Sure Screening for Transelliptical Graphical Models

Abstract

We propose transelliptical graphical sure screening for recovering the structure of a transelliptical graphical model in the high dimensional setting. The graphical sure screening estimate of the partial correlation graph is obtained by thresholding the elements of an estimator of the sample correlation matrix obtained using Kendall's tau statistic. Under a simple assumption on the relationship between the correlation and partial correlation graphs, we shows that with high probability, the estimated edge set contains the true edge set, and the size of the estimated edge set is controlled. For some special transelliptical distributions, we provide a closed form threshold value that can control the expected false positive rate. For more general transelliptical distributions for which no closed form threshold value is available, we also provide a choice of the threshold value based on a Jackknife estimator for the variance of Kendall's tau. In simulation, we show transelliptical graphical sure screening performs quite competitively with more computationally demanding techniques for graph estimation.

some key words: Partial correlation; High Dimensionality; Undirected Graph; Kendall's tau; Sparsity

1 Introduction

Graphical modeling has been a topic of great interest in both the statistical and scientific literature for the past decade. For example, graphical models have been extensively used to model gene regulatory networks, composed of tens of thousands of genes. Inference on the structure of the graph is made under the high dimensional setting where the number of observations for which gene expression measurements are available is much smaller than the number of features.

Consider the random vector $X = (X_1, \ldots, X_p)^T$, and an undirected graph denoted by $G = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V} = \{1, \ldots, p\}$ is the set of nodes, and $\mathcal{E}$ is the set of edges describing the conditional independence relationships among $X = (X_1, \ldots, X_p)^T$. A pair (j, k) is contained in the edge set $\mathcal{E}$ if and only if X_j is conditionally independent with X_k, given all remaining variables $X_{\mathcal{V}\backslash\{j,k\}} = \{X_i; i \in \mathcal{V}\backslash\{j,k\}\}$. In the Gaussian graphical model, recovering the edge set $\mathcal{E}$ is equivalent to recovering the sparsity pattern of the precision matrix $\mathbf{\Sigma}^{-1}$ (Lauritzen, 1996; Mardia et al., 1980).

In recent years, much work has been done for recovering the structure of different classes of graphical models. A great number of literatures have studied the estimation of a sparse precision matrix for Gaussian graphical models (Yuan and Lin (2007); Friedman et al. (2008); Rothman et al. (2008); Ravikumar et al. (2009) and references therein). Liu et al. (2009) and Liu et al. (2012a) introduced the notion of a nonparanormal distribution, which results from univariate monotonic transformations of the Gaussian distribution, and showed that the structural properties of the inverse covariance matrix of the Gaussian distribution carry over to the corresponding nonparanormal distribution. Ravikumar et al. (2010) and Anandkumar et al. (2012) have considered recovering structure of Ising graphical model in high dimensional setting. Yang et al. (2012) have studied a class of graphical models based on generalized linear models by assuming that node-wise conditional distributions arise from exponential families; they argued that the neighborhood of their graphical models can be recovered exactly with high probability.

Luo et al. (2014) proposed a computationally-efficient screening approach for Gaussian graphcial models, *Graphical Sure Screening* (GRASS), which possesses desirable statistical properties. Motivated by the fact that jth column of the precision matrix Σ^{-1} can be obtained by regressing the jth feature onto the $p-1$ other features (Mardia et al., 1980), they simply threshold the sample correlations of the jth feature with the $p-1$ other features, and they show that under certain simple assumptions, the set of nodes for which the sample correlation with the jth node exceeds some threshold is guaranteed to contain the true neighborhood, $\mathcal{E}_j$, with very high probability. Such property holds when the dimension p grows as an exponential function of the sample size n. In contrast to other existing methods for estimating a sparse precision matrix, which typically require $\mathcal{O}(p^3)$ computations (Friedman et al., 2008), GRASS requires only $\mathcal{O}(p^2)$ operations (Luo et al., 2014).

Although Gaussian graphical models are useful, a reliance on exact normality is not desirable in many cases. So our goal in this paper is to extend the GRASS procedure to a more general family named *transelliptical graphical model*, introduced by Liu et al. (2012b). We call our algorithm transelliptical GRASS, and we show that under a certain set of simple assumptions, the desirable statistical properties held by GRASS are also held by transelliptical GRASS. However, for most transelliptical graphical models, the graphs only represent the partial correlation, rather than conditional dependence, among variables. Therefore, by extending the Gaussian graphical model to the transelliptical graphical model, the gain in modeling flexibility is at a cost of the strength of inference.

The rest of this paper is organized as follows. In section 2, we provide the background on transelliptical graphical models and present a useful property of Kendall's tau statistic. In section 3, we establish the theoretical properties for transelliptical sure screening, which include the sure screening property, size control of the selected edge set, and the control of the expected false positive rate; we provide a choice of the threshold value that leads to the aforementioned desirable properties. Simulation studies are presented in section 4, and we close with a discussion in section 5.

2 Transelliptical Graphical Models & Background on Kendall's τ

The transelliptical distribution, first introduced by Liu et al. (2012b), is a generalization of the nonparanormal distribution proposed by Liu et al. (2009, 2012a). The transelliptical distribution extends the elliptical distribution by mimicking the way how the nonparanormal extends the normal distribution. So before presenting the definition of the transelliptical distribution, we first provide the following definition of the elliptical distribution.

Definition 2.1 (elliptical distribution): Let $\mu \in \mathbb{R}^p$ and $\Sigma \in \mathbb{R}^{p\times p}$ with $\mathrm{rank}(\Sigma) = q \leq p$. A p-dimensional random vector X has an elliptical distribution, denoted by $X \sim \mathrm{EC}_p(\mu, \Sigma, r)$, if it has a stochastic representation:

$$X \overset{d}{=} \mu + rAU, \tag{2.1}$$

where U is a random vector uniformly distributed on the unit sphere in $\mathbb{R}^q$, $r \geq 0$ is a scalar random variable independent of U, $A \in \mathbb{R}^{p\times q}$ satisfies $\Sigma = AA^T$, and $X \overset{d}{=} Y$ indicates that X and Y have the same distribution.

Remark 2.1: Many multivariate distributions belong to the elliptical distribution family. For example, the multivariate normal and the multivariate t-distribution belong to the elliptical family. In addition, if $X = (X_1, \ldots, X_p)^T$ follows a certain p-dimensional elliptical distribution, any pair of variables (X_j, X_k) follows the same bivariate elliptical distribution.

Definition 2.2 (transelliptical distribution): A continuous random vector $X = (X_1, \ldots, X_p)^T$ follows a p-dimensional transelliptical distribution if there exists a vector of monotone univariate functions $\{f_j\}_{j=1}^p$ such that the functional transformation $Z = f(X) = (f_1(X_1), \ldots, f_p(X_p))^T$ follows a p-dimensional elliptical distribution $\mathrm{EC}_p(\mu, \Sigma, r)$, denoted by $\mathrm{TE}_p(\mu, \Sigma, r; f_1, \ldots, f_p)$, where the covariance matrix Σ has unit diagonal elements. From now on, we denote $Z = (Z_1, \ldots, Z_p)^T = (f_1(X_1), \ldots, f_p(X_p))^T$ as the latent variables of $X = (X_1, \ldots, X_p)^T$.

Remark 2.2: A random vector $X = (X_1, \ldots, X_p)^T$ follows a nonparanormal distribution (Liu et al., 2009, 2012a) if there exists a vector of monotone univariate functions $\{f_j\}_{j=1}^p$ such that $Z = f(X) = (f_1(X_1), \ldots, f_p(X_p))^T \sim N_p(0, \Sigma)$. Therefore, transelliptical is a strict extension of the nonparanormal distribution.

Given a transelliptical distribution $\mathrm{TE}_p(\mu, \Sigma, r; f_1, \ldots, f_p)$, we can define an undirected graph $G = (\mathcal{V}, \mathcal{E})$. $\mathcal{V}$ contains the nodes corresponding to the p variables in X, and the edge set $\mathcal{E}$ satisfies that $(X_j, X_k) \in \mathcal{E}$ if and only if $\Sigma^{-1}_{jk} \neq 0$ for $j, k = 1, \ldots, p$ (Liu et al., 2012b).

In the special case of a nonparanormal graphical model, whose latent variables $(Z_1, \ldots, Z_p)^T$ conform to a multivariate normal distribution $N(\mu, \Sigma)$, the zero entry of the precision matrix Σ^{-1} implies not only the pairwise conditionally uncorrelatedness, but also conditionally independence between the corresponding random variables.

From now on, we assume Σ is positive definite and denote the (j, k)th entry of Σ as ρ_{jk}. Therefore, the matrix A in (2.1) is a non-degenerate $p \times p$ square matrix, since $\mathrm{rank}(\Sigma) = \mathrm{rank}(AA^T) \leq \mathrm{rank}(A)$. To simplify the presentation, in the rest of this article we assume that Σ has been standardized in such a way that the diagonal elements are equal to 1.

Next, we present the definition and some theoretical properties of of Kendall's tau, which are relevant to the transelliptical GRASS.

Definition 2.3 (Kendall's τ): Consider a random vector $X = (X_1, \ldots, X_p)^T$, each variable having n observations so that X_{ij} denotes the value of the jth variable in the ith observation. Then

$$\hat{\tau}_{jk} = \frac{2}{n(n-1)} \sum_{1 \leq i \leq i' \leq n} \mathrm{sign}((X_{ij} - X_{i'j})(X_{ik} - X_{i'k})). \tag{2.2}$$

It is an unbiased estimator for the population-level Kendall's tau equals to τ_{jk}, i.e.,

$$\tau_{jk} = \mathbb{E}[\mathrm{sign}((X_{ij} - X_{i'j})(X_{ik} - X_{i'k}))].$$

Now consider a random vector $X = (X_1, \ldots, X_p)^T$ following $\mathrm{TE}_p(\mu, \Sigma, r; f_1, \ldots, f_p)$, and each variable contains n observations. The following lemma presents the connection between the Kendall's tau and the correlation matrix Σ.

***Lemma* 1 (Liu et al., 2012b)** Assuming a random vector $X = (X_1, \ldots, X_p)^T$ following

$\mathrm{TE}_p(\mu, \Sigma, r; f_1, \dots, f_p)$, denote τ_{jk} to be the population Kendall's τ statistic between X_j and X_k, then we have $\rho_{jk} = \sin(\frac{\pi}{2}\tau_{jk})$.

Motivated by this lemma, (Liu et al., 2012b) use the following estimator $\hat{S}^\tau$ for the unknown correlation matrix Σ:

$$\hat{S}^\tau_{jk} = \begin{cases} \sin(\frac{\pi}{2}\hat{\tau}_{jk}) & \text{if } j \neq k \\ 1 \quad \text{if } j = k \end{cases}$$

We will use this $\hat{S}^\tau$ as an estimator of Σ in our transelliptical sure screening algorithm.

The following lemma about the asymptotic variance of Kendall's tau statistic is used to derive the asymptotic distribution of $\hat{S}^\tau_{jk}$:

***Lemma* 2 (Vaart, 2000).** For any i.i.d pairs (X_{ij}, X_{ik}), $i = 1, \dots, n$, we have that $\sqrt{n}(\hat{\tau}_{jk} - \tau_{jk})$ is asymptotically normally distributed with mean equal to 0 and variance $\sigma^2_{\tau_{jk}}$, where

$$\sigma^2_{\tau_{jk}} = 4\mathrm{Var}[\mathbb{E}[\mathrm{sign}(X_{1j} - X_{2j})\mathrm{sign}(X_{1k} - X_{2k})|X_{1j}, X_{1k}]]. \tag{2.3}$$

A direct use of delta method shows that $\sqrt{n}[\sin(\frac{\pi}{2}\hat{\tau}_{jk}) - \sin(\frac{\pi}{2}\tau_{jk})]$ is also asymptotically normal with mean zero and variance $\frac{\pi^2}{4}\sigma^2_{\tau_{jk}}\cos^2(\frac{\pi}{2}\tau_{jk})$.

3 Transelliptical Graphical Sure Screening

3.1 Proposed Approach

Suppose $X_1, \dots, X_p \overset{i.i.d.}{\sim} \mathrm{TE}_p(\mu, \Sigma, r; f_1, \dots, f_p)$. Let $\gamma_{n,jk} > 0$ be some pre-specified threshold for the (j,k)th entry of the estimated correlation matrix.

We propose to estimate the candidate edge set $\hat{\mathcal{E}}_{\gamma_n}$ and the candidate neighborhood $\hat{\mathcal{E}}_{j,\gamma_n}$ for the jth node as follows,

$$\hat{\mathcal{E}}_{\gamma_n} = \{(j,k) : j < k, |\hat{S}^\tau_{jk}| > \gamma_{n,jk}\} \tag{3.1}$$

and

$$\hat{\mathcal{E}}_{j,\gamma_n} = \{k : k \neq j, |\hat{S}^{\tau}_{jk}| > \gamma_{n,jk}\}. \tag{3.2}$$

We refer to $\hat{\mathcal{E}}_{\gamma_n}$ and $\hat{\mathcal{E}}_{j,\gamma_n}$ as the transelliptical graphical sure screening estimators.

We will show in section 3.2 that for an appropriate choice of $\gamma_{n,jk}$, $\mathcal{E}_{j,\gamma_n}$ is contained in $\hat{\mathcal{E}}_{j,\gamma_n}$ with very high probability, when p grows exponentially with n. This is referred to as the sure screening property for the graphical model. Furthermore, under certain conditions, we can choose $\gamma_{n,jk}$ to control the expected false positive rate.

3.2 Theoretical Properties

We now present some theoretical properties of transelliptical graphical sure screening. Proofs are in the Appendix.

***Assumption* 1.** For some constant $C_1 > 0$ and $0 < \kappa < 1/2$,

$$\min_{(j,k)\in\mathcal{E}} |\rho_{jk}| \geq C_1 n^{-\kappa}.$$

Assumption 1 imposes conditions on the underlying correlation matrix that the elements in the edge set $\mathcal{E}$ correspond to sufficiently large values in correlation matrix. Since the precision matrix Σ^{-1} contains the information about the structure of partial correlation graph (Liu et al., 2012b), it will also be natural to put conditions directly on the precision matrix Σ^{-1}. Let λ_1, λ_p be the maximum and minimum eigenvalues of Σ^{-1} respectively, $\alpha \triangleq \lambda_1/\lambda_p$ and $\nu = \frac{2}{\lambda_1+\lambda_p}$. The following proposition develops a connection between the precision matrix and the correlation matrix satisfying the Assumption 1.

***Proposition* 1.** Let M be an integer that satisfies $M \geq [\kappa \log n - \log \lambda_p - \log \frac{C_1}{2}]/\log(\frac{\alpha+1}{\alpha-1})$, where C_1 is as in Assumption 1. If

$$\min_{(j,k)\in\mathcal{E}} |\sum_{i=1}^{M-1} (I - \nu\Sigma^{-1})^i|_{jk} \geq 2C_1 n^{-\kappa}/\nu, \tag{3.3}$$

then Assumption 1 holds.

Furthermore, if Assumption 1 holds, then

$$\min_{(j,k)\in\mathcal{E}} |\sum_{i=1}^{M-1}(I-\nu\Sigma^{-1})^i|_{jk} \geq \frac{C_1 n^{-\kappa}}{2\nu}. \tag{3.4}$$

This proposition does not provide a necessary and sufficient condition on the precision matrix that leads to a correlation matrix that satisfies Assumption 1. Assumption 1 implies (3.4), which is true for a larger set of precision matrices than (3.3). But we can say (3.3) is sufficient and almost necessary for Assumption 1.

We now present the sure screening property in Theorem 1.

***Theorem* 1.** Assume that Assumption 1 holds,and that $\log(p) = C_3 n^{\xi}$ for some constants $C_3 > 0$ and $\xi \in (0, 1-2\kappa)$. Let $\gamma_{n,jk} \equiv \gamma_n = \frac{2}{3}C_1 n^{-\kappa}$ in (3.1) and (3.2). Then there exists constants C_4 and C_5 such that

$$\Pr(\mathcal{E} \subseteq \hat{\mathcal{E}}_{\gamma_n}) \geq 1 - C_4 \exp(-C_5 n^{1-2\kappa})$$

and

$$\Pr(\mathcal{E}_j \subseteq \hat{\mathcal{E}}_{j,\gamma_n}) \geq 1 - C_4 \exp(-C_5 n^{1-2\kappa}).$$

Theorem 1 guarantees that the candidate edge set obtained from the transelliptical screening method contains the true edge set with high probability, which means the screening method will not result in false negatives with high probability. Our next theorem will provide a bound on the size of $\hat{\mathcal{E}}_{j,\gamma_n}$. This requires an additional assumption.

Assumption 2. There exist constants $\alpha \geq 0$ and $C_2 > 0$ such that $\lambda_{\max}(\Sigma) \leq C_2 n^{\alpha}$, where $\lambda_{\max}(\Sigma)$ is the maximal eigenvalue of Σ.

Assumption 2 allows the largest eigenvalue of the population covariance matrix Σ to diverge as n grows, but it cannot diverge too quickly. This condition holds in many cases; the covariance matrix of a stationary time series is one example (Fan and Lv, 2008).

***Theorem* 2.** Let $\gamma_{n,jk} \equiv \gamma_n = \frac{2}{3}C_1 n^{-\kappa}$ in (3.1) and (3.2). Under Assumptions 1-2, if $\log(p) = C_3 n^{\xi}$ for some constants $C_3 > 0$ and $\xi \in (0, 1-2\kappa)$, then $\Pr(|\hat{\mathcal{E}}_{j,\gamma_n}|\leq O(n^{2\kappa+\alpha}))\geq 1 - C_4 \exp(-C_5 n^{1-2\kappa})$, where the constants C_4 and C_5 are as in Theorem 1.

Next we propose a choice of the threshold $\gamma_{n,jk}$ that enables us to control the expected false positive rate at a prespecified value. We define the false positive rate as $\frac{|\hat{\mathcal{E}}_{\gamma_n} \cap \mathcal{E}^C|}{\mathcal{E}^C}$. First we assign a fixed value f as the number of false positives that we are willing to tolerate, then the false positive rate will be $f/|\mathcal{E}^C|$. Two assumptions and one lemma are introduced below to facilitate the theorem.

***Assumption* 3.** For the same ξ as in Theorem 1,

$$\max_{(j,k)\notin\mathcal{E}} |\rho_{jk}| = o(n^{-\frac{1-\xi}{2}}).$$

Remark 3.1: Since Lemma 2 defines $\rho_{jk} = \sin(\frac{\pi}{2}\tau_{jk})$, the Assumption 3 is equivalent to the following in terms of τ_{jk}:

$$\max_{(j,k)\notin\mathcal{E}} |\tau_{jk}| = o(n^{-\frac{1-\xi}{2}}).$$

Another weak assumption on the underlying elliptical distribution is needed to control the false positive rate. To simplify the presentation, we define

$$Y = (Y_1, \ldots, Y_p)^T \overset{d}{=} rU, \tag{3.5}$$

where r, U are defined in (2.1). The random vector Y is said to possess a spherical distribution. The use of the representation (3.5) and the following Assumption 4 appear in the proof of Theorem 3 in the Appendix.

***Assumption* 4.** The joint c.d.f of the random vector Y defined in (3.5) is continuous.

Then we can introduce Theorem 3 as follows.

***Theorem* 3.** Assume that assumptions 1-4 hold. If $\log(p) = C_3 n^{\xi}$ for ξ defined in Theorem 1, then we can control the asymptotic expected false positive rate at $f/|\mathcal{E}^C|$ by choosing $\gamma_{n,jk} \equiv \gamma_n = \frac{\pi}{2}\sigma_0 \Phi^{-1}(1 - \frac{f}{p(p-1)})/\sqrt{n}$ where σ_0 is defined in (2.3) with $\tau_{jk} = 0$. Furthermore, with this threshold, the sure screening property of Theorem 1 still holds.

Remark 3.2: In Theorem 3, We need to know the value of σ_0^2 in order to get the threshold value $\gamma_{n,jk}$. For some special transelliptical distribution, σ_0^2 has closed form value. For example, if it is nonparanormal distribution, then σ_0^2 is $\frac{4}{9}$ (Vaart, 2000); if it is transelliptical t distribution with degree of freedom equal to ν, then σ_0^2 is $\frac{32\Gamma(\frac{3v}{2})}{\pi^2\Gamma^3(\frac{v}{2})} \int_0^{\infty} u^{v-1} \arctan^2 u \int_0^1 t^{v-1}(1-t)^{v-1}(u^2+t)^{-v} dt du$ (Dengler, 2010). However, for some more general transelliptical distribution, no closed form value of σ_0^2 is available; this leads to the following corollary:

Corollary. Assume all the conditions in Theorem 3 hold. We can control the asymptotic expected false positive rate at $f/|\mathcal{E}^C|$ by choosing $\gamma_{n,jk} = \frac{\pi}{2}\hat{\sigma}_{\tau_{jk}} \Phi^{-1}(1 - \frac{f}{p(p-1)})/\sqrt{n}$, where $\hat{\sigma}^2_{\tau_{jk}}$ is a Jackknife estimator of the $\sigma^2_{\tau_{jk}}$ in (2.3) and has the form as

$$\hat{\sigma}^2_{\tau_{jk}} = \frac{4(n-1)}{(n-2)^2} \sum_{i'=1}^{n} \left[\left\{ \frac{1}{n-1} \sum_{i=1, i \neq i'}^{n} \operatorname{sign}((X_{ij} - X_{i'j})(X_{ik} - X_{i'k})) \right\} - \hat{\tau}_{jk} \right]^2, \qquad (3.6)$$

where $\hat{\tau}_{jk}$ is defined in (2.2).

This Jackknife estimator for the variance of Kendall's tau follows from the Jackknife estimator for the variance of U-statistics proposed by Arvesen (1969). It has been used and recommended as an estimator for the variance of Kendall's tau in many literatures (Sen, 1977; Callaert and Veraverbeke, 1981; Fligner and Rust, 1983; Lee, 1985).

4 Simulation Studies

4.1 Data Generation

Let p be the number of features, and n be the number of observations. We considered four ways of generating the edge set $\mathcal{E}$:

Simulation A: For all $j < k$, we set $(j, k) \in \mathcal{E}$ with probability 0.1. Once the edge set $\mathcal{E}$ was generated, we created a precision matrix via the following steps:

Step 1: We generated a $p \times p$ matrix A, where

$$A_{jk} = A_{kj} = \begin{cases} 1 & \text{for } j = k \\ \text{Unif}[-0.3, 0.7] & \text{for } (j,k) \in \mathcal{E} \\ 0 & \text{otherwise} \end{cases} \cdot$$

Step 2: We created a positive definite matrix $\mathbf{\Sigma}^{-1}$:

$$\Sigma^{-1} = A + (0.1 - \lambda_{\min}(A))I,$$

where $\lambda_{\min}(A)$ is the smallest eigenvalue of A, and I denotes the $p \times p$ identity matrix. Then we took the inverse of Σ^{-1} to get Σ.

Simulation B: We partitioned the p features into 10 equally-sized and non-overlapping sets $\{C_1, ..., C_{10}\}$ such that $C_i = \{(i-1)p/10+1, ..., ip/10\}$, $|C_i| = p/10$ for $i = 1, ..., 10$, and $C_i \cap C_{i'} = \emptyset$ for $i \neq i'$. For all $j \in C_i, k \in C_i, j < k$, we set $(j, k) \in \mathcal{E}$. This constructed a graph with ten densely connected components. Therefore, Σ^{-1} is a block diagonal matrix with each block matrix being a $p/10 \times p/10$ matrix. Each block matrix of Σ^{-1} was constructed by following Step 1 and Step 2 in Simulation A. Then we took the inverse of Σ^{-1} to get Σ.

Simulation C: For all $j \leq k$, we set $\Sigma_{jk} = 0.3^{|j-k|}$. All other elements of Σ were set to zero. This was a setup of autocorrelation with correlation equal to 0.3.

Simulation D: We partitioned the features into $p/10$ equally-sized and non-overlapping sets $\{C_1, ..., C_{p/10}\}$ such that $C_i = \{10(i-1)+1, ..., 10i\}$, $|C_i| = 10$ for $i = 1, ..., p/10$, and $C_i \cap C_{i'} = \emptyset$ for $i \neq i'$. Then for all $j \in C_i, k \in C_i$, we set $(\Sigma^{-1})_{jk} = 0.9^{|j-k|}$. All other elements of Σ^{-1} were set to zero. Therefore, Σ^{-1} is a block diagonal matrix with $p/10$ block matrices, and each block matrix is 10×10. After all, we took the inverse of Σ^{-1} to get Σ.

In all simulations, Σ was rescaled to have diagonal elements equal to 1. To generate nonparanormal distributed data, we first generated n observations i.i.d. from a $N(0, \Sigma)$ distribution. Then we applied four monotonic functions, $exp(x), x^3, x^5, (x-1)^3$, to observations randomly with equal probability in order to generate n i.i.d nonparanormal distributed observations. To generate transelliptical t distribution with degree of freedom = ν, we first generated n observations i.i.d. from multivariate t distribution with degree of freedom = ν and correlation matrix Σ, and then applied same monotonic functions to get transelliptical t distributed observations.

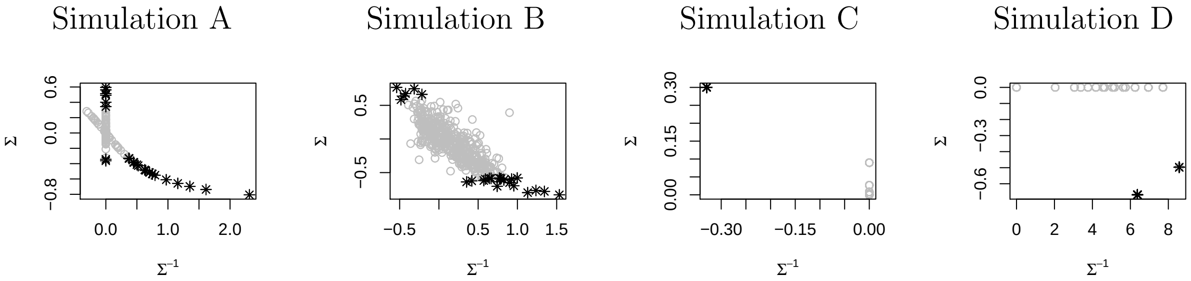
4.2 Control of False Positive Rate

Theorem 3 and Remark 3.2 state that under certain conditions, performing Transelliptical GRASS with $\gamma_{n,jk} \equiv \gamma_n = \frac{\pi}{3\sqrt{n}}\Phi^{-1}(1-q/2)$ leads to control of the expected false positive rate (FPR) at level $q \equiv f/|\mathcal{E}^c|$ if the distribution is nonparanormal. Simulation results of controlling false positive rates for Simulations A-D are shown in Table 1. The false positive rates are controlled very well in all four simulation settings.

Assumption 3, required for Theorem 3 to hold, states that $\max_{(j,k)\notin\mathcal{E}} |\rho_{jk}| \to 0$ as $n \to \infty$, which implies that $\max_{(j,k)\notin\mathcal{E}} |\tau_{jk}| \to 0$ as $n \to \infty$. In Simulation B, Σ^{-1} is block diagonal with ten completely dense blocks, thus the same is true for Σ. So Σ and Σ^{-1} have same sparsity patterns, and the assumption 3 holds in Simulation B. In Simulation D, all the zero elements in the precision matrix Σ^{-1} correspond to zero elements in Σ, thus the assumption 3 holds exactly in Simulation D. As expected, the FPRs are controlled successfully in Simulation B and D (Table 1).

Nonetheless, in Simulations A and C, not all the zero elements in the precision matrix Σ^{-1} correspond to zero elements in the correlation matrix Σ. But Table 1 reveals that the FPRs are still controlled well in these settings. The reason is that the assumption 3 only requires the elements in $\mathcal{E}^c$ to correspond to small, though not necessarily zero, elements of Σ. This assumption holds for most of the elements in $\mathcal{E}^c$ and $\mathcal{E}$ for Simulation A and C, as can be seen in Fig.1. Therefore, the FPRs are also well-controlled in Simulation A and C.

We then performed transelliptical GRASS on transelliptical t distributed data with degree of freedom = 5. We chose both $\gamma_{n,jk} \equiv \gamma_n = \frac{\pi}{2}\sigma_0\Phi^{-1}(1 - \frac{f}{p(p-1)})/\sqrt{n}$ where σ_0 was derived by the formula in Remark 3.2 (Dengler, 2010), and $\gamma_{n,jk} = \frac{\pi}{2}\hat{\sigma}_{\tau_{jk}}\Phi^{-1}(1 - \frac{f}{p(p-1)})/\sqrt{n}$ where $\hat{\sigma}_{\tau_{jk}}$ was defined in (3.6) in Corollary. We see that in either case, the false positive rates are well-controlled as Table 2 & 3 shows.


Simulation A

Simulation B

Simulation C

Simulation D

Figure 1: For Simulations A-D using nonparanormal data with $p = 100$, the off-diagonal elements of Σ^{-1} (*x-axis*) and Σ (*y-axis*) are shown. The 0.5% of largest absolute off-diagonal elements of Σ are shown in black; the rest are in grey. For all setups, the vast majority of large off-diagonal elements of Σ correspond to non-zero elements of Σ^{-1}, and the vast majority of zero elements in Σ^{-1} correspond to small elements in Σ. The pronounced relationship seen for Simulation A is due to the extreme sparsity of Σ^{-1}.

			Simulation A			Simulation B		
n	p	q	$\lvert\widehat{\mathcal{E}}_{\gamma_n}\rvert$	FPR	FNR	$\lvert\widehat{\mathcal{E}}_{\gamma_n}\rvert$	FPR	FNR
100	1000	1e-04	366.08	0	0.940	1257.892	0	0.975
		0.001	1219.876	0.0012	0.875	2968.876	0.00087	0.948
		0.01	6992.152	0.012	0.758	10513.404	0.01	0.879
		0.1	56942.056	0.11	0.538	62707.764	0.105	0.684
		0.2	108340.424	0.213	0.433	114150.056	0.206	0.568
		0.3	159132.392	0.315	0.356	164522.116	0.308	0.474
		0.5	257967.544	0.514	0.236	262013.268	0.508	0.321
			Simulation C			Simulation D		
n	p	q	$\lvert\widehat{\mathcal{E}}_{\gamma_n}\rvert$	FPR	FNR	$\lvert\widehat{\mathcal{E}}_{\gamma_n}\rvert$	FPR	FNR
100	1000	1e-04	163.86	0	0.87	834.96	0	0.822
		0.001	770.792	0.0009	0.673	1310.952	0.0009	0.806
		0.01	5694.66	0.0102	0.373	5922.472	0.0101	0.793
		0.1	53206.996	0.105	0.0998	53073.068	0.105	0.716
		0.2	103881.508	0.206	0.05	103705.392	0.206	0.634
		0.3	154620.912	0.308	0.03	154404.724	0.308	0.553
		0.5	254199.664	0.508	0.012	253970.268	0.508	0.393

Table 1: This table is for Nonparanormal distribution. The false positive rate (FPR; defined as FP/(FP+TN)) and false negative rate (FNR; defined as FN/(TP+FN)) are reported for various values of the level of desired FPR control, q. The value of $|\widehat{\mathcal{E}}_{\gamma_n}|$ is also reported. We set $n = 100$ and $p = 1000$ for all Simulations A-D. Results are averaged over 250 simulated data sets.

			Simulation A			Simulation B		
n	p	q	$\lvert\widehat{\mathcal{E}}_{\gamma_n}\rvert$	FPR	FNR	$\lvert\widehat{\mathcal{E}}_{\gamma_n}\rvert$	FPR	FNR
100	1000	1e-04	253.02	0	0.959	940.428	0	0.981
		0.001	980.66	0.001	0.903	2417.696	0.0008	0.958
		0.01	6435.608	0.011	0.791	9542.504	0.0096	0.895
		0.1	55497.032	0.11	0.568	60941.968	0.103	0.705
		0.2	107330.096	0.212	0.458	112922.612	0.205	0.587
		0.3	157562.256	0.312	0.378	162780.472	0.306	0.493
		0.5	256236.104	0.511	0.252	260198.724	0.505	0.335
			Simulation C			Simulation D		
n	p	q	$\lvert\widehat{\mathcal{E}}_{\gamma_n}\rvert$	FPR	FNR	$\lvert\widehat{\mathcal{E}}_{\gamma_n}\rvert$	FPR	FNR
100	1000	1e-04	107.624	0	0.92	736.5	0	0.842
		0.001	627.18	0.0008	0.759	1214.632	0.0008	0.814
		0.01	5364.504	0.0097	0.467	5690.736	0.0096	0.795
		0.1	52321.2	0.103	0.146	52272.248	0.103	0.718
		0.2	103551.524	0.206	0.078	103427.08	0.206	0.636
		0.3	153672.364	0.306	0.049	153514.02	0.306	0.555
		0.5	252959.084	0.505	0.022	252769.824	0.505	0.396

Table 2: This table is for Transelliptical T distribution with degree of freedom = 5. We choose $\gamma_{n,jk}$ according to Theorem 3 here. The false positive rate (FPR; defined as FP/(FP+TN)) and false negative rate (FNR; defined as FN/(TP+FN)) are reported for various values of the level of desired FPR control, q. The value of $|\widehat{\mathcal{E}}_{\gamma_n}|$ is also reported. We set $n = 100$ and $p = 1000$ for all Simulations A-D. Results are averaged over 250 simulated data sets.

			Simulation A			Simulation B		
n	p	q	$\lvert\widehat{\mathcal{E}}_{\gamma_n}\rvert$	FPR	FNR	$\lvert\widehat{\mathcal{E}}_{\gamma_n}\rvert$	FPR	FNR
100	1000	1e-04	448.836	0	0.941	1325.172	0	0.975
		0.001	1447.58	0.0018	0.889	3039.336	0.0015	0.952
		0.01	7261.952	0.013	0.788	10340.736	0.011	0.892
		0.1	53507.624	0.104	0.575	58758.12	0.099	0.713
		0.2	102902.456	0.203	0.468	108367.908	0.197	0.598
		0.3	152190.388	0.302	0.387	157358.224	0.295	0.504
		0.5	250987.548	0.50	0.258	254984.984	0.494	0.344
			Simulation C			Simulation D		
n	p	q	$\lvert\widehat{\mathcal{E}}_{\gamma_n}\rvert$	FPR	FNR	$\lvert\widehat{\mathcal{E}}_{\gamma_n}\rvert$	FPR	FNR
100	1000	1e-04	240.52	0	0.87	859.488	0	0.833
		0.001	1012.12	0.0015	0.721	1564.668	0.0015	0.813
		0.01	6123.648	0.0112	0.466	6438.42	0.0111	0.794
		0.1	50330.904	0.099	0.158	50297.052	0.099	0.721
		0.2	99069.432	0.197	0.085	98909.556	0.197	0.642
		0.3	148208.564	0.295	0.053	148021.74	0.295	0.563
		0.5	247579.316	0.495	0.023	247400.096	0.494	0.404

Table 3: This table is for Transelliptical T distribution with degree of freedom = 5. We choose $\gamma_{n,jk}$ according to Corollary here. The false positive rate (FPR; defined as FP/(FP+TN)) and false negative rate (FNR; defined as FN/(TP+FN)) are reported for various values of the level of desired FPR control, q. The value of $|\widehat{\mathcal{E}}_{\gamma_n}|$ is also reported. We set $n = 100$ and $p = 1000$ for all Simulations A-D. Results are averaged over 250 simulated data sets.

4.3 Comparison to Existing Approaches

We first compare the performances of the graphical lasso (Friedman et al., 2008), neighborhood selection (Meinshausen and Bühlmann, 2006), and Transelliptical GRASS on nonparanormal data generated from Simulations A-D, with $n = 50$ and $p = 750$. For each method, we use Kendall's tau estimator $\hat{S}^{\tau}$ to estimate Σ. On purpose for further comparison, we also include the simulation results for graphical lasso, neighborhood selection, and GRASS (Luo et al., 2014) without using Kendall's tau estimator $\hat{S}^{\tau}$ to estimate Σ. Results are displayed in Figure 2.

Recall that in Simulation B the sparsity patterns of Σ and Σ^{-1} are identical, thus it is the ideal scenario for applying Transelliptical GRASS. In this setting, the Transelliptical GRASS outperforms the graphical lasso and neighborhood selection that also use Kendall's tau estimator of Σ, since it correctly assumes that the sparsity patterns of the correlation matrix and precision matrix are similar.

In Simulations A and C, the Assumption 3 holds for most of the elements as we discussed in the previous section. The Assumption 1 also holds in these two simulation setups because most of the elements in the edge set $\mathcal{E}$ correspond to large values in correlation matrix Σ. Therefore, the performance of Transelliptical GRASS is quite competitive compared to graphical lasso and neighborhood selection using Kendall's tau estimator of Σ in Simulations A and C.

We designed Simulation D to violate Assumption 1 on purpose; most of the elements in the edge set $\mathcal{E}$ (which are nonzero in the precision matrix Σ^{-1}) correspond to zero values in correlation matrix Σ. Therefore we can see that Transelliptical GRASS does not perform well in Simulation D. But even in this undesirable setting, the performance of Transelliptical GRASS is still close to the graphical lasso and neighborhood selection using Kendall's tau estimator of Σ. It reveals that the graphical lasso and neighborhood selection using Kendall's tau estimator of Σ are also not good in such setting.

Not surprisingly, the graphical lasso, neighborhood selection, and GRASS without using Kendall's tau estimator $\hat{S}^{\tau}$ of Σ perform poorly, because they are designed for Gaussian distributed data while we use nonparanormal data here.

We then compare the performance of Transelliptical GRASS (assuming nonparanormal data) to the performance of regular GRASS on Gaussian distributed data generated from Simulations A-D, with $n = 50$ and $p = 750$. Figure 3 shows that the performance of regular GRASS is only slightly better than the performance of the Transelliptical GRASS on Gaussian data . It suggests that if we are not sure whether the data is Gaussian or nonparanormal, we can directly apply the Transelliptical GRASS (assuming nonparanormal data) without losing much power if the distribution is exactly normal but gaining much more power if the distribution is nonparanormal.

Overall, Figure 2 and Figure 3 suggest that in these four settings the Transelliptical GRASS performs competitively compared to some specialized and computationally-intensive procedures for estimating a precision matrix.

For further comparison, we also did the same simulations by using Spearman's rho estimator to estimate Σ suggested by Liu et al. (2012a). It gave almost same results for Transelliptical GRASS using Kendall's tau estimator $\hat{S}^{\tau}$ of Σ, but it gave slightly different results for graphical lasso and neighborhood selection using Kendall's tau estimator of Σ.

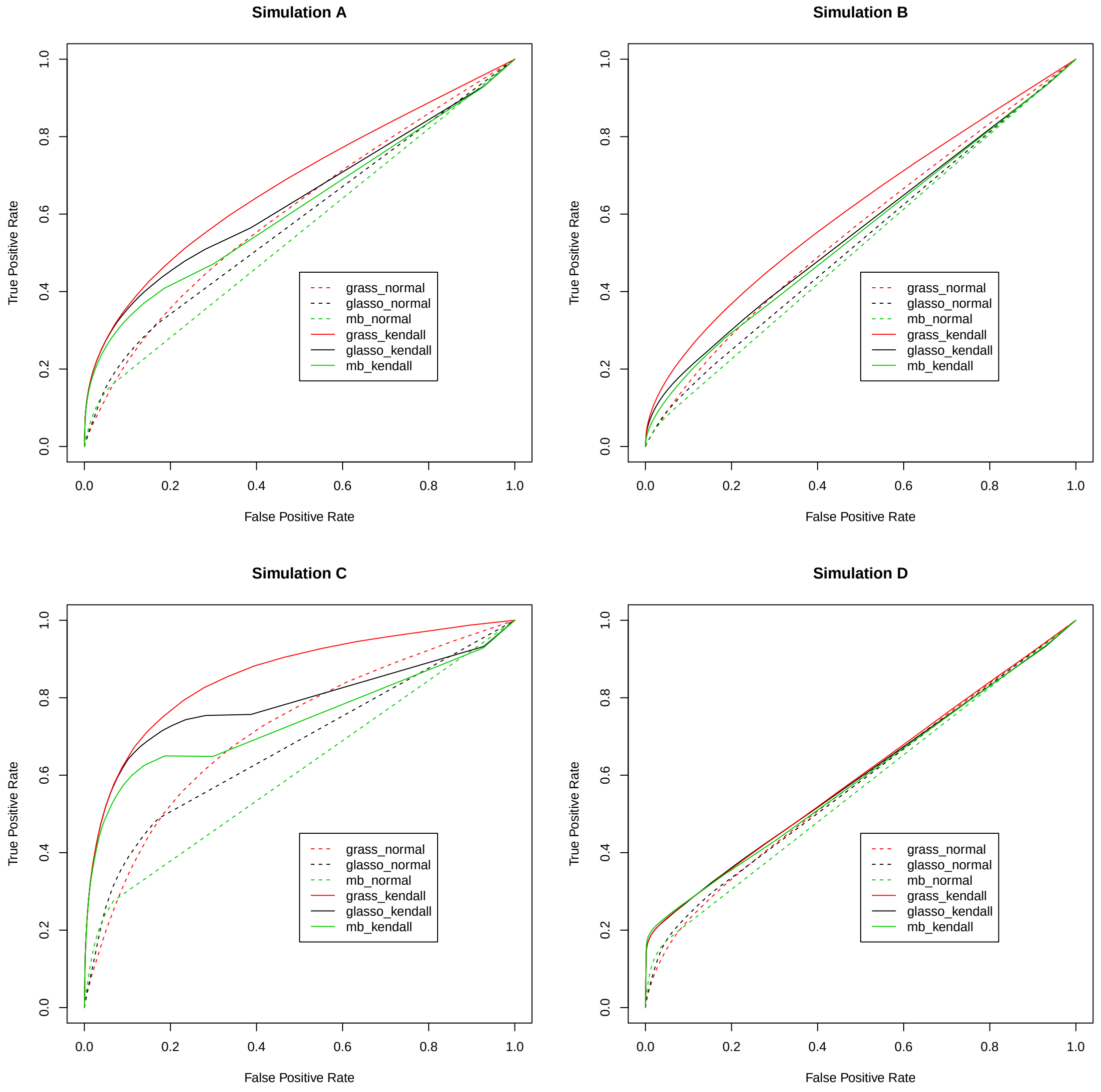


Figure 2: For Simulations A-D using nonparanormal distributed data with $p = 750$ and $n = 50$, the true positive and false positive rates are shown. Curves are obtained by varying the tuning parameter for each method and the results are averaged over 20 simulated data sets.

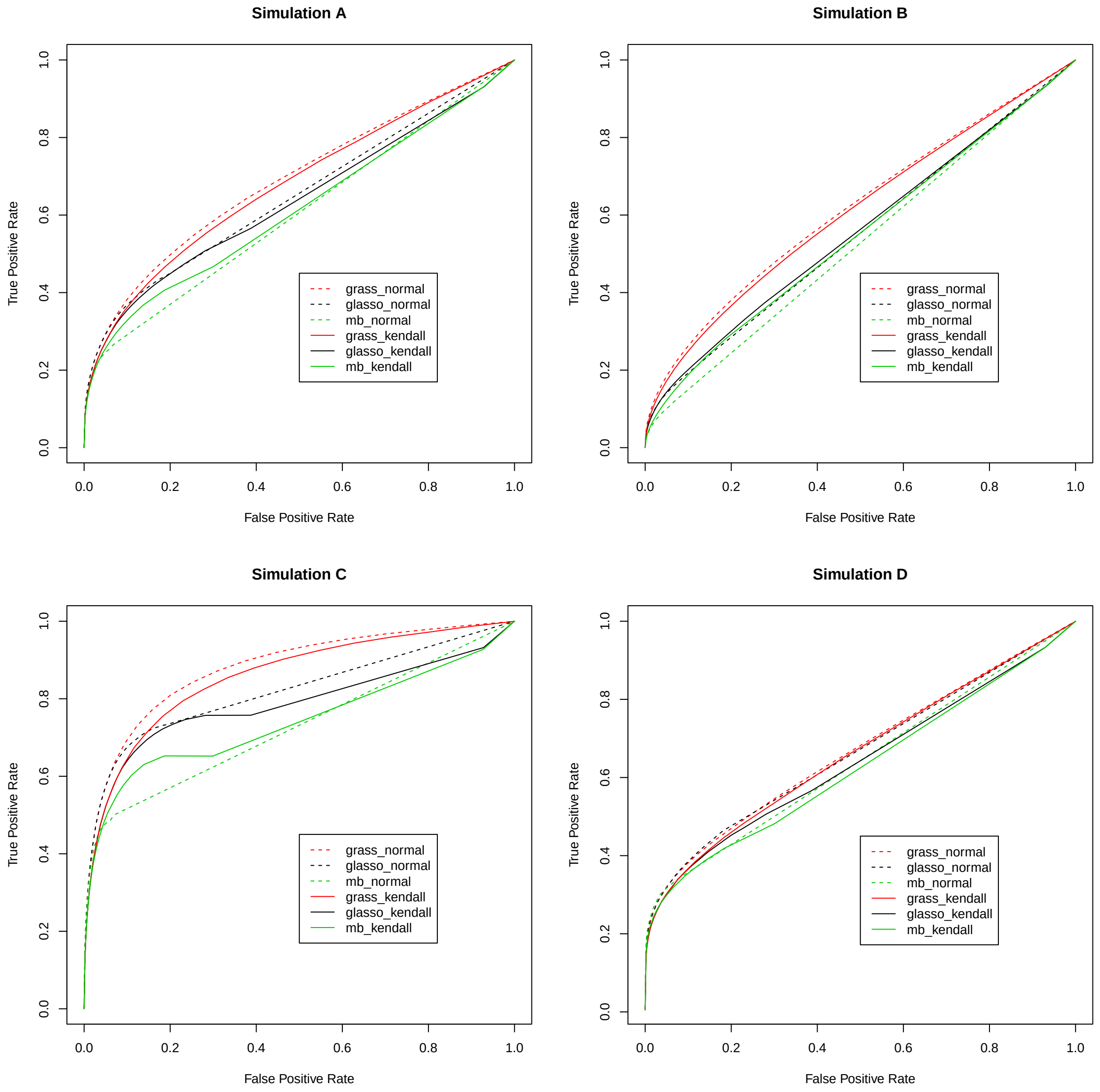


Figure 3: For Simulations A-D using Gaussian distributed data with $p = 750$ and $n = 50$, the true positive and false positive rates are shown. Curves are obtained by varying the tuning parameter for each method and the results are averaged over 20 simulated data sets.

5 Discussion

In this paper, we have proposed the Transelliptical GRASS, which is a simple and efficient procedure for recovering the structure of a high-dimensional transelliptical graphical model. It is a natural extension of the graphical sure screening proposed by Luo et al. (2014). For some special transelliptical graphical models, we provide closed-form threshold values to use in our algorithm. For more general transelliptical graphical models, we provide a choice of threshold values based on a Jackknife estimator; this makes our algorithm more applicable to many real life cases. In addition, the assumptions required for the transelliptical GRASS are as simple as the GRASS. Also, the transelliptical GRASS inherits the computational advantages of the GRASS; it requires only $\mathcal{O}(p^2)$ operations.

For nonparanormal distribution, the estimated graph presents the conditional independence among variables, since the underlying elliptical distribution is Gaussian. But for other transelliptical graphical models, the graphs only present the partial correlation, not conditional independence, among variables. Therefore, by extending the Gaussian graphical model to the transelliptical graphical models, a loss in the strength of inference is accompanied when we gain modeling flexibility.

Future work could consider further extensions of graphical sure screening to other models, or consider to deal with the loss in the strength of inference while keeping the modeling flexibility.

6 Appendix

6.1 Proof of Proposition 1

Since Ω is a positive definite matrix, letting $\nu = \frac{2}{\lambda_1+\lambda_p}$, we have

$$||I - \nu\Sigma^{-1}||_2 = \max(|\nu\lambda_1 - 1|, |\nu\lambda_p - 1|) = \frac{\lambda_1 - \lambda_p}{\lambda_1 + \lambda_p} = \frac{\alpha - 1}{\alpha + 1} < 1, \tag{6.1}$$

where $|| \cdot ||_2$ stands for the L_2 norm of a matrix.

By Neumann series,

$$\Sigma = \nu \sum_{i=0}^{M-1} (I - \nu\Sigma^{-1})^i + \nu \sum_{i=M}^{\infty} (I - \nu\Sigma^{-1})^i \triangleq B_M + A_M,$$

now it follows (6.1) that

$$||A_M||_2 \leq |\nu| \sum_{i=M}^{\infty} ||(I - \nu\Sigma^{-1})^i||_2 = \frac{2}{\lambda_1 + \lambda_p} \frac{(\frac{\lambda_1 - \lambda_p}{\lambda_1 + \lambda_p})^M}{1 - \frac{\lambda_1 - \lambda_p}{\lambda_1 + \lambda_p}} = \frac{1}{\lambda_p}(\frac{\alpha - 1}{\alpha + 1})^M.$$

When $M \geq [\kappa \log n - \log \lambda_p - \log \frac{C_1}{2}] / \log(\frac{\alpha+1}{\alpha-1})$, which suggests

$$|A_{Mjk}| = |e_j^T A_M e_k| \leq \sqrt{|e_j^T A_M A_M^T e_j|} \sqrt{|e_k^T e_k|} \leq ||A_M||_2 = \frac{1}{\lambda_p}(\frac{\alpha - 1}{\alpha + 1})^M \leq \frac{C_1}{2} n^{-\kappa}.$$

Together with (3.3), we have

$$|\rho_{jk}| \geq |B_{Mjk}| - |A_{Mjk}| \geq 2C_1 n^{-\kappa} - \frac{C_1}{2} n^{-\kappa} \geq C_1 n^{-\kappa}, \forall (j,k) \in \mathcal{E}.$$

Conversely, when assumption 1 holds, for $M \geq [\kappa \log n - \log \lambda_p - \log(C_1/2)] / \log(\frac{\alpha+1}{\alpha-1})$, similarly we have

$$|A_{Mjk}| \leq ||A_M||_2 = \frac{1}{\lambda_p}(\frac{\alpha - 1}{\alpha + 1})^M \leq \frac{C_1}{2} n^{-\kappa}.$$

Therefore,

$$|B_{Mjk}| \geq |\rho_{jk}| - |A_{Mjk}| \geq C_1 n^{-\kappa} - \frac{C_1}{2} n^{-\kappa} = \frac{C_1}{2} n^{-\kappa}, \forall (j,k) \in \mathcal{E},$$

which verifies (3.4). This completes the proof.

6.2 Proof of Theorem 1

The following lemma is used for the proof of sure screening property in Theorem 1.

Lemma (Hoeffding, 1963). Let $g = g(x_1, \ldots, x_m)$ be a symmetric kernel of the U-statistic, U, with $a \leq h(x_1, \ldots, x_m) \leq b$. For any $t > 0$ and $m \leq n$, we have

$$\Pr\{|U - \mathbb{E}(U)| > t\} \leq 2\exp\{\frac{-2\lfloor n/m \rfloor t^2}{(b-a)^2}\}.$$

Proof of Theorem 1. For the expected value of Kendall's tau statistic

$$\begin{aligned}\tau_{jk} &= \mathbb{E}(\hat{\tau}_{jk}) \\ &= \frac{2}{n(n-1)} \sum_{1 \leq i \leq i' \leq n} \Pr((X_{ij} - X_{i'j})(X_{ik} - X_{i'k}) > 0) - \Pr((X_{ij} - X_{i'j})(X_{ik} - X_{i'k}) < 0) \\ &= \Pr((X_{1j} - X_{2j})(X_{1k} - X_{2k}) > 0) - \Pr((X_{1j} - X_{2j})(X_{1k} - X_{2k}) < 0).\end{aligned}$$

For the Kendall's tau statistic $\hat{\tau}_{jk}$, it can be written as the structure of U-statistic $\hat{\tau}_{jk} = \frac{2}{n(n-1)} \sum_{1 \leq i \leq i' \leq n} g(W_i, W_{i'})$, where $g(W_i, W_{i'}) = sign((X_{ij} - X_{i'j})(X_{ik} - X_{i'k}))$ is symmetric and taking the values -1 and 1. Therefore we apply lemma 1 into this case setting $n = n$, $m = 2$ and $h(x_1, \ldots, x_m) \in [-1, 1]$,

$$\Pr(|\hat{\tau}_{jk} - \tau_{jk}| > \frac{2}{3\pi} C_1 n^{-\kappa}) \leq 2e^{-\frac{1}{9\pi^2} C_1^2 n^{1-2\kappa}}.$$

Next we notice that

$$\begin{aligned}\Pr(|\hat{S}^{\tau}_{jk} - \rho_{jk}| > \frac{1}{3} C_1 n^{-\kappa}) &= \Pr(|\sin(\frac{\pi}{2}\hat{\tau}_{jk}) - \sin(\frac{\pi}{2}\tau_{jk})| > \frac{1}{3} C_1 n^{-\kappa}) \\ &\leq \Pr(|\hat{\tau}_{jk} - \tau_{jk}| > \frac{2}{3\pi} C_1 n^{-\kappa}) \\ &\leq 2e^{-\frac{1}{9\pi^2} C_1^2 n^{1-2\kappa}}.\end{aligned}$$

The first equation results from directly applying *Lemma 2* and our definition of $\hat{S}^{\tau}_{jk}$. The first inequality results from applying mean value theorem.

We also notice that $\Pr(\mathcal{E} \not\subseteq \hat{\mathcal{E}}_{\gamma_n}) = \Pr(\bigcup_{(j,k) \in \mathcal{E}} |\hat{S}^{\tau}_{jk}| < \gamma_n) \leq \sum_{(j,k) \in \mathcal{E}} \Pr(|\hat{S}^{\tau}_{jk}| < \frac{2}{3} C_1 n^{-\kappa})$. And the number of pairs in edge set $\mathcal{E}$ is $|\mathcal{E}| < p^2$, then the assumption 1 above implies that $\Pr(|\hat{S}^{\tau}_{jk}| < \frac{2}{3} C_1 n^{-\kappa}) \leq \Pr(|\hat{S}^{\tau}_{jk} - \rho_{jk}| \geq \frac{1}{3} C_1 n^{-\kappa})$.

Therefore, we have concluded that

$$\Pr(\mathcal{E} \not\subseteq \hat{\mathcal{E}}_{\gamma_n}) \leq \sum_{(j,k)\in\mathcal{E}} \Pr(|\hat{S}^{\tau}_{jk}| < \frac{2}{3}C_1 n^{-\kappa}) \leq \sum_{(j,k)\in\mathcal{E}} \Pr(|\hat{S}^{\tau}_{jk} - \rho_{jk}| \geq \frac{1}{3}C_1 n^{-\kappa})$$
$$< 2p^2 \cdot \exp(-\frac{1}{9\pi^2}C_1^2 n^{1-2\kappa})$$

So that $\Pr(\mathcal{E} \subseteq \hat{\mathcal{E}}_{\gamma_n}) \geq 1 - 2p^2 \cdot \exp(-\frac{1}{9\pi^2}C_1^2 n^{1-2\kappa})$. And similarly, we can prove that $\Pr(\mathcal{E}_j \subseteq \hat{\mathcal{E}}_{j,\gamma_n}) \geq 1 - 2p^2 \cdot \exp(-\frac{1}{9\pi^2}C_1^2 n^{1-2\kappa})$.

6.3 Proof of Theorem 2

Let

$$L_j = \{k : k \neq j, |\rho_{jk}| \geq \frac{1}{3}C_1 n^{-\kappa}\}$$

and

$$\Gamma_{j,\gamma_n} = \cap_{k:k\neq j}\{|\hat{S}^{\tau}_{jk} - \rho_{jk}| \leq \frac{1}{3}C_1 n^{-\kappa}\}$$

By definition, $\hat{\mathcal{E}}_{j,\gamma_n} = \{k : k \neq j, |\hat{S}^{\tau}_{jk}| > \frac{2}{3}C_1 n^{-\kappa}\}$. Then on the set Γ_{j,γ_n}, if k belongs to $\hat{\mathcal{E}}_{j,\gamma_n}$, it has to belong to L_j. Thus, we conclude that $\Pr(\hat{\mathcal{E}}_{j,\gamma_n} \subseteq L_j) \geq \Pr(\Gamma_{j,\gamma_n})$. Then an argument similar to that in the proof of Theorem 1 can be used to show that

$$\Pr(\Gamma_{j,\gamma_n}) \geq 1 - C_4 \exp(-C_5 n^{1-2\kappa}).$$

This implies that

$$\Pr(\hat{\mathcal{E}}_{j,\gamma_n} \subseteq L_j) \geq 1 - C_4 \exp(-C_5 n^{1-2\kappa}).$$

Define $S = \sum_{k\in L_j} \rho_{jk}^2$, we have $S \geq \frac{1}{9}C_1^2|L_j|n^{-2\kappa}$ on one hand by the definition of L_j. On the other hand, $S \leq \sum_k \rho_{jk}^2 = ||\Sigma e_j||_2^2$, where e_j is the vector with the jth entry equal to 1, and others 0. Further we have

$$S \leq ||\Sigma e_j||_2^2 \leq \lambda_{\max}(\Sigma) e_j^T \Sigma e_j = \lambda_{\max}(\Sigma),$$

the last equality results from the fact that the diagonal elements of Σ is equal to 1.

Therefore, the cardinality of L_j, $|L_j|$, is less than or equal to $\lambda_{max}(\Sigma)/(\frac{1}{9}C_1^2 n^{-2\kappa}) = 9C_1^{-2}n^{2\kappa}\lambda_{max}(\Sigma)$. This in conjunction with Assumption 2 yields the desired result stated in Theorem 2.

6.4 Proof of Theorem 3

We first show that the assumptions of Theorem 1 are satisfied so that the sure screening property holds for choosing $\gamma_{n,jk} \equiv \gamma_n = \frac{\pi}{2}\sigma_0\Phi^{-1}(1 - \frac{f}{p(p-1)})/\sqrt{n}$. It is equivalent to show that

$$\gamma_{n,jk} \equiv \gamma_n = \frac{\pi}{2}\sigma_0\Phi^{-1}(1 - \frac{f}{p(p-1)})/\sqrt{n} \leq \frac{2}{3}C_1 n^{-\kappa} \tag{6.2}$$

which is equivalent to show

$$\frac{f}{p(p-1)} \geq 1 - \Phi(\frac{4}{3\pi\sigma_0}C_1 n^{\frac{1}{2}-\kappa}). \tag{6.3}$$

That is, from the fact that $1 - \Phi(x) \leq \frac{1}{\sqrt{3\pi}}x^{-1}\exp(-x^2/2)$, we have

$$1 - \Phi(\frac{4}{3\pi\sigma_0}C_1 n^{\frac{1}{2}-\kappa}) \leq C_7 n^{-\frac{1}{2}+\kappa}\exp(-C_8 n^{1-2\kappa}) \tag{6.4}$$

Since $\log(p) = C_3 n^{\xi}$, we have that $f/\{p(p-1)\} \geq C_9\exp(-C_{10}n^{\xi})$. Using the fact that $\xi < 1 - 2\kappa$, we can prove that the above inequality holds. Therefore, $\gamma_{n,jk} \equiv \gamma_n$ is no greater than the threshold in Theorem 1, $\frac{2}{3}C_1 n^{-\kappa}$. Then with the threshold value defined in Theorem 3, sure screening property holds.

Next, we show the control of false positive rate. Recall that the false positive rate is defined as

$$FPR_n = \frac{1}{|\mathcal{E}^C|}\sum_{(j,k)\notin\mathcal{E}} 1(|\hat{S}^{\tau}_{jk}| > \gamma_n) = \frac{1}{\mathcal{E}^C}\sum_{(j,k)\notin\mathcal{E}} 1(|\sin(\frac{\pi}{2}\hat{\tau}_{jk})| > \gamma_n).$$

Recall that our Assumption 3 guarantees that when n approaches infinity, $\max_{(j,k)\notin\mathcal{E}}|\rho_{jk}|$ approaches to 0. Following from Lemma 2 in Section 2 and a direct use of delta method, $\sqrt{n}(\hat{S}^{\tau}_{jk}-\rho_{jk}) = \sqrt{n}(\sin(\frac{\pi}{2}\hat{\tau}_{jk})-\sin(\frac{\pi}{2}\tau_{jk}))$ is asymptotically normal with mean zero and variance $\frac{\pi^2}{4}\sigma^2_{\tau_{jk}}\cos^2(\frac{\pi}{2}\tau_{jk})$, where $\sigma^2_{\tau_{jk}}$ is defined in (2.3).

Now we want to show that $\sigma^2_{\tau_{jk}} \to \sigma^2_0$ as $\tau_{jk}\to 0$ or equivalently, as $\rho_{jk}\to 0$. After showing this, we can have

$$\sqrt{n}(\hat{S}^{\tau}_{jk}-\rho_{jk}) \xrightarrow{d} N(0,(\frac{\pi}{2})^2\sigma_0^2\cos^2(\frac{\pi}{2}\tau_{jk})) = N(0,(\frac{\pi\sigma_0}{2})^2(1-\rho^2_{jk})), \quad n\to\infty. \tag{6.5}$$

We write matrix A in (2.1) as $A=(A_1^T,\dots,A_p^T)^T$. By the definition of transelliptical distribution, the latent variable $Z_{ij}=f_j(X_{ij})$ and $Z_{ik}=f_k(X_{ik})$ can be presented as

$$Z_{ij}=rA_j^TU, \quad Z_{ik}=rA_k^TU,$$

for $i=1,...,n$. Note that

$$\begin{pmatrix} A_j^TA_j & A_j^TA_k \\ A_k^TA_j & A_k^TA_k \end{pmatrix} = \begin{pmatrix} 1 & \rho_{jk} \\ \rho_{jk} & 1 \end{pmatrix},$$

so there exists an orthogonal matrix H such that $H^TA_j=(\sqrt{1-\rho^2_{jk}},\rho_{jk},0,\dots,0)^T$ and $H^TA_k=(0,1,0,\dots,0)$. Rewrite $U=(U_1,\dots,U_p)$, it follows from Fang et al. (1990) that $H^TU\overset{d}{=}U$. Therefore, for $i=1,...,n$

$$Z_{ij}=r(H^TA_j)^TH^TU\overset{d}{=}\sqrt{1-\rho^2_{jk}}rU_1+\rho_{jk}rU_2=Y_j, \quad Z_{ik}=r(H^TA_k)^TH^TU\overset{d}{=}rU_2=Y_k \tag{6.6}$$

Since the transformation functions f_j, f_k are monotone, (2.3) is equivalent to

$$\begin{aligned}\sigma^2_{\tau_{jk}} &= 4\mathrm{Var}[\mathbb{E}[\mathrm{sign}(Z_{1j}-Z_{2j})\mathrm{sign}(Z_{1k}-Z_{2k})|Z_{1j},Z_{1k}]] \\ &= 4\mathbb{E}[\mathbb{E}[\mathrm{sign}(Z_{1j}-Z_{2j})\mathrm{sign}(Z_{1k}-Z_{2k})|Z_{1j},Z_{1k}]]^2-4\tau^2_{jk}.\end{aligned}$$

So we want to show that as $\rho_{jk} \to 0$,

$$4\mathbb{E}[\mathbb{E}[\text{sign}(Z_{1j} - Z_{2j})\text{sign}(Z_{1k} - Z_{2k})|Z_{1j}, Z_{1k}]]^2 \to \sigma_0^2. \tag{6.7}$$

Denote F to be the cdf of r, it follows by (6.6) that (6.7) is equivalent to

$$\begin{aligned} &\frac{1}{|S|}\int_0^\infty \int_S \left(2\Pr[(Y_j - \sqrt{1-\rho_{jk}^2}r'x_1 - \rho_{jk}r'x_2)(Y_k - r'x_2) > 0] - 1\right)^2 dSdF(r') \qquad (6.8)\\ \to\ &\frac{1}{|S|}\int_0^\infty \int_S \left(2\Pr[(rU_1 - r'x_1)(rU_2 - r'x_2) > 0] - 1\right)^2 dSdF(r'), \end{aligned}$$

as $\rho_{jk} \to 0$, where S refers to the p-dimensional spherical region $x_1^2 + x_2^2 + \cdots + x_p^2 = 1$, $|S|$ the surface area of S. As $\rho_{jk} \to 0$, denote $\xrightarrow{P}$ as convergence in probability, we have

$$(Y_j, Y_k) \xrightarrow{P} (rU_1, rU_2).$$

By Assumption 4, the joint cdf of (rU_1, rU_2) is continuous, therefore, for fixed r', x_1 and x_2, we have

$$\begin{aligned} \Pr(Y_j > \sqrt{1-\rho_{jk}^2}r'x_1 + \rho_{jk}r'x_2, Y_k > r'x_2) &\to \Pr(rU_1 > r'x_1, rU_2 > r'x_2),\\ \Pr(Y_j < \sqrt{1-\rho_{jk}^2}r'x_1 + \rho_{jk}r'x_2, Y_k < r'x_2) &\to \Pr(rU_1 < r'x_1, rU_2 < r'x_2), \end{aligned}$$

which implies

$$\begin{aligned} &\left(2\Pr[(Y_j - \sqrt{1-\rho_{jk}^2}r'x_1 - \rho_{jk}r'x_2)(Y_k - r'x_2) > 0] - 1\right)^2\\ &\to (2\Pr[(rU_1 - r'x_1)(rU_2 - r'x_2) > 0] - 1)^2, \end{aligned}$$

pointwisely. Note that the integrand in (6.8) is bounded, it follows from Dominated Convergence Theorem that

$$\begin{aligned}
&\int_0^M \int_S \Big(2\Pr[(Y_j - \sqrt{1-\rho_{jk}^2}r'x_1 - \rho_{jk}r'x_2)(Y_k - r'x_2) > 0] - 1\Big)^2 dSdF(r') \quad (6.9)\\
\to\quad &\int_0^M \int_S (2\Pr[(rU_1 - r'x_1)(rU_2 - r'x_2) > 0] - 1)^2 dSdF(r'),
\end{aligned}$$

for any fixed $M > 0$. For any arbitrarily small ε, there exists an M_0 such that $1-F(M_0) < \varepsilon$, together with the fact that

$$\int_S \Pr[(Y_j - \sqrt{1-\rho_{jk}^2}r'x_1 - \rho_{jk}r'x_2)(Y_k - r'x_2) > 0]dS < |S|$$

and

$$\int_S \Pr[(rU_1 - r'x_1)(rU_2 - r'x_2) > 0]dS < |S|,$$

for all r', we have

$$\begin{aligned}
&|\frac{1}{|S|}\int_{M_0}^\infty \int_S \Big(2\Pr[(Y_j - \sqrt{1-\rho_{jk}^2}r'x_1 - \rho_{jk}r'x_2)(Y_k - r'x_2) > 0] - 1\Big)^2 dSdF(r')\\
&-\frac{1}{|S|}\int_{M_0}^\infty \int_S (2\Pr[(rU_1 - r'x_1)(rU_2 - r'x_2) > 0] - 1)^2 dSdF(r')| < 2\varepsilon,
\end{aligned}$$

for arbitrarily small ε. Combining this together with (6.9) proves (6.8). Hence (6.5) is satisfied. Therefore, for any $(j,k) \notin \mathcal{E}$ and $\gamma_{n,jk} \equiv \gamma_n = \frac{\pi}{2}\sigma_0\Phi^{-1}(1-\frac{f}{p(p-1)})/\sqrt{n} \le \frac{2}{3}C_1 n^{-\kappa}$,

$$\begin{aligned}
\Pr(|\hat{S}_{jk}^\tau| > \gamma_{n,jk}) &= \Pr(\frac{\sqrt{n}(\hat{S}_{jk}^\tau - \rho_{jk})}{\frac{\pi\sigma_0}{2}\sqrt{1-\rho_{jk}^2}} > \frac{\sqrt{n}(\gamma_n - \rho_{jk})}{\frac{\pi\sigma_0}{2}\sqrt{1-\rho_{jk}^2}})\\
&\quad + \Pr(\frac{\sqrt{n}(\hat{S}_{jk}^\tau - \rho_{jk})}{\frac{\pi\sigma_0}{2}\sqrt{1-\rho_{jk}^2}} < -\frac{\sqrt{n}(\gamma_n + \rho_{jk})}{\frac{\pi\sigma_0}{2}\sqrt{1-\rho_{jk}^2}})\\
&= 1 - \Phi(\frac{\sqrt{n}(\gamma_n - \rho_{jk})}{\frac{\pi\sigma_0}{2}\sqrt{1-\rho_{jk}^2}}) + 1 - \Phi(\frac{\sqrt{n}(\gamma_n + \rho_{jk})}{\frac{\pi\sigma_0}{2}\sqrt{1-\rho_{jk}^2}})\\
&\asymp 2 - 2\Phi(\frac{\sqrt{n}\gamma_n}{\frac{\pi\sigma_0}{2}}).
\end{aligned}$$

The asymptotic equivalence in the last row is given by the fact that $\sqrt{n}\gamma_n$ is of the same order as $n^{\frac{\xi}{2}}$, combined with Assumption 3 implying that $(1-\rho_{jk}^2) \asymp 1$ asymptotically.

Consequently, the expectation for FPR_n is controlled as desired,

$$\begin{aligned} E(FPR_n) &= \frac{1}{\mathcal{E}^C} \sum_{(j,k)\notin\mathcal{E}} \Pr(|\hat{S}^{\tau}_{jk}| > \gamma_{n,jk}) \\ &\asymp \frac{\sum_{(j,k)\notin\mathcal{E}}[2-2\Phi(\frac{\sqrt{n}\gamma_n}{\frac{\pi\sigma_0}{2}})]}{|\mathcal{E}^C|} = 2f/(p(p-1)) \leq f/|\mathcal{E}^C|, \end{aligned}$$

where the last inequality is given by the fact that $|\mathcal{E}^C| \leq \frac{p(p-1)}{2}$.

6.5 Proof of the Corollary

To show the sure screening property holds, it is enough to show that any $\gamma_{n,jk}$ satisfies

$$\gamma_{n,jk} = \frac{\pi}{2}\hat{\sigma}_{\tau_{jk}}\Phi^{-1}(1-\frac{f}{p(p-1)})/\sqrt{n} \leq \frac{2}{3}C_1 n^{-\kappa} \tag{6.10}$$

Replacing σ_0 in (6.3) and (6.4) by $\hat{\sigma}_{\tau_{jk}}$, the same arguments beneath the equation (6.4) hold, and thus the sure screening property holds here.

Next, we show that using these $\gamma_{n,jk}$ leads to the control of the asymptotic expected false positive rate at $\frac{f}{p(1-p)}$. The following lemma is used here.

Lemma: Consider two random variables X_j and X_k and each random variable has n i.i.d observations. We have that $\sqrt{n}(\hat{\tau}_{jk}-\tau_{jk})$ is asymptotically normally distributed with mean equal to 0 and variance consistently estimated by $\hat{\sigma}^2_{\tau_{jk}}$, where

$$\hat{\sigma}^2_{\tau_{jk}} = \frac{4(n-1)}{(n-2)^2}\sum_{i'=1}^{n}\left[\left\{\frac{1}{n-1}\sum_{i=1,i\neq i'}^{n} \mathrm{sign}((X_{ij}-X_{i'j})(X_{ik}-X_{i'k}))\right\} - \hat{\tau}_{jk}\right]^2 \tag{6.11}$$

and $\hat{\tau}_{jk} = \frac{2}{n(n-1)}\sum_{1\leq i \leq i' \leq n}\mathrm{sign}((X_{ij}-X_{i'j})(X_{ik}-X_{i'k}))$. This lemma follows from Arvesen (1969).

A direct use of delta method shows $\sqrt{n}[\sin(\frac{\pi}{2}\hat{\tau}_{jk}) - \sin(\frac{\pi}{2}\tau_{jk})] = \sqrt{n}[\hat{S}^{\tau}_{jk} - \rho_{jk}]$ is also asymptotically normal with mean zero and variance consistently estimated by $\frac{\pi^2}{4}\hat{\sigma}^2_{\tau_{jk}}(1-\rho^2_{jk})$.

Therefore, for any $(j,k) \notin \mathcal{E}$, we have

$$\begin{aligned}
\Pr(|\hat{S}^{\tau}_{jk}| > \gamma_{n,jk}) &= \Pr(\frac{\sqrt{n}(\hat{S}^{\tau}_{jk} - \rho_{jk})}{\frac{\pi\hat{\sigma}_{\tau_{jk}}}{2}\sqrt{1-\rho^2_{jk}}} > \frac{\sqrt{n}(\gamma_{n,jk} - \rho_{jk})}{\frac{\pi\hat{\sigma}_{\tau_{jk}}}{2}\sqrt{1-\rho^2_{jk}}}) \\
&\quad + \Pr(\frac{\sqrt{n}(\hat{S}^{\tau}_{jk} - \rho_{jk})}{\frac{\pi\hat{\sigma}_{\tau_{jk}}}{2}\sqrt{1-\rho^2_{jk}}} < -\frac{\sqrt{n}(\gamma_{n,jk} + \rho_{jk})}{\frac{\pi\hat{\sigma}_{\tau_{jk}}}{2}\sqrt{1-\rho^2_{jk}}}) \\
&= 1 - \Phi(\frac{\sqrt{n}(\gamma_{n,jk} - \rho_{jk})}{\frac{\pi\hat{\sigma}_{\tau_{jk}}}{2}\sqrt{1-\rho^2_{jk}}}) + 1 - \Phi(\frac{\sqrt{n}(\gamma_{n,jk} + \rho_{jk})}{\frac{\pi\hat{\sigma}_{\tau_{jk}}}{2}\sqrt{1-\rho^2_{jk}}}) \\
&\asymp 2 - 2\Phi(\frac{\sqrt{n}\gamma_{n,jk}}{\frac{\pi\hat{\sigma}_{\tau_{jk}}}{2}}).
\end{aligned}$$

The asymptotic equivalence in the last row is given by the fact that $\sqrt{n}\gamma_{n,jk}$ is of the same order as $n^{\frac{\xi}{2}}$, combined with Assumption 3 implying that $(1-\rho^2_{jk}) \asymp 1$ asymptotically.

Consequently, the expectation for FPR_n is controlled as desired,

$$\begin{aligned}
E(FPR_n) &= \frac{1}{\mathcal{E}^C}\sum_{(j,k)\notin\mathcal{E}} \Pr(|\hat{S}^{\tau}_{jk}| > \gamma_{n,jk}) \\
&\asymp \frac{\sum_{(j,k)\notin\mathcal{E}}[2 - 2\Phi(\frac{\sqrt{n}\gamma_{n,jk}}{\frac{\pi\hat{\sigma}_{\tau_{jk}}}{2}})]}{|\mathcal{E}^C|} = 2f/(p(p-1)) \le f/|\mathcal{E}^C|,
\end{aligned}$$

where the last inequality is given by the fact that $|\mathcal{E}^C| \le \frac{p(p-1)}{2}$.

References

Anandkumar, A., Tan, V. Y. F., Huang, F., and Willsky, A. S. (2012). High-dimensional Structure Estimation in Ising Models: Local Separation Criterion. *The Annals of Statistics*, 40:1346–1375.

Arvesen, J. N. (1969). Jackknifing U-Statistics. *The Annals of Mathematical Statistics*, 40:2076–2100.

Callaert, H. and Veraverbeke, N. (1981). The Order of the Normal Approximation for a Studentized U-Statistic. *The Annals of Statistics*, 9:194–200.

Dengler, B. (2010). *On the Asymptotic Behaviour of the Estimator of Kendall's Tau*. PhD thesis, Vienna University of Technology.

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

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

Fligner, M. A. and Rust, S. W. (1983). On the independence problem and Kendall's tau. *Communications in Statistics - Theory and Methods*, 12:1597–1607.

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

Hoeffding, W. (1963). Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58:13–30.

Lauritzen, S. L. (1996). *Graphical Models*. Oxford University Press.

Lee, A. J. (1985). On estimating the variance of a U-statistic. *Communications in Statistics - Theory and Methods*, 14:289–302.

Liu, H., Han, F., Yuan, M., Lafferty, J., and Wasserman, L. (2012a). High-dimensional semiparametric Gaussian copula graphical models. *The Annals of Statistics*, 40:2293–2326.

Liu, H., Han, F., and Zhang, C. (2012b). Transelliptical graphical models. In *Advances in Neural Information Processing Systems 25*, pages 809–817.

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.

Luo, S., Song, R., and Witten, D. M. (2014). Sure Screening for Gaussian Graphical Models. *to appear*.

Mardia, K. V., Kent, J. T., and Bibby, J. M. (1980). *Multivariate Analysis*. Academic Press.

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

Ravikumar, P., Lafferty, J., Liu, H., and Wasserman, L. (2009). Sparse additive models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71:1009–1030.

Ravikumar, P., Wainwright, M. J., and Lafferty, J. D. (2010). High-dimensional Ising Model Selection Using L1-Regularized Logistic Regression. *The Annals of Statistics*, 38:1287–1319.

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

Sen, P. K. (1977). Some Invariance Principles Relating to Jackknifing and Their Role in Sequential Analysis. *The Annals of Statistics*, 5:316–329.

Vaart, A. W. v. d. (2000). *Asymptotic statistics*. Cambridge series in statistical and probabilistic mathematics. Cambridge University Press.

Yang, E., Allen, G., Liu, Z., and Ravikumar, P. (2012). Graphical models via generalized linear models. In *Advances in Neural Information Processing Systems 25*, pages 1367–1375.

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