the rank-based adaptive Dantzig selector with* $\mathbf{w}_k = \mathbf{w}_k^d$ *as in* (16) *finds the unique optimal solution* $\hat{\boldsymbol{\beta}}_k^{s.nad} = (\hat{\boldsymbol{\beta}}_{\mathcal{A}_k}^{s.nad}, \hat{\boldsymbol{\beta}}_{\mathcal{A}_k^c}^{s.nad})$ *with* $\operatorname{sign}(\hat{\boldsymbol{\beta}}_{\mathcal{A}_k}^{s.nad}) = \operatorname{sign}(\boldsymbol{\beta}^*_{\mathcal{A}_k})$ *and* $\hat{\boldsymbol{\beta}}_{\mathcal{A}_k^c}^{s.nad} = \mathbf{0}$, *and thus the rank-based neighborhood adaptive Dantzig selector is consistent for the graphical model selection.*

The sign-consistency of the adaptive Dantzig selector is similar to that of the adaptive lasso [van de Geer, Bühlmann and Zhou (2011)]. Based on Theorem 2 we construct the adaptive weights in (16) which is critical for the success of the rank-based adaptive Dantzig selector in the high-dimensional setting. It is important to mention that the rank-based adaptive Dantzig selector does not require the strong irrepresentable condition for the rank-based graphical lasso to have the sparsity recovery property. Our treatment of the adaptive Dantzig selector is fundamentally different from Dicker and Lin (2009). Dicker and Lin (2009) focused on the canonical linear regression model and constructed the adaptive weights as the inverse of the absolute values of ordinary least square estimator. Their theoretical results only hold in the classical fixed p setting. In our problem p can be much bigger than n. The choice of adaptive weights in (16) plays a critical role in establishing the graphical model selection consistency for the adaptive Dantzig selector under the high-dimensional setting where p is at a nearly exponential rate to n. Our technical analysis uses some key ideas such as the strong duality and the complementary slackness from the linear optimization theory [Bertsimas and Tsitsiklis (1997), Boyd and Vandenberghe (2004)].

3.3. *Rank-based CLIME.* Compared to the graphical lasso, the CLIME can enjoy nice theoretical properties without assuming the irrepresentable condition [Cai, Liu and Luo (2011)]. This continues to hold when comparing the rank-based graphical lasso and the rank-based CLIME.

THEOREM 4. *Recall that* $M = \|\mathbf{\Theta}^*\|_{\ell_1}$. *Pick a regularizing parameter* λ *such that* $n\lambda \ge 12\pi M$. *With a probability at least* $1 - p^2 \exp(-\frac{c_0}{M^2} n\lambda^2)$,

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

Moreover, assume that $n \gg d^2 \log p$, *and suppose* M *is a fixed constant. Pick a regularization parameter* λ *satisfying* $\lambda = O((\log p/n)^{1/2})$. *Then we have*

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

Theorem 4 is parallel to Theorem 6 in Cai, Liu and Luo (2011) which can be used to establish the rate of convergence of the oracle CLIME.

To improve graphical model selection performance, Cai, Liu and Luo (2011) suggested an additional thresholding step by applying the element-wise hard-thresholding rule to $\hat{\mathbf{\Theta}}_c^s$,

$$\mathrm{HT}(\hat{\mathbf{\Theta}}_c^s) = \big(\hat{\theta}_{ij}^{s,c} \cdot I_{\{|\hat{\theta}_{ij}^{s,c}| \geq \tau_n\}}\big)_{1 \leq i,j \leq p}, \tag{17}$$

where $\tau_n \geq 2M\lambda$ is the threshold, and λ is given in Theorem 4. Here we show that consistent graphical model selection can be achieved by an adaptive version of the rank-based CLIME. Given an adaptive weight matrix $\mathbf{W}$ we define the rank-based adaptive CLIME as follows:

$$\hat{\mathbf{\Theta}}_{ac}^s = \arg\min_{\mathbf{\Theta}} \|\mathbf{W} \circ \mathbf{\Theta}\|_1 \qquad \text{subject to } \big|\hat{\mathbf{R}}^s \mathbf{\Theta} - \mathbf{I}\big| \leq \lambda \mathbf{W}, \tag{18}$$

where $\mathbf{A}_{p\times p} \leq \mathbf{B}_{p\times p}$ is a simplified expression for the set of inequalities $a_{ij} \leq b_{ij}$ (for all $1 \leq i, j \leq p$). Write $\mathbf{W} = (\mathbf{w}_1, \ldots, \mathbf{w}_p)$. By Lemma 1 in Cai, Liu and Luo (2011) the above linear programming problem in (18) is exactly equivalent to p vector minimization subproblems,

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

In both our theory and implementation, we utilize the rank-based CLIME's optimal solution $\hat{\mathbf{\Theta}}_c^s$ to construct an adaptive weight matrix $\mathbf{W}$ by

$$\mathbf{W}^c = \left(|\hat{\mathbf{\Theta}}_c^s| + \frac{1}{n}\right)^{-1}. \tag{19}$$

We now prove the graphical model selection consistency of the rank-based adaptive CLIME. Denote $\mathbf{\Theta}^*$ as $(\boldsymbol{\theta}_1^*, \ldots, \boldsymbol{\theta}_p^*)$, and define $\tilde{\mathcal{A}}_k = \mathcal{A}_k \cup \{k\}$. Then we can organize $\boldsymbol{\theta}_k^*$ and $\mathbf{\Sigma}^*$ with respect to $\tilde{\mathcal{A}}_k$ and $\tilde{\mathcal{A}}_k^c$. For $k = 1, 2, \ldots, p$, define $\tilde{G}_k = \|(\mathbf{\Sigma}^*_{\tilde{\mathcal{A}}_k \tilde{\mathcal{A}}_k})^{-1}\|_{\ell_\infty}$ and $\tilde{H}_k = \|\mathbf{\Sigma}^*_{\tilde{\mathcal{A}}_k^c \tilde{\mathcal{A}}_k} (\mathbf{\Sigma}^*_{\tilde{\mathcal{A}}_k \tilde{\mathcal{A}}_k})^{-1}\|_{\ell_\infty}$.

THEOREM 5. *Recall* $\psi_{\min} = \min_{(i,j)\in\mathcal{A}} |\theta_{ij}^*|$. *For each* k *pick* $\lambda = \lambda_c$ *as in* (12) *such that* $\min\{\frac{\psi_{\min}}{4M}, \frac{1}{4M(\psi_{\min} + 2\tilde{G}_k)d} - \frac{2}{Mn}\} \geq \lambda_c \geq \frac{12\pi M}{n}$ *and* $d\lambda_c = o(1)$, *and we further pick the regularization parameter* $\lambda = \lambda_{ac}$ *as in* (18) *satisfying that* $\frac{\psi_{\min}^2}{8\tilde{G}_k} \geq \lambda_{ac} \geq \max\{12\pi/n, (2M\lambda_c + \frac{1}{n})\frac{\tilde{H}_k \psi_{\min}}{\tilde{G}_k}\}$ *and* $o(1) = d\lambda_{ac} \leq \min\{\lambda_{\min}(\mathbf{\Sigma}^*_{\mathcal{A}_k \mathcal{A}_k}), \frac{1}{2\tilde{G}_k}, \frac{\psi_{\min}}{4\tilde{G}_k(\psi_{\min} + \tilde{G}_k)}\}$. *In addition we choose* $\mathbf{W} = \mathbf{W}^c$ *as in* (19). *With a probability at least* $1 - p^2 \exp(-c_0 n \min\{\lambda_{ac}^2, \frac{1}{M^2}\lambda_c^2\})$, *the rank-based adaptive CLIME's optimal solution* $\hat{\mathbf{\Theta}}_{ac}^s$ *is sign consistent, that is,* $\operatorname{sign}(\hat{\theta}_{ij}^{s,ac}) = \operatorname{sign}(\theta_{ij}^*)$ *for* $(i, j) \in \mathcal{A}$ *and* $\operatorname{sign}(\hat{\theta}_{ij}^{s,ac}) = 0$ *for* $(i, j) \in \mathcal{A}^c$.

COROLLARY 4. *Suppose* $\psi_{\min}$, M, $\tilde{G}_k$ *and* $\tilde{H}_k$ $(1 \le k \le p)$ *are all constants. Assume that* $n \gg d^2 \log p$ *and* $\lambda_{\min}(\mathbf{\Sigma}^*_{\tilde{\mathcal{A}}_k \tilde{\mathcal{A}}_k}) \gg d(\log p/n)^{1/2}$. *Pick the regularization parameters* λ_c *and* λ_{ac} *such that* $\frac{1}{d} \ge \lambda_c = O((\log p/n)^{1/2})$, *and* $\min\{\lambda_{\min}(\mathbf{\Sigma}^*_{\tilde{\mathcal{A}}_k \tilde{\mathcal{A}}_k})/d, \frac{1}{d}\} \gg \lambda_{ac} \gg \lambda_c$. *Let* $\mathbf{W} = \mathbf{W}^c$ *as in* (19). *Then with probability tending to* 1, *the rank-based adaptive CLIME's optimal solution* $\hat{\mathbf{\Theta}}^s_{ac}$ *is sign consistent for the graphical model selection, that is,* $\operatorname{sign}(\hat{\theta}^{s,ac}_{ij}) = \operatorname{sign}(\theta^*_{ij})$ *for* $(i, j) \in \mathcal{A}$ *and* $\operatorname{sign}(\hat{\theta}^{s,ac}_{ij}) = 0$ *for* $(i, j) \in \mathcal{A}^c$.

The nice theoretical property of the rank-based CLIME allows us to construct the adaptive weights in (19), which is critical for establishing the graphical model selection consistency for the rank-based adaptive CLIME estimator in the high-dimensional setting without the strong ir-representable condition.

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

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

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

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

Model 3: Randomly choose 16 nodes to be the hub nodes in $\mathbf{\Theta}$, and each of them connects with 5 distinct nodes with $\mathbf{\Theta}_{ij} = 0.2$. Elements, not associated with hub nodes, are set as 0 in $\mathbf{\Theta}$. The diagonal element σ is chosen similarly as that in the previous model.

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

In models 1b–4b we first generate n independent data from $N_p(0, \mathbf{\Sigma})$ and then transfer the normal data using transformation functions

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

where $f_1(x) = x$, $f_2(x) = \log(x)$, $f_3(x) = x^{\frac{1}{3}}$, $f_4(x) = \log(\frac{x}{1-x})$ and $f_5(x) = f_2(x)I_{\{x<-1\}} + f_1(x)I_{\{-1\le x\le 1\}} + (f_4(x-1)+1)I_{\{x>1\}}$. In all cases we let $n = 300$ and $p = 100$.

Table 2 summarizes all the estimators investigated in our study. For each estimator, the tuning parameter is chosen by cross-validation. Estimation accuracy is

TABLE 2
List of all estimators in the numercial study

Notation	Details
GLASSO	Penalized likelihood estimation via graphical lasso
MB	Neighborhood lasso [Meinshausen and Bühlmann (2006)]
MB.au (or MB.ai)	MB + aggregation by union (or by intersection)
NDS	Neighborhood selection via Dantzig selector
NDS.au (or NDS.ai)	NDS + aggregation by union (or by intersection)
CLIME	Constrained ℓ_1 minimization estimator
LLW	The "plug-in" extension of GLASSO [Liu, Lafferty and Wasserman (2009)]
R-GLASSO	Proposed rank-based extension of GLASSO
R-NDS	Proposed rank-based extension of NDS
R-NDS.au (or R-NDS.ai)	R-NDS + aggregation by union (or by intersection)
R-NADS	Proposed rank-based adaptive extension of R-NDS
R-NADS.au (or R-NADS.ai)	R-NADS + aggregation by union (or by intersection)
R-CLIME	Proposed rank-based extension of CLIME
R-ACLIME	Proposed rank-based adaptive extension of CLIME

measured by the average matrix ℓ_2-norm over 100 independent replications, and selection accuracy is evaluated by the average false positive/negative.

The simulation results are summarized in Tables 3–6. First of all, we can see that the graphical lasso, neighborhood selection and CLIME do not have satisfac-

TABLE 3
Estimation performance in the Gaussian model. Estimation accuracy is measured by the matrix ℓ_2-norm with standard errors in the bracket

Method	Model 1	Model 2	Model 3	Model 4
GLASSO	0.74	1.23	0.67	0.63
	(0.01)	(0.02)	(0.01)	(0.01)
LLW	0.84	1.28	0.68	0.67
	(0.01)	(0.02)	(0.01)	(0.01)
R-GLASSO	0.81	1.30	0.64	0.70
	(0.01)	(0.02)	(0.01)	(0.01)
NDS	0.78	1.25	0.61	0.57
	(0.01)	(0.02)	(0.01)	(0.01)
R-NDS	0.81	1.28	0.63	0.62
	(0.01)	(0.02)	(0.01)	(0.01)
CLIME	0.71	1.19	0.54	0.59
	(0.01)	(0.02)	(0.01)	(0.01)
R-CLIME	0.79	1.27	0.58	0.61
	(0.01)	(0.02)	(0.01)	(0.01)

TABLE 4
Estimation performance in the nonparanormal model. Estimation accuracy is measured by the matrix ℓ_2-norm with standard errors in the bracket

Method	Model 1b	Model 2b	Model 3b	Model 4b
GLASSO	1.77	2.68	1.31	1.28
	(0.01)	(0.06)	(0.02)	(0.01)
LLW	0.84	1.28	0.68	0.67
	(0.01)	(0.01)	(0.01)	(0.01)
R-GLASSO	0.81	1.30	0.64	0.70
	(0.01)	(0.02)	(0.01)	(0.01)
NDS	1.41	2.42	1.16	1.13
	(0.01)	(0.03)	(0.02)	(0.02)
R-NDS	0.81	1.28	0.63	0.62
	(0.01)	(0.02)	(0.01)	(0.01)
CLIME	1.22	2.51	1.24	1.03
	(0.02)	(0.03)	(0.02)	(0.01)
R-CLIME	0.79	1.27	0.58	0.61
	(0.01)	(0.02)	(0.01)	(0.01)

tory performance under models 1b–4b due to the lack of ability to handle nonnormality. Second, the three rank-based estimators perform similarly to their oracle counterparts. Note that in models 1b–4b the oracle graphical lasso, the oracle neighborhood Dantzig and the oracle CLIME are actually the graphical lasso, the neighborhood Dantzig and the CLIME in models 1–4. In terms of precision matrix estimation the rank-based CLIME seems to be the best, while the rank-based neighborhood adaptive Dantzig selector has the best graphical model selection performance. We have also obtained the simulation results under the matrix ℓ_1-norm. The conclusions stay the same. For space consideration we leave these ℓ_1-norm results to the technical report version of this paper.

4.2. *Applications to gene expression genomics*. We illustrate our proposed rank-based estimators on a real data set to recover the isoprenoid genetic regulatory network in Arabidposis thaliana [Wille et al. (2004)]. This dataset contains the gene expression measurements of 39 genes (excluding protein GGPPS7 in the MEP pathway) assayed on $n = 118$ Affymetrix GeneChip microarrays.

We used seven estimators (GLASSO, MB, CLIME, LLW, R-GLASSO, R-NADS and R-ACLIME) to reconstruct the regulatory network. The first three estimators are performed after taking the log-transformation of the original data, and the other four estimators are directly applied to the original data. To be more conservative, we only considered the integration by union for the neighborhood selec-

TABLE 5

Selection performance in the Gaussian model. Selection accuracy is measured by counts of false negative (#FN) or false positive (#FP) with standard errors in the bracket

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

tion procedures. We generated 100 independent Bootstrap samples and computed the frequency of each edge being selected by each estimator. The final model by each method only includes edges selected by at least 80 times over 100 Bootstrap samples. We report the number of selected edges by each estimator in Table 7. The rank-based graphical lasso performs similarly to the LLW method. The rank-based adaptive CLIME produces the sparsest graphs. We also compared pairwise intersections of the selected edges among different estimators. More than 70% of the selected edges by GLASSO, MB or CLIME turn out to be validated by both LLW and R-GLASSO, and more than 40% of the selected edges by GLASSO, MB or CLIME are justified by R-NADS and R-ACLIME. The selected models support the biological arguments that the interactions between the pathways do exist although they operate independently under normal conditions [Laule et al. (2003), Rodríguez-Concepción et al. (2004)].

TABLE 6
Selection performance in the nonparanormal model. Selection accuracy is measured by counts of false negative (#FN) or false positive (#FP) with standard errors in the bracket

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

5. Discussion. Using ranks of the raw data for statistical inference is a powerful and elegant idea in the nonparametric statistics literature; see Lehmann (1998) for detailed treatment and discussion. Some classical rank-based statistical meth-

TABLE 7
The isoprenoid genetic regulatory network: counts of stable edges

	GLASSO	**Neighborhood LASSO**	**CLIME**	
# of stable edges	100	101	67	
	LLW	**R-GLASSO**	**R-NADS**	**R-ACLIME**
# of stable edges	87	88	50	52

ods include Friedman's test in analysis of variance and Wilcoxon signed-rank test. This work is devoted to the rank-based estimation of $\mathbf{\Sigma}^{-1}$ of the nonparanormal model under a strong sparsity assumption that $\mathbf{\Sigma}^{-1}$ has only a few nonzero entries, and our results show that rank-based estimation is still powerful and elegant in the new setting of high-dimensional nonparametric graphical modeling. In a separate paper, Xue and Zou (2011a) also studied the problem of optimal estimation of $\mathbf{\Sigma}$ of the nonparanormal model under a weak sparsity assumption that $\mathbf{\Sigma}$ belongs to some weak ℓ_q ball and showed that a rank-based thresholding estimator is adaptive minimax optimal under the matrix ℓ_1 norm and ℓ_2 norm.

APPENDIX: TECHNICAL PROOFS

PROOF OF THEOREM 1. Using Lemma 3 in Ravikumar et al. (2011), $\hat{\mathbf{\Theta}}_g^s \succ 0$ is uniquely characterized by the sub-differential optimality condition that $\hat{\mathbf{R}}^s - (\hat{\mathbf{\Theta}}_g^s)^{-1} + \lambda\hat{\mathbf{Z}} = \mathbf{0}$, where $\hat{\mathbf{Z}}$ is the sub-differential with respect to $\hat{\mathbf{\Theta}}_g^s$. Define the "oracle" estimator $\tilde{\mathbf{\Theta}}_g^s$ by

$$\tilde{\mathbf{\Theta}}_g^s = \mathop{\arg\min}_{\mathbf{\Theta}\succ 0, \mathbf{\Theta}_{\mathcal{A}^c}=\mathbf{0}} -\log\det(\mathbf{\Theta}) + \operatorname{tr}\big(\hat{\mathbf{\Sigma}}^o\mathbf{\Theta}\big) + \lambda\sum_{i\neq j}|\theta_{ij}|.$$

Then we can construct $\tilde{\mathbf{Z}}$ to satisfy that $\hat{\mathbf{R}}^s - (\tilde{\mathbf{\Theta}}_g^s)^{-1} + \lambda\tilde{\mathbf{Z}} = \mathbf{0}$. As in Ravikumar et al. (2011) the rest of the proof depends on a exponential-type concentration bound concerning the accuracy of the sample estimator of the correlation matrix under the entry-wise ℓ_∞ bound. Our Lemma 1 fulfills that role. With Lemma 1, Theorem 1 can be proved by following the line of the proof in Ravikumar et al. (2011). For the sake of space, we move the rest of proof to the technical report version of this paper [Xue and Zou (2011b)].

We now prove Lemma 1. First, Spearman's rank correlation $\hat{r}_{ij}$ can be written in terms of the Hoeffding decomposition [Hoeffding (1948)]

$$\hat{r}_{ij} = \frac{n-2}{n+1}u_{ij} + \frac{3}{n+1}d_{ij}, \tag{20}$$

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

$$u_{ij} = \frac{3}{n(n-1)(n-2)}\sum_{k\neq l, k\neq m, l\neq m}\operatorname{sign}(x_{ki}-x_{li})\cdot\operatorname{sign}(x_{kj}-x_{mj}). \tag{21}$$

Direct calculation yields that $\mathrm{E}(u_{ij}) = \frac{6}{\pi}\sin^{-1}(\frac{\sigma_{ij}}{2})$ [Kendall (1948)]. Then we can obtain that $\sigma_{ij} = 2\sin(\frac{\pi}{6}E(u_{ij}))$. By definition $\hat{r}_{ij}^s = 2\sin(\frac{\pi}{6}\hat{r}_{ij})$. Note that $2\sin(\frac{\pi}{6}\cdot)$ is a Lipschitz function with the Lipschitz constant $\pi/3$. Then we have

$$\Pr\big(|\hat{r}_{ij}^s - \sigma_{ij}| > \varepsilon\big) \le \Pr\bigg(|\hat{r}_{ij} - E(u_{ij})| > \frac{3\varepsilon}{\pi}\bigg).$$

Applying (20) and (21) yields $\hat{r}_{ij} - E(u_{ij}) = u_{ij} - E(u_{ij}) + \frac{3}{n+1}d_{ij} - \frac{3}{n+1}u_{ij}$. Note $|u_{ij}| \le 3$ and $|d_{ij}| \le 1$. Hence, $|u_{ij}| \le \frac{\varepsilon}{4\pi}(n+1)$ and $|d_{ij}| \le \frac{\varepsilon}{4\pi}(n+1)$ always hold

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

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

Finally, we observe that u_{ij} is a function of independent samples $(\mathbf{x}_1, \ldots, \mathbf{x}_n)$. Now we make a claim that if we replace the tth sample by some $\hat{\mathbf{x}}_t$, the change in u_{ij} will be bounded as

$$\sup_{\mathbf{x}_1,\ldots,\mathbf{x}_n,\hat{\mathbf{x}}_t} \bigl|u_{ij}(\mathbf{x}_1, \ldots, \mathbf{x}_n) - u_{ij}(\mathbf{x}_1, \ldots, \mathbf{x}_{t-1}, \hat{\mathbf{x}}_t, \mathbf{x}_{t+1}, \ldots, \mathbf{x}_n)\bigr| \le \frac{15}{n}. \tag{22}$$

Then we can apply the McDiarmid's inequality [McDiarmid (1989)] to conclude the desired concentration bound for some absolute constant $c_0 > 0$,

$$\Pr\bigl(|\hat{r}_{ij}^s - \sigma_{ij}| > \varepsilon\bigr) \le \Pr\biggl(|u_{ij} - E(u_{ij})| \ge \frac{3\varepsilon}{2\pi}\biggr) \le 2\exp\bigl(-c_0 n\varepsilon^2\bigr).$$

Now it remains to verify (22) to complete the proof of Lemma 1. We provide a brief proof for this claim. Assume that $\mathbf{x}_t = (x_{1t}, \ldots, x_{pt})'$ is replaced by $\tilde{\mathbf{x}}_t = (\tilde{x}_{1t}, \ldots, \tilde{x}_{pt})'$, and we want to prove that the change of u_{ij} is at most $15/n$. Without loss of generality we may assume that $n_i = \#\{s : \operatorname{sign}(\tilde{x}_{ti} - x_{si}) = -\operatorname{sign}(x_{ti} - x_{si}), s \ne t\}$ and also assume that $n_j = \#\{s : \operatorname{sign}(\tilde{x}_{tj} - x_{sj}) = -\operatorname{sign}(x_{tj} - x_{sj}), s \ne t\}$. Then we have

$$\begin{aligned}
&\bigl|u_{ij}(\mathbf{x}_1, \ldots, \mathbf{x}_n) - u_{ij}(\mathbf{x}_1, \ldots, \mathbf{x}_{t-1}, \tilde{\mathbf{x}}_t, \mathbf{x}_{t+1}, \ldots, \mathbf{x}_n)\bigr| \\
&\quad\le \Biggl| \sum_{k\ne t, k\ne m, m\ne t} \bigl\{\operatorname{sign}(x_{ki} - x_{ti}) - \operatorname{sign}(x_{ki} - \tilde{x}_{ti})\bigr\} \cdot \operatorname{sign}(x_{kj} - x_{mj}) \\
&\qquad + \sum_{k\ne t, k\ne l, l\ne t} \bigl(\operatorname{sign}(x_{kj} - x_{tj}) - \operatorname{sign}(x_{kj} - \tilde{x}_{tj})\bigr) \cdot \operatorname{sign}(x_{ki} - x_{li}) \\
&\qquad + \sum_{l\ne t, m\ne t, l\ne m} \bigl(\operatorname{sign}(x_{ti} - x_{li}) \cdot \operatorname{sign}(x_{tj} - x_{mj}) - \operatorname{sign}(\tilde{x}_{ti} - x_{li}) \\
&\qquad\qquad \times \operatorname{sign}(\tilde{x}_{tj} - x_{mj})\bigr)\Biggr| \\
&\qquad \times \frac{3}{n(n-1)(n-2)} \\
&\quad\le \frac{3\cdot 2\cdot [n_i(n-2) + n_j(n-2) + n_j(n-1-n_i) + n_i(n-1-n_j)]}{n(n-1)(n-2)} \\
&\quad\le \frac{12}{n}\biggl(1 + \frac{1}{4}\frac{1}{(n-1)(n-2)}\biggr) \\
&\quad\le \frac{15}{n},
\end{aligned}$$

where the third inequality holds if and only if $n_i = n_j = n - \frac{3}{2}$. $\square$

PROOF OF THEOREM 2. For space of consideration, we only show the sketch of the proof, and the detailed proof is relegated to the supplementary file [Xue and Zou (2012)] and also the technical report version of this paper [Xue and Zou (2011b)]. We begin with an important observation that we only need to prove the risk bound for $\hat{\mathbf{\Theta}}^s_{nd}$ because

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

To bound the difference between $\hat{\mathbf{\Theta}}^s_{nd}$ and $\mathbf{\Theta}^*$ under the matrix ℓ_1-norm, we only need to bound $|\hat{\theta}^{s.nd}_{kk} - \theta^*_{kk}|$ and $\|\hat{\boldsymbol{\theta}}^{s.nd}_{(k)} - \boldsymbol{\theta}^*_{(k)}\|_{\ell_1}$ for each $k = 1, \ldots, p$. To this end, we consider the probability event $\{\|\hat{\mathbf{R}}^s - \mathbf{\Sigma}^*\|_{\max} \le \frac{b}{M}\lambda\}$, and under this event, we can show that for $k = 1, \ldots, p$,

$$\|\hat{\mathbf{R}}^s_{(k)}\boldsymbol{\beta}^*_k - \hat{\mathbf{r}}^s_{(k)}\|_{\ell_\infty} \le \lambda \quad \text{and} \quad \|\hat{\boldsymbol{\beta}}^{s.nd}_k - \boldsymbol{\beta}^*_k\|_{\ell_1} \le C_0 d\lambda, \tag{23}$$

where C_0 is some quantity depending on b, B and M only.

Now we can use (23) to further bound $|\hat{\theta}^{s.nd}_{kk} - \theta^*_{kk}|$ under the same event. To this end, we first derive an upper bound for $|(\hat{\theta}^{s.nd}_{kk})^{-1} - (\theta^*_{kk})^{-1}|$ as

$$|(\hat{\theta}^{s.nd}_{kk})^{-1} - (\theta^*_{kk})^{-1}| \le \left(1 + \frac{b}{M}\right) \cdot \lambda\|\boldsymbol{\beta}^*_k\|_{\ell_1} + \|\hat{\boldsymbol{\beta}}^{s.nd}_k - \boldsymbol{\beta}^*_k\|_{\ell_1}. \tag{24}$$

Notice that $|\hat{\theta}^{s.nd}_{kk} - \theta^*_{kk}| = |(\hat{\theta}^{s.nd}_{kk})^{-1} - (\theta^*_{kk})^{-1}| \cdot |\hat{\theta}^{s.nd}_{kk}| \cdot |\theta^*_{kk}|$ and also $|\hat{\theta}^{s.nd}_{kk}| \le |\hat{\theta}^{s.nd}_{kk} - \theta^*_{kk}| + |\theta^*_{kk}|$. Then $|\hat{\theta}^{s.nd}_{kk} - \theta^*_{kk}|$ can be upper bounded by

$$\begin{aligned} |\hat{\theta}^{s.nd}_{kk} - \theta^*_{kk}| &\le \frac{|(\hat{\theta}^{s.nd}_{kk})^{-1} - (\theta^*_{kk})^{-1}| \cdot |\theta^*_{kk}|^2}{1 - |(\hat{\theta}^{s.nd}_{kk})^{-1} - (\theta^*_{kk})^{-1}| \cdot |\theta^*_{kk}|} \\ &\le \frac{B^2[(1 + b/M)(M/b)\lambda + C_0 d\lambda]}{1 - B[(1 + b/M)(M/b)\lambda + C_0 d\lambda]}. \end{aligned}$$

Since $d\lambda = o(1)$, we denote the right-hand side as $C_1 d\lambda$ for some $C_1 > 0$.

Next, we can further obtain a bound for $\|\hat{\boldsymbol{\theta}}^{s.nd}_{(k)} - \boldsymbol{\theta}^*_{(k)}\|_{\ell_1}$.

$$\begin{aligned} \|\hat{\boldsymbol{\theta}}^{s.nd}_{(k)} - \boldsymbol{\theta}^*_{(k)}\|_{\ell_1} &\le \|(\hat{\theta}^{s.nd}_{kk} - \theta^*_{kk})\hat{\boldsymbol{\beta}}^{s.nd}_k\|_{\ell_1} + \|\theta^*_{kk}(\hat{\boldsymbol{\beta}}^{s.nd}_k - \boldsymbol{\beta}^*_k)\|_{\ell_1} \\ &\le C_1 d\lambda \cdot b^{-1}M + B \cdot C_0 d\lambda. \end{aligned} \tag{25}$$

Thus we can combine (24) and (25) to derive the desired upper bound under the same event. This completes the proof of Theorem 2. $\square$

PROOF OF THEOREM 3. Throughout the proof, we consider the event

$$\left\{\|\hat{\mathbf{R}}^s - \mathbf{\Sigma}^*\|_{\max} \le \min\left(\lambda_{ad}, \frac{b}{M}\lambda_d\right)\right\}. \tag{26}$$

For ease of notation, define $\lambda_d = \lambda_0$ and $\lambda_{ad} = \lambda_1$. We focus on the proof of the sign consistency of $\hat{\boldsymbol{\beta}}^{s.nad}_k$ in the sequel.

Under event (26), $\hat{\mathbf{R}}^s_{\mathcal{A}_k\mathcal{A}_k}$ is always positive-definite. To see this, the Weyl's inequality yields $\lambda_{\min}(\hat{\mathbf{R}}^s_{\mathcal{A}_k\mathcal{A}_k}) + \lambda_{\max}(\hat{\mathbf{R}}^s_{\mathcal{A}_k\mathcal{A}_k} - \boldsymbol{\Sigma}^*_{\mathcal{A}_k\mathcal{A}_k}) \geq \lambda_{\min}(\boldsymbol{\Sigma}^*_{\mathcal{A}_k\mathcal{A}_k})$, and then we can bound the minimal eigenvalue of $\hat{\mathbf{R}}^s_{\mathcal{A}_k\mathcal{A}_k}$,

$$\lambda_{\min}\big(\hat{\mathbf{R}}^s_{\mathcal{A}_k\mathcal{A}_k}\big) \geq \lambda_{\min}\big(\boldsymbol{\Sigma}^*_{\mathcal{A}_k\mathcal{A}_k}\big) - \big\|\hat{\mathbf{R}}^s_{\mathcal{A}_k\mathcal{A}_k} - \boldsymbol{\Sigma}^*_{\mathcal{A}_k\mathcal{A}_k}\big\|_F \geq \lambda_1\big(d - \sqrt{d(d-1)}\big) > 0.$$

For each k we introduce the dual variables $\boldsymbol{\alpha}^+_k = (\alpha^+_j)_{j\neq k} \in \mathbb{R}^{p-1}_+$ and $\boldsymbol{\alpha}^-_k = (\alpha^-_j)_{j\neq k} \in \mathbb{R}^{p-1}_+$. Then the Lagrange dual function is defined as

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

where $\circ$ denotes the Hadamard product. Due to the strong duality of linear programming [Boyd and Vandenberghe (2004)], the complementary slackness condition holds for the primal problem with respect to any primal and dual solution pair $(\boldsymbol{\beta}, \boldsymbol{\alpha}^+_k, \boldsymbol{\alpha}^-_k)$, which implies that $\alpha^+_j[(\hat{\mathbf{R}}^s_{(k)}\boldsymbol{\beta} - \hat{\mathbf{r}}^s_{(k)})_j - \lambda_1 w^d_j] = 0$ and $\alpha^-_j[-(\hat{\mathbf{R}}^s_{(k)}\boldsymbol{\beta} - \hat{\mathbf{r}}^s_{(k)})_j - \lambda_1 w^d_j] = 0$ for any $j \neq k$. Observe that only one of α^+_j and α^-_j can be zero since only one of $(\hat{\mathbf{R}}^s_{(k)}\boldsymbol{\beta} - \hat{\mathbf{r}}^s_{(k)})_j = \lambda_1 w^d_j$ and $(\hat{\mathbf{R}}^s_{(k)}\boldsymbol{\beta} - \hat{\mathbf{r}}^s_{(k)})_j = -\lambda_1 w^d_j$ can hold indeed, and thus we can uniquely define $\boldsymbol{\alpha}_k = \boldsymbol{\alpha}^+_k - \boldsymbol{\alpha}^-_k$. Then we can rewrite the Lagrange dual function as

$$L(\boldsymbol{\beta}; \boldsymbol{\alpha}_k) = \big(\mathbf{w}^d_k \circ \operatorname{sign}(\boldsymbol{\beta}) - \hat{\mathbf{R}}^s_{(k)}\boldsymbol{\alpha}_k\big)^T \boldsymbol{\beta} - \lambda_1 \big\|\mathbf{w}^d_k \circ \boldsymbol{\alpha}_k\big\|_{\ell_1} - \boldsymbol{\alpha}^T_k \hat{\mathbf{r}}^s_{(k)}.$$

By the Lagrange duality, the dual problem of (15) is

$$\max_{\boldsymbol{\alpha}\in\mathbb{R}^{p-1}} -\lambda_1 \big\|\mathbf{w}^d_k \circ \boldsymbol{\alpha}_k\big\|_{\ell_1} - \big\langle \boldsymbol{\alpha}_k, \hat{\mathbf{r}}^s_{(k)}\big\rangle \qquad \text{subject to } \big|\hat{\mathbf{R}}^s_{(k)}\boldsymbol{\alpha}_k\big| \leq \mathbf{w}^d_k.$$

Now we shall construct an optimal primal and dual solution pair $(\tilde{\boldsymbol{\beta}}_k, \tilde{\boldsymbol{\alpha}}_k)$ to the rank-based adaptive Dantzig selector. In addition, we show that $(\tilde{\boldsymbol{\beta}}_k, \tilde{\boldsymbol{\alpha}}_k)$ is actually the unique solution pair to the rank-based adaptive Dantzig selector, and $\tilde{\boldsymbol{\beta}}_k$ is exactly supported in the true active set $\mathcal{A}_k$. To this end, we construct $(\tilde{\boldsymbol{\beta}}_k, \tilde{\boldsymbol{\alpha}}_k)$ as $\tilde{\boldsymbol{\alpha}}_k = (\tilde{\boldsymbol{\alpha}}_{\mathcal{A}_k}, \tilde{\boldsymbol{\alpha}}_{\mathcal{A}^c_k}) = (\tilde{\boldsymbol{\alpha}}_{\mathcal{A}_k}, \mathbf{0})$ and $\tilde{\boldsymbol{\beta}}_k = (\tilde{\boldsymbol{\beta}}_{\mathcal{A}_k}, \tilde{\boldsymbol{\beta}}_{\mathcal{A}^c_k}) = (\tilde{\boldsymbol{\beta}}_{\mathcal{A}_k}, \mathbf{0})$ where $\tilde{\boldsymbol{\alpha}}_{\mathcal{A}_k} = -(\hat{\mathbf{R}}^s_{\mathcal{A}_k\mathcal{A}_k})^{-1}\mathbf{w}^d_{\mathcal{A}_k} \circ \operatorname{sign}(\boldsymbol{\beta}^*_{\mathcal{A}_k})$ and $\tilde{\boldsymbol{\beta}}_{\mathcal{A}_k} = (\hat{\mathbf{R}}^s_{\mathcal{A}_k\mathcal{A}_k})^{-1}(\hat{\mathbf{r}}^s_{\mathcal{A}_k} + \lambda_1 \mathbf{w}^d_{\mathcal{A}_k} \circ \operatorname{sign}(\tilde{\boldsymbol{\alpha}}_{\mathcal{A}_k}))$.

In what follows, we first show that $(\tilde{\boldsymbol{\beta}}_k, \tilde{\boldsymbol{\alpha}}_k)$ satisfies four optimality conditions, and then we will use these four optimality conditions to prove that $(\tilde{\boldsymbol{\beta}}_k, \tilde{\boldsymbol{\alpha}}_k)$ is indeed a unique optimal solution pair. The four optimality conditions are stated as follows:

$$\hat{\mathbf{R}}^s_{\mathcal{A}_k\mathcal{A}_k}\tilde{\boldsymbol{\beta}}_{\mathcal{A}_k} - \hat{\mathbf{r}}^s_{\mathcal{A}_k} = \lambda_1 \mathbf{w}^d_{\mathcal{A}_k} \circ \operatorname{sign}(\tilde{\boldsymbol{\alpha}}_{\mathcal{A}_k}), \tag{27}$$

$$\hat{\mathbf{R}}^s_{\mathcal{A}_k\mathcal{A}_k}\tilde{\boldsymbol{\alpha}}_{\mathcal{A}_k} = -\mathbf{w}^d_{\mathcal{A}_k} \circ \operatorname{sign}(\tilde{\boldsymbol{\beta}}_{\mathcal{A}_k}), \tag{28}$$

$$|\hat{\mathbf{R}}^s_{\mathcal{A}_k^c\mathcal{A}_k}\tilde{\boldsymbol{\beta}}_{\mathcal{A}_k} - \hat{\mathbf{r}}^s_{\mathcal{A}_k^c}| < \lambda_1 \mathbf{w}^d_{\mathcal{A}_k^c}, \tag{29}$$

$$|\hat{\mathbf{R}}^s_{\mathcal{A}_k^c\mathcal{A}_k}\tilde{\boldsymbol{\alpha}}_{\mathcal{A}_k}| < \mathbf{w}^d_{\mathcal{A}_k^c}, \tag{30}$$

where (27) and (29) are primal constraints, and (28) and (30) are dual constraints.

Note (27) can be easily verified by substituting $\tilde{\boldsymbol{\alpha}}_{\mathcal{A}_k}$ and $\tilde{\boldsymbol{\beta}}_{\mathcal{A}_k}$. Under (26), we can derive upper bounds for $K_1 = \|(\hat{\mathbf{R}}^s_{\mathcal{A}_k\mathcal{A}_k})^{-1} - (\boldsymbol{\Sigma}^*_{\mathcal{A}_k\mathcal{A}_k})^{-1}\|_{\ell_\infty}$, and $K_2 = \|\hat{\mathbf{R}}^s_{\mathcal{A}_k^c\mathcal{A}_k}(\hat{\mathbf{R}}^s_{\mathcal{A}_k\mathcal{A}_k})^{-1} - \boldsymbol{\Sigma}^*_{\mathcal{A}_k^c\mathcal{A}_k}(\boldsymbol{\Sigma}^*_{\mathcal{A}_k\mathcal{A}_k})^{-1}\|_{\ell_\infty}$.

Note $K_1 = (\hat{\mathbf{R}}^s_{\mathcal{A}_k\mathcal{A}_k})^{-1} \cdot (\hat{\mathbf{R}}^s_{\mathcal{A}_k\mathcal{A}_k} - \boldsymbol{\Sigma}^*_{\mathcal{A}_k\mathcal{A}_k}) \cdot (\boldsymbol{\Sigma}^*_{\mathcal{A}_k\mathcal{A}_k})^{-1}$, and then we have

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

Some simple calculation shows $K_1 \le \frac{d\lambda_1 G_k^2}{1-d\lambda_1 G_k}$. On the other hand,

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

Under probability event (26), we claim about $\mathbf{w}^d_k$ that

$$\|\mathbf{w}^d_{\mathcal{A}_k^c}\|_{\min} \ge \frac{d\lambda_1 G_k + H_k}{2\lambda_1 G_k}\psi_k + \frac{1+dG_k}{1-d\lambda_1 G_k}, \tag{31}$$

$$\|\mathbf{w}^d_{\mathcal{A}_k}\|_\infty \le \frac{1-d\lambda_1 G_k}{2\lambda_1 G_k}\psi_k - dG_k - 1. \tag{32}$$

This claim is very useful to prove the other three optimality conditions (28), (29) and (30), and their proofs will be provided later.

Now we are ready to prove (28), (29) and (30) for the solution pair $(\tilde{\boldsymbol{\beta}}_k, \tilde{\boldsymbol{\alpha}}_k)$. To prove (28), it is equivalent to show the sign consistency that $\operatorname{sign}(\boldsymbol{\beta}^*_{\mathcal{A}_k}) = \operatorname{sign}(\tilde{\boldsymbol{\beta}}_{\mathcal{A}_k})$ since we have $\hat{\mathbf{R}}^s_{\mathcal{A}_k\mathcal{A}_k}\tilde{\boldsymbol{\alpha}}_{\mathcal{A}_k} = -\mathbf{w}^d_{\mathcal{A}_k} \circ \operatorname{sign}(\boldsymbol{\beta}^*_{\mathcal{A}_k})$ if we plug in $\tilde{\boldsymbol{\alpha}}_{\mathcal{A}_k}$ to its left-hand side. Recall that $\boldsymbol{\beta}^*_{\mathcal{A}_k} = (\boldsymbol{\Sigma}^*_{\mathcal{A}_k\mathcal{A}_k})^{-1}\boldsymbol{\sigma}^*_{\mathcal{A}_k}$, and then we consider the difference between $\tilde{\boldsymbol{\beta}}_{\mathcal{A}_k}$ and $\boldsymbol{\beta}^*_{\mathcal{A}_k}$,

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

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

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

where the last inequality obviously holds by claim (31). Then by the above upper bound, $\operatorname{sign}(\boldsymbol{\beta}^*_{\mathcal{A}_k}) = \operatorname{sign}(\tilde{\boldsymbol{\beta}}_{\mathcal{A}_k})$ will be immediately satisfied.

Next, we can easily obtain (30) via the triangular inequality

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

where the last inequality can be easily shown by combining (31) and (32).

Now it remains to prove (29). Using the facts that $\boldsymbol{\theta}^*_{\mathcal{A}_k^c} = \mathbf{0}$ and $\boldsymbol{\Sigma}^*\boldsymbol{\Theta}^* = \mathbf{I}$, simple calculation yields that $\boldsymbol{\Sigma}^*_{\mathcal{A}_k^c\mathcal{A}_k}(\boldsymbol{\Sigma}^*_{\mathcal{A}_k\mathcal{A}_k})^{-1}\boldsymbol{\sigma}^*_{\mathcal{A}_k} = \boldsymbol{\sigma}^*_{\mathcal{A}_k^c}$. Then we can rewrite the left-hand side of (29) as

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

Again we apply the triangle inequality to obtain an upper bound as follows:

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

where the last inequality is due to (31) and (32).

So far, the four optimality conditions have been verified for $(\tilde{\boldsymbol{\beta}}_k, \tilde{\boldsymbol{\alpha}}_k)$. In the sequel, we shall show that $\tilde{\boldsymbol{\beta}}_k$ is indeed a unique optimal solution. First, due to (29)

and (30), $(\tilde{\boldsymbol{\beta}}_k, \tilde{\boldsymbol{\alpha}}_k)$ are feasible solutions to the primal and dual problems, respectively. Then (27) and (28) show that $(\tilde{\boldsymbol{\beta}}_k, \tilde{\boldsymbol{\alpha}}_k)$ satisfy the complementary-slackness conditions for both the primal and the dual problems. Thus, $(\tilde{\boldsymbol{\beta}}_k, \tilde{\boldsymbol{\alpha}}_k)$ are optimal solutions to these problems by Theorem 4.5 in Bertsimas and Tsitsiklis (1997). Now it remains to show the uniqueness. Suppose there exists another optimal solution $\breve{\boldsymbol{\beta}}_k$, and we have $\|\mathbf{w}_k^d \circ \breve{\boldsymbol{\beta}}_k\|_{\ell_1} = \|\mathbf{w}_k^d \circ \tilde{\boldsymbol{\beta}}_k\|_{\ell_1}$. Let Γ_k denote the support of $\breve{\boldsymbol{\beta}}_k$, and then $\breve{\boldsymbol{\beta}}_k = (\breve{\boldsymbol{\beta}}_{\Gamma_k}, \mathbf{0})$. By the strong duality we have

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

Thus $L(\breve{\boldsymbol{\beta}}_k; \tilde{\boldsymbol{\alpha}}_k^+, \tilde{\boldsymbol{\alpha}}_k^-) = \|\mathbf{w}_k^d \circ \breve{\boldsymbol{\beta}}_k\|_{\ell_1}$, which immediately implies that the complementary slackness condition holds for the primal problem, that is, $(\hat{\mathbf{R}}_{(k)}^s \breve{\boldsymbol{\beta}}_k - \hat{\mathbf{r}}_{(k)}^s - \lambda_1 \mathbf{w}_k^d)^T \tilde{\boldsymbol{\alpha}}_k^+ = 0$ and $(-\hat{\mathbf{R}}_{(k)}^s \breve{\boldsymbol{\beta}}_k + \hat{\mathbf{r}}_{(k)}^s - \lambda_1 \mathbf{w}_k^d)^T \tilde{\boldsymbol{\alpha}}_k^- = 0$. Now let $\breve{\boldsymbol{\beta}}_k^+ = \max(\breve{\boldsymbol{\beta}}_k, \mathbf{0})$ and $\breve{\boldsymbol{\beta}}_k^- = \min(\breve{\boldsymbol{\beta}}_k, \mathbf{0})$. Besides, we can similarly show that the complementary slackness condition also holds for the dual problem, that is, $(\hat{\mathbf{R}}_{(k)}^s \tilde{\boldsymbol{\alpha}}_k - \mathbf{w}_k^d)^T \breve{\boldsymbol{\beta}}_k^+ = 0$ and $(-\hat{\mathbf{R}}_{(k)}^s \tilde{\boldsymbol{\alpha}}_k - \mathbf{w}_k^d)^T \breve{\boldsymbol{\beta}}_k^- = 0$. Notice that $\tilde{\boldsymbol{\alpha}}_{\mathcal{A}_k} \neq \mathbf{0}$ and $\tilde{\boldsymbol{\alpha}}_{\mathcal{A}_k^c} = \mathbf{0}$ by definition, and then we have

$$\hat{\mathbf{R}}_{\mathcal{A}_k \Gamma_k}^s \breve{\boldsymbol{\beta}}_{\Gamma_k} - \hat{\mathbf{r}}_{\mathcal{A}_k}^s = \lambda_1 \mathbf{w}_{\mathcal{A}_k}^d \circ \operatorname{sign}(\tilde{\boldsymbol{\alpha}}_{\mathcal{A}_k}), \tag{33}$$

$$\hat{\mathbf{R}}_{\Gamma_k \mathcal{A}_k}^s \tilde{\boldsymbol{\alpha}}_{\mathcal{A}_k} = -\mathbf{w}_{\Gamma_k}^d \circ \operatorname{sign}(\breve{\boldsymbol{\beta}}_{\Gamma_k}). \tag{34}$$

Observe that for any $j \in \Gamma_k$ but $j \notin \mathcal{A}_k$, $\hat{\mathbf{R}}_{j\mathcal{A}_k}^s \tilde{\boldsymbol{\alpha}}_{\mathcal{A}_k} = -w_j^d \operatorname{sign}(\breve{\beta}_j)$ in (34) cannot hold since it contradicts with (30). Then it is easy to see that $\Gamma_k \subset \mathcal{A}_k$ obviously holds, which immediately implies that $\hat{\boldsymbol{\beta}}_{\mathcal{A}_k}$ and $\breve{\boldsymbol{\beta}}_{\Gamma_k}$ satisfy the same optimality condition (27). Thus the uniqueness follows from (27), (33) and the nonsingularity of $\hat{\mathbf{R}}_{\mathcal{A}_k \mathcal{A}_k}^s$.

Now it remains to verify the claims (31) and (32) under event (26). Under the event $\|\hat{\mathbf{R}}^s - \boldsymbol{\Sigma}^*\|_{\max} \le b\lambda_0/M$, it has been shown in Theorem 2 that for some $C_0 = 4B^2(2 + \frac{b}{M}) > 0$, we have $\|\hat{\boldsymbol{\beta}}_k^{s.nd} - \boldsymbol{\beta}_k^*\|_{\ell_1} \le C_0 d\lambda_0$. Then we can derive a lower bound for $\|\mathbf{w}_{\mathcal{A}_k^c}^d\|_{\min}$,

$$\|\mathbf{w}_{\mathcal{A}_k^c}^d\|_{\min} = \frac{1}{\max_{j \in \mathcal{A}_k^c} |\hat{\beta}_j^{s.nd}| + 1/n} \ge \frac{1}{C_0 d\lambda_0 + 1/n},$$

which immediately yields the desired lower bound by noting that

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

where both inequalities follow from the proper choices of tuning parameters λ_0 and λ_1 as stated in Theorem 3. On the other hand,

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

where the last inequality follows from the proper choice of λ_1 as stated in Theorem 3. Likewise we can prove the second claim (32) by noticing that

$$\|\mathbf{w}^d_{\mathcal{A}_k}\|_\infty \le \frac{1}{\min_{j \in \mathcal{A}_k} |\hat{\beta}^{s.nd}_j|} \le \frac{1}{\psi_k - C_0 d\lambda_0} \le \frac{2}{\psi_k} \le \frac{\psi_k}{4G_k \cdot \lambda_1},$$

where we use facts that $\psi_k \ge 2C_0 d\lambda_0$ and $\psi_k^2 \ge 8G_k\lambda_1$. The two claims are proved, which completes the proof of Theorem 3. □

PROOF OF THEOREM 4. To bound the difference between $\hat{\mathbf{\Theta}}^s_c$ and $\mathbf{\Theta}^*$ under the entry-wise ℓ_∞-norm, we consider the event $\{\|\hat{\mathbf{R}}^s - \mathbf{\Sigma}\|_{\max} \le \lambda/M\}$. First, we show that $\mathbf{\Theta}^*$ is always a feasible solution under the above event,

$$\|\hat{\mathbf{R}}^s \mathbf{\Theta}^* - \mathbf{I}\|_{\max} \le \|(\hat{\mathbf{R}}^s - \mathbf{\Sigma}^*)\mathbf{\Theta}^*\|_{\max} \le \|\hat{\mathbf{R}}^s - \mathbf{\Sigma}\|_{\max} \cdot \|\mathbf{\Theta}^*\|_{\ell_1} \le \lambda.$$

Note that $\hat{\mathbf{\Theta}}^s_c$ is the optimal solution, and then $\|\hat{\mathbf{R}}^s \hat{\mathbf{\Theta}}^s_c - \mathbf{I}\|_{\max} \le \lambda$ obviously holds. Moreover, it is easy to see that by definition $\|\hat{\mathbf{\Theta}}^s_c\|_{\ell_1} \le \|\mathbf{\Theta}^*\|_{\ell_1}$ always holds. Now we can obtain the desired bound under the entry-wise ℓ_∞-norm.

$$\begin{aligned}
\|\hat{\mathbf{\Theta}}^s_c - \mathbf{\Theta}^*\|_{\max} &\le \|\mathbf{\Theta}^*\|_{\ell_1} \cdot \|\mathbf{\Sigma}^* \hat{\mathbf{\Theta}}^s_c - \mathbf{I}\|_{\max} \\
&= M \cdot \|(\mathbf{\Sigma}^* - \hat{\mathbf{R}}^s)\hat{\mathbf{\Theta}}^s + \hat{\mathbf{R}}^s \hat{\mathbf{\Theta}}^s_c - \mathbf{I}\|_{\max} \\
&\le M \cdot \|\mathbf{\Sigma}^* - \hat{\mathbf{R}}^s\|_{\max} \cdot \|\hat{\mathbf{\Theta}}^s_c\|_{\ell_1} + M \cdot \|\hat{\mathbf{R}}^s \hat{\mathbf{\Theta}}^s_c - \mathbf{I}\|_{\max} \\
&\le \lambda \cdot \|\mathbf{\Theta}^*\|_{\ell_1} + M\lambda \\
&= 2M\lambda.
\end{aligned}$$

□

PROOF OF THEOREM 5. The techniques we use are similar to these for the proof of Theorem 3. The detailed proof of Theorem 5 is relegated to the supplementary material [Xue and Zou (2012)] and also the technical report version of this paper [Xue and Zou (2011b)] for the sake of space. □

Acknowledgments. We thank the Editor, the Associate Editor and three referees for their helpful comments.

SUPPLEMENTARY MATERIAL

Supplement material for "Regularized rank-based estimation of high-dimensional nonparanormal graphical models" (DOI: 10.1214/12-AOS1041SUPP; .pdf). In this supplementary note, we give the complete proofs of Theorems 2 and 5.

REFERENCES

BANERJEE, O., EL GHAOUI, L. and D'ASPREMONT, A. (2008). Model selection through sparse maximum likelihood estimation for multivariate Gaussian or binary data. *J. Mach. Learn. Res.* **9** 485–516. MR2417243

BERTSIMAS, D. and TSITSIKLIS, J. N. (1997). *Introduction to Linear Optimization*. Athena Scientific, Belmont, MA.

BICKEL, P. J., RITOV, Y. and TSYBAKOV, A. B. (2009). Simultaneous analysis of lasso and Dantzig selector. *Ann. Statist.* **37** 1705–1732. MR2533469

BOYD, S. and VANDENBERGHE, L. (2004). *Convex Optimization*. Cambridge Univ. Press, Cambridge. MR2061575

CAI, T., LIU, W. and LUO, X. (2011). A constrained ℓ_1 minimization approach to sparse precision matrix estimation. *J. Amer. Statist. Assoc.* **106** 594–607. MR2847973

CANDES, E. and TAO, T. (2007). The Dantzig selector: Statistical estimation when p is much larger than n. *Ann. Statist.* **35** 2313–2351. MR2382644

CHEN, X. and FAN, Y. (2006). Estimation of copula-based semiparametric time series models. *J. Econometrics* **130** 307–335. MR2211797

CHEN, X., FAN, Y. and TSYRENNIKOV, V. (2006). Efficient estimation of semiparametric multivariate copula models. *J. Amer. Statist. Assoc.* **101** 1228–1240. MR2328309

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

DEVLIN, S. J., GNANADESIKAN, R. and KETTENRING, J. R. (1975). Robust estimation and outlier detection with correlation coefficients. *Biometrika* **62** 531–545.

DICKER, L. and LIN, X. (2009). Variable selection using the Dantzig selector: Asymptotic theory and extensions. Unpublished manuscript.

DOBRA, A., EICHER, T. S. and LENKOSKI, A. (2010). Modeling uncertainty in macroeconomic growth determinants using Gaussian graphical models. *Stat. Methodol.* **7** 292–306. MR2643603

DRTON, M. and PERLMAN, M. D. (2004). Model selection for Gaussian concentration graphs. *Biometrika* **91** 591–602. MR2090624

DRTON, M. and PERLMAN, M. D. (2007). Multiple testing and error control in Gaussian graphical model selection. *Statist. Sci.* **22** 430–449. MR2416818

EDWARDS, D. (2000). *Introduction to Graphical Modelling*, 2nd ed. Springer, New York. MR1880319

FAN, J. and LI, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.* **96** 1348–1360. MR1946581

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

FRIEDMAN, J., HASTIE, T. and TIBSHIRANI, R. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics* **9** 432–441.

HOEFFDING, W. (1948). A class of statistics with asymptotically normal distribution. *Ann. Math. Statistics* **19** 293–325. MR0026294

JAMES, G. M., RADCHENKO, P. and LV, J. (2009). DASSO: Connections between the Dantzig selector and lasso. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **71** 127–142. MR2655526

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

KRUSKAL, W. H. (1958). Ordinal measures of association. *J. Amer. Statist. Assoc.* **53** 814–861. MR0100941

LAM, C. and FAN, J. (2009). Sparsistency and rates of convergence in large covariance matrix estimation. *Ann. Statist.* **37** 4254–4278. MR2572459

LAULE, O., FÜRHOLZ, A., CHANG, H. S., ZHU, T., WANG, X., HEIFETZ, P. B., GRUISSEM, W. and LANGE, M. (2003). Crosstalk between cytosolic and plastidial pathways of isoprenoid biosynthesis in Arabidopsis thaliana. *Proc. Natl. Acad. Sci. USA* **100** 6866–6871.

LAURITZEN, S. L. (1996). *Graphical Models*. *Oxford Statistical Science Series* **17**. The Clarendon Press Oxford Univ. Press, New York. MR1419991

LEHMANN, E. L. (1998). *Nonparametrics*: *Statistical Methods Based on Ranks*. Prentice Hall Upper Saddle River, New Jersey.

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

LIU, H., LAFFERTY, J. and WASSERMAN, L. (2009). The nonparanormal: Semiparametric estimation of high dimensional undirected graphs. *J. Mach. Learn. Res.* **10** 2295–2328. MR2563983

LIU, H., HAN, F., YUAN, M., LAFFERTY, J. and WASSERMAN, L. (2012). High dimensional semiparametric Gaussian copula graphical models. Technical report, Johns Hopkins Univ.

MCDIARMID, C. (1989). On the method of bounded differences. In *Surveys in Combinatorics*, 1989 (*Norwich*, 1989). *London Mathematical Society Lecture Note Series* **141** 148–188. Cambridge Univ. Press, Cambridge. MR1036755

MEINSHAUSEN, N. and BÜHLMANN, P. (2006). High-dimensional graphs and variable selection with the lasso. *Ann. Statist.* **34** 1436–1462. MR2278363

PENG, J., WANG, P., ZHOU, N. and ZHU, J. (2009). Partial correlation estimation by joint sparse regression models. *J. Amer. Statist. Assoc.* **104** 735–746. MR2541591

RAVIKUMAR, P., WAINWRIGHT, M. J., RASKUTTI, G. and YU, B. (2011). High-dimensional covariance estimation by minimizing ℓ_1-penalized log-determinant divergence. *Electron. J. Statist.* **5** 935–980.

RODRÍGUEZ-CONCEPCIÓN, M., FORÉS, O., MARTINEZ-GARCÍA, J. F., GONZÁLEZ, V., PHILLIPS, M. A., FERRER, A. and BORONAT, A. (2004). Distinct light-mediated pathways regulate the biosynthesis and exchange of isoprenoid precursors during Arabidopsis seedling development. *Plant Cell* **16** 144–156.

ROTHMAN, A. J., BICKEL, P. J., LEVINA, E. and ZHU, J. (2008). Sparse permutation invariant covariance estimation. *Electron. J. Stat.* **2** 494–515. MR2417391

SONG, P. X.-K. (2000). Multivariate dispersion models generated from Gaussian copula. *Scand. J. Stat.* **27** 305–320. MR1777506

TIBSHIRANI, R. (1996). Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B* **58** 267–288. MR1379242

VAN DE GEER, S., BÜHLMANN, P. and ZHOU, S. (2011). The adaptive and the thresholded Lasso for potentially misspecified models (and a lower bound for the Lasso). *Electron. J. Stat.* **5** 688–749. MR2820636

WILLE, A., ZIMMERMANN, P., VRANOVÁ, E., FÜRHOLZ, A., LAULE, O., BLEULER, S., HENNIG, L., PRELIC, A., VON ROHR, P., THIELE, L. et al. (2004). Sparse graphical Gaussian modeling of the isoprenoid gene network in Arabidopsis thaliana. *Genome Biology* **5** 1–13.

XUE, L. and ZOU, H. (2011a). On estimating sparse correlation matrices of semiparametric Gaussian copulas. Technical report, Univ. Minnesota.

XUE, L. and ZOU, H. (2011b). Regularized rank-based estimation of high-dimensional nonparanormal graphical models. Technical report, Univ. Minnesota.

XUE, L. and ZOU, H. (2012). Supplement to "Regularized rank-based estimation of high-dimensional nonparanormal graphical models." DOI:10.1214/12-AOS1041SUPP.

YUAN, M. (2010). High dimensional inverse covariance matrix estimation via linear programming. *J. Mach. Learn. Res.* **11** 2261–2286. MR2719856

YUAN, M. and LIN, Y. (2007). Model selection and estimation in the Gaussian graphical model. *Biometrika* **94** 19–35. MR2367824

ZHOU, S., RÜTIMANN, P., XU, M. and BÜHLMANN, P. (2011). High-dimensional covariance estimation based on Gaussian graphical models. *J. Mach. Learn. Res.* **12** 2975–3026. MR2854354

ZOU, H. (2006). The adaptive lasso and its oracle properties. *J. Amer. Statist. Assoc.* **101** 1418–1429. MR2279469

SCHOOL OF STATISTICS
UNIVERSITY OF MINNESOTA
MINNEAPOLIS, MINNESOTA 55455
USA
E-MAIL: lzxue@stat.umn.edu
zouxx019@umn.edu