Transelliptical Graphical Modeling under A Hierarchical Latent Variable Framework

Han Liu, Fang Han, and Cun-hui Zhang

May 28, 2013

Abstract

We advocate the use of a semiparametric distribution family—the transelliptical—for robust inference of high dimensional graphical models. The transelliptical graphical model has a three-layer hierarchical latent variable representation. We provide interpretations of the inferred graph for variables at different layers: (i) For the first layer, the absence of an edge between two variables means the absence of a certain rank-based association of the pair given other variables; (ii) For the second layer, the absence of an edge means the conditional uncorrelatedness of the pair; (iii) For the third layer, the absence of an edge means the conditional independence of the pair. We propose a tuning-insensitive, rank-based method that is invariant within the whole transelliptical family and achieves parametric rates of convergence for both graph recovery and parameter estimation. This result suggests that the extra robustness and flexibility gained by semiparametric transelliptical modeling comes with almost no cost. We also report numerical results on synthetic and real datasets to support the theoretical analysis.

Keyword: High dimensional statistics; Multivariate analysis; Undirected graphical models; Transelliptical family; Robust inference Semiparametric inference.

1 Introduction

We consider the problem of learning high dimensional graphical models: given independent observations from a d-dimensional random vector $\boldsymbol{X} := (X_1, ..., X_d)^T$, we want to estimate an undirected graph $G := (V, E)$, where V contains nodes corresponding to the d variables in $\boldsymbol{X}$ and the edge set E describes certain relationships between $X_1, \ldots, X_d$. In particular, $(j, k) \notin E$ implies the absence of a certain notion of association between X_j and X_k conditionally on the rest of variables.

For Gaussian data $\boldsymbol{X} \sim N_d(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, this problem reduces to the covariance selection problem (Dempster, 1972): we want to infer a graph G that encodes the conditional independence relationship between $X_1, ..., X_d$. Let $\boldsymbol{X}_{\setminus\{j,k\}} := \{X_\ell : \ \ell \neq j, k\}$. We say that

the joint distribution of $\boldsymbol{X}$ is Markovian in the graph $G = (V, E)$ when $(j, k) \notin E$ is equivalent to the conditional independence of X_j and X_k given $\boldsymbol{X}_{\setminus\{j,k\}}$. Under this normality assumption, the graph G is encoded by the precision matrix $\boldsymbol{\Theta} := \boldsymbol{\Sigma}^{-1}$. More specifically, no edge connects X_j and X_k if and only if $\boldsymbol{\Theta}_{jk} = 0$. In the low dimensional case of $d < n$, Drton and Perlman (2007, 2008) develop a multiple testing procedure for identifying the sparsity pattern of the precision matrix. In the high dimensional case of $d \gg n$, Meinshausen and Bühlmann (2006) proposes a neighborhood pursuit approach of estimating Gaussian graphical models by solving a collection of sparse regression problems using the Lasso (Tibshirani, 1996; Chen et al., 1998). Such an approach can be viewed as a pseudo-likelihood approximation of the full likelihood. In contrast, Banerjee et al. (2008), Yuan and Lin (2007) and Friedman et al. (2008) propose penalized likelihood approaches to directly estimate $\boldsymbol{\Theta}$. Lam and Fan (2009) and Shen et al. (2012) propose to maximize the non-concave penalized likelihood to obtain an estimator with less bias than the traditional L_1-regularized estimator. Jalali et al. (2012) proposes to estimate the precision matrix via a greedy algorithm. Under certain conditions, Ravikumar et al. (2011) and Rothman et al. (2008) study the theoretical properties of the penalized likelihood methods. Yuan (2010) and Cai et al. (2011) propose the graphical Dantzig selector and CLIME respectively, which can be solved by linear programming and are more amenable to theoretical analysis than the penalized likelihood approach. More recently, Liu and Luo (2012) and Sun and Zhang (2012) propose the SCIO and scaled-Lasso methods, which estimate the sparse precision matrix in a column-by-column fashion and have good theoretical properties.

For non-Gaussian data, Liu et al. (2009) proposes a semiparametric Gaussian copula model named *nonparanormal*. Instead of imposing a normality condition on $\boldsymbol{X}$, the nonparanormal model assumes the existence of a set of monotone functions $f_1, \ldots, f_d$ such that the transformed data $f(\boldsymbol{X}) := (f_1(X_1), \ldots, f_d(X_d))^T$ is Gaussian. More details about this model can be found in Liu et al. (2012), Lafferty et al. (2012), and Xue and Zou (2012). Zhao et al. (2012) develops a scalable software package to implement the nonparanormal algorithms. Other nonparametric methods include forest graphical models (Liu et al., 2011) and conditional graphical models (Liu et al., 2010a). In a different line of research, Vogel and Fried (2011) considers the elliptical graphical models. The elliptical family contains many multivariate distributions, including multivariate Gaussian, multivariate t-distribution, Cauchy, logistic, Kotz, symmetric Pearson type-II and type-VII distributions. The inferred graph is named the *generalized partial correlation graph*, which represents conditional uncorrelatedness among variables. Conditional uncorrelatedness is a weaker notion than conditional independence. Therefore, by extending the Gaussian to elliptical family, the gain in modeling flexibility is traded off with a loss in the strength of inference. The analysis and algorithm of Vogel and Fried (2011) are mainly for low dimensional settings and are not straightforwardly extendable to high dimensional settings where $d > n$. In a related work, Finegold and Drton (2009) studies the t-graphical model and propose an EM-

type algorithm for model fitting in high dimensions. However, no asymptotic properties are shown for this algorithm.

In this paper, we introduce a new graphical modeling strategy based on a newly defined model family named *transelliptical distribution*. The transelliptical modeling strategy extends the idea of nonparanormal modeling from Liu et al. (2009): by mimicking how the nonparanormal extends the normal family, the transelliptical extends the elliptical family in the same way. More specifically, we say that a random vector $\boldsymbol{X} = (X_1, \ldots, X_d)^T \in \mathbb{R}^d$ follows a transelliptical distribution if there exists a set of strictly increasing functions $f := \{f_j\}_{j=1}^d$ such that $f(\boldsymbol{X}) := (f_1(X_1), \ldots, f_d(X_d))^T$ follows an elliptical distribution with non-degenerate marginals (More details will be given in later sections).

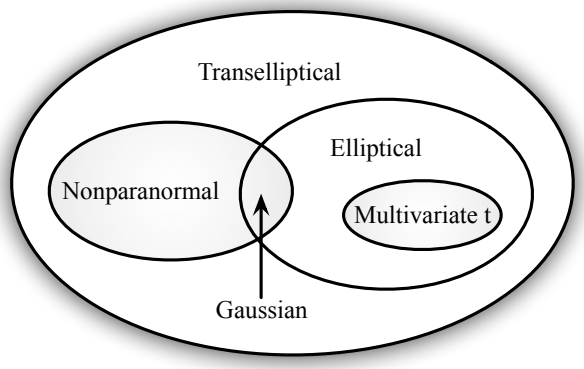


Figure 1: The Venn diagram illustrating the relationships of the transelliptical, elliptical, and nonparanormal families: The nonparanormal and elliptical distributions are proper subsets of the transelliptical family, and the intersection between the nonparanomral and elliptical families is Gaussian (More details of this diagram are provided in later sections).

Figure 1 illustrates the relationships of the transelliptical, elliptical, and nonparanormal families. Both the nonparanormal and elliptical distributions are proper subsets of the transelliptical family[1], and the only distribution that simultaneously belongs to the nonparanormal and elliptical families is Gaussian. More details about these points will be elaborated in later sections.

In Section 3, we construct the transelliptical graphical model based on the transelliptical family. In particular, as is illustrated in Figure 2, we provide a three-layer hierarchical latent variable representation of the transelliptical graphical model. The observable vector, denoted by $\boldsymbol{X} := (X_1, \ldots, X_d)^T$ and presented in the first-layer, has a transelliptical distribution, and a latent random vector, $\boldsymbol{Z} := (Z_1, \ldots, Z_d)^T$ in the second-layer, is elliptically distributed. Variables in the first and second layers are related through the transformation

[1]Whenever we say the transelliptical family contains elliptical family, we only consider elliptical distributions with non-degenerate marginal distributions.

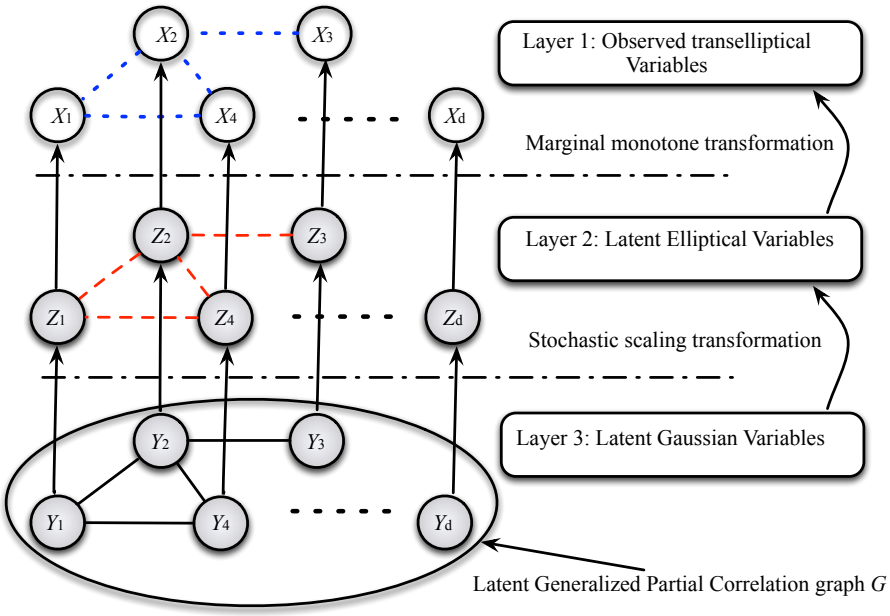


Figure 2: Hierarchal latent variable representation of the transelliptical graphical model with the latent variables gray-colored. The first layer is composed of observed variables X_j, and the second and third layers are composed of latent variables Z_j and Y_j. The solid undirected lines in the third layer encode the conditional independence graph of $Y_1, \ldots, Y_d$. The same graph is drawn using dashed lines and dotted lines for the second layer latent variables and the first layer observed variables. The transformation from Layer 2 to Layer 1 is through deterministic marginal monotone transformation. The transformation from Layer 3 to Layer 2 is through stochastic scaling (with more details in Section 3).

$Z_j := f_j(X_j)$ with f_j being an unknown monotone function. In the next section, we show that the latent random vector $\boldsymbol{Z}$ can be further represented by a third-layer latent random vector $\boldsymbol{Y} := (Y_1, \ldots, Y_d)^T$ which has a multivariate Gaussian distribution with a correlation matrix $\boldsymbol{\Sigma}$ and precision matrix $\boldsymbol{\Theta} := \boldsymbol{\Sigma}^{-1}$. We define the *latent generalized partial correlation graph* $G := (V, E)$ with the vertex set $V = \{1, \ldots, d\}$ and the edge set E encoding the nonzero entries of $\boldsymbol{\Theta}$. We provide interpretations of graph G for variables in different layers: (i) For the observed variables in the first layer, the absence of an edge between two variables means the absence of a certain rank-based association of the pair given other variables; (ii) For the latent variables in the second layer, the absence of an edge means the absence of the conditional Pearson's correlation of the pair; (iii) For the latent variables in the third layer, the absence of an edge means the conditional independence of the pair. Therefore, compared with the Gaussian graphical model, the transelliptical graphical model has much richer structures and interpretations. We have the same graphical structure for all three layers of variables. However, the interpretations of these graphs are different.

To infer the graph structure, we propose a new rank-based estimating procedure named

scaled CLIME, which can be formulated as a linear program. This procedure is adaptive over the whole transelliptical family in the sense that it is invariant to the generating variable of the latent elliptical distributions. Detailed definition of the generating variable will be introduced later. This procedure is tuning-insensitive non-asymptotically and tuning-free asymptotically: it requires little effort to choose the tuning parameter in finite sample settings. Moreover, it has the same computational complexity as the CLIME estimator. Theoretically, the new procedure achieves the same parametric rates of convergence as the CLIME does for graph recovery and parameter estimation, even though the transelliptical family is much larger than the nonparanormal and elliptical families. These results suggest that the transelliptical graphical model can be used as a safe replacement of the nonparanormal and elliptical graphical models. We also provide thorough numerical results on both synthetic and real datasets to back up our theory. Some of the results in this paper were first stated without proof in a conference version: `http://books.nips.cc/papers/files/nips25/NIPS2012_0380.pdf`.

The rest of this paper is organized as follows. In Section 2, we review the elliptical distribution. In Section 3 we introduce the transelliptical graphical model and study its relationship with the nonparanormal graphical model. In Section 4 we discuss the invariant property of Kendall's tau statistic within the transelliptical family and propose our rank based estimation procedures motivated by this result. We also propose a new rank-based graph estimator, named scaled CLIME, which can be formulated as a linear program. In Section 5 we present asymptotic properties of the proposed procedure in both parameter estimation and graph recovery. In Section 6 we provide experimental results on both synthetic and real-world datasets. Some discussions and conclusions are summarized in the last section. All the proofs are put in the appendix.

2 Background

In this section, we introduce our notation and discuss the elliptical distribution family. Let $\boldsymbol{v} := (v_1, \ldots, v_d)^T \in \mathbb{R}^d$ denote vectors, $\mathbf{A} = (\mathbf{A}_{jk})$ matrices, and $I(\cdot)$ the indicator function. Let $\|\boldsymbol{v}\|_q$ denote the ℓ_q norm, with $\|\boldsymbol{v}\|_q := \left(\sum_{j=1}^d |v_j|^q\right)^{1/q}$ for $0 < q < \infty$, $\|\boldsymbol{v}\|_0 := \sum_{j=1}^d I(v_j \neq 0)$ and $\|\boldsymbol{v}\|_\infty := \max_j |v_j|$. For a symmetric matrix $\mathbf{A} \in \mathbb{R}^{d\times d}$ and index sets I and J in $\{1, \ldots, d\}$, we denote by $\mathbf{A}_{I,J}$ the submatrix of $\mathbf{A}$ with row and column indices in I and J, $\mathbf{A}_{*j}$ the j^{th} column of $\mathbf{A}$, and $\mathbf{A}_{*\backslash j}$ the submatrix of $\mathbf{A}$ with the j^{th} column $\mathbf{A}_{*j}$ removed. We use the following notation for matrix norms:

$$\|\mathbf{A}\|_q := \max_{\|v\|_q=1} \|\mathbf{A}\boldsymbol{v}\|_q, \quad \|\mathbf{A}\|_{\max} := \max_{jk} |\mathbf{A}_{jk}|, \text{ and } \quad \|\mathbf{A}\|_{\mathsf{F}} := \Big(\sum_{j,k} |\mathbf{A}_{jk}|^2\Big)^{1/2}.$$

It is easy to see that $\|\mathbf{A}\|_\infty = \|\mathbf{A}\|_1$ for symmetric $\mathbf{A}$. We denote by $\Lambda_{\max}(\mathbf{A})$ and $\Lambda_{\min}(\mathbf{A})$ the largest and smallest eigenvalues of $\mathbf{A}$. For any univariate function $f : \mathbb{R} \to \mathbb{R}$, let

$f(\mathbf{A}) = [f(\mathbf{A}_{jk})]$ denote the d by d matrix with an application of the function f to each individual entry of $\mathbf{A}$.

2.1 Background on Elliptical Distribution

We denote by $\boldsymbol{X} \stackrel{d}{=} \boldsymbol{Y}$ if random vectors $\boldsymbol{X}$ and $\boldsymbol{Y}$ have the same distribution. The elliptical distribution is defined as follows.

Definition 2.1 (Elliptical distribution (Fang et al., 1990)). *Let $\boldsymbol{\mu} \in \mathbb{R}^d$ and $\boldsymbol{\Sigma} \in \mathbb{R}^{d\times d}$ with* $\operatorname{rank}(\boldsymbol{\Sigma}) = q \leq d$. *A d-dimensional random vector $\boldsymbol{X}$ has an elliptical distribution, denoted by $\boldsymbol{X} \sim EC_d(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \xi)$, if it has a stochastic representation*

$$\boldsymbol{X} \stackrel{d}{=} \boldsymbol{\mu} + \xi \mathbf{A}\boldsymbol{U},$$

where $\boldsymbol{U}$ is a uniform random vector on the unit sphere in $\mathbb{R}^q$, $\xi \geq 0$ is a scalar random variable independent of $\boldsymbol{U}$, $\mathbf{A} \in \mathbb{R}^{d\times q}$ is a deterministic matrix satisfying $\mathbf{A}\mathbf{A}^T = \boldsymbol{\Sigma}$.

An equivalent definition of the elliptical distribution is that its characteristic function can be written as $\exp(i\boldsymbol{t}^T\boldsymbol{\mu})\psi(\boldsymbol{t}^T\boldsymbol{\Sigma}\boldsymbol{t})$, where ψ is a properly-defined characteristic function uniquely determined by the generating variable ξ in Definition 2.1. When $\operatorname{rank}(\boldsymbol{\Sigma}) = q$, ψ also uniquely determines the distribution of ξ. This justifies the alternative notation, $\boldsymbol{X} \sim EC_d(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \psi)$. Another stochastic representation of a random vector $\boldsymbol{X}$ with the elliptical distribution is provided in Theorem 3.12.

An elliptical distribution does not necessarily have a density. One example is rank-defficient Gaussian. Another example is discrete ξ. More examples can be found in Halmos (1974). However, when the random variable ξ is absolute continuous with respect to the Lebesgue measure and $\boldsymbol{\Sigma}$ is non-singular, the joint density of $\boldsymbol{X}$ exists and has the form

$$p(\boldsymbol{x}) = |\boldsymbol{\Sigma}|^{-1/2} g\left((\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right), \tag{2.1}$$

where $g(\cdot)$, called the scale function, is uniquely determined by the distribution of ξ. In this case, we also write $\boldsymbol{X} \sim EC_d(\boldsymbol{\mu}, \boldsymbol{\Sigma}, g)$. Many multivariate distributions belong to the elliptical family. For example, when $g(x) = (2\pi)^{-d/2}\exp\{-x/2\}$, $\boldsymbol{X}$ is d-dimensional Gaussian. Another important subclass of the elliptical family is the multivariate t-distribution with v degrees of freedom, corresponding to

$$g(x) = c_v \frac{\Gamma\left(\frac{v+d}{2}\right)}{(v\pi)^{\frac{d}{2}}\Gamma\left(\frac{v}{2}\right)}\left(1+\frac{c_v^2 x}{v}\right)^{-\frac{v+d}{2}},$$

where c_v is a normalizing constant.

2.2 Identifiability Condition

The quantities used to define the elliptical family in Definition 2.1 are not completely identifiable. For example, given $\boldsymbol{X} \sim EC_d(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \xi)$ with $\text{rank}(\boldsymbol{\Sigma}) = q$, there exist multiple matrices $\mathbf{A}$ corresponding to the same $\boldsymbol{\Sigma}$: i.e. $\mathbf{A}_1 \neq \mathbf{A}_2 \in \mathbb{R}^{d\times q}$ but $\mathbf{A}_1\mathbf{A}_1^T = \mathbf{A}_2\mathbf{A}_2^T = \boldsymbol{\Sigma}$. Therefore we parameterize the distribution by $\boldsymbol{\Sigma}$ instead of $\mathbf{A}$. Moreover, $\boldsymbol{\Sigma}$ is unique only up to a constant scaling: i.e. $\xi\mathbf{A}\boldsymbol{U} = \xi^*\mathbf{A}^*\boldsymbol{U}$ with $\xi^* = \xi/c$ and $\mathbf{A}^* = c\mathbf{A}$ for all $c > 0$. To make the model identifiable, we impose the following condition:

Definition 2.2 (Identifiability condition)**.** *For identifiability purpose, we require the condition that* $\max_{1\leq i\leq d} \boldsymbol{\Sigma}_{ii} = 1$. *Such* $\boldsymbol{\Sigma}$ *is called the generalized covariance matrix.* $\boldsymbol{\Sigma}^{-1}$ *is called the generalized inverse covariance matrix.*

More discussions about the identifiability issue can be found in Fang et al. (1990).

3 Transelliptical Graphical Models

In this paper we only consider distributions with non-degenerate marginals. We introduce the transelliptical graphical model in analogy to the nonparanormal graphical model (Liu et al., 2009) as follows. Let $\mathbf{A}$ be a symmetric positive definite matrix. We denote by $\text{diag}(\mathbf{A})$ the matrix $\mathbf{A}$ with off-diagonal elements replaced by zero, and by $\mathbf{A}^{1/2}$ a squared root matrix of $\mathbf{A}$, i.e., $\mathbf{A}^{1/2}(\mathbf{A}^{1/2})^T = \mathbf{A}$. We also denote by $\mathbf{I}_d$ the d-dimensional identity matrix. In the sequel, we denote the class of correlation matrices by

$$\mathcal{R}_d^+ := \{\boldsymbol{\Sigma} \in \mathbb{R}^{d\times d} : \boldsymbol{\Sigma}^T = \boldsymbol{\Sigma}, \text{diag}(\boldsymbol{\Sigma}) = \mathbf{I}_d, \Lambda_{\min}(\boldsymbol{\Sigma}) > 0\}.$$

3.1 Definition of Transelliptical Distribution

The Transelliptical distribution is defined as follows:

Definition 3.1 (Transelliptical distribution)**.** *A random vector* $\boldsymbol{X} = (X_1, \ldots, X_d)^T$ *follows a transelliptical distribution, denoted by*

$$\boldsymbol{X} \sim TE_d(\boldsymbol{\Sigma}, \xi; f_1, \ldots, f_d),$$

if there exists a set of strictly increasing functions $f_1, \ldots, f_d$ *and a nonnegative random variable* ξ *satisfying* $\mathbb{P}(\xi = 0) = 0$, *such that* $(f_1(X_1), \ldots, f_d(X_d))^T \sim EC_d(\mathbf{0}, \boldsymbol{\Sigma}, \xi)$, *where* $\boldsymbol{\Sigma}$, *called the latent generalized correlation matrix, satisfies the identifiability condition that* $\text{diag}(\boldsymbol{\Sigma}) = \mathbf{I}_d$.

Remark 3.2. *From the above definition, we can easily recover any elliptical distribution with nondegenerate marginals by choosing suitable linear functions* f_j. *Therefore, the*

transelliptical family is a strict extension of the elliptical family[2]*. Since* $\mathrm{diag}(\mathbf{\Sigma}) = \mathbf{I}_d$ *in Definition 3.1,* $f_j(X_j)$ *all have marginal densities. The transelliptical family of distributions is closed under sign change of individual variables, so that* $f_j(X_j)$ *with decreasing* f_j *can be viewed as* $-f_j(-X_j)$*. However, by assuming that* f_j *are all increasing functions, we gain the identifiability of the sign of the latent generalized correlation.*

We discuss in the rest of this subsection the relationship of the transelliptical family with the nonparanormal (Liu et al., 2012) and meta-elliptical (Fang et al., 2002) families, which are defined as follows:

Definition 3.3 (Nonparanormal distribution)**.** *A ramdom vector* $\boldsymbol{X} = (X_1, \ldots, X_d)^T$ *follows a nonparanormal distribution, denoted by*

$$\boldsymbol{X} \sim NPN_d(\mathbf{\Sigma}; f_1, \ldots, f_d),$$

if there exists a set of monotone functions $f_1, \ldots, f_d$ *such that* $(f_1(X_1), \ldots, f_d(X_d))^T \sim N_d(\mathbf{0}, \mathbf{\Sigma})$*, where* $\mathbf{\Sigma} \in \mathcal{R}_d^+$ *is called the latent correlation matrix.*

Definition 3.4 (Meta-elliptical distribution)**.** *Let* $\boldsymbol{X} = (X_1, \ldots, X_d)^T$ *be a random vector with marginal distribution functions* $F_1, \ldots, F_d$ *and a joint density. We say that* $\boldsymbol{X}$ *follows a meta-elliptical distribution, denoted by* $\boldsymbol{X} \sim ME_d(\mathbf{\Sigma}; Q_g, F_1, \ldots, F_d)$*, if there exists a continuous elliptical random vector* $\boldsymbol{Z} \sim EC_d(\mathbf{0}, \mathbf{\Sigma}, g)$ *with the marginal distribution function* Q_g *and* $\mathbf{\Sigma} \in \mathcal{R}_d^+$*, such that*

$$(Q_g^{-1}(F_1(X_1)), \ldots, Q_g^{-1}(F_d(X_d)))^T \stackrel{d}{=} \boldsymbol{Z}.$$

Here, $\mathbf{\Sigma}$ *is called the latent generalized correlation matrix.*

From Definitions 3.1 and 3.3, we see that the transelliptical is a strict extension of the nonparanormal. Both families assume that there exits a set of univariate transformations such that the transformed data follow a base distribution: the nonparanormal exploits a normal base distribution; While the transelliptical exploits an elliptical base distribution. For nonparanormal distributions, $\mathbf{\Sigma}$ is the correlation matrix of the latent normal, therefore it is called latent correlation matrix; For transelliptical distributions, $\mathbf{\Sigma}$ is the generalized correlation matrix of the latent elliptical distribution, therefore it is called latent generalized correlation matrix.

The comparison between the transelliptical and meta-elliptical families is subtler. Lemma 3.5 shows that the transelliptical family contains the meta-elliptical family. This lemma illustrates the intrinsic connection between the transelliptical and meta-elliptical families. One thing to note is that, even though they are equivalent when density exists, the way these

[2]Again, whenever we say the transelliptical family contains elliptical family, we only consider continuous elliptical distributions.

two families are defined are fundamentally different. The transelliptical family is defined by characterizing varaible transformations while the meta-elliptical family is defined by characterizing the density function. Later, we will show that the way we define transelliptical family brings new insights in both theoretical analysis and model interpretation.

Lemma 3.5. *Let $\boldsymbol{X} \sim ME_d(\boldsymbol{\Sigma}; Q_g, F_1, \ldots, F_d)$, where $\boldsymbol{\Sigma} \in \mathcal{R}_d^+$. There exists a scalar random variable $\xi \geq 0$ whose density exists and a set of univariate monotone functions $f_1, \ldots, f_d$, such that $\boldsymbol{X} \sim TE_d(\boldsymbol{\Sigma}, \xi; f_1, \ldots, f_d)$. Moreover, if $\boldsymbol{X} \sim TE_d(\boldsymbol{\Sigma}, \xi; f_1, \ldots, f_d)$, its joint density exists, and ξ is absolute continuous, then it is also meta-elliptically distributed.*

The transelliptical family is strictly larger than the meta-elliptical family. This can be seen from two perspectives: (i) For a transelliptical variable $\boldsymbol{X} \sim TE_d(\boldsymbol{\Sigma}, \xi; f_1, \ldots, f_d)$, the generating variable ξ is not necessarily absolute continuous with respect to the Lebesgue measure. In contrast, the meta-elliptical family requires the existence of a joint density, which implies the absolute continuity of the underlying generating variable ξ; (ii) The marginals of a transelliptical distribution do not necessarily possess density, while the marginal densities of a meta-elliptical distribution must exist.

3.2 Transelliptical Graphical Models

We now define the transelliptical graphical model. Let $\boldsymbol{X} \sim TE_d(\boldsymbol{\Sigma}, \xi; f_1, \ldots, f_d)$, where $\boldsymbol{\Sigma} \in \mathcal{R}_d^+$ is the latent generalized correlation matrix. We define $\boldsymbol{\Theta} := \boldsymbol{\Sigma}^{-1}$ to be the *latent generalized concentration matrix*. Let $\boldsymbol{\Theta}_{jk}$ be the element of $\boldsymbol{\Theta}$ on the j-th row and k-th column. We define the *latent generalized partial correlation matrix* $\boldsymbol{\Gamma}$ as

$$\boldsymbol{\Gamma}_{jk} := -\frac{\boldsymbol{\Theta}_{jk}}{\sqrt{\boldsymbol{\Theta}_{jj} \cdot \boldsymbol{\Theta}_{kk}}}.$$

It is easy to see that $\boldsymbol{\Gamma}$ has the same nonzero pattern as $\boldsymbol{\Theta} := \boldsymbol{\Sigma}^{-1}$ and

$$\boldsymbol{\Gamma} = -[\mathrm{diag}(\boldsymbol{\Sigma}^{-1})]^{-1/2} \boldsymbol{\Sigma}^{-1} [\mathrm{diag}(\boldsymbol{\Sigma}^{-1})]^{-1/2}. \tag{3.1}$$

We then define an undirected graph $G = (V, E)$: the vertex set V contains nodes corresponding to the d variables in $\boldsymbol{X}$, and the edge set E satisfies

$$(X_j, X_k) \in E \;\text{ if and only if }\; \boldsymbol{\Gamma}_{jk} \neq 0 \;\text{ for }\; j, k = 1, \ldots, d. \tag{3.2}$$

Given a graph G, we define $\mathcal{R}_d^+(G)$ to be

$$\mathcal{R}_d^+(G) := \big\{ \boldsymbol{\Sigma} \in \mathcal{R}_d^+ : \Lambda_{\min}(\boldsymbol{\Sigma}) > 0, G \text{ characterizes zero entries of } \boldsymbol{\Sigma}^{-1} \big\}.$$

In another word, $\boldsymbol{\Sigma} \in \mathcal{R}_d^+(G)$ implies that: $[\boldsymbol{\Sigma}^{-1}]_{jk} = 0$ whenever the edge (j, k) is absent in G. The transelliptical graphical model induced by a graph G is defined as:

Definition 3.6 (transelliptical graphical model)**.** *The transelliptical graphical model induced by a graph G, denoted by $\mathcal{P}(G)$, is defined to be the set of distributions:*

$$\mathcal{P}(G) := \Big\{ \textit{all the transelliptical distributions } TE_d(\mathbf{\Sigma}, \xi; f_1, \ldots, f_d) \textit{ satisfying } \mathbf{\Sigma} \in \mathcal{R}_d^+(G) \Big\}.$$

This graph G is called latent generalized partial correlation graph.

In the rest of this section, we prove some properties of the transelliptical distribution family and discuss the interpretation of the graph G. First, we show that the transelliptical family is closed under marginalization and conditioning. This result suggests that the conditional transelliptical graph given $\{X_j, j \in J\}$ is the subgraph of the marginal graph G with all vertices in J and edges connected to vertices not in J removed.

Lemma 3.7. *Let $\boldsymbol{X} := (X_1, \ldots, X_d)^T \sim TE_d(\mathbf{\Sigma}, \xi; f_1, \ldots, f_d)$. For any nontrivial subset J of $\{1, \ldots, d\}$, $\boldsymbol{X}_J$ is transelliptical with the latent generalized correlation matrix $\mathbf{\Sigma}_{J,J}$. Moreover, if $\mathbf{\Sigma}$ is of full rank, then conditionally on $\boldsymbol{X}_{J^c}$, $\boldsymbol{X}_J$ is transelliptical with the latent generalized correlation matrix* $\mathrm{diag}^{-1/2}\left([\mathbf{\Theta}_{J,J}]^{-1}\right)[\mathbf{\Theta}_{J,J}]^{-1}\mathrm{diag}^{-1/2}\left([\mathbf{\Theta}_{J,J}]^{-1}\right)$. *In particular, the marginal and conditional distributions of $(X_1, X_2)^T$ given the remaining variables are still transellipitical.*

From (3.1), we see that $\mathbf{\Gamma}$ and $\mathbf{\Theta}$ have the same nonzero pattern, therefore, they encode the same graph G. Let $\boldsymbol{X} \sim TE_d(\mathbf{\Sigma}, \xi; f_1, \ldots, f_d)$ belong to a transelliptical graphical model $\mathcal{P}(G)$. The next lemma shows that the absence of an edge in the graph G is equivalent to the pairwise conditional uncorrelatedness of the two corresponding latent variables. Note that we do not need the second moment condition $\mathbb{E}\xi^2 < \infty$ to make the conditional uncorrelatedness well-defined.

Lemma 3.8. *Let $\boldsymbol{X} := (X_1, \ldots, X_d)^T \sim TE_d(\mathbf{\Sigma}, \xi; f_1, \ldots, f_d)$ belong to a transelliptical graphical model $\mathcal{P}(G)$, and*

$$Z_j := f_j(X_j) \textit{ for } j = 1, \ldots, d.$$

Then, for $|J| < d$, $\mathbf{\Theta}_{J,J}$ is diagonal if and only if $\{Z_j, j \in J\}$ are uncorrelated given $\{Z_k, k \notin J\}$. In particular, for $d > 2$, $\mathbf{\Gamma}_{jk} = 0$ if and only if Z_j and Z_k are conditionally uncorrelated given $\boldsymbol{Z}_{\setminus\{j,k\}}$.

Let A, B, C be disjoint subsets of $\{1, \ldots, d\}$. We say C separates A and B in the graph G if any path from a node in A to a node in B goes through at least one node in C. We denote by $\boldsymbol{X}_A$ the subvector of $\boldsymbol{X}$ indexed by A. The next lemma reveals the connection between the pairwise and global conditional uncorrelatedness of the latent variables for the transelliptical graphical models. This theorem connects the graph theory with probability theory.

Theorem 3.9. *Let $\boldsymbol{X} \sim TE_d(\boldsymbol{\Sigma}, \xi; f_1, \ldots, f_d)$ belong to a transelliptical graphical model $\mathcal{P}(G)$. Let $\boldsymbol{Z} := (Z_1, \ldots, Z_d)^T$ with $Z_j = f_j(X_j)$ and $A, B, C \subset \{1, \ldots, d\}$ with $|C| > 0$. Then, if C separates A and B in G, $\boldsymbol{Z}_A$ and $\boldsymbol{Z}_B$ are conditionally uncorrelated given $\boldsymbol{Z}_C$. Conversely, if $A \cup B \cup C = \{1, \ldots, d\}$ and $\boldsymbol{Z}_A$ and $\boldsymbol{Z}_B$ are conditionally uncorrelated given $\boldsymbol{Z}_C$, then C separates A and B in G.*

Compared with the nonparanormal graphical model, the transelliptical graphical model gains a lot on modeling flexibility, but at the price of inferring a weaker notion of graphs: a missing edge in the graph only represents the conditional uncorrelatedness of the latent variables $Z_j = f_j(X_j)$. The next lemma shows that the nonparanormal is the only subfamily in which a transelliptical graph encodes the conditional independence relationships among the observed variables $X_1, \ldots, X_d$.

Theorem 3.10. *Let $\boldsymbol{X} \sim TE_d(\boldsymbol{\Sigma}, \xi; f_1, \ldots, f_d)$ be a member of the transelliptical graphical model $\mathcal{P}(G)$. Then the graph G encodes the conditional independence relationship of $\boldsymbol{X}$ (In other words, the distribution of $\boldsymbol{X}$ is Markov to G) if and only if $\boldsymbol{X}$ is nonparanormal.*

3.3 Relationships between Different Distribution Families

In this subsection, we present Theorem 3.11, which describes the relationship among the nonparanormal, the elliptical and the transelliptical families in detail. This theorem justifies Figure 1 in the introduction section.

Theorem 3.11. *Both the nonparanormal and elliptical families belong to the transelliptical family. Furthermore, if $\boldsymbol{X} \sim NPN_d(\boldsymbol{\Sigma}; f_1, \ldots, f_d)$ with $\mathrm{rank}(\boldsymbol{\Sigma}) > 1$ and $\boldsymbol{X}$ is also elliptically distributed, then $\boldsymbol{X}$ has a Gaussian distribution.*

Next, we present Theorem 3.12, which provides a latent Gaussian representation of the elliptical distribution. This theorem, together with the definition of transelliptical distribution, justifies Figure 2 in the introduction section.

Theorem 3.12. *Let $\boldsymbol{Z} \sim EC_d(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \xi)$ be an elliptical distribution with $\boldsymbol{\Sigma} = \mathbf{A}\mathbf{A}^T$. It takes another stochastic representation:*

$$\boldsymbol{Z} \stackrel{d}{=} \boldsymbol{\mu} + \xi \boldsymbol{Y} / \|\mathbf{A}^\dagger \boldsymbol{Y}\|_2,$$

where $\boldsymbol{Y} \sim N_d(\mathbf{0}, \boldsymbol{\Sigma})$, $\xi \geq 0$ is independent of $\boldsymbol{Y}/||\mathbf{A}^\dagger \boldsymbol{Y}||_2$ and $\mathbf{A}^\dagger$ is the Moore-Penrose pseudoinverse of $\mathbf{A}$.

3.4 Conditional Uncorrelatedness with Respect to Rank Association

Let $\widetilde{\boldsymbol{X}} = (\widetilde{X}_1, \ldots, \widetilde{X}_d)^T$ be an independent copy of $\boldsymbol{X}$. The population version of the Kendall's tau statistic is

$$\tau_{jk} := \mathrm{Corr}\big(\mathrm{sign}(X_j - \widetilde{X}_j), \mathrm{sign}(X_k - \widetilde{X}_k)\big).$$

Let $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n \in \mathbb{R}^d$ be n observed data points with $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{id})^T$. The sample version Kendall's tau statistic is defined as:

$$\widehat{\tau}_{jk} := \frac{2}{n(n-1)} \sum_{1 \le i < i' \le n} \operatorname{sign}\left(x_{ij} - x_{i'j}\right)\left(x_{ik} - x_{i'k}\right),$$

which is a monotone transformation-invariant measure of association between the empirical realizations of two random variables X_j and X_k. It is easy to verify that $\mathbb{E}\widehat{\tau}_{jk} = \tau_{jk}$. Another interpretation of the Kendall's tau statistic is based on the notion of concordance. We call two pairs of real numbers (s,t) and (u,v) concordant if $(s-t)(u-v) > 0$ and disconcordant if $(s-t)(u-v) < 0$. Kruskal (1958) shows that

$$\tau_{jk} = \mathbb{P}\big((Z_j - \widetilde{Z}_j)(Z_k - \widetilde{Z}_k) > 0\big) - \mathbb{P}\big((Z_j - \widetilde{Z}_j)(Z_k - \widetilde{Z}_k) < 0\big). \tag{3.3}$$

Let $\boldsymbol{X} \sim TE_d(\boldsymbol{\Sigma}, \xi; f_1, \ldots, f_d)$, the following theorem illustrates an important relationship between the population Kendall's tau statistic τ_{jk} and the latent generalized correlation coefficient $\boldsymbol{\Sigma}_{jk}$. A similar result held for the meta-elliptical distribution family can be seen in Fang et al. (2002).

Theorem 3.13. *Let $\boldsymbol{X} := (X_1, \ldots, X_d)^T \sim TE_d(\boldsymbol{\Sigma}, \xi; f_1, \ldots, f_d)$ and τ_{jk} be the population Kendall's tau statistic between X_j and X_k. Then, $\boldsymbol{\Sigma}_{jk} = \sin\left(\frac{\pi}{2}\tau_{jk}\right)$.*

From Lemma 3.7, we know that $(X_j, X_k)^T \,|\, \boldsymbol{X}_C$ follows a transelliptical distribution for any nontrivial subset C of $\{1, \ldots, d\}$. Let $\tau(X_j, X_k \,|\, \boldsymbol{X}_C)$ be the population Kendall's tau correlation under this conditional distribution and $\tau(\boldsymbol{X}_J \,|\, \boldsymbol{X}_{J^c})$ the $|J| \times |J|$ matrix of such conditional Kendall's tau with elements $[\tau(\boldsymbol{X}_J|\boldsymbol{X}_{J^c})]_{jk} = \tau(X_{J_j}, X_{J_k}|\boldsymbol{X}_{J^c})$. The next lemma shows that the graph G obtained in (3.2) characterizes pairwise conditional uncorrelatedness with respect to the rank correlation graph. For such an interpretation, we again do not need the second moment condition $\mathbb{E}\xi^2 < \infty$.

Lemma 3.14. *Let $\boldsymbol{X} \sim TE_d(\boldsymbol{\Sigma}, \xi; f_1, \ldots, f_d)$ belong to the transelliptical graphical model $\mathcal{P}(G)$. Then, for $|J| < d$, $\boldsymbol{\Theta}_{J,J}$ is diagonal if and only if $\tau(\boldsymbol{X}_J|\boldsymbol{X}_{J^c})$ is diagonal. In particular, we have $\boldsymbol{\Theta}_{jk} = \boldsymbol{\Gamma}_{jk} = 0$ if and only if $\tau(X_j, X_k \,|\, \boldsymbol{X}_{\setminus\{j,k\}}) = 0$.*

For disjoint subsets A, B and C of $\{1, 2, \ldots, d\}$, we define the population Kendall's tau correlation matrix conditionally on $\boldsymbol{X}_C$, $\tau(\boldsymbol{X}_A, \boldsymbol{X}_B|\boldsymbol{X}_C)$, as the matrix with elements $\tau(X_j, X_k|\boldsymbol{X}_C), j \in A, k \in B$. The next theorem characterizes the connection between the graph G and global conditional uncorrelatedness with respect to the rank correlation graph.

Theorem 3.15. *Let $\boldsymbol{X} \sim TE_d(\boldsymbol{\Sigma}, \xi; f_1, \ldots, f_d)$ belong to the transelliptical graphical model $\mathcal{P}(G)$. Let A, B, C be three disjoint subsets of $\{1, 2, \ldots, d\}$. If C separates A and B in the graph G, then $\tau(\boldsymbol{X}_A, \boldsymbol{X}_B|\boldsymbol{X}_C) = \mathbf{0}$. Conversely, if $A \cup B \cup C = \{1, \ldots, d\}$ and $\tau(\boldsymbol{X}_A, \boldsymbol{X}_B|\boldsymbol{X}_C) = \mathbf{0}$, then C separates A and B in G.*

4 Parameter and Graph Estimations

In this section, we propose two nonparametric rank-based regularization estimators which achieve parametric rates of convergence for both graph recovery and parameter estimation. The main idea of our procedure is to treat the marginal transformation functions f_j and the generating variable ξ as nuisance parameters, and exploit the nonparametric Kendall's tau statistic to directly estimate the latent generalized correlation matrix $\mathbf{\Sigma}$. The obtained correlation matrix estimate is then plugged into either the CLIME estimator or its variant–named *Scaled CLIME*—to estimate the sparse latent generalized concentration matrix $\mathbf{\Theta}$. From the previous discussion, we know that the graph G is encoded by the nonzero pattern of $\mathbf{\Theta}$. We then get a graph estimator by thresholding the estimated $\widehat{\mathbf{\Theta}}$. Theoretically, we show that even though the transelliptical family is larger than the nonparanormal and elliptical families, our procedures achieve the same parametric rates of convergence as CLIME for graph recovery and parameter estimation. Moreover, the rank-based scaled CLIME estimator can be shown to be non-asymptotically tuning-insensitive and asymptotically tuning-free. Computationally, it has the same computational complexity as the CLIME estimator.

4.1 Rank-based estimation of latent generalized correlation matrix

Let $I(\cdot)$ be the indicator function. Motivated by the explicit relationship between the Kendall's tau and the generalized correlation in Theorem 3.13, we define a raw estimate of the latent generalized correlation matrix $\mathbf{\Sigma}$ as

$$\widehat{\mathbf{S}} = [\widehat{\mathbf{S}}_{jk}] \in \mathbb{R}^{d\times d}, \text{ where } \widehat{\mathbf{S}}_{jk} = \sin\left(\tfrac{\pi}{2}\widehat{\tau}_{jk}\right) \cdot I(j \neq k) + I(j = k). \tag{4.1}$$

For the transelliptical family, this Kendall's tau-based raw estimator plays a role parallel to that of the sample correlation matrix for the multivariate Gaussian family. For the estimation of a fixed $\mathbf{\Sigma}$, $\widehat{\mathbf{S}}$ is $n^{-1/2}$ consistent and asymptotically normal. For large sparse $\mathbf{\Sigma}$, $\widehat{\mathbf{S}}$ is not sparse but it can be thresholded to produce a sparse estimator of $\mathbf{\Sigma}$. Similar to Bickel and Levina (2008a,b), it can be shown that such a thresholded estimator is consistent in the spectral norm when $\max_{j\leq d}\sum_{k=1}^{d}\min(|\mathbf{\Sigma}_{jk}|, \sqrt{(\log d)/n}) \to 0$. When the thresholded $\widehat{\mathbf{S}}$ is consistent in the spectral norm and the spectral norm of the latent generalized concentration matrix $\mathbf{\Theta} = \mathbf{\Sigma}^{-1}$ is bounded, the thresholded estimator can be inverted to yield a consistent estimator of $\mathbf{\Theta}$. However, thresholding the estimated correlation does not guarantee consistency when the sparsity assumption is imposed on the latent generalized concentration matrix of the transelliptical graphical model.

4.2 Rank-based CLIME Estimator

When a suitable sparsity assumption on the latent generalized concentration matrix $\mathbf{\Theta} = \mathbf{\Sigma}^{-1}$ is reasonable, we propose to estimate $\mathbf{\Theta}$ by plugging the Kendall's tau-based raw

estimator (4.1) into the CLIME estimator (Cai et al., 2011). More specifically, the latent generalized concentration matrix $\boldsymbol{\Theta}$ can be estimated by solving

$$\widehat{\boldsymbol{\Theta}} = \arg\min_{\boldsymbol{\Omega}} \sum_{j,k} |\boldsymbol{\Omega}_{jk}| \quad \text{subject to } \|\widehat{\mathbf{S}}\boldsymbol{\Omega} - \mathbf{I}_d\|_{\max} \leq \lambda, \tag{4.2}$$

where $\lambda > 0$ is a tuning parameter.

By Cai et al. (2011), this optimization problem can be decomposed into d vector minimization problems and solved in parallel. For each j, we can directly estimates the j^{th} column of $\boldsymbol{\Theta}$ by solving

$$\widehat{\boldsymbol{\Theta}}_{*j} = \underset{\boldsymbol{\Omega}_{*j}}{\operatorname{argmin}} \big\|\boldsymbol{\Omega}_{*j}\big\|_1 \ \text{ subject to } \ \big\|\widehat{\mathbf{S}}\boldsymbol{\Omega}_{*j} - \boldsymbol{e}_j\big\|_\infty \leq \lambda, \ \text{ for } j = 1, \ldots, d, \tag{4.3}$$

where $\boldsymbol{e}_j$ is the j^{th} canonical vector and δ_j is a tuning parameter. This decomposability of the CLIME indicates its potential scalability to large problems.

Once $\widehat{\boldsymbol{\Theta}}$ is obtained, we can apply an additional thresholding step to estimate the graph G. To this end, we define a graph estimator

$$\widehat{G} = (V, \widehat{E}), \quad \text{where } (j,k) \in \widehat{E} \text{ if and only if } |\widehat{\boldsymbol{\Theta}}_{jk}| \geq \gamma. \tag{4.4}$$

Here γ is another tuning parameter. Empirically, we found the rank-based CLIME procedure works very effectively in graph estimation even without this hard-thresholding step (i.e., simply setting $\gamma = 0$ already delivers good graph estimates). Thus, this hard-thresholding step is more of theoretical interest and may not be necessary in applications. For all the empirical results illustrated in this paper, we set $\gamma = 0$.

4.3 Rank-based Scaled CLIME Estimator

As will be explained in Theorem 5.1, the above rank-based CLIME estimator requires a tuning parameter λ depending on the unknown quantity $\|\boldsymbol{\Theta}\|_1$. Therefore, even though the algorithm runs in a column-by-column fashion, all the regression subproblems in (4.3) must use the same tuning parameter λ. Such a restriction may lead to the use of an overly conservative tuning-parameter with the rank-based CLIME, especially in cases where different subproblems require different levels of regularization. In this section, we propose a new estimator named rank-based scaled CLIME which is more tuning-insensitive. In another word, the tuning parameter λ does not depend on the unknown quantity $\|\boldsymbol{\Theta}\|_1$. As will be reported in the experimental section, the rank-based scaled CLIME outperforms the rank-based original CLIME in many situations in terms of graph estimation.

Recall that any real number a takes the decomposition $a = a^+ - a^-$, where $a^+ = a \cdot I(a \geq 0)$ and $a^- = -a \cdot I(a < 0)$. For any vector $\boldsymbol{v} = (v_1, \ldots, v_d)^T \in \mathbb{R}^d$, let $\boldsymbol{v}^+ := (v_1^+, \ldots, v_d^+)^T$ and $\boldsymbol{v}^- = (v_1^-, \ldots, v_d^-)$. We have $\boldsymbol{v} = \boldsymbol{v}^+ - \boldsymbol{v}^-$ and $||\boldsymbol{v}||_1 = \mathbf{1}_d^T(\boldsymbol{v}^+ + \boldsymbol{v}^-)$, where $\mathbf{1}_d := (1, \ldots, 1)^T$. We say that $\boldsymbol{v} \geq 0$ if $\min_{j \leq d} v_j \geq 0$, $\boldsymbol{v}_1 \geq \boldsymbol{v}_2$ if $\boldsymbol{v}_1 - \boldsymbol{v}_2 \geq 0$, and $\boldsymbol{v}_1 \leq \boldsymbol{v}_2$

if $\boldsymbol{v}_2 - \boldsymbol{v}_1 \geq 0$. Letting $\mathbb{R}^d_{0+}$ represent the d dimensional real space where each entry is nonnegative, the scaled CLIME estimator is defined as the following convex program which can be efficiently computed by a linear program solver:

Rank-based Scaled CLIME Estimator

For $j = 1, \ldots, d$, we calculate $\widehat{\boldsymbol{\Theta}}_{*j} := \widehat{\boldsymbol{\beta}}_j^+ - \widehat{\boldsymbol{\beta}}_j^-$ where $\widehat{\boldsymbol{\beta}}_j^+$ and $\widehat{\boldsymbol{\beta}}_j^-$ are solved by the following convex program

$$\{\widehat{\boldsymbol{\beta}}_j^+, \widehat{\boldsymbol{\beta}}_j^-\} = \underset{\boldsymbol{\beta}_j^+, \boldsymbol{\beta}_j^- \in \mathbb{R}^d_{0+}}{\operatorname{argmin}} \ \mathbf{1}_d^T(\boldsymbol{\beta}_j^+ + \boldsymbol{\beta}_j^-), \tag{4.5}$$

$$\text{subject to } \ \left\|\widehat{\mathbf{S}}(\boldsymbol{\beta}_j^+ - \boldsymbol{\beta}_j^-) - \boldsymbol{e}_j\right\|_\infty \leq \lambda_0 \mathbf{1}_d^T(\boldsymbol{\beta}_j^+ + \boldsymbol{\beta}_j^-). \tag{4.6}$$

We next present more insights on the formulation of the rank-based scaled CLIME estimator. Motivated by the rank-based CLIME estimator (4.3), we consider the modification:

$$\widehat{\boldsymbol{\Theta}}_{*j} = \underset{\boldsymbol{\Omega}_{*j}}{\operatorname{argmin}} \left\|\boldsymbol{\Omega}_{*j}\right\|_1 \ \text{ subject to } \ \left\|\widehat{\mathbf{S}}\boldsymbol{\Omega}_{*j} - \boldsymbol{e}_j\right\|_\infty \leq \lambda_0 \left\|\boldsymbol{\Omega}_{*j}\right\|_1, \ \text{ for } j = 1, \ldots, d, \tag{4.7}$$

The main difference between (4.7) and (4.3) is the replacement of the penalty level λ by $\lambda_0\left\|\boldsymbol{\Omega}_{*j}\right\|_1$, which leads to a non-convex problem. However, (4.5) and (4.6) provide a convex relaxation of (4.7) in a linear programming form. More specifically, letting $\mathbf{C} = [\lambda_0] \in \mathbb{R}^{d\times d}$, we may write (4.5) and (4.6) as:

$$\{\widehat{\boldsymbol{\beta}}_j^+, \widehat{\boldsymbol{\beta}}_j^-\} = \underset{\boldsymbol{\beta}_j^+, \boldsymbol{\beta}_j^-}{\operatorname{argmin}} \mathbf{1}_d^T(\boldsymbol{\beta}_j^+ + \boldsymbol{\beta}_j^-) \tag{4.8}$$

$$\text{subject to } \begin{cases} \widehat{\mathbf{S}}\boldsymbol{\beta}_j^+ - \widehat{\mathbf{S}}\boldsymbol{\beta}_j^- - \boldsymbol{e}_j \leq \mathbf{C}(\boldsymbol{\beta}_j^+ + \boldsymbol{\beta}_j^-), \\ -\widehat{\mathbf{S}}\boldsymbol{\beta}_j^+ + \widehat{\mathbf{S}}\boldsymbol{\beta}_j^- + \boldsymbol{e}_j \leq \mathbf{C}(\boldsymbol{\beta}_j^+ + \boldsymbol{\beta}_j^-), \\ \boldsymbol{\beta}_j^+ \geq 0 \\ \boldsymbol{\beta}_j^- \geq 0. \end{cases}$$

Letting $\mathbf{R} = \widehat{\mathbf{S}} + \mathbf{C}$ and $\mathbf{W} = \widehat{\mathbf{S}} - \mathbf{C}$, we write (4.8) as

$$\omega = \underset{\omega}{\operatorname{argmin}} \mathbf{1}_{2d}^T \omega \ \text{ subject to } \boldsymbol{\theta} + \mathbf{A}\omega \geq 0, \text{ and } \omega \geq 0, \tag{4.9}$$

where

$$\omega = \begin{pmatrix} \boldsymbol{\beta}_j^+ \\ \boldsymbol{\beta}_j^- \end{pmatrix}, \boldsymbol{\theta} = \begin{pmatrix} \boldsymbol{e}_j \\ -\boldsymbol{e}_j \end{pmatrix}, \text{ and } \mathbf{A} = \begin{bmatrix} -\mathbf{W} & \mathbf{R} \\ \mathbf{R} & -\mathbf{W} \end{bmatrix}.$$

Equation (4.9), however, is a linear programming problem. In this paper, we use the simplex algorithm to compute both the rank-based CLIME and rank-based scaled CLIME estimators.

5 Theoretical Properties

In this section we analyze the theoretical properties of the rank-based original CLIME and scaled CLIME estimators proposed in Section 4. Our main results show that: under the same conditions on $\mathbf{\Sigma}$ that ensure the parameter estimation and graph recovery consistency of the original CLIME estimator under Gaussian graphical models, our rank-based regularization procedures achieve exactly the same parametric rates of convergence for both parameter estimation and graph recovery for the much larger transelliptical family. This result suggests that the transelliptical graphical model can be used as a safe replacement of the Gaussian graphical models, the nonparanormal graphical models, and the elliptical graphical models.

We start with some additional notation. Let M_d be a quantity which may scale with the dimensionality d, we define

$$\mathcal{S}_d(q, s, M_d) := \Big\{\mathbf{\Theta} : \|\mathbf{\Theta}\|_1 \leq M_d \text{ and } \|\mathbf{\Theta}\|_q \leq s\Big\}. \tag{5.1}$$

For $q = 0$, the class $\mathcal{S}_d(0, s, M)$ contains all the s-sparse matrices.

Theorem 5.1 presents the parameter estimation and graph estimation consistency results for the rank-based CLIME estimator defined in (4.2).

Theorem 5.1. *Let $\boldsymbol{X} \sim TE_d(\mathbf{\Sigma}, \xi; f_1, \ldots, f_d)$ with $\mathbf{\Sigma} \in \mathcal{R}_d^+$ and $\mathbf{\Theta} := \mathbf{\Sigma}^{-1} \in \mathcal{S}_d(q, s, M_d)$ with $0 \leq q < 1$. Let $\widehat{\mathbf{\Theta}}$ be defined in (4.2). There exist constants C_0 and C_1 only depending on q, such that, whenever we choose the tuning parameter*

$$\lambda = C_0 M_d \sqrt{\frac{\log d}{n}}, \tag{5.2}$$

with probability no less than $1 - d^{-2}$, we have

$$\text{(Parameter estimation)} \qquad \|\widehat{\mathbf{\Theta}} - \mathbf{\Theta}\|_2 \leq C_1 M_d^{2-2q} \cdot s \cdot \left(\frac{\log d}{n}\right)^{(1-q)/2}.$$

Let $\widehat{G}$ be the graph estimator defined in (4.4) with the second step tuning parameter $\gamma = 4M_d\lambda$. If we further assume $\mathbf{\Theta} \in \mathcal{S}_d(0, s, M_d)$ and $\min_{j,k:|\mathbf{\Theta}_{jk}|\neq 0} |\mathbf{\Theta}_{jk}| \geq 2\gamma$, then

$$\text{(Graph recovery)} \qquad \mathbb{P}\left(\widehat{G} \neq G\right) \geq 1 - o(1),$$

where G is the graph determined by the nonzero pattern of $\mathbf{\Theta}$.

Similar rank-based procedures have been discussed in Liu et al. (2009, 2012). Unlike our work, they focus on the more restrictive nonparanormal family and discuss several rank-based procedures using the normal-score, Spearman's rho, and Kendall's tau. Moreover, they advocate the use of the Spearman's rho and normal-score correlation coefficients. Their main concern is that, within the more restrictive nonparanormal family, the Spearman's rho

and normal-score correlations are slightly easier to compute and have smaller asymptotic variance. In contrast to their line of thinking, the new insight here is the invariance property of the Kendall's tau within the much larger transelliptical family, which has led to our advocate of using the Kendall's tau. In fact, it can be shown that the Spearman's rho is not invariant within the transelliptical family unless the true distribution is nonparanormal. More details on this issue can be found in Fang et al. (1990).

An important issue arising from Theroem 5.1 is the dependence of the tuning parameter λ in (5.3) on M_d. Since M_d is an upper bound of $\|\mathbf{\Theta}\|_1$ and the tuning parameter λ is uniformly set for all the regression subproblems in (4.3), such a choice of the tuning parameter tends to be overly conservative. In particular, it is not adaptive to cases where different subproblems require different levels of regularization. Theorem 5.2 presents the parameter estimation and graph estimation consistency results for the rank-based scaled CLIME estimator defined in (4.5) and (4.6). The result shows that the rank-based scaled CLIME achieves the same theoretical results as the rank-based CLIME estimator with a tuning parameter λ_0 not dependent on the knowledge of $\|\mathbf{\Theta}\|_1$.

Theorem 5.2. *Let* $\boldsymbol{X} \sim TE_d(\mathbf{\Sigma}, \xi; f_1, \ldots, f_d)$ *with* $\mathbf{\Sigma} \in \mathcal{R}_d^+$ *and* $\mathbf{\Theta} := \mathbf{\Sigma}^{-1} \in \mathcal{S}_d(q, s, M_d)$ *with* $0 \le q < 1$. *Let* $\widehat{\mathbf{\Theta}}$ *be the rank-based scaled CLIME estimator defined in* (4.5) *and* (4.6). *When we choose the tuning parameter*

$$\lambda_0 = 2.45\pi\sqrt{\frac{\log d}{n}}, \tag{5.3}$$

with probability no less than $1 - d^{-2} - d^{-1}$, *we have*

$$\text{(Parameter estimation)} \qquad \|\widehat{\mathbf{\Theta}} - \mathbf{\Theta}\|_2 \le \frac{8}{2^q} \cdot M_d^{2-2q} \cdot s \cdot \left(\frac{\log d}{n}\right)^{(1-q)/2}.$$

Let $\widehat{G}$ *be the graph estimator defined in Section 4 with the second step tuning parameter* $\gamma = 4M_d^2\lambda_0$. *If we further assume* $\mathbf{\Theta} \in \mathcal{S}_d(0, s, M_d)$ *and* $\min_{j,k:|\mathbf{\Theta}_{jk}|\neq 0} |\mathbf{\Theta}_{jk}| \ge 2\gamma$, *then*

$$\text{(Graph recovery)} \qquad \mathbb{P}\left(\widehat{G} \neq G\right) \ge 1 - o(1),$$

where G *is the graph determined by the nonzero pattern of* $\mathbf{\Theta}$.

From (5.3), we see that the tuning parameter λ_0 for the rank-based scaled CLIME is asymptotically tuning free. However, the default value $\lambda_0 = 2.45\pi\sqrt{\frac{\log d}{n}}$ could be overly conservative to achieve the best finite sample performance. To achieve the best empirical performance, we still need to find a better tuning parameter λ_0 within the same order.

6 Numerical Experiments

We investigate the empirical performance of the proposed rank-based regularization estimators. We compare them with the following methods: (1) Pearson: the CLIME/scaled

CLIME using the Pearson sample correlation; (2) Kendall: the CLIME/scaled CLIME using the Kendall's tau; (3) Spearman: the CLIME/scaled CLIME using the Spearman's rho; (4) NPN: the CLIME/scaled CLIME using the original nonparanormal correlation estimator proposed by Liu et al. (2009); (5) NS: the CLIME/scaled CLIME using the normal score correlation. The later three methods are discussed under the nonparanormal graphical model and we refer to Liu et al. (2012) for more detailed descriptions.

6.1 Simulation Studies

We adopt the same data generating procedure as in Liu et al. (2012). To generate a d dimensional sparse graph $G = (V, E)$ where $V = \{1, \ldots, d\}$ correspond to variables $\boldsymbol{X} = (X_1, \ldots, X_d)^T$, we associate each index $j \in \{1, \ldots, d\}$ with a bivariate data point $(Y_j^{(1)}, Y_j^{(2)}) \in [0,1]^2$ where $Y_1^{(k)}, \ldots, Y_n^{(k)} \sim \text{Uniform}[0,1]$ for $k = 1, 2$. Each pair of vertices (i, j) is included in the edge set E with probability $\mathbb{P}((i,j) \in E) = \exp(-\|y_i - y_j\|_2^2/0.25)/\sqrt{2\pi}$, where $y_i := (y_i^{(1)}, y_i^{(2)})$ is the empirical observation of $(Y_i^{(1)}, Y_i^{(2)})$ and $\|\cdot\|_2$ represents the Euclidean distance. We restrict the maximum degree of the graph to be 4 and build the inverse correlation matrix $\boldsymbol{\Omega}$ according to $\boldsymbol{\Omega}_{jk} = 1$ if $j = k$, $\boldsymbol{\Omega}_{jk} = 0.145$ if $(j,k) \in E$, and $\boldsymbol{\Omega}_{jk} = 0$ otherwise. The value 0.145 guarantees the positive definiteness of $\boldsymbol{\Omega}$ because the diagonal values dominate. Let $\boldsymbol{\Sigma} = \boldsymbol{\Omega}^{-1}$. To obtain the correlation matrix, we rescale $\boldsymbol{\Sigma}$ so that all its diagonal elements are 1.

In the simulated study we randomly sample n data points from a certain transelliptical distribution $\boldsymbol{X} \sim TE_d(\boldsymbol{\Sigma}, \xi; f_1, \ldots, f_d)$. We set $d = 100$. To determine the transelliptical distribution, we first generate $\boldsymbol{\Sigma}$ as discussed in the previous paragraph. Secondly, three types of ξ are considered. Here we remind that χ_d denotes the chi-distribution: For any random variable $Y \in \mathbb{R}^+$, $Y \sim \chi_d$ if and only if $Y^2 \sim \chi_d^2$. (1) $\xi^{(1)} \sim \chi_d$, i.e., ξ follows a chi-distribution with degree of freedom d; (2) $\xi^{(2)} \overset{d}{=} \xi_1^*/\xi_2^*$, $\xi_1^* \sim \chi_d$, $\xi_2^* \sim \chi_1$, ξ_1^* is independent of ξ_2^*; (3) $\xi^{(3)} \sim F(d, 1)$, i.e., ξ follows an F-distribution with degree of freedom d and 1. Thirdly, two type of transformation functions $f = \{f_j\}_{j=1}^d$ are considered: (1) **linear transformation:** $f^{(1)} = \{f_0, \ldots, f_0\}$ with $f_0(x) = x$; (2) **nonlinear transformation:** $f^{(2)} = \{f_1, \ldots, f_d\} = \{h_1, h_2, h_3, h_4, h_5, h_1, h_2, h_3, h_4, h_5, \ldots\}$, where

$$h_1^{-1}(x) := x, \quad h_2^{-1}(x) := \frac{\text{sign}(x)|x|^{1/2}}{\sqrt{\int |t|\phi(t)dt}}, \quad h_3^{-1}(x) := \frac{x^3}{\sqrt{\int t^6\phi(t)dt}},$$

$$h_4^{-1}(x) := \frac{\Phi(x) - \int \Phi(t)\phi(t)dt}{\sqrt{\int \left(\Phi(y) - \int \Phi(t)\phi(t)dt\right)^2 \phi(y)dy}},$$

$$h_5^{-1}(x) := \frac{\exp(x) - \int \exp(t)\phi(t)dt}{\sqrt{\int \left(\exp(y) - \int \exp(t)\phi(t)dt\right)^2 \phi(y)dy}}. \tag{6.1}$$

Here ϕ and Φ are the density and distribution functions of a standard Gaussian distribution. We then consider the following four data generating models:

- **Model 1:** $\boldsymbol{X} \sim TE_d(\boldsymbol{\Sigma}, \xi^{(1)}; f^{(1)})$, i.e., $\boldsymbol{X} \sim N_d(0, \boldsymbol{\Sigma})$.
- **Model 2:** $\boldsymbol{X} \sim TE_d(\boldsymbol{\Sigma}, \xi^{(2)}; f^{(1)})$, i.e., $\boldsymbol{X}$ follows the multivariate Cauchy.
- **Model 3:** $\boldsymbol{X} \sim TE_d(\boldsymbol{\Sigma}, \xi^{(3)}; f^{(1)})$, i.e., the distribution is highly related to the multivariate t.
- **Model 4:** $\boldsymbol{X} \sim TE_d(\boldsymbol{\Sigma}, \xi^{(3)}; f^{(2)})$.

To evaluate the robustness of different methods, let $r \in [0, 1)$ represent the proportion of samples being contaminated. For each component of $\boldsymbol{X}$, we randomly select $\lfloor nr \rfloor$ entries and replace them with either 5 or -5 with equal probability. The final data matrix we obtained is $\mathbf{X} \in \mathbb{R}^{n \times d}$. Here we pick $r = 0, 0.02$ or 0.05. Let $\widehat{G}^\lambda = (V, \widehat{E}^\lambda)$ be an estimated graph using the regularization parameter λ and γ. We further define the false negative rate (FNR) and false positive rate (FPR) as

$$\mathrm{FNR}(\lambda) := \frac{\text{the number of edges in } \widehat{E}^\lambda \text{ but not in } E}{|E|} \quad \text{and} \tag{6.2}$$

$$\mathrm{FPR}(\lambda) := \frac{\text{the number of edges in } E \text{ but not in } \widehat{E}^\lambda}{\binom{d}{2} - |E|}. \tag{6.3}$$

Under Model 1 to Model 4 with different levels of contamination ($r = 0, 0.02$ or 0.05), we repeatedly generate the data matrix $\mathbf{X}$ 100 times and compute the averaged False Positive Rates $\overline{\mathrm{FPR}}(\lambda)$ and averaged False Negative Rates $\overline{\mathrm{FNR}}(\lambda)$ using a path of tuning parameters λ from 0.01 to 0.5 and $\gamma = 0$. The feature selection performances of different methods are evaluated by plotting $(\overline{\mathrm{FPR}}(\lambda), 1 - \overline{\mathrm{FNR}}(\lambda))$. The corresponding ROC curves are presented in Figures 3 and 4.

To further evaluate the performance of different methods, we define the oracle regularization parameter λ^* to be:

$$\lambda^* := \arg\min_{\lambda \in \Lambda} \left\{ \mathrm{FNR}(\lambda) + \mathrm{FPR}(\lambda) \right\},$$

Here Λ denotes by a full path of regularization parameters we use. We define FPR and FNR to be:

$$\mathrm{FPR} = \mathrm{FPR}(\lambda^*) \quad \text{and} \quad \mathrm{FNR} = \mathrm{FNR}(\lambda^*).$$

Table 1 provides numerical comparisons of the five methods on datasets with different models, where we repeat the experiments 100 times and report the average FPR and FNR values with the corresponding standard errors in the parenthesis.

Table 1: Quantitative comparison of the five methods on simulated datasets with different models. The graphs are estimated using the CLIME algorithm with random data contamination.

			Pearson		Kendall		Spearman		NPN		NS	
Model	r	n	FPR(%)	FNR	FPR	FNR	FPR	FNR	FPR	FNR	FPR	FNR
Model1	0.00	200	14(2.0)	26(3.0)	17(2.0)	27(1.8)	15(2.0)	30(1.4)	15(3.6)	26(4.1)	14(2.3)	26(3.2)
		400	9(1.5)	10(1.9)	10(1.6)	12(2.1)	10(1.6)	11(1.9)	10(1.8)	10(2.2)	9(1.4)	10(1.9)
		800	2(0.6)	2(0.8)	3(0.8)	3(1.1)	3(0.7)	3(1.1)	3(0.8)	2(0.7)	3(0.7)	2(0.7)
	0.02	200	23(5.0)	44(5.8)	20(3.7)	30(2.2)	20(2.9)	30(3.2)	18(2.6)	32(3.1)	16(2.4)	35(3.3)
		400	19(3.0)	27(3.4)	11(2.0)	15(2.2)	11(1.8)	15(2.3)	11(1.7)	15(2.7)	12(1.9)	17(2.6)
		800	10(1.4)	14(2.1)	4(1.0)	4(1.2)	4(1.0)	4(1.3)	5(0.9)	5(1.4)	5(1)	6(1.5)
	0.05	200	30(7.7)	53(7.2)	21(4.9)	37(5.4)	18(5.2)	41(5.7)	20(4.3)	42(4.1)	21(3.6)	44(3.7)
		400	27(5.2)	45(5.5)	14(2.3)	21(3.4)	14(2.5)	21(3.5)	16(2.4)	24(3.3)	18(3.2)	27(3.9)
		800	19(3.2)	33(4.2)	7(1.5)	8(1.5)	7(1.5)	8(1.6)	9(1.8)	10(1.8)	11(1.8)	14(2.3)
Model2	0.00	200	25(3.6)	61(3.6)	21(3.4)	42(2.8)	18(2.0)	46(1.8)	24(2.4)	43(2.3)	23(2.5)	44(2.1)
		400	28(3.5)	59(3.7)	15(1.4)	24(1.8)	17(1.3)	23(1.5)	22(1.9)	26(1.9)	22(1.7)	27(1.9)
		800	28(4.0)	58(4.0)	8(0.9)	10(1.0)	9(0.9)	11(1.0)	13(1.2)	15(1.3)	14(1.3)	16(1.6)
	0.02	200	27(3.9)	62(3.3)	19(1.6)	47(1.9)	20(2.0)	48(2.9)	25(1.7)	45(2.9)	27(1.5)	43(2.3)
		400	25(2.9)	63(3.4)	16(1.2)	27(1.7)	18(1.5)	25(1.5)	22(2.0)	28(2.2)	24(2.2)	28(2.5)
		800	29(5.1)	59(4.7)	10(1.1)	12(1.2)	10(1.1)	13(1.3)	14(1.3)	17(1.5)	16(1.4)	18(1.7)
	0.05	200	22(3.9)	68(4.3)	18(2.1)	50(2.1)	16(2.0)	52(2.6)	28(3.3)	42(3.4)	26(2.1)	45(2.5)
		400	23(2.9)	65(3.2)	17(1.3)	31(1.6)	19(1.9)	30(2.1)	23(1.7)	32(1.9)	25(2.0)	32(2.2)
		800	26(3.2)	62(3.6)	11(1.1)	14(1.4)	12(1.1)	16(1.2)	18(1.5)	19(1.8)	18(1.8)	20(2)
Model3	0.00	200	28(5.7)	65(6.2)	25(3.5)	44(3)	23(3.3)	48(3.6)	32(2.3)	43(2.5)	28(2.0)	46(2.0)
		400	29(8.2)	65(8.8)	15(1.1)	31(1.6)	18(1.3)	32(1.5)	22(1.7)	35(2.0)	25(2.4)	34(2.7)
		800	24(6.1)	70(6.6)	11(1.0)	13(1.1)	13(1.3)	15(1.3)	17(1.5)	21(1.9)	19(1.6)	21(1.7)
	0.02	200	24(6.6)	70(7.2)	24(3.4)	44(2.6)	28(3.4)	44(3.1)	25(2.7)	48(3.3)	27(2.6)	47(2.7)
		400	25(6.3)	69(6.4)	17(1.5)	33(1.6)	20(1.6)	33(1.8)	23(1.9)	37(2.2)	25(2.0)	37(2.2)
		800	22(7.1)	73(7.6)	12(1.2)	15(1.5)	14(1.1)	17(1.5)	18(1.3)	24(1.7)	20(1.5)	24(1.9)
	0.05	200	25(4.9)	69(5.8)	26(3.3)	47(2.6)	28(4.4)	47(3.4)	21(3.8)	58(4.7)	29(3.5)	50(4.2)
		400	24(6.1)	70(6.5)	18(2.3)	38(2.5)	20(2.1)	38(2.3)	27(2.7)	39(3.1)	27(2.6)	40(3.0)
		800	23(6.5)	72(7.0)	14(1.2)	19(1.4)	15(1.2)	22(1.6)	21(1.8)	27(1.9)	21(1.7)	28(1.8)
Model4	0.00	200	21(3.8)	75(4)	20(2.6)	49(3.2)	23(2.0)	49(2.8)	34(2.9)	40(2.4)	30(3.4)	44(3.4)
		400	21(5.6)	74(6.0)	16(1.7)	31(2.0)	20(1.7)	30(2.1)	23(1.9)	34(2.2)	23(1.9)	35(2.3)
		800	21(6.0)	74(6.6)	10(0.8)	13(1.3)	13(1.0)	15(1.4)	17(1.1)	21(1.5)	17(1.1)	23(1.4)
	0.02	200	11(3.7)	85(4.3)	22(3.3)	49(3.3)	22(3.0)	53(3.7)	31(4.4)	48(5.0)	28(3.9)	51(4.5)
		400	11(3.3)	86(3.7)	17(1.6)	33(1.8)	21(1.8)	32(2.0)	22(2.2)	39(2.6)	24(2.5)	38(2.6)
		800	14(3.9)	82(4.4)	12(1.3)	14(1.5)	14(1.3)	17(1.5)	17(1.5)	25(1.7)	18(1.6)	25(1.7)
	0.05	200	12(3.9)	84(4.5)	29(2.7)	45(3.1)	28(3.8)	50(4.4)	27(2.9)	54(2.4)	26(3.0)	56(3.4)
		400	9(2.8)	87(3.3)	19(1.8)	35(2.1)	21(2.0)	37(1.9)	25(2.3)	41(2.8)	26(2.6)	42(3.2)
		800	11(3.4)	86(4.2)	14(1.1)	18(1.3)	16(1.3)	22(1.5)	20(1.7)	28(1.9)	21(1.9)	30(2.0)

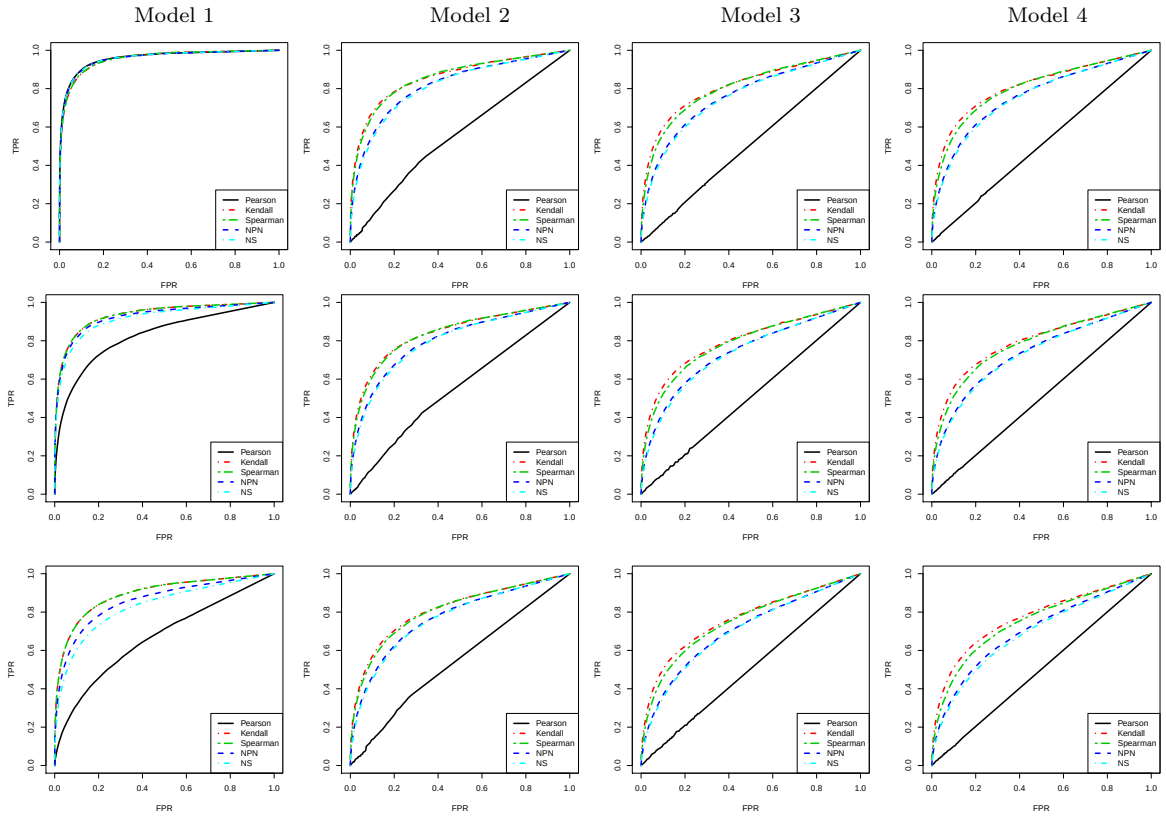


Figure 3: ROC curves for different methods in models 1 to 4 and different contamination level $r = 0, 0.02, 0.05$ (top, middle, bottom) using the CLIME. Here $n = 400$ and $d = 100$. Here TPR and FPR stand for the true positive rate and false positive rate.

Next we proceed to show the results for the scaled CLIME. To best visualize the difference between CLIME and scaled CLIME, a slightly different graph structure, called Erdös-Rényi random graph, are utilized. In other words, we add an edge between each pair of codes with probability 0.02 independently. We are still using the same four models as discussed before. We set $d = 100, n = 400$ to best illustrate the comparisons among different methods. The corresponding ROC curves are shown in Figures 5 and numerical comparisons are presented in Table 2. It can be observed that the scaled CLIME has a better performance than the CLIME in most settings.

We summarize our discovery as follows: (1) when the data are perfectly Gaussian without contamination, all five methods perform well; (2) when the data are Gaussian, but contaminated by outliers, rank-based methods (Kendall, Spearman, NPN, NS) perform better than Pearson; (3) when the data are non-Gaussian, outliers existing or not, rank-based methods perform better than Pearson; (5) when the data are non-Gaussian, outliers existing or not, Kendall performs the best; (6) the scaled CLIME can have a generally better performance than the CLIME.

Table 2: Quantitative comparison of the six methods on simulated datasets with different models. The Erdös-Rènry random graph are estimated using the scaled CLIME and CLIME algorithms with random data contamination. Note: "CLIME-K" applies the CLIME and Kendall's tau matrix. All the others use the scaled CLIME.

				Pearson		Kendall		CLIME-K	
Model	d	n	r	FPR(%)	FNR	FPR	FNR	FPR	FNR
Model1	100	400	0.00	0.4(0.23)	0.0(0.00)	0.6(0.37)	0.1(0.20)	0.9(0.40)	0.0(0.00)
	100	400	0.02	5.7(1.11)	4.1(1.45)	0.8(0.36)	0.3(0.40)	0.9(0.31)	0.4(0.44)
	100	400	0.05	14.5(2.13)	20.3(2.97)	2.2(0.57)	1.3(0.79)	2.6(0.65)	0.9(0.65)
Model2	100	400	0.00	14.3(1.47)	61.8(3.58)	4.2(0.37)	4.3(0.50)	6.3(0.54)	4.9(0.66)
	100	400	0.02	17.8(1.79)	57.1(3.17)	5.3(0.39)	4.9(0.62)	6.3(0.50)	5.7(0.68)
	100	400	0.05	16.2(1.75)	61.2(3.77)	6.7(0.66)	8.3(0.78)	7.2(0.69)	9.4(0.90)
Model3	100	400	0.00	16.0(3.50)	73.3(3.55)	3.3(0.28)	3.2(0.43)	3.7(0.33)	3.2(0.46)
	100	400	0.02	9.2(0.98)	79.6(1.89)	4.2(0.43)	4.5(0.51)	4.9(0.46)	4.4(0.57)
	100	400	0.05	12.2(3.11)	76.8(3.39)	4.4(0.36)	6.3(0.74)	4.7(0.45)	5.8(0.71)
Model4	100	400	0.00	45.6(5.15)	47.2(4.71)	3.3(0.28)	3.2(0.43)	3.7(0.33)	3.2(0.46)
	100	400	0.02	12.2(1.67)	80.6(2.23)	4.0(0.40)	4.3(0.57)	5.0(0.51)	3.9(0.52)
	100	400	0.05	19.4(2.06)	72.2(2.39)	4.9(0.47)	5.6(0.72)	4.8(0.41)	5.6(0.63)
				Spearman		NPN		NS	
Model	d	n	r	FPR(%)	FNR	FPR	FNR	FPR	FNR
Model1	100	400	0.00	0.9(0.39)	0.0(0.00)	0.5(0.27)	0.0(0.00)	0.5(0.23)	0.0(0.00)
	100	400	0.02	0.9(0.31)	0.4(0.44)	0.6(0.20)	0.6(0.51)	1.0(0.33)	0.6(0.51)
	100	400	0.05	2.5(0.60)	1.1(0.71)	3.8(0.93)	1.5(0.86)	4.4(1.01)	2.9(1.33)
Model2	100	400	0.00	4.3(0.53)	4.3(0.56)	8.7(0.65)	7.0(0.84)	9.6(0.60)	7.7(0.88)
	100	400	0.02	5.6(0.48)	4.9(0.75)	9.6(0.70)	9.0(1.01)	10.7(0.70)	9.3(0.87)
	100	400	0.05	7.1(0.71)	9.0(0.90)	11(0.79)	12.4(1.21)	10.9(0.83)	13.7(1.16)
Model3	100	400	0.00	4.8(0.40)	3.6(0.47)	7.3(0.54)	7.2(0.87)	7.9(0.63)	8.2(0.87)
	100	400	0.02	5.1(0.44)	5.5(0.64)	7.5(0.50)	9.4(0.86)	8.5(0.64)	9.7(0.90)
	100	400	0.05	6.3(0.50)	6.9(0.78)	10.0(0.60)	11.3(0.89)	10(0.52)	12.3(0.96)
Model4	100	400	0.00	4.8(0.40)	3.6(0.47)	7.3(0.54)	7.2(0.87)	7.9(0.63)	8.2(0.87)
	100	400	0.02	5.2(0.51)	5.4(0.69)	7.2(0.58)	10.1(0.94)	7.9(0.64)	10.4(0.95)
	100	400	0.05	6.3(0.55)	7.5(0.76)	9.9(0.80)	11.9(0.94)	10.2(0.70)	13.8(0.99)

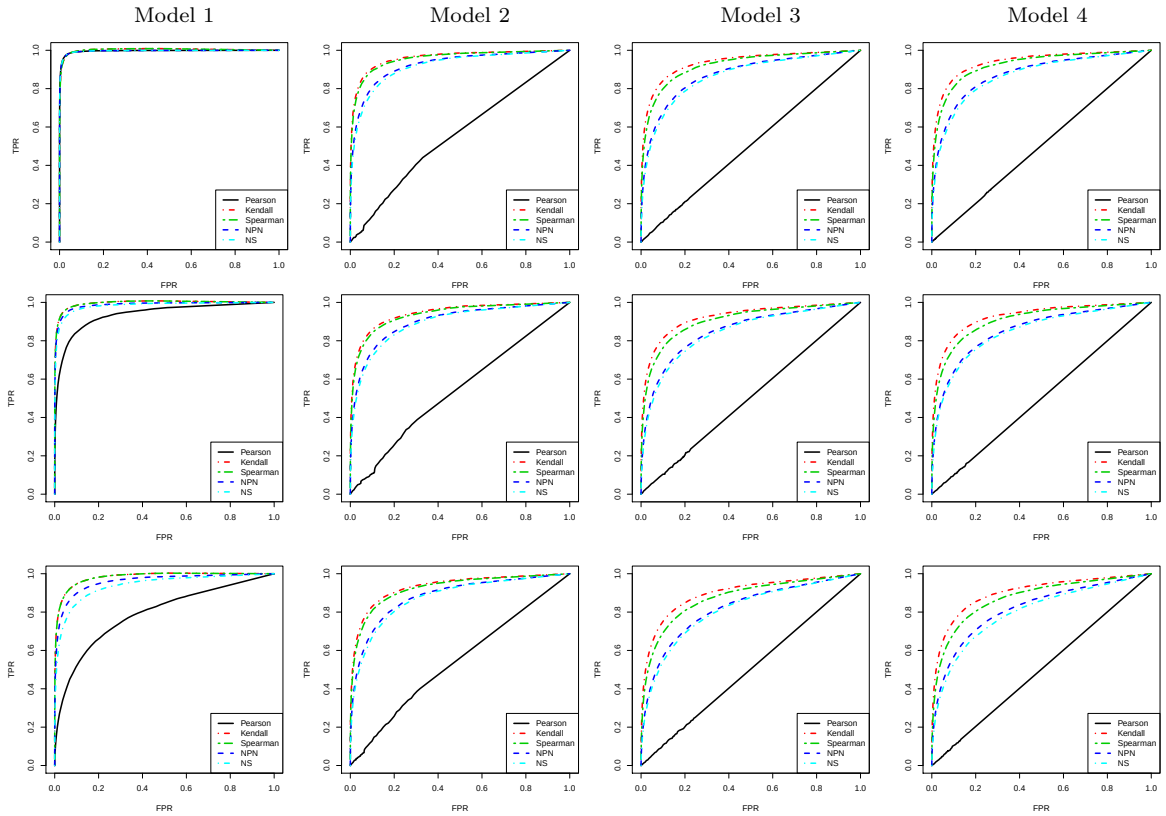


Figure 4: ROC curves for different methods in models 1 to 4 and different contamination level $r = 0, 0.02, 0.05$ (top, middle, bottom) using the CLIME. Here $n = 800$ and $d = 100$. Here TPR and FPR stand for the true positive rate and false positive rate.

6.2 Equities Data

We compare different methods on the stock price data from Yahoo! Finance (`finance.yahoo.com`). We collect the daily closing prices for 452 stocks that are consistently in the S&P 500 index between January 1, 2003 through January 1, 2008. This gives us altogether 1,257 data points, each data point corresponding to the vector of closing prices on a trading day. With $S_{t,j}$ denoting the closing price of stock j on day t, we consider the variables $X_{tj} = \log(S_{t,j}/S_{t-1,j})$ and build graphs over the indices j. Though this is a time series, we treat the instances X_t as independent replicates.

The 452 stocks are categorized into 10 Global Industry Classification Standard (GICS) sectors, including `Consumer Discretionary` (70 stocks), `Consumer Staples` (35 stocks), `Energy` (37 stocks), `Financials` (74 stocks), `Health Care` (46 stocks), `Industrials` (59 stocks), `Materials` (29 stocks), `Information Technology` (64 stocks) `Telecommunications Services` (6 stocks), and `Utilities` (32 stocks).

Figure 6(a) shows the histogram and qq-norm plot of the 25th stock. Here we observe that the marginal distribution is quite non-Gaussian. We plot the data points for the first

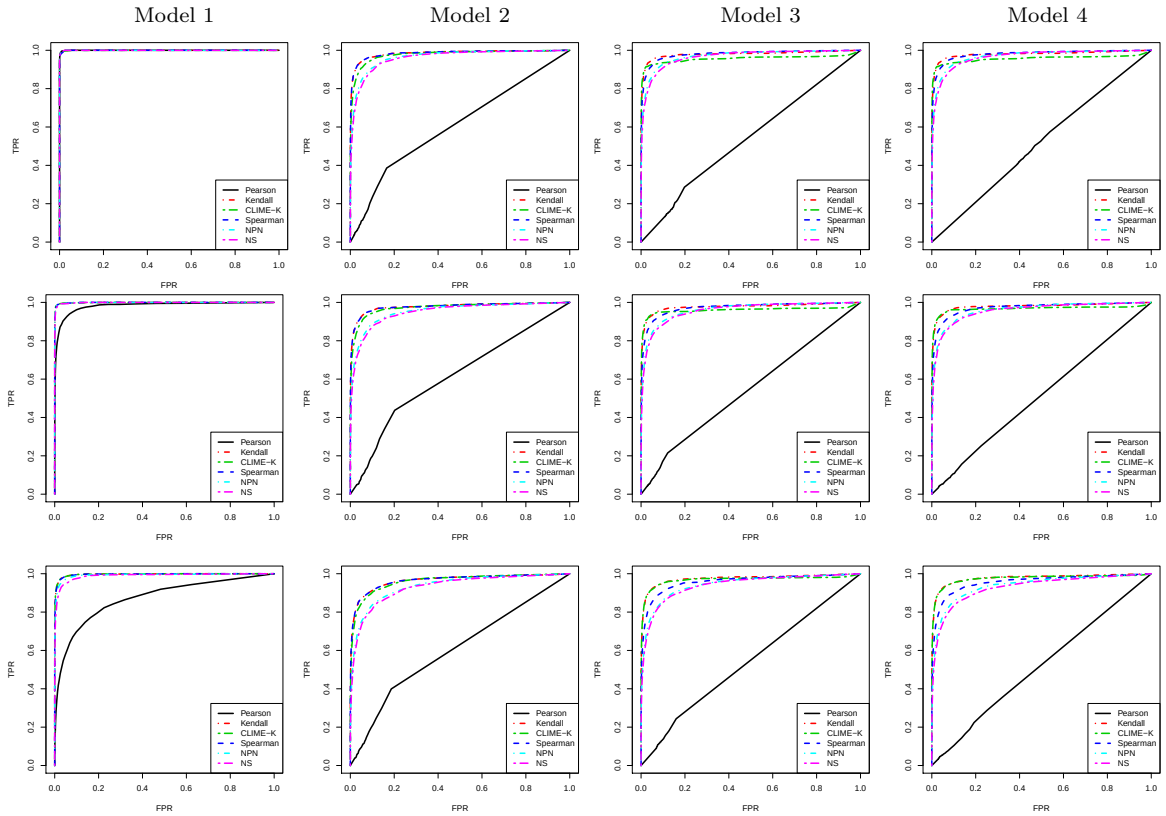
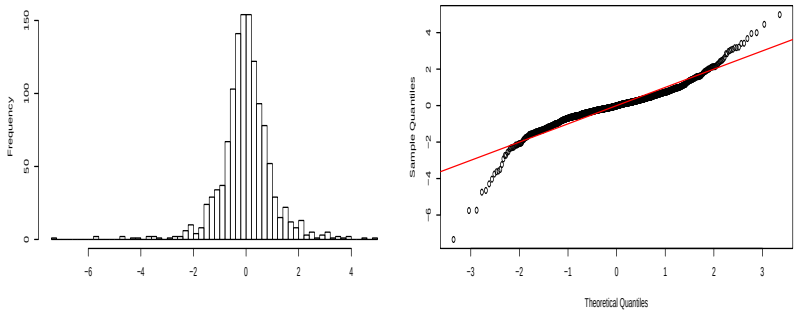


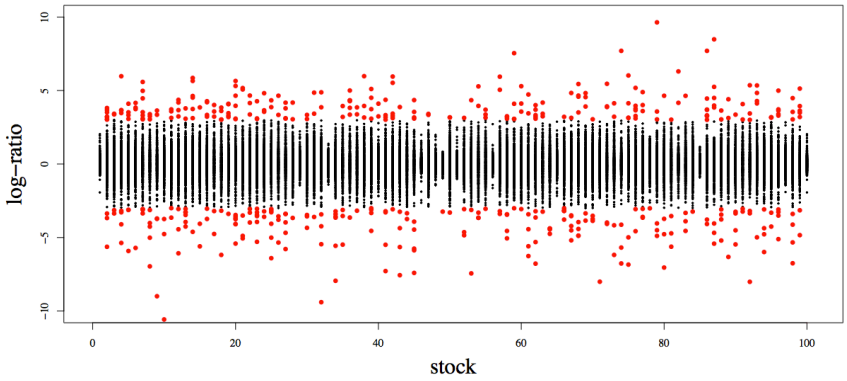
Figure 5: ROC curves for different methods in models 1 to 4 and different contamination level $r = 0, 0.02, 0.05$ (top, middle, bottom). "CLIME-K" applies the CLIME on Kendall's tau matrix. All the others are using the scaled CLIME. Here $n = 400$ and $d = 100$ and the generating graph model is Erdos-Renyi random graph. Here TPR and FPR stand for the true positive rate and false positive rate.

100 stocks in Figure 6(b). Here we highlight data points in red if their absolute values are greater than 3. We can see that there is non-Gaussian issue and a large number of potential outliers exist. Both of them may affect the quality of the estimated graph.

In this section we evaluate different variants of the scale CLIME algorithm. Since Kendall, Spearman, NPN, NS are rank-based, they are much more robust to outliers. Moreover, Kendall is more adaptive to model assumptions than all the other methods. The tuning parameter is selected such that there are near 1% of the edges remained. Moreover, the numbers of edges obtained by using different methods are close to each other. We find that the estimated graphs are similar with the graphs learnt by using a stability based approach named StARS (Liu et al., 2010b). Some summary statistics of the estimated graphs are presented in Table 3. From Table 3, we see that the estimated graph of Pearson is significantly different from the other rank-based methods, suggesting that the data are highly non-Gaussian. We further compare the Kendall with NPN and NS and find that there are



(a)Histogram and Quantile-quantile Plot



(b) Outlier Plot

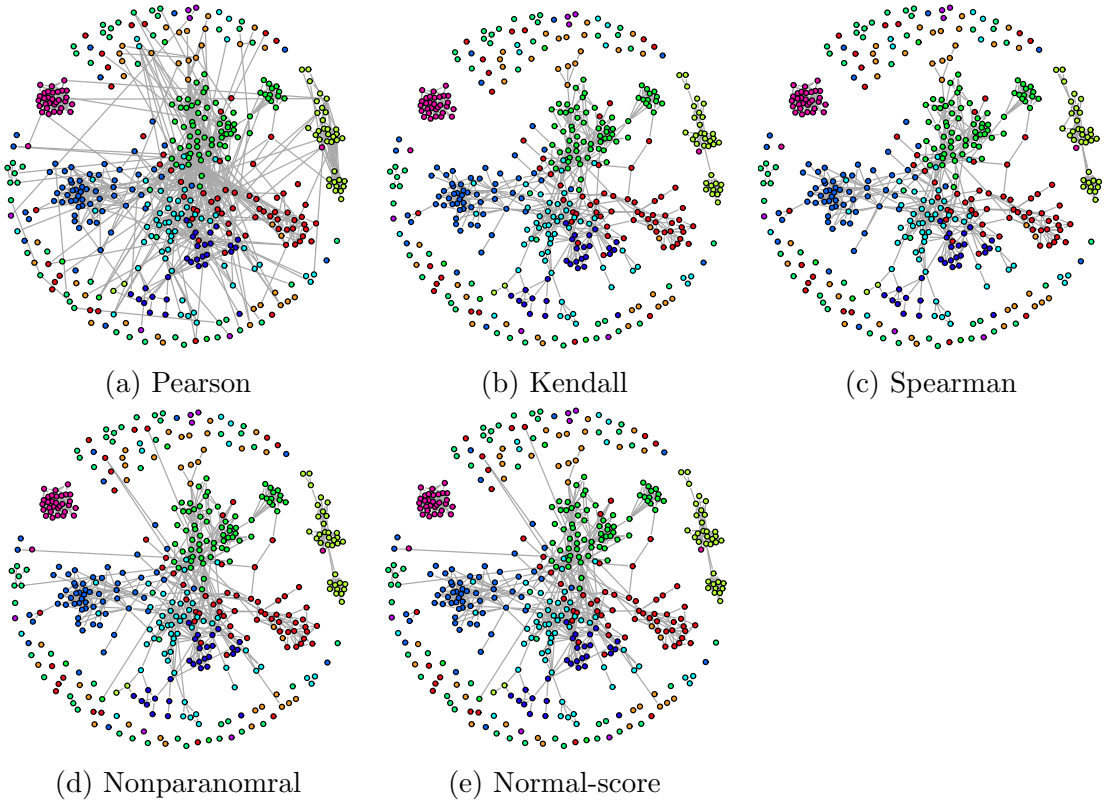
Figure 6: Stock Market Dataset. (a) We can see that the marginal distribution is away from the Gaussian; (b) We can see a large amounted of the outliers (Red dots). The existence of outliers and non-Gaussian data greatly affect the quality of the estimated graph

around 10% edges that are not present in the Kendall graph, suggesting that this data may contain high levels of outliers. Finally, we find that around 5% edges present in the Spearman graph but not in the Kendall graph, suggesting that the data are not nonparanormal. This conclusion is consistent with common finance theory, in which the log-return data are heavy tailed and have tail dependency. Since the nonparnormal is a subfamily of Gaussian copula, it is well known that a non-degenerate Gaussian copula can not capture any tail dependency.

For better visualization, we plot the five different graphs in Figure 7. Figure 7 illustrates the estimated graphs using different methods (Pearson, Kendall, Spearman, NPN and NS). The nodes are colored according to the GICS sectors of the stocks. Here the common layout is drawn by a force-based algorithm using the estimated graph from the Kendall. We see that different methods deliver slightly different graphs. We see the stocks from the same GICS sector tends to be grouped with each other in the Kendall graph, pink and red points

Table 3: Summary statistics of the stock data networks estimated. Note: In the edge difference part, the number corresponding to the method A(row) and method B(column) represents the number of edges appears in the estimated graph of A, but not in that of B

		Edge diff					
Method	Edge No.		Pearson	Kendall	Spearman	NPN	NS
Pearson	1015	Pearson	0	519	519	467	457
Kendall	1008	Kendall	512	0	48	96	111
Spearman	1008	Spearman	512	48	0	93	109
NPN	1059	NPN	511	147	144	0	37
NS	1052	NS	494	155	153	30	0



(a) Pearson (b) Kendall (c) Spearman

(d) Nonparanomral (e) Normal-score

Figure 7: The graph estimated from the S&P 500 stock data from Jan. 1, 2003 to Jan. 1, 2008 using Pearson, Kendall, Spearman, NPN, NS (left to right). The nodes are colored according to their GICS sector categories.

in particular. This result suggests that Kendall delivers an informative graph estimate.

7 Dicussion

In this paper, we advocate the use of a new distribution family, named transelliptical, for robust estimation of high dimensional semiparametric graphical models. The main contributions of this paper include the following: (i) We generalize the nonparanormal and elliptical distribution families to the larger transelliptical family; (ii) We construct the transelliptical graphical model and provide a three-layer hierarchical latent variable interpretation of the inferred graph; (iii) We provide sharp characterization of the relationships between nonparnaormal, elliptical, meta-elliptical, and transelliptical families; (iv) We propose a new rank-based scaled CLIME method which is simultaneously tuning-insensitive and adaptive over the whole transelliptical family; (v) Theoretically, our method achieves parametric rates in both graph and latent generalized concentration matrix estimation. These results suggest that the extra robustness and flexibility gained by the semiparametric transelliptical modeling come with little cost. We also provide numerical experiments on synthetic and real datasets. By Lemma 3.5, our method also works for the meta-elliptical family. To the best of our knowledge, no effective method has been proposed to estimate high dimensional graphical models for the meta-elliptical family.

A Proof of the Results in Section 5

In this appendix we present the technical proofs.

A.1 Proof of Lemma 3.5

Proof. If $\boldsymbol{X} \sim ME_d(\boldsymbol{\Sigma}; Q_g, F_1, \ldots, F_d)$, we know that $\boldsymbol{X}$ has density and there exist continuous marginal distribution functions $F_1, \ldots, F_d$ of $\boldsymbol{X}$. In addition, there exists a a continuous random vector $\boldsymbol{Z} \sim EC_d(\boldsymbol{0}, \boldsymbol{\Sigma}, g)$ such that

$$(Q_g^{-1}(F_1(X_1)), \ldots, Q_g^{-1}(F_d(X_d)))^T \stackrel{d}{=} \boldsymbol{Z}.$$

We define $f_j(x) = Q_g^{-1}(F_j(x))$ for $j = 1, \ldots, d$. It is obvious that f_j is monotone. Therefore, by the definition of elliptical distribution, there must exist a ξ such that $\boldsymbol{X} \sim TE_d(\boldsymbol{\Sigma}, \xi; Q_g^{-1}(F_1), \ldots, Q_g^{-1}(F_d))$. The absolute continuousness of ξ has been proved by Theorem 2.9 in the page 35 of Fang et al. (1990). Also, ξ must be nondegenerate due to the continuity of $\boldsymbol{X}$.

On the other hand, if $\boldsymbol{X} \sim TE_d(\boldsymbol{\Sigma}, \xi; f_1, \ldots, f_d)$ and its joint density exists, then $\boldsymbol{\Sigma}$ must be positive definite. Combining with the fact that ξ is absolute continuous, we have that $\boldsymbol{Z} := (f_1(X_1), \ldots, f_d(X_d))^T$ possesses density and can be represented as $\boldsymbol{Z} \sim EC_d(\boldsymbol{0}, \boldsymbol{\Sigma}, g)$ for some g. From the definition of the transelliptical distribution and Theorem 2.16 of Fang et al. (1990), we know that the marginal distributions of $Z_1, \ldots, Z_d$ are exactly the same. We denote this common marginal cumulative distribution function to be Q_g where $g(\cdot)$ is

the scale function uniquely determined by the distribution of ξ (See (2.1) for more details). Recalling that F_j is the cumulative distribution function of X_j and using the fact that f_j is strictly increasing, we have for any $z \in \mathbb{R}$,

$$f_j(X_j) = Z_j \Rightarrow \mathbb{P}(f_j(X_j) < z) = \mathbb{P}(Z_j < z) \Rightarrow F_j(f_j^{-1}(z)) = Q_g(z).$$

Since the last equality holds for any $z \in \mathbb{R}$, we have $f_j = Q_g^{-1}(F_j)$. Thus,

$$(Q_g^{-1}(F_1(X_1)), \ldots, Q_g^{-1}(F_d(X_d)))^T \sim EC_d(\mathbf{0}, \mathbf{\Sigma}, g).$$

Using Definition 3.4, we have $\boldsymbol{X}$ is meta-elliptically distributed. □

A.2 Proof of Lemma 3.7

To prove Lemma 3.7, we need the following lemma, which states the independence property of Gaussian distributions.

Lemma A.1. *Let $\boldsymbol{Z} \sim N_d(\mathbf{0}, \mathbf{I}_d)$ and $\mathbf{A}$ be any matrix with d columns, we have $||\boldsymbol{Z}||_2$ is independent of $\mathbf{A}\boldsymbol{Z}/||\mathbf{A}\boldsymbol{Z}||_2$. In particular, $||\boldsymbol{Z}||_2$ is independent of $\boldsymbol{Z}/||\boldsymbol{Z}||_2$.*

Proof. It is easy to see that $\boldsymbol{Z}$ is spherically distributed. Therefore, $\boldsymbol{Z}$ has the stochastic representation $\boldsymbol{Z} \overset{d}{=} \xi^* \boldsymbol{U}^*$ for some random variable $\xi^* \in \mathbb{R}$ and random vector $\boldsymbol{U}^*$ uniformly distributed on a unit sphere. Here ξ^* and $\boldsymbol{U}^*$ are independent and $\mathbb{P}(\xi^* = 0) = 0$. Therefore,

$$\begin{pmatrix} ||\boldsymbol{Z}||_2 \\ \mathbf{A}\boldsymbol{Z}/||\mathbf{A}\boldsymbol{Z}||_2 \end{pmatrix} \overset{d}{=} \begin{pmatrix} \xi^* \\ \mathbf{A}\boldsymbol{U}^*/||\mathbf{A}\boldsymbol{U}^*||_2 \end{pmatrix}$$

are independent of each other. This completes the proof. □

With lemma begging, we now proceed to prove Lemma 3.7.

Proof of Lemma 3.7. Since $\boldsymbol{X} \sim TE_d(\mathbf{\Sigma}, \xi; f_1, \ldots, f_d)$, we have $(f_1(X_1), \ldots, f_d(X_d))^T \sim EC_d(\mathbf{0}, \mathbf{\Sigma}, \xi)$. Let $Z_j := f_j(X_j)$ for $j = 1, \ldots, d$. By Definition 2.1, we can write $\boldsymbol{Z} := (Z_1, \ldots, Z_d)^T \overset{d}{=} \xi \mathbf{A}\boldsymbol{\epsilon}/||\boldsymbol{\epsilon}||_2$ with a certain $\boldsymbol{\epsilon} \sim N_q(0, \mathbf{I}_q)$, reminding that $q := \text{rank}(\mathbf{\Sigma})$. Let $\mathbf{P} = \mathbf{A}_{J*}^{\dagger}\mathbf{A}_{J*}$, where $\mathbf{A}_{J*}^{\dagger}$ is the the Moore-Penrose pseudoinverse of $\mathbf{A}_{J*}$. We rewrite

$$\boldsymbol{Z}_J \overset{d}{=} \xi \mathbf{A}_{J*}\boldsymbol{\epsilon}/||\boldsymbol{\epsilon}||_2 = (\xi \|\mathbf{P}\boldsymbol{\epsilon}\|_2/\|\boldsymbol{\epsilon}\|_2)(\mathbf{A}_{J*}\mathbf{P})(\mathbf{P}\boldsymbol{\epsilon}/\|\mathbf{P}\boldsymbol{\epsilon}\|_2). \tag{A.1}$$

Because $\mathbf{P}$ is a symmetric projection matrix, its singular value decomposition can be written as $\mathbf{P} = \mathbf{H}\mathbf{H}^T$, where $\mathbf{H} \in \mathbb{R}^{q \times r}$ has r orthonormal columns and $r := \text{rank}(\mathbf{A}_{J*}) = \text{rank}(\mathbf{P})$. By algebra, we have

$$\begin{aligned} \boldsymbol{Z}_J &\overset{d}{=} (\xi ||\mathbf{P}\boldsymbol{\epsilon}||_2/||\boldsymbol{\epsilon}||_2)(\mathbf{A}_{J*}\mathbf{P})(\mathbf{P}\boldsymbol{\epsilon}/||\mathbf{P}\boldsymbol{\epsilon}||_2) \\ &= (\xi ||\mathbf{H}^T\boldsymbol{\epsilon}||_2/||\boldsymbol{\epsilon}||_2)(\mathbf{A}_{J*}\mathbf{H})(\mathbf{H}^T\boldsymbol{\epsilon}/||\mathbf{H}^T\boldsymbol{\epsilon}||_2). \end{aligned}$$

Because $\mathbf{H}^T\boldsymbol{\epsilon} \sim N_r(\mathbf{0}, \mathbf{I}_r)$, $\mathbf{H}^T\boldsymbol{\epsilon}/||\mathbf{H}^T\boldsymbol{\epsilon}||_2$ is uniformly distributed on the unit sphere in $\mathbb{R}^r$. Moreover, $||\boldsymbol{\epsilon}||_2^2 - ||\mathbf{H}^T\boldsymbol{\epsilon}||_2^2 = ||\mathbf{P}^\perp\boldsymbol{\epsilon}||_2^2$ is independent of $\mathbf{H}^T\boldsymbol{\epsilon} = \mathbf{H}^T\mathbf{P}\boldsymbol{\epsilon}$. Therefore, the independence of $\{||\boldsymbol{\epsilon}||_2, ||\mathbf{H}^T\boldsymbol{\epsilon}||_2\}$ and $\mathbf{H}^T\boldsymbol{\epsilon}/||\mathbf{H}^T\boldsymbol{\epsilon}||_2$ follows from Lemma A.1 in Appendix A.2. Accordingly, letting $|J|$ be the cardinality of J, we have

$$\boldsymbol{Z}_J \sim EC_{|J|}(\mathbf{0}, (\mathbf{A}_{J*}\mathbf{H})(\mathbf{A}_{J*}\mathbf{H})^T, \xi||\mathbf{H}^T\boldsymbol{\epsilon}||_2/||\boldsymbol{\epsilon}||_2).$$

This can be written as

$$\boldsymbol{Z}_J \sim EC_{|J|}(\mathbf{0}, \boldsymbol{\Sigma}_{J,J}, \xi\|\mathbf{P}\boldsymbol{\epsilon}\|_2/\|\boldsymbol{\epsilon}\|_2) \tag{A.2}$$

due to $||\mathbf{H}^T\boldsymbol{\epsilon}||_2 = ||\mathbf{P}\boldsymbol{\epsilon}||_2$ and $(\mathbf{A}_{J*}\mathbf{H})(\mathbf{A}_{J*}\mathbf{H})^T = \mathbf{A}_{J*}\mathbf{A}_{J*}^T = \boldsymbol{\Sigma}_{J,J}$. By definition, the marginal distribution of $\boldsymbol{X}_J$ is transelliptically distributed.

Now consider the conditional distribution of $\boldsymbol{X}_J$. Let $\mathbf{Q} = \mathbf{A}_{J^c*}^\dagger\mathbf{A}_{J^c*}$, $\mathbf{Q}^\perp$ be the projection matrix perpendicular to $\mathbf{Q}$, $\widetilde{\boldsymbol{\mu}} = \mathbf{A}_{J*}\mathbf{A}_{J^c*}^\dagger\boldsymbol{Z}_{J^c}$ and $\widetilde{\xi} = \xi||\mathbf{Q}^\perp\boldsymbol{\epsilon}||_2/||\boldsymbol{\epsilon}||_2$. By algebra,

$$\begin{aligned}
\boldsymbol{Z}_J &\overset{d}{=} \mathbf{A}_{J*}(\mathbf{Q}^\perp + \mathbf{Q})(\xi\boldsymbol{\epsilon}/\|\boldsymbol{\epsilon}\|_2) \\
&= \mathbf{A}_{J*}\mathbf{A}_{J^c*}^\dagger\boldsymbol{Z}_{J^c} + \xi(\|\mathbf{Q}^\perp\boldsymbol{\epsilon}\|_2/\|\boldsymbol{\epsilon}\|_2)(\mathbf{A}_{J*}\mathbf{Q}^\perp)\mathbf{Q}^\perp\boldsymbol{\epsilon}/\|\mathbf{Q}^\perp\boldsymbol{\epsilon}\|_2 \\
&= \widetilde{\boldsymbol{\mu}} + \widetilde{\xi}(\mathbf{A}_{J*}\mathbf{Q}^\perp)\mathbf{Q}^\perp\boldsymbol{\epsilon}/\|\mathbf{Q}^\perp\boldsymbol{\epsilon}\|_2.
\end{aligned}$$

Accordingly, by a similar argument as in the proof of the first part, we find that $\boldsymbol{Z}_J \,|\, \boldsymbol{Z}_{J^c}$ is transelliptical distributed. In particular, it follows from the independence of $\mathbf{Q}^\perp\boldsymbol{\epsilon}/\|\mathbf{Q}^\perp\boldsymbol{\epsilon}\|_2$ and $\{\boldsymbol{Z}_{J^c}, \xi, \|\boldsymbol{\epsilon}\|_2, \|\mathbf{Q}^\perp\boldsymbol{\epsilon}\}$ that $\mathbf{Q}^\perp\boldsymbol{\epsilon}/\|\mathbf{Q}^\perp\boldsymbol{\epsilon}\|_2$ is independent of $\boldsymbol{Z}_{J^c}$ and $\widetilde{\xi}$. Since $\boldsymbol{\Sigma}$ is of full rank and $(\mathbf{A}_{J*}\mathbf{Q}^\perp)(\mathbf{A}_{J*}\mathbf{Q}^\perp)^T = [\boldsymbol{\Theta}_{J,J}]^{-1}$ (thinking about the Gaussian case), we have

$$\boldsymbol{Z}_J|\boldsymbol{Z}_{J^c} \sim EC_{|J|}(\widetilde{\boldsymbol{\mu}}, [\boldsymbol{\Theta}_{J,J}]^{-1}, \widetilde{\xi}).$$

Since $Z_j = f_j(X_j)$ with strictly increasing f_j, conditioning on $\boldsymbol{X}_{J^c}$ is the same as conditioning on $\boldsymbol{Z}_{J^c}$. Thus, $\boldsymbol{X}_J \,|\, \boldsymbol{X}_{J^c}$ follows a transelliptical distribution. □

A.3 Proof of Lemma 3.8

Proof. Let $\boldsymbol{Z} := (Z_1, \ldots, Z_d)^T$. Since $\boldsymbol{X} \sim TE_d(\boldsymbol{\Sigma}, \xi; f_1, \ldots, f_d)$, we have $\boldsymbol{Z} \sim EC_d(\mathbf{0}, \boldsymbol{\Sigma}, \xi)$. Using the proof of Lemma 3.7, we have

$$\boldsymbol{Z}_J|\boldsymbol{Z}_{J^c} \sim EC_{|J|}(\widetilde{\boldsymbol{\mu}}, [\boldsymbol{\Theta}_{J,J}]^{-1}, \widetilde{\xi}),$$

where $\widetilde{\boldsymbol{\mu}} = \mathbf{A}_{J*}\mathbf{A}_{J^c*}^\dagger\boldsymbol{Z}_{J^c}$ and $\widetilde{\xi} = \xi||\mathbf{Q}^\perp\boldsymbol{\epsilon}||_2/||\boldsymbol{\epsilon}||_2$. Therefore, when $\mathbb{E}(\widetilde{\xi}^2|\boldsymbol{Z}_{J^c}) < \infty$, $\boldsymbol{\Theta}_{J,J}$ is diagonal if and only if $\mathrm{Cov}(\boldsymbol{Z}_J|\boldsymbol{Z}_{J^c}) = \left(\mathbb{E}(\widetilde{\xi}^2|\boldsymbol{Z}_{J^c})\right)[\boldsymbol{\Theta}_{J,J}]^{-1}$ is diagonal, or equivalently $\boldsymbol{Z}_J$ are pairwise uncorrelated given $\boldsymbol{Z}_{J^c}$.

It remains to prove that conditioning on $\boldsymbol{Z}_{J^c} = \boldsymbol{v}$ for any vector $\boldsymbol{v}$, $\mathbb{E}(\widetilde{\xi}^2|\boldsymbol{Z}_{J^c} = \boldsymbol{v}) < \infty$ or equivalently $\mathbb{E}(\xi^2|\boldsymbol{Z}_{J^c} = \boldsymbol{v}) < \infty$. We note that if this holds, then for any $j, k \in J$, the

correlation between Z_j and Z_k given $\boldsymbol{Z}_{J^c}$ is well-defined even if $\mathbb{E}\xi^2$ is unbounded. Using the proof of Lemma 3.7 again, we have $\mathbf{A}^{\dagger}_{J^c*} Z_{J^c} = (\xi||\mathbf{Q}\boldsymbol{\epsilon}||_2/||\boldsymbol{\epsilon}||_2)\,\mathbf{Q}\boldsymbol{\epsilon}/||\mathbf{Q}\boldsymbol{\epsilon}||_2^2$. Therefore,

$$Y_0 := \xi^2\|\mathbf{Q}\boldsymbol{\epsilon}\|_2^2/\|\boldsymbol{\epsilon}\|_2^2 = ||\mathbf{A}^{\dagger}_{J^c*}\boldsymbol{Z}_{J^c}||_2^2$$

is a constant conditionally on $\boldsymbol{Z}_{J^c}$. Let F_{ξ^2} be the marginal distribution function of ξ^2, $\alpha_1 = (d - |J|)/2$, and $\alpha_2 = |J|/2$. Since $\|\mathbf{Q}\boldsymbol{\epsilon}\|_2^2/\|\boldsymbol{\epsilon}\|_2^2 \sim \text{beta}(\alpha_1, \alpha_2)$ distribution, the marginal distribution of Y_0 can be derived as:

$$\begin{aligned}
\mathbb{P}(Y_0 \le c_0) &= \mathbb{P}(\xi^2||\mathbf{Q}\boldsymbol{\epsilon}||_2^2/||\boldsymbol{\epsilon}||_2^2 \le c_0) \\
&= \int_0^\infty \mathbb{P}(\xi^2||\mathbf{Q}\boldsymbol{\epsilon}||_2^2/||\boldsymbol{\epsilon}||_2^2 \le c_0 \,|\, \xi^2 = t) F_{\xi^2}(dt) \\
&= \int_0^\infty \int_0^{c_0/t} \text{beta}(z; \alpha_1, \alpha_2) dz F_{\xi^2}(dt) \quad (\text{letting } y = zt) \\
&= \int_0^{c_0} \left(\int_y^\infty t^{-1} \text{beta}(y/t; \alpha_1, \alpha_2) F_{\xi^2}(dt) \right) dy.
\end{aligned}$$

Accordingly, the marginal density of Y_0 is $\displaystyle\int_y^\infty t^{-1}\text{beta}(y/t; \alpha_1, \alpha_2) F_{\xi^2}(dt)$. Thus, the conditional expectation of ξ^2 given $\xi^2\|\mathbf{Q}\boldsymbol{\epsilon}\|_2^2/\|\boldsymbol{\epsilon}\|_2^2 = c_0 > 0$ is

$$\mathbb{E}(\xi^2 \,|\, Y = c_0) = \frac{\displaystyle\int_{c_0}^\infty \text{beta}(c_0/t; \alpha_1, \alpha_2)\ F_{\xi^2}(dt)}{\displaystyle\int_{c_0}^\infty t^{-1}\text{beta}(c_0/t; \alpha_1, \alpha_2)\ F_{\xi^2}(dt)} < \infty. \tag{A.3}$$

The proof is complete since $\widetilde{\xi}^2 = \xi^2||\mathbf{Q}^{\perp}\boldsymbol{\epsilon}||_2^2/||\boldsymbol{\epsilon}||_2^2 \le \xi^2$. □

A.4 Proof of Theorem 3.9

Proof. Let $J = \{1, \ldots, d\} \setminus C$ and $\widetilde{A}, \widetilde{B}$ be two disjoint subsets of J and form a partition of J. Let $G_J = (J, V_J)$ be the subgraph of G composed of vertices J and all edges in G connecting vertices in J.

Using Lemma 3.7, we have $\boldsymbol{Z}_J \,|\, \boldsymbol{Z}_C \sim EC_{|J|}(\widetilde{\boldsymbol{\mu}}, [\boldsymbol{\Theta}_{J,J}]^{-1}, \widetilde{\xi})$ and G_J is the graph of $\boldsymbol{Z}_J \,|\, \boldsymbol{Z}_C$. Accordingly, by definition, the following statements are equivalent: (i) C separates $\widetilde{A}$ and $\widetilde{B}$ in G; (ii) $\widetilde{A}$ and $\widetilde{B}$ are not connected in G_J; (iii) $\boldsymbol{\Theta}_{J,J}$ is composed of diagonal blocks $\boldsymbol{\Theta}_{\widetilde{A},\widetilde{A}}$ and $\boldsymbol{\Theta}_{\widetilde{B},\widetilde{B}}$; (iv) $[\boldsymbol{\Theta}_{J,J}]^{-1}$ is composed of diagonal blocks $[\boldsymbol{\Theta}_{\widetilde{A},\widetilde{A}}]^{-1}$ and $[\boldsymbol{\Theta}_{\widetilde{B},\widetilde{B}}]^{-1}$; (v) For any $j \in \widetilde{A}$ and $k \in \widetilde{B}$, $[[\boldsymbol{\Theta}_{J,J}]^{-1}]_{jk} = \text{Cov}(Z_j, Z_k \,|\, \boldsymbol{Z}_C) = 0$.

Suppose C separates A and B. Then let $\widetilde{A}$ be all vertices in J connected to A, and let $\widetilde{B} = J \setminus \widetilde{A}$. Since A and B are not connected, we have $A \subset \widetilde{A}$ and $B \subset \widetilde{B}$. Moreover, C must separate $\widetilde{A}$ and $\widetilde{B}$ in G by the definition of $\widetilde{A}$ and the fact that $\widetilde{A}, \widetilde{B}$ form a partition of J. Therefore, by the equivalence relationship stated in the previous paragraph, $\boldsymbol{Z}_{\widetilde{A}}$ and

$\boldsymbol{Z}_{\widetilde{B}}$ are conditionally uncorrelated given $\boldsymbol{Z}_C$. Consequently, $\boldsymbol{Z}_A$ and $\boldsymbol{Z}_B$ are conditionally uncorrelated given $\boldsymbol{Z}_C$.

Conversely, let $\{A, B, C\}$ form a partition of $\{1, \ldots, d\}$. Set $\widetilde{A} = A$ and $\widetilde{B} = B$. Then, by the same argument, $\boldsymbol{Z}_{\widetilde{A}}$ and $\boldsymbol{Z}_{\widetilde{B}}$ are conditionally uncorrelated given $\boldsymbol{Z}_C$ if and only if C separates $\widetilde{A}$ and $\widetilde{B}$. □

A.5 Proof of Theorem 3.10

Proof. Let $\boldsymbol{Z} := (Z_1, \ldots, Z_d)^T = (f_1(X_1), \ldots, f_d(X_d))^T$. If $\boldsymbol{X}$ is nonparanormal, then $\boldsymbol{Z}$ is Gaussian distributed and the zero entries in $\boldsymbol{\Theta}$ encode the conditional independence of $\boldsymbol{Z}$. Because $\{f_j\}_{j=1}^d$ are marginal monotone transformations, G also encodes the conditional independence of $\boldsymbol{X}$.

On the other hand, if G encodes the conditional independence of $\boldsymbol{X}$, then it also encodes the conditional independence of $\boldsymbol{Z}$. Suppose that $\boldsymbol{Z}_j$ and $\boldsymbol{Z}_k$ are conditionally independent given $\boldsymbol{Z}_{\setminus\{j,k\}}$. Then using the proof of Lemma 3.7, letting $J = \{j, k\}$, $\boldsymbol{Z}_J \,|\, \boldsymbol{Z}_{J^c} \sim EC_{|J|}(\widetilde{\mu}, [\boldsymbol{\Theta}_{J,J}]^{-1}, \widetilde{\xi})$. However, using a similar argument as in the proof of case 2 in Theorem 3.11, we have $\boldsymbol{Z}_J | \boldsymbol{Z}_{J^c} \in \mathbb{R}^2$ must be Gaussian distributed. It then can be proved that $\boldsymbol{Z}$ is Gaussian distributed using a similar argument as in the proof of case 3 in Theorem 3.11. □

A.6 Proof of Theorem 3.11

Proof. The first part of this lemma is obvious. For the second part, since $\boldsymbol{X}$ is simultaneously elliptical and nonparanormal, there must exist $\xi \geq 0$, $\mathbf{A} \in \mathbb{R}^{d \times d}$, and $\boldsymbol{\mu} \in \mathbb{R}^d$, such that $\boldsymbol{X} := (X_1, \ldots, X_d)^T \stackrel{d}{=} \boldsymbol{\mu} + \xi \mathbf{A}\boldsymbol{U} \sim NPN_d(\boldsymbol{\Sigma}; f_1, \ldots, f_d)$ where $\boldsymbol{U}$ is a uniform random vector on the d-dimensional unit sphere and is independent of ξ. Since $\boldsymbol{X}$ has a nonparanormal distribution, $\mathrm{diag}(\boldsymbol{\Sigma}) = \mathbf{I}_d$. Since f_j's are strictly increasing, by Theorem 3.13 we know that the Kendall's tau correlation matrices of $f(\boldsymbol{X}) := (f_1(X_1), \ldots, f_d(X_d))^T$ and $\boldsymbol{X} := (X_1, \ldots, X_d)^T$ are exactly the same. Therefore, without loss of generality, we assume $\boldsymbol{\mu} = \mathbf{0}$ and $\mathbf{A}\mathbf{A}^T = \mathbf{A}^T\mathbf{A} = \boldsymbol{\Sigma}$. Since $\mathrm{diag}(\boldsymbol{\Sigma}) = \mathbf{I}_d$, it follows from (A.2) that $\boldsymbol{X}$ has identical marginal distributions. Thus, $f_1, \ldots, f_d$ are identical to some strictly increasing function f_0. The desired result follows by considering the following three cases.

Case 1: $d = 2$ and $\boldsymbol{\Sigma}_{12} \neq 0$. Let $\rho = \boldsymbol{\Sigma}_{12}$. Since $\mathrm{rank}(\boldsymbol{\Sigma}) > 1$, $|\rho| \in (0, 1)$. Let $\mathbf{A}_{j*}$ be the j^{th} row of $\mathbf{A}$, $Z_j := \mathbf{A}_{j*}\boldsymbol{U}$ and $\delta = I\{Z_2 > \rho Z_1\}$. Since $\boldsymbol{\Sigma}_{11} = \boldsymbol{\Sigma}_{22} = 1$, $\max_{\|\boldsymbol{u}\|_2=1}(\pm\mathbf{A}_{1*}\boldsymbol{u}) = 1$ is attained when $\boldsymbol{u} = \pm\mathbf{A}_{1*}$ and $\mathbf{A}_{2*}\boldsymbol{u} = \pm\rho$. Similarly $\max_{\|\boldsymbol{u}\|_2=1}(\pm\mathbf{A}_{2*}\boldsymbol{u}) = 1$ is attained when $\boldsymbol{u} = \pm\mathbf{A}_{2*}$ and $\mathbf{A}_{1*}\boldsymbol{u} = \pm\rho$. Thus, $\mathbf{A} : \boldsymbol{u} \to \boldsymbol{z} = \mathbf{A}\boldsymbol{u}$ maps the unit circle $\{\boldsymbol{u} : \|\boldsymbol{u}\|_2 = 1\}$ to the ellipse inscribing the square $\{\boldsymbol{z} : \|\boldsymbol{z}\|_\infty = 1\}$ at four points $(1, \rho)$, $(-1, -\rho)$, $(\rho, 1)$, $(-\rho, -1)$ in the $\boldsymbol{z}$-space. Let $t \in (-1, 1)$, we consider three lines $\ell_1 = \{\boldsymbol{z} : z_1 = 1\}$, $\ell_2 = \{\boldsymbol{z} : z_1 = t\}$ and $\ell_3 = \{\boldsymbol{z} : z_2 = \rho z_1\}$. Let $\mathbf{A}^{-1}\ell_k$ be corresponding lines in the $\boldsymbol{u}$-space as the set $\left\{\mathbf{A}^{-1}\boldsymbol{z} \,|\, \boldsymbol{z} \in \ell_k\right\}$, $k = 1, 2, 3$. Let us say that

$\boldsymbol{u}$ lies "above" $\mathbf{A}^{-1}\ell_3$ when $\mathbf{A}\boldsymbol{u}$ lies above ℓ_3. Since ℓ_1 and ℓ_2 are parallel and the ellipse is tangent to ℓ_1 at their intersection with ℓ_3, $\mathbf{A}^{-1}\ell_1$ and $\mathbf{A}^{-1}\ell_2$ are parallel and perpendicular to $\mathbf{A}^{-1}\ell_3$. Since $\mathbf{A}^{-1}\ell_3$ passes $(0,0)$ and $\boldsymbol{U}$ is uniformly distributed in the unit circle, we have

$$\begin{aligned} & \mathbb{P}\big(\delta = 1 \big| t \le Z_1 \le 1\big) \\ = \; & \mathbb{P}(\boldsymbol{U} \text{ lies "above" } \mathbf{A}^{-1}\ell_3 \big| \boldsymbol{U} \text{ lies between } \mathbf{A}^{-1}\ell_1 \text{ and } \mathbf{A}^{-1}\ell_2\big) \\ = \; & 1/2. \end{aligned}$$

Since t is arbitrary, δ is independent of Z_1. Consequently,

$$\mathbb{P}\big(X_2 > \rho X_1 | X_1 = x_1\big) = \mathbb{E}\Big[\mathbb{P}\big(\delta = 1 | Z_1, \xi\big)\Big| Z_1\xi = x_1\Big] = 1/2,$$

where we have used the identity $\boldsymbol{X} = \xi\boldsymbol{Z}$ and the independence of $\boldsymbol{Z}$ and ξ. This implies that the conditional median of X_2 is ρX_1 given $X_1 = x_1$. Since $f(\boldsymbol{X}) \sim N(0, \boldsymbol{\Sigma})$, the conditional median of $f_2(X_2)$ given $f_1(X_1) = f_1(x_1)$ is $\rho f(x_1)$. Since f_1 and f_2 are both strictly increasing and $f_1 = f_2 = f_0$, comparison of the conditional medians yield $\rho f_0(x_1) = f_0(\rho x_1)$ for all real x_1, so that $f_0(x) = a_0 x$ for a constant $a_0 \neq 0$. This implies that $\boldsymbol{X} = f(\boldsymbol{X})/a_0 \sim N(0, \boldsymbol{\Sigma}/a_0^2)$.

Case 2: $d = 2$ and $\boldsymbol{\Sigma}_{12} = 0$. In this case X_1 and X_2 are i.i.d.. Recall that $\boldsymbol{X}$ is elliptically distributed. Suppose $\boldsymbol{\mu} = 0$, so that $\boldsymbol{X}$ has the characteristic function $\psi(\boldsymbol{t}^T\boldsymbol{\Sigma}\boldsymbol{t})$ for some properly defined function ψ. Let $i = \sqrt{-1}$. Since $\boldsymbol{X} \sim NPN_2(\boldsymbol{\Sigma}; f_1, f_2)$ and $\boldsymbol{\Sigma}_{12} = 0$, X_1 and X_2 are independent. Thus,

$$\psi(\boldsymbol{t}^T\boldsymbol{\Sigma}\boldsymbol{t}) = \mathbb{E}\exp(i\boldsymbol{t}^T\boldsymbol{X}) = \mathbb{E}\exp(it_1X_1)\mathbb{E}\exp(it_2X_2) = \psi(t_1^2\boldsymbol{\Sigma}_{11})\psi(t_2^2\boldsymbol{\Sigma}_{22}).$$

Accordingly, since $\boldsymbol{\Sigma}_{11} = \boldsymbol{\Sigma}_{22} = 1$, we have

$$\psi(t_1^2 + t_2^2) = \psi(t_1^2)\psi(t_2^2).$$

This equation is known as Hamel's equation and has the solution $\psi(t^2) = \exp(kt^2)$ for some constant k (Kuczma, 2008). Since $\psi(t^2)$ is a characteristic function, it is bounded in t and $\psi(t^2) \to 0$ as $t^2 \to \infty$. Consequently, $k < 0$ and $\boldsymbol{X}$ is Gaussian.

Case 3: $d > 2$. Since $\text{rank}(\boldsymbol{\Sigma}) > 1$, one of the off-diagonal elements of $\boldsymbol{\Sigma}$ is in $(-1, 1)$. Assume $\boldsymbol{\Sigma}_{12} \in (-1, 1)$ without loss of generality. Let $J = \{1, 2\}$. It follows from (A.2) in the proof of Lemma 3.7 that

$$\boldsymbol{X}_J \sim EC_2\left(0, \boldsymbol{\Sigma}_{J,J}, \frac{\xi\|\mathbf{P}\boldsymbol{\epsilon}\|_2}{\|\boldsymbol{\epsilon}\|_2}\right)$$

with a standard normal $\boldsymbol{\epsilon}$ independent of ξ and a projection matrix $\mathbf{P}$ of rank two. Moreover, since the conclusion holds for $d = 2$, the membership of the distribution of $\boldsymbol{X}_J$ in both the elliptical and nonparanormal families implies that $\boldsymbol{X}_J$ follows a Gaussian distribution.

Thus, $(\xi||\mathbf{P}\boldsymbol{\epsilon}||_2/||\boldsymbol{\epsilon}||_2)^2 = \|\boldsymbol{\Sigma}_{J,J}^{-1/2}\boldsymbol{X}_J\|_2^2$ has the chi-square distribution with two degrees of freedom, χ_2^2. Since $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \mathbf{I}_d)$, $\|\mathbf{P}\boldsymbol{\epsilon}\|_2^2/\|\boldsymbol{\epsilon}\|_2^2$ has the beta$(1,(d-2)/2)$ distribution and is independent of $\|\boldsymbol{\epsilon}\|_2^2$. By the definition of the elliptical distribution, $\|\mathbf{P}\boldsymbol{\epsilon}\|_2^2/\|\boldsymbol{\epsilon}\|_2^2$ is also independent of ξ^2. It follows that

$$\mathbb{E}\xi^{2k}\mathbb{E}\Big(\frac{\|\mathbf{P}\boldsymbol{\epsilon}\|_2^2}{\|\boldsymbol{\epsilon}\|_2^2}\Big)^k = \mathbb{E}\Big(\xi^2\frac{\|\mathbf{P}\boldsymbol{\epsilon}\|_2^2}{\|\boldsymbol{\epsilon}\|_2^2}\Big)^k = \mathbb{E}\chi_2^{2k} = \mathbb{E}\|\mathbf{P}\boldsymbol{\epsilon}\|_2^{2k} = \mathbb{E}\|\boldsymbol{\epsilon}\|_2^{2k}\mathbb{E}\Big(\frac{\|\mathbf{P}\boldsymbol{\epsilon}\|_2^2}{\|\boldsymbol{\epsilon}\|_2^2}\Big)^k.$$

Canceling $\mathbb{E}(\|\mathbf{P}\boldsymbol{\epsilon}\|_2^2/\|\boldsymbol{\epsilon}\|_2^2)^k$ from both sides above, we find $\mathbb{E}(\xi^2)^k = \mathbb{E}(\|\boldsymbol{\epsilon}\|_2^2)^k$ for all positive integers k. Since $\|\boldsymbol{\epsilon}\|_2^2 \sim \chi_d^2$ and the chi-square distributions are uniquely identified by moments, ξ^2 must have the χ_d^2 distribution. This gives the normality of $\boldsymbol{X}$ and completes the proof. □

A.7 Proof of Theorem 3.12

Proof. Let $\boldsymbol{Z} = \boldsymbol{\mu} + \xi\mathbf{A}\boldsymbol{U}$ and $q := \text{rank}(\boldsymbol{\Sigma}) = \text{rank}(\mathbf{A})$ as in Definition 2.1. Let $\boldsymbol{U} = \boldsymbol{\epsilon}/\|\boldsymbol{\epsilon}\|_2$ with a standard normal vector $\boldsymbol{\epsilon}$ in $\mathbb{R}^q$. Note that if $\mathbf{A} = \mathbf{V}_1\mathbf{D}\mathbf{V}_2^T$ is the singular value decomposition of $\mathbf{A}$, then $\mathbf{A}^\dagger = \mathbf{V}_2\mathbf{D}^{-1}\mathbf{V}_1^T$. Since $\text{rank}(\mathbf{A}) = q$, $\mathbf{A}^\dagger\mathbf{A} = \mathbf{I}_q$. Let $\boldsymbol{Y} = \mathbf{A}\boldsymbol{\epsilon} \sim N(0, \boldsymbol{\Sigma})$. It follows that

$$\boldsymbol{Z} - \boldsymbol{\mu} = \xi\mathbf{A}\boldsymbol{U} = \xi\boldsymbol{Y}/\|\boldsymbol{\epsilon}\|_2 = \xi\boldsymbol{Y}/\|\mathbf{A}^\dagger\boldsymbol{Y}\|_2.$$

The proof is complete. □

A.8 Proof of Theorem 3.13

Proof. We define $Z_j := f_j(X_j)$ and $Z_k := f_k(X_k)$. Using Lemma 3.7, we have $(Z_j, Z_k)^T \sim EC_2(\mathbf{0}, \boldsymbol{\Sigma}^{'}, \xi)$, where $\boldsymbol{\Sigma}_{11}^{'} = \boldsymbol{\Sigma}_{22}^{'} = 1$ and $\boldsymbol{\Sigma}_{12}^{'} = \boldsymbol{\Sigma}_{21}^{'} = \boldsymbol{\Sigma}_{jk}$. Since the Kendall's tau statistic is monotone transformation-invariant, we have $\tau_{jk} := \tau(X_j, X_k) = \tau(Z_j, Z_k)$. Let $(\widetilde{Z}_j, \widetilde{Z}_k)^T$ be an independent copy of $(Z_j, Z_k)^T$. By the equivalent definition, there exists a characteristic function ψ uniquely determined by ξ, such that $\boldsymbol{Z} \sim EC_d(\mathbf{0}, \boldsymbol{\Sigma}, \psi)$ and $\widetilde{\boldsymbol{Z}} \sim EC_d(\mathbf{0}, \boldsymbol{\Sigma}, \psi)$. Since $\boldsymbol{Z}$ and $\widetilde{\boldsymbol{Z}}$ are independent, we have $\mathbb{E}\exp(i\boldsymbol{t}^T(\boldsymbol{Z} - \widetilde{\boldsymbol{Z}})) = \mathbb{E}\exp(i\boldsymbol{t}^T Z)\mathbb{E}\exp(i\boldsymbol{t}^T\widetilde{Z}) = \psi^2(\boldsymbol{t}^T\boldsymbol{\Sigma}t)$, which implies $\boldsymbol{Z} - \widetilde{\boldsymbol{Z}} \sim EC_d(\mathbf{0}, \boldsymbol{\Sigma}, \psi^2)$. Again, by definition, there exists a nonnegative random variable $\xi^{'}$ uniquely determined by ψ^2, such that $\boldsymbol{Z} - \widetilde{\boldsymbol{Z}} \sim EC_d(\mathbf{0}, \boldsymbol{\Sigma}, \xi')$. Since $\boldsymbol{Z} - \widetilde{\boldsymbol{Z}}$ is continuous, we have $\mathbb{P}(\xi^{'} = 0) = 0$. Using the stochastic representation of an elliptical random variable, we further have $\boldsymbol{Z} - \widetilde{\boldsymbol{Z}} \overset{d}{=} \xi^{'}\mathbf{A}\boldsymbol{U}$, where $\mathbf{A}$ and $\boldsymbol{U}$ are only determined by $\boldsymbol{\Sigma}$. Then

$$\mathbb{P}((Z_j - \widetilde{Z}_j)(Z_k - \widetilde{Z}_k) > 0) \quad = \quad \mathbb{P}((\xi^{'})^2(\mathbf{A}\boldsymbol{U})_j(\mathbf{A}\boldsymbol{U})_k > 0) = \mathbb{P}((\mathbf{A}\boldsymbol{U})_j(\mathbf{A}\boldsymbol{U})_k > 0)$$

is invariant to $\xi^{'}$. From (3.3), we know that $\tau(Z_j, Z_k)$ is invariant to ξ. To prove the final result, we first verify that it holds for the trivial cases where $\boldsymbol{\Sigma}_{jk} = 1$ or $\boldsymbol{\Sigma}_{jk} = -1$. For other cases, we define $(Y_j, Y_k)^T \sim N_2(\mathbf{0}, \boldsymbol{\Sigma}^{'})$, where $\boldsymbol{\Sigma}^{'} \succ \mathbf{0}$. We have

$$\tau_{jk} \quad := \quad \tau(X_j, X_k) = \tau(Z_j, Z_k) = \tau(Y_j, Y_k)$$

$$= \quad \frac{2}{\pi}\arcsin(\boldsymbol{\Sigma}_{12}^{'}) = \frac{2}{\pi}\arcsin(\boldsymbol{\Sigma}_{jk}).$$

The second to last equality is due to the relationship between the Kendall's tau and Pearson's correlation coefficient in Gaussian distributions (Kruskal, 1958). The last equality holds since $\boldsymbol{\Sigma}_{12}^{'} = \boldsymbol{\Sigma}_{jk}$. □

A.9 Proof of Lemma 3.14

Proof. Using Lemma 3.7, we have $\boldsymbol{X}_J \,|\, \boldsymbol{X}_{J^c} \sim EC_d(\widetilde{\boldsymbol{\mu}}, [\boldsymbol{\Theta}_{J,J}]^{-1}, \widetilde{\xi})$. Accordingly, using Theorem 3.13, $\tau(\boldsymbol{X}_J \,|\, \boldsymbol{X}_{J^c}) = \frac{2}{\pi}\arcsin([\boldsymbol{\Theta}_{J,J}]^{-1})$. Because $\arcsin(x) = 0$ if and only if $x = 0$, we have $\boldsymbol{\Theta}_{J,J}$ is diagonal if and only if $\tau(\boldsymbol{X}_J|\boldsymbol{X}_{J^c}) = \mathrm{diag}(\tau(\boldsymbol{X}_J|\boldsymbol{X}_{J^c}))$. □

A.10 Proof of Theorem 3.15

Proof. Let $D := A \cup B$ and $\boldsymbol{\Sigma}^* := \boldsymbol{\Sigma}_{D,D} - \boldsymbol{\Sigma}_{D,C}\boldsymbol{\Sigma}_{C,C}^{-1}\boldsymbol{\Sigma}_{C,D}$. Using Lemma 3.7, we have $(\boldsymbol{X}_A, \boldsymbol{X}_B \,|\, \boldsymbol{X}_C)$ is transelliptically distributed with generalized latent correlation matrix $\boldsymbol{\Sigma}^*$. Using Theorem 3.13, $\tau(\boldsymbol{X}_A, \boldsymbol{X}_B \,|\, \boldsymbol{X}_C) = \frac{2}{\pi}\arcsin(\boldsymbol{\Sigma}^*_{A,B})$, here the $\arcsin(\cdot)$ transformation is applied on each element of the matrix $\boldsymbol{\Sigma}^*_{A,B}$. On the other hand, by the proof of Theorem 3.9, if C separates A and B in G, $\boldsymbol{\Sigma}^*_{A,B} = \mathbf{0}$, or equivalently $\tau(\boldsymbol{X}_A, \boldsymbol{X}_B \,|\, \boldsymbol{X}_C) = \mathbf{0}$. Moreover, if $A \cup B \cup C = \{1, \ldots, d\}$, then using Theorem 3.9 again, C separates A and B if and only if $\boldsymbol{\Sigma}^*_{A,B} = 0$, or equivalently $\tau(\boldsymbol{X}_A, \boldsymbol{X}_B \,|\, \boldsymbol{X}_C) = \mathbf{0}$. □

A.11 Proof of Theorem 5.1

Proof. The only difference between the rank-based CLIME and the original CLIME is that we replace the Pearson correlation coefficient matrix $\widehat{\mathbf{R}}$ by the Kendall's tau matrix $\widehat{\mathbf{S}}$. By examining the proofs of Theorem 1 and Theorem 7 in Cai et al. (2011), the only property needed of the Pearson correlation matrix $\widehat{\mathbf{R}}$ is an exponential concentration inequality

$$\mathbb{P}\left(|\widehat{\mathbf{R}}_{jk} - \boldsymbol{\Sigma}_{jk}| > t\right) \leq c_1 \exp(-c_2 n t^2).$$

Therefore, it suffices if we can prove a similar concentration inequality for $|\widehat{\mathbf{S}}_{jk} - \boldsymbol{\Sigma}_{jk}|$. Since $\widehat{\mathbf{S}} = \sin\left(\frac{\pi}{2}\widehat{\tau}_{jk}\right)$ and $\boldsymbol{\Sigma}_{jk} = \sin\left(\frac{\pi}{2}\tau_{jk}\right)$, we have $|\widehat{\mathbf{S}}_{jk} - \boldsymbol{\Sigma}_{jk}| \leq |\widehat{\tau}_{jk} - \tau|$. Therefore, we only need to prove

$$\mathbb{P}\left(|\widehat{\tau}_{jk} - \tau_{jk}| > t\right) \leq \exp\left(-nt^2/(2\pi)\right). \tag{A.4}$$

This result has been proved in Theorem 4.2 of Liu et al. (2012) using the Hoeffding's inequality for U-statistic Hoeffding (1963). □

A.12 Proof of Theorem 5.2

Proof. Let $\widehat{\mathbf{S}}$ be defined in (4.1) and $\widehat{\mathbf{\Theta}}$ be the rank-based scaled CLIME estimator defined in (4.5) and (4.6). First, from Theorem 4.2 of Liu et al. (2012), we have that, with probability at least $1 - 1/d$,

$$\big\|\widehat{\mathbf{S}} - \mathbf{\Sigma}\big\|_{\max} \leq \lambda_0.$$

Let $\boldsymbol{\theta}_j^+ \geq \mathbf{0}$ and $\boldsymbol{\theta}_j^- \geq \mathbf{0}$ be the positive and negative parts of $\mathbf{\Theta}_{*j}$ such that $\mathbf{\Theta}_{*j} = \boldsymbol{\theta}_j^+ - \boldsymbol{\theta}_j^-$ and $\|\mathbf{\Theta}_{*j}\|_1 = \mathbf{1}_d^T(\boldsymbol{\theta}_j^+ + \boldsymbol{\theta}_j^-)$. Suppose, for any $j \in \{1, \ldots, d\}$, $\boldsymbol{\theta}_j^+$ and $\boldsymbol{\theta}_j^-$ are feasible in the sense of satisfying the constraint (4.6). Then,

$$\mathbf{1}_d^T(\widehat{\boldsymbol{\beta}}_j^+ + \widehat{\boldsymbol{\beta}}_j^-) \leq \mathbf{1}_d^T(\boldsymbol{\theta}_j^+ + \boldsymbol{\theta}_j^-) \ \text{ for all } j = 1, \ldots, d.$$

We then have

$$\begin{aligned}
\|\mathbf{\Theta} - \widehat{\mathbf{\Theta}}\|_{\max} &= \|\mathbf{\Theta}(\widehat{\mathbf{S}}\widehat{\mathbf{\Theta}} - \mathbf{I}_d) + (\mathbf{I}_d - \mathbf{\Theta}\widehat{\mathbf{S}})\widehat{\mathbf{\Theta}}\|_{\max} \\
&\leq \|\mathbf{\Theta}\|_1 \|\widehat{\mathbf{S}}\widehat{\mathbf{\Theta}} - \mathbf{I}_d\|_{\max} + \|\mathbf{I}_d - \mathbf{\Theta}\widehat{\mathbf{S}}\|_{\max} \max_j \|\widehat{\mathbf{\Theta}}_{*j}\|_1 \\
&\leq \lambda_0 \cdot \mathbf{1}_d^T(\widehat{\boldsymbol{\beta}}_j^+ + \widehat{\boldsymbol{\beta}}_j^-) \cdot \|\mathbf{\Theta}\|_1 + \lambda_0 \|\mathbf{\Theta}\|_1 \cdot \max_j \|\widehat{\mathbf{\Theta}}_{*j}\|_1 \\
&\leq 2\lambda_0 \cdot \mathbf{1}_d^T(\widehat{\boldsymbol{\beta}}_j^+ + \widehat{\boldsymbol{\beta}}_j^-) \cdot \|\mathbf{\Theta}\|_1 \\
&\leq 2\lambda_0 \|\mathbf{\Theta}\|_1^2. \qquad \text{(A.5)}
\end{aligned}$$

Let λ_1 be a threshold level and we define

$$s_1 = \max_{1 \leq j \leq d} \sum_{1 \leq i \leq d} \min\Big\{|\mathbf{\Theta}_{ij}|/\lambda_1, 1\Big\},$$
$$T_j = \Big\{i : |\mathbf{\Theta}_{ij}| \geq \lambda_1\Big\}.$$

Since $\|\widehat{\mathbf{\Theta}}_{*j}\|_1 \leq \mathbf{1}_d^T(\widehat{\boldsymbol{\beta}}_j^+ + \widehat{\boldsymbol{\beta}}_j^-) \leq \mathbf{1}_d^T(\boldsymbol{\theta}_j^+ + \boldsymbol{\theta}_j^-) = \|\mathbf{\Theta}_{*j}\|_1$, we have

$$\begin{aligned}
\|\widehat{\mathbf{\Theta}}_{*j} - \mathbf{\Theta}_{*j}\|_1 &\leq \|\widehat{\mathbf{\Theta}}_{T_j^c,j}\|_1 + \|\mathbf{\Theta}_{T_j^c,j}\|_1 + \|\widehat{\mathbf{\Theta}}_{T_j,j} - \mathbf{\Theta}_{T_j,j}\|_1 \\
&= \|\widehat{\mathbf{\Theta}}_{*j}\|_1 - \|\widehat{\mathbf{\Theta}}_{T_j,j}\|_1 + \|\mathbf{\Theta}_{T_j^c,j}\|_1 + \|\widehat{\mathbf{\Theta}}_{T_j,j} - \mathbf{\Theta}_{T_j,j}\|_1 \\
&\leq \|\mathbf{\Theta}_{*j}\|_1 - \|\widehat{\mathbf{\Theta}}_{T_j,j}\|_1 + \|\mathbf{\Theta}_{T_j^c,j}\|_1 + \|\widehat{\mathbf{\Theta}}_{T_j,j} - \mathbf{\Theta}_{T_j,j}\|_1 \\
&\leq 2\|\mathbf{\Theta}_{T_j^c,j}\|_1 + 2\|\widehat{\mathbf{\Theta}}_{T_j,j} - \mathbf{\Theta}_{T_j,j}\|_1 \\
&\leq 2\|\mathbf{\Theta}_{T_j^c,j}\|_1 + 4\lambda_0\|\mathbf{\Theta}\|_1^2 |T_j| \\
&\leq \big(2\lambda_1 + 4\lambda_0\|\mathbf{\Theta}\|_1^2\big) s_1,
\end{aligned}$$

where the second to last inequality follows from (A.5) and the last inequality follows from the definition of λ_1. Suppose $\max_j \sum_i |\mathbf{\Theta}_{ij}|^q \leq s$ and $\lambda_1 = 2\lambda_0\|\mathbf{\Theta}\|_1^2$. Then,

$$\lambda_1 s_1 = \max_{1 \leq j \leq d} \sum_{1 \leq i \leq d} \min\big\{|\mathbf{\Theta}_{ij}|, \lambda_1\big\} \leq \lambda_1^{1-q} s.$$

It follows that

$$\|\widehat{\mathbf{\Theta}}_{*j} - \mathbf{\Theta}_{*j}\|_1 \leq 4\lambda_1 s_1 \leq 4(2\lambda_0 \|\mathbf{\Theta}\|_1^2)^{1-q} s.$$

Recall that we still need to prove the feasibility of $\boldsymbol{\theta}_j^+$ and $\boldsymbol{\theta}_j^-$, i.e., we need to show that

$$\left\|\widehat{\mathbf{S}}(\boldsymbol{\theta}_j^+ - \boldsymbol{\theta}_j^-) - \boldsymbol{e}_j\right\|_\infty \leq \lambda_0 \mathbf{1}_d^T(\boldsymbol{\theta}_j^+ + \boldsymbol{\theta}_j^-) \text{ for all } j = 1, \ldots, d,$$

with high probability. This is true since $\boldsymbol{\theta}_j = \boldsymbol{\theta}_j^+ - \boldsymbol{\theta}_j^-$ and $\mathbf{\Sigma}\boldsymbol{\theta}_j = \mathbf{e_j}$,

$$\|\widehat{\mathbf{S}}\boldsymbol{\theta}_j - \boldsymbol{e}_j\|_\infty = \|(\widehat{\mathbf{S}} - \mathbf{\Sigma})\boldsymbol{\theta}_j\|_\infty \leq \|\widehat{\mathbf{S}} - \mathbf{\Sigma}\|_{\max} \|\boldsymbol{\theta}\|_1.$$

Accordingly, the feasibility condition is satisfied with large probability following the fact that $||\widehat{\mathbf{S}} - \mathbf{\Sigma}|| \leq \lambda_0$ with large probability. Thus, we have $\|\widehat{\mathbf{\Theta}}_{*j} - \mathbf{\Theta}_{*j}\|_1 \leq 2^{3-q} \|\mathbf{\Theta}\|_1^{2-2q} \lambda_0^{1-q} s$.

The desired results then follow from the same analysis as in Theorem 1 and Theorem 7 in Cai et al. (2011). □

References

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

Bickel, P. and Levina, E. (2008a). Covariance regularization by thresholding. *Annals of Statistics*, 36(6):2577–2604.

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

Cai, T., Liu, W., and Luo, X. (2011). A constrained ℓ_1 minimization approach to sparse precision matrix estimation. *Journal of the American Statistical Association*, 106(494):594–607.

Chen, S., Donoho, D., and Saunders, M. (1998). Atomic decomposition by basis pursuit. *SIAM Journal on Scientific Computing*, 20(1):33–61.

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

Drton, M. and Perlman, M. (2007). Multiple testing and error control in Gaussian graphical model selection. *Statistical Science*, 22(3):430–449.

Drton, M. and Perlman, M. (2008). A SINful approach to Gaussian graphical model selection. *Journal of Statistical Planning and Inference*, 138(4):1179–1200.

Fang, H., Fang, K., and Kotz, S. (2002). The meta-elliptical distributions with given marginals. *Journal of Multivariate Analysis*, 82(1):1–16.

Fang, K., Kotz, S., and Ng, K. (1990). Symmetric multivariate and related distributions. *Chapman&Hall, London*.

Finegold, M. A. and Drton, M. (2009). Robust graphical modeling with t-distributions. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, UAI '09, pages 169–176.

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

Halmos, P. (1974). *Measure theory*, volume 18. Springer.

Hoeffding, W. (1963). Probability Inequalities for Sums of Bounded Random Variables. *Journal of the American Statistical Association*, 58(301):13–30.

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

Kruskal, W. (1958). Ordinal measures of association. *Journal of the American Statistical Association*, pages 814–861.

Kuczma, M. (2008). *An introduction to the theory of functional equations and inequalities: Cauchy's equation and Jensen's inequality*. Birkhäuser Basel.

Lafferty, J., Liu, H., and Wasserman, L. (2012). Sparse Nonparametric Graphical Models. *Statistical Science*. To appear.

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

Liu, H., Chen, X., Lafferty, J., and Wasserman, L. (2010a). Graph-valued regression. In *Proceedings of the Twenty-Third Annual Conference on Neural Information Processing Systems (NIPS)*.

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

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

Liu, H., Roeder, K., and Wasserman, L. (2010b). Stability approach to regularization selection (stars) for high dimensional graphical models. In *Proceedings of the Twenty-Third Annual Conference on Neural Information Processing Systems (NIPS)*.

Liu, H., Xu, M., Gu, H., Gupta, A., Lafferty, J., and Wasserman, L. (2011). Forest density estimation. *Journal of Machine Learning Research*, 12:907–951. A short version has appeared in the 23rd Annual Conference on Learning Theory (COLT).

Liu, W. and Luo, X. (2012). High-dimensional sparse precision matrix estimation via sparse column inverse operator. *arXiv/1203.3896*.

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

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

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

Shen, X., Pan, W., and Zhu, Y. (2012).). likelihood-based selection and sharp parameter estimation. *Journal of the American Statistical Association*. to appear.

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

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

Vogel, D. and Fried, R. (2011). Elliptical graphical modelling. *Biometrika*, 98(4):935–951.

Xue, L. and Zou, H. (2012). Regularized rank-based estimation of high-dimensional nonparanormal graphical models.

Yuan, M. (2010). High dimensional inverse covariance matrix estimation via linear programming. *Journal of Machine Learning Research*, 11(8):2261–2286.

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

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