Robust Estimation of High-Dimensional Graphical Models under Total Positivity

Yuhao Wang, Caroline Uhler

March 15, 2019

Abstract

We consider the problem of estimating high-dimensional graphical models for distributions that are multivariate totally positive of order two (MTP_2). Such distributions are for example of interest in financial econometrics for the analysis of stock prices, which often show positive dependence. We here propose a regularized maximum likelihood estimator (MLE) for MTP_2 distributions and establish graphical structure recovery guarantees (*sparsistency*) of the proposed estimator in the high-dimensional setting without requiring the incoherence assumption. Instead we require a strong form of *diagonal dominance*, which is often naturally satisfied by high-dimensional MTP_2 distributions. Interestingly, we prove that the graphical structure recovery guarantees are robust with respect to the choice of the tuning parameter and that the MLE of the graphical model is decreasing for increasing tuning parameter. In particular, we prove that the maximum weight spanning forest of the sample correlation matrix is a subgraph of the true graphical model for high-dimensional MTP_2 distributions. Finally, we provide corollaries of our work to the estimation of high-dimensional graphical models without the MTP_2 constraint and analyze the performance of the proposed estimator in simulations and on stock pricing data.

1 Introduction

Consistent support recovery of precision matrices has been a central topic in high-dimensional estimation and a number of methods have been proposed for this purpose. One of the most prominent is ℓ_1-regularized maximum likelihood estimation [37], also known as *graphical lasso*. Given the sample correlation matrix S as well as a tuning parameter $\lambda > 0$, graphical lasso estimates the precision matrix by solving the following optimization problem:

$$\underset{K}{\text{maximize}} \ \log\det(K) - \text{trace}(KS) - \lambda \sum_{i \neq j} |K_{ij}|. \tag{1}$$

Theoretical analysis [30] shows that the estimated precision matrix is *sparsistent* with respect to the true precision matrix, meaning: (a) the estimate converges in ℓ_∞ norm to the true precision matrix K^* as the number of nodes p and data size n goes to infinity; and (b) the support of the estimated precision matrix is the same as the support of K^*. These sparsistency guarantees of graphical lasso require strong assumptions, including the *incoherence assumption*, which cannot be verified from the data. To relax the incoherence assumption, [12, 19, 24] proposed replacing the ℓ_1-regularization by non-convex penalties; CLIME [7] and ACLIME [8] focused on estimating precision matrices via moment matching. However, all these methods require a careful selection of the tuning parameter for consistent estimation, which has a significant impact on the performance of these methods in practice. In particular, [14] showed that when $p \gg n$, minor changes

of λ could lead to the appearance or disappearance of up to $\frac{(p-n)(p-n+1)}{2}+1$ edges in the graph estimated from graphical lasso, thereby indicating that graphical lasso is in general sensitive with respect to the choice of the tuning parameter. It is therefore of interest to propose consistent estimators that are robust with respect to parameter tuning. However, work on this problem is very limited; related studies include the connection between graphical lasso and the robust thresholding estimator [14, 33].

All of the above methods are for general high-dimensional graphical models. In this paper, we instead focus on the problem of estimating precision matrices under the constraint that the distribution is *multivariate totally positive of order 2* (MTP_2). MTP_2 is the strongest form of positive dependence and implies that all partial correlations are non-negative. Such distribution constraint has been used in many applications ranging from finance and psychology (more details will be given in Section 2). MTP_2 was first studied by [4, 18] and by [11] in the context of graphical models. Our main contributions in this paper are as follows.

- We have proposed a new constrained maximum likelihood based estimator for learning the precision matrices of high-dimensional MTP_2 distributions. Our new estimator is *sparsistent* in estimating high-dimensional Gaussian and transelliptical graphical models under an assumption that seems more intuitive than the incoherence condition. Moreover, our simulation analysis shows that such assumption is usually naturally satisfied for high-dimensional MTP_2 distributions;

- Our estimator is robust with respect to the choice of tuning parameters. As a consequence, even if we cannot discover the correct tuning parameter, our estimator is still guaranteed to recover a subset of edges in the underlying graphical model, which is usually not satisfied by the previous estimators;

- As a corollary, we developed a new estimator that is provably consistent in estimating graphical models for general high-dimensional distributions. Moreover, the robustness guarantee still holds for this new estimator;

- We evaluated our new method on simulation and real data.

2 Problem setup and notation

A distribution with density function f w.r.t. a product measure μ on $\mathbb{R}^p$ is *multivariate totally positive of order 2* (MTP_2) if it satisfies

$$f(x)f(y) \leq f(x \wedge y)f(x \vee y) \qquad \text{for all } x, y \in \mathbb{R}^p,$$

where $\wedge$ and $\vee$ denote the elementwise maximum and minimum, respectively. A Gaussian distribution is MTP_2 if and only if the underlying precision matrix K^* is an M-matrix, i.e., $(K^*)_{ij} \leq 0$ for all $i \neq j$, or equivalently, all partial correlations are non-negative [4, 15].

MTP_2 distributions have for example been applied in psychology as an alternative to single factor model [20], and in economics for auction [16] and portfolio management [29]. Given that MTP_2 is a strong form of positive dependence, another natural application is to modeling global stock prices, which are known to be strongly positively correlated. For example, consider the following 2016 monthly correla-

tions of global stock markets[1]:

$$S = \begin{array}{c} \\ \text{Nasdaq} \\ \text{Canada} \\ \text{Europe} \\ \text{UK} \\ \text{Australia} \end{array} \begin{array}{c} \begin{array}{ccccc} \text{Nasdaq} & \text{Canada} & \text{Europe} & \text{UK} & \text{Australia} \end{array} \\ \begin{pmatrix} 1.000 & 0.606 & 0.731 & 0.618 & 0.613 \\ 0.606 & 1.000 & 0.550 & 0.661 & 0.598 \\ 0.731 & 0.550 & 1.000 & 0.644 & 0.569 \\ 0.618 & 0.661 & 0.644 & 1.000 & 0.615 \\ 0.613 & 0.598 & 0.569 & 0.615 & 1.000 \end{pmatrix} \end{array}$$

One can easily check that S^{-1} is an M-matrix. This is quite surprising considering that if you sample a correlation matrix uniformly at random the probability of it being MTP_2 is less than 10^{-6}, thereby suggesting MTP_2 as an interesting constraint for financial econometrics applications to the modelling of stock prices. In [32], Slawski and Hein also applied this constraint to modelling the association of students' performance on different math subjects.

In this paper, we consider the problem of estimating the precision matrix of a Gaussian MTP_2 distribution in the high-dimensional setting. Our new estimator is a *regularized* maximum likelihood estimator (MLE) with MTP_2 constraint. Preliminary results in [32, 20] show that adding MTP_2 constraint to MLE can provide us stronger results than without exploiting such additional structure (for example, [20] shows that the MLE with MTP_2 constraint exists whenever $n > 2$, while without MTP_2 constraint the existence of MLE requires that $n > p$). However, a major limitation of these research is that both MLE and MLE with MTP_2 constraint are not provably consistent in the high-dimensional regime. It is therefore of interest to understand if adding MTP_2 constraint into the other estimators that are provably consistent in the high-dimensional setting can still give us stronger results. In this paper, we focus on adding MTP_2 constraint to the *regularized* maximum likelihood estimator, which is provably consistent in the high-dimensional setting. Our new estimator is as follows.

$$\begin{aligned} \hat{K}^{\lambda} := \underset{K}{\text{argmax}} \ \ & \log\det(K) - \text{trace}(KS) + \lambda \sum_{i,j} K_{ij} \\ \textit{s.t. } & K_{ij} \leq 0 \quad \forall i \neq j \end{aligned} \tag{2}$$

(the constraint that K is an M-matrix corresponds to MTP_2 in the Gaussian setting). Just as graphical lasso, our estimator is also a maximum likelihood based estimator with ℓ_1 regularization. However, compared with graphical lasso, one key advantage of our estimator is that the consistency guarantee does not require the rather unintuitive incoherence assumption. Instead, it requires that the underlying precision matrix K^* satisfies *strictly diagonal dominance* (SDD):

Definition 2.1. Let $\gamma \geq 0$; a symmetric matrix $A \in \mathbb{R}^{p \times p}$ is γ-strictly diagonally dominant if for all $i \in [p]$,

$$|A_{ii}| > \gamma \sum_{j \neq i} |A_{ij}|.$$

Slightly abusing notation, we say that a Gaussian distribution satisfies the γ-SDD assumption if its underlying precision matrix is γ-SDD. Strictly diagonally dominant precision matrices have also been studied in [25, 36]. For example, in [36], the authors show that SDD with respect to the underlying precision matrix is a sufficient condition for the convergence of loopy belief propagation. In addition, the SDD assumption is also an interpretable condition in many real world applications. For example, in electrical circuits, K^*_{ij}

[1] taken from investmentfrontier.com

corresponds to the strength of the capacitance of the line connecting nodes i and j [34], then intuitively the γ-SDD assumption means that for any node in the circuit, the total capacitance strength between this node and all other nodes is weak.

When the SDD assumption with respect to K^* is not satisfied, we also prove that the sparsistency guarantees still hold by instead imposing assumptions on the *inverse Isserlis matrix* of the underlying distribution. The Isserlis matrix H^* is a $p^2 \times p^2$ matrix where the rows and columns are indexed by node pairs, such that each entry in the Isserlis matrix $H^*_{(i,j),(\ell,m)}$ corresponds to the covariance between the random variables X_iX_j and $X_\ell X_m$, i.e., $H^*_{(i,j),(\ell,m)} := \text{cov}(X_iX_j, X_\ell X_m)$. Using standard results in matrix derivatives and exponential families [6, 5], we also have that $H^* = \Sigma^* \otimes \Sigma^*$, where $\otimes$ denote the Kronecker matrix product and Σ^* is the true underlying correlation matrix. In addition, the inverse $\Gamma^* := (H^*)^{-1}$ also satisfies that $\Gamma^* = K^* \otimes K^*$.

In addition to the consistency guarantees, we also give robustness guarantees of $\hat{K}^\lambda$ for support recovery, which is also the *main contribution* of this paper. More specifically, let $\lambda_{n,p}$ denote the tuning parameter that depends on n and p such that our estimator (2) is *sparsistent*, we show that even if we cannot discover $\lambda_{n,p}$, by choosing any $\lambda \geq \lambda_{n,p}$, we always have that $\hat{E}^\lambda \subseteq E^*$, where E^* denotes the support of K^*. In this case, we have provided theoretical guarantees of our estimator not only for a particular choice of tuning parameter $\lambda_{n,p}$, but also for all tuning parameters $\lambda \geq \lambda_{n,p}$. This implies that our estimator is robust with respect to the choice of tuning parameters, which is usually not satisfied by the existing estimators. Such robustness guarantee is of interest for example in estimating graphical models of extremely high-dimensional distributions, where the parameter tuning procedure to correctly specify $\lambda_{n,p}$ is usually computationally infeasible.

Beyond Gaussian distributions, in this paper we also show that our estimator (2) can be applied to elliptical and transelliptical distributions [13, 23]. Elliptical distribution is a generalization of Gaussian distribution that allows for heavy-tailed dependence between random variables. The transellitpical distribution is a nonparametric distribution family that combines elliptical distribution and marginal transformations. Theoretical guarantees under such distribution families allow us to apply our estimator to a wider range of applications where Gaussianity does not hold, such as the stock daily changes [23]. We end this section with the formal definitions of elliptical and transelliptical distributions, respectively.

Definition 2.2 (elliptical distribution [13])**.** Let $\mu^* \in \mathbb{R}^p$, $\Sigma^* \in \mathbb{R}^{p\times p}$ with $\text{rank}(\Sigma^*) = q \leq p$. A random vector X follows an elliptical distribution, denoted by $X \sim \text{EC}(\mu^*, \Sigma^*, \xi)$, if and only if X has a stochastic representation $X \overset{d}{=} \mu^* + \xi AU$, where U is a random vector uniformly distributed on the unit sphere in $\mathbb{R}^q$, $\xi \geq 0$ is a scalar random variable independent of U, A is a deterministic matrix with $AA^T = \Sigma^*$.

Note that elliptical distribution does not necessarily have a density. In this paper, we consider a particular form of elliptical distribution where the random variable ξ is absolutely continuous with respect to the Lebesgue measure and Σ^* is non-singular. In this setting, the density exists and has the form

$$p(x) = \det(\Sigma^*)^{1/2} g\big((x-\nu^*)^T(\Sigma^*)^{-1}(x-\nu^*)\big),$$

where $g(\cdot)$ is a scalar function that depends on the distribution of ξ. Hence, we can denote a elliptical distribution by $X \sim \text{EC}(\mu^*, \Sigma^*, g)$.

Definition 2.3 (transelliptical distribution [23])**.** A continuous random vector is transelliptical, denoted by $X \sim \text{TE}(\Sigma^*, \xi; f_1, \cdots, f_p)$, if there exists a set of monotone univariate functions $f_1, \cdots, f_p$ and a nonnegative random variable ξ with $\mathbb{P}(\xi = 0) = 0$ such that

$$(f_1(X_1), \cdots, f_p(X_p))^T \sim \text{EC}(0, \Sigma^*, \xi), \quad \text{where } \text{diag}(\Sigma^*) = \mathbf{1}_p.$$

Here Σ^* is called the latent generalized correlation matrix.

Equivalently, we can denote a transelliptical distribution by $X \sim \text{TE}(\Sigma^*, g; f_1, \cdots, f_p)$. Note that since transelliptical distribution family is a semiparametric model where the underlying parameters are encoded in the latent generalized precision matrix $K^* := (\Sigma^*)^{-1}$ [23], the main goal is to instead estimate the latent generalized precision matrix $K^* := (\Sigma^*)^{-1}$ [23].

3 Consistency guarantees

In this section, we investigate the statistical consistency of our estimator and establish the rates of convergence in the high-dimensional setting. Our main result is given in Theorem 3.1, which considers the statistical consistency of for estimating precision matrices for high-dimensional Gaussian MTP_2 distributions. In Section 3.3, we further extend to transelliptical MTP_2 distributions.

3.1 Statistical consistency of $\hat{K}^\lambda$

In this section, we provide consistency guarantees of $\hat{K}^\lambda$ in the high-dimensional Gaussian setting. Let $M^* := \max_i K^*_{ii}$, then $\|\hat{K}^\lambda - K^*\|_\infty$, i.e., the elementwise maximal difference between $\hat{K}^\lambda$ and K^*, satisfies the following theorem:

Theorem 3.1. *(consistency guarantee) Consider a Gaussian MTP_2 distribution satisfying either of the following two conditions for some $0 < \alpha < 1$:*

- *the inverse Isserlis matrix Γ^* is $\frac{1+\alpha}{1-\alpha}$-SDD;*
- *the precision matrix K^* is $\frac{1+\alpha+\sqrt{2(1+\alpha)}}{1-\alpha}$-SDD.*

Then there exists some positive constants C and τ that depend on α such that by choosing tuning parameter $\lambda_{n,p} := C\sqrt{\frac{\log p}{n}}$, the solution to (2) satisfies that

$$\|\hat{K}^{\lambda_{n,p}} - K^*\|_\infty \leq 8M^{*2}C\sqrt{\frac{\log p}{n}}$$

with probability $1 - p^{-\tau}$. Moreover, the support of $\hat{K}^{\lambda_{n,p}}$ is a subset of the support of K^* and includes all the edges such that $|K^*_{ij}| > 8M^{*2}C\sqrt{\frac{\log p}{n}}$.

The proof is given in the appendix. Since for all $0 < \alpha < 1$, $\frac{1+\alpha}{1-\alpha} > 1$ and $\frac{1+\alpha+\sqrt{2(1+\alpha)}}{1-\alpha} > \sqrt{2}+1$, apparently the minimal requirement for $\hat{K}^\lambda$ to be sparsistent is that K^* is $(\sqrt{2}+1)$-SDD or that Γ^* is 1-SDD.

Remark 3.2. Compared with graphical lasso, the main improvement of our estimator is that the sparsistency guarantee does not require the unintuitive incoherence assumption, but rather the more intuitive γ-SDD assumption. Since incoherence assumption has also been relaxed by many other methods [37, 30, 24, 19, 12, 7, 8, 27], compared to these methods, just by looking at Theorem 3.1 it seems that our estimator does not have too much improvement. However, a common limitation of the previous estimators is that they are usually not robust with respect to the choice of tuning parameters. We fill this gap by giving robustness guarantee in Section 4, which significantly distinguishes our estimator from the previous methods.

The proof of Theorem 3.1 is inspired by the existence of a homeomorphism between the space of canonical parameters and the space of sufficient statistics of the maximum likelihood estimator with zero constraints [1, 6, 35]. In particular, our proof relies on the following lemma, which is of independent interest in matrix algebra:

Lemma 3.3. *Suppose a $p \times p$ symmetric matrix A is γ-SDD for some $\gamma > 1$, then the Schur complement of its first k entries, denoted as $A' \in \mathbb{R}^{k \times k}$, is also γ-SDD. Moreover, the matrix L_1 norm of A', i.e., $\|A'\|_{L_1} := \max_j \sum_i |A'_{ij}|$, satisfies that $\|A'\|_{L_1} \leq \|A\|_{L_1}$.*

The proof is given in the appendix. Lemma 3.3 shows that the γ-SDD matrices are closed under Schur complement. Similar results include [9, 21], where the authors show that the closure property holds for 1-SDD matrices. In this lemma, we further extend the closure property into γ-SDD matrices for any $\gamma \geq 1$.

3.2 γ-SDD assumption for high-dimensional M-matrices

Since our consistency guarantee relies heavily on the γ-SDD assumption, it is important to understand how restrictive the γ-SDD assumption is for MTP_2 distributions. In this section, we perform a simulation study to show that the γ-SDD assumption is usually naturally satisfied for high-dimensional sparse M-matrices. This implies that the γ-SDD assumption is usually not a very restrictive assumption for Gaussian MTP_2 distributions in the high-dimensional setting.

In this simulation, we randomly generated 100 M-matrices with dimension $p = 100$ and analyzed the proportion of randomly generated M-matrices satisfying the γ-SDD assumption. We generated each M-matrix according to the following two steps:

1. Uniformly generate a random correlation matrix S using the uniform sampler from [17];

2. Take the randomly generated S as input to the estimator (2) with $\lambda = 0.19$, then take the $\hat{K}^\lambda$ estimated from (2) as the randomly generated M-matrix.

Note that since the estimate $\hat{K}^\lambda$ is usually sparser as we increase λ, our main purpose of choosing $\lambda = 0.19$ is to control the maximum degree size of the randomly generated M-matrices to be around 7, thereby controlling all randomly generated M-matrices to be sparse matrices.

Figure 1 depicts the proportion of γ-SDD M-matrices among the 100 randomly generated M-matrices for different γ. In particular, for $\gamma = \sqrt{2} + 1$, i.e., the minimal requirement for the sparsistency guarantee, Figure 1 shows that around 40% of the 100 randomly generated M-matrices satisfy the γ-SDD assumption. This result implies that the γ-SDD assumption for $\gamma = \sqrt{2} + 1$ is usually not a very restrictive assumption for high-dimensional sparse M-matrices.

3.3 Extension to transelliptical distributions

So far we have provided consistency guarantees of our estimator for high-dimensional Gaussian MTP_2 distributions. In this section, we further show how to apply our estimator for transelliptical MTP_2 distributions. Before illustrating how our estimator is applied in transelliptical MTP_2 distributions, we first need to briefly review preliminary research in estimating latent generalized precision matrices when the distribution follows a general transelliptical distribution, i.e., the transelliptical distribution without MTP_2 constraint.

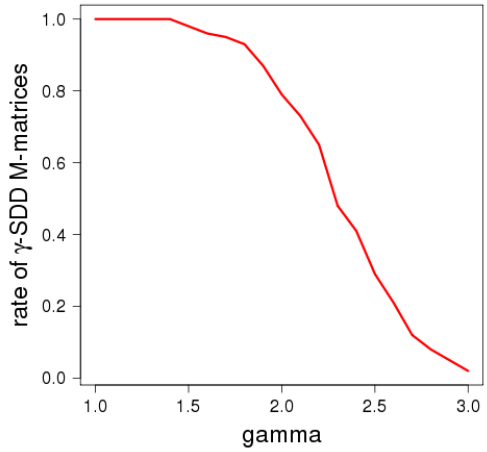


Figure 1: The proportion of γ-SDD M-matrices as a function $\gamma \in [1, 3]$.

3.3.1 Preliminaries

A major challenge of the transelliptical distribution is its heavy tailed property. Due to such heavy tailed property, existing estimators that use sample correlation matrices as data statistics are usually not provably consistent in the high-dimensional setting. To solve this challenge, given i.i.d. samples $x^1, \cdots, x^n \in \mathbb{R}^p$ from a transelliptical distribution, Liu et al. [23] proposed the rank-based data statistics matrix $S^\tau \in \mathbb{R}^{p \times p}$ to better approximate the latent generalized correlation matrix.

$$(S^\tau)_{ij} := \begin{cases} \sin\left(\frac{\pi}{2}\tau_{ij}\right) & i \neq j \\ 1 & i = j \end{cases}$$

where $\sin(\cdot)$ is the sine function, τ_{ij} is the empirical estimate of the kendall's tau statistic, i.e.,

$$\tau_{ij} := \frac{2}{n(n-1)} \sum_{1 \leq k < k' \leq n} \text{sign}(x_i^k - x_i^{k'}) \text{sign}(x_j^k - x_j^{k'}).$$

Based on such new rank-based data statistics, Liu et al. [23] shows that we can consistently estimate the latent generalized precision matrix by taking S^τ as a plug-in to existing estimators, such as graphical lasso or CLIME. For example, when using graphical lasso, instead of solving the original program in (1), we estimate the latent generalized precision matrix by solving the following convex program

$$\underset{K}{\text{maximize}} \ \log\det(K) - \text{trace}(KS^\tau) - \lambda \sum_{i \neq j} |K_{ij}|.$$

3.3.2 Main results

Our main result relies on the following proposition that characterizes the latent generalized precision matrices of transelliptical MTP_2 distributions.

Proposition 3.4. *Consider a transelliptical MTP$_2$ distribution $X \sim TE(\Sigma^*, g; f_1, \cdots, f_p)$. If g is monotonically decreasing and that $f_1, \cdots, f_p$ are differentiable functions with non-zero derivatives, then the latent generalized precision matrix $K^* := (\Sigma^*)^{-1}$ is an M-matrix.*

For a large class of transellitpical distribution, g is monotonically decreasing. Such as p-dimensional Gaussian distribution, where $g(x) = (2\pi)^{p/2}\exp(-x/2)$, and p-dimensional t-distribution with the degree of freedom ν, in which

$$g(x) = c_\nu \frac{\Gamma\left(\frac{\nu+p}{2}\right)}{(\nu\pi)^{\frac{p}{2}}\Gamma\left(\frac{\nu}{2}\right)}\left(1+\frac{c_\nu^2 x}{\nu}\right)^{-\frac{\nu+p}{2}},$$

where c_ν is the normalizing constant and $\Gamma(\cdot)$ is the gamma function. The assumption that the f_i's are differentiable and have non-zero derivatives ensure the smoothness and existence of the transelliptical density. The proof of Proposition 3.4 is given in the appendix. Proposition 3.4 shows that beyond Gaussian MTP$_2$ distributions, the M-matrix constraint still holds for transelliptical MTP$_2$ distributions. This allows us to estimate the latent generalized precision matrices of transelliptical MTP$_2$ distributions by taking S^τ as plug-in to our estimator (2). More specifically, we can estimate the latent generalized precision matrices by solving the following objective function

$$\begin{aligned}\hat{K}^\lambda := \underset{K}{\text{argmax}} \ \log\det(K) - \text{trace}(KS^\tau) + \lambda\sum_{i,j} K_{ij} \\ \text{s.t. } K_{ij} \le 0 \quad \forall i \neq j.\end{aligned}$$

Proposition 3.5. *Consider a transelliptical MTP$_2$ distribution TE$(\Sigma^*, \xi; f_1, \cdots, f_p)$ where $K^* := (\Sigma^*)^{-1}$ satisfies either that K^* is $\frac{1+\alpha+\sqrt{2(1+\alpha)}}{1-\alpha}$-SDD or that $\Gamma^* := K^* \otimes K^*$ is $\frac{1+\alpha}{1-\alpha}$-SDD for some $0 < \alpha < 1$. Then if we take S^τ as plug-in to (2), there exists a tuning parameter $\lambda \asymp \sqrt{\log p/n}$ such that with high probability, the estimate $\hat{K}^\lambda$ satisfies that*

$$\|\hat{K}^\lambda - K^*\|_\infty = \mathcal{O}\left(M^2\sqrt{\frac{\log p}{n}}\right).$$

Moreover, the support of $\hat{K}^\lambda$ is a subset of the support of K^*.

The proof is given in the appendix. The proof follows closely with the proof of Theorem 3.1.

4 Robustness guarantees

In Section 3, we provided statistical consistency of our estimator in the high-dimensional setting. In this section, we focus on showing that our estimator is robust with respect to the choice of tuning parameters, which, as also mentioned in the introduction section, is usually not satisfied by the other estimators.

Our main result is given in Theorem 4.1, which is explained in more detail in Section 4.1. As a corollary, in Section 4.2 we further give a characterization of the maximum weight spanning forest (MWSF) of the inverse of diagonally dominant M-matrices, which can be applied to discover a subset of edges in E^* without any tuning parameter.

4.1 Robustness of $\hat{K}^\lambda$ with respect to tuning parameter

Let $\hat{\Gamma}^\lambda := \hat{K}^\lambda \otimes \hat{K}^\lambda$, our robustness guarantee is as follows:

Theorem 4.1. *(robustness guarantee) Assume for some $\lambda > 0$, the solution to (2) with tuning parameter λ satisfies that $\hat{K}^\lambda$ is $(\sqrt{2}+1)$-SDD or that $\hat{\Gamma}^\lambda$ is SDD. Then for all $\tilde{\lambda}$ such that $\lambda < \tilde{\lambda} < 1$, $\hat{E}^{\tilde{\lambda}} \subseteq \hat{E}^\lambda$. In addition, $\hat{\Gamma}^{\tilde{\lambda}}$ is also 1-SDD (or $\hat{K}^{\tilde{\lambda}}$ is also $(\sqrt{2}+1)$-SDD).*

The proof is given in the appendix. To understand how restrictive the assumption in Theorem 4.1 is, we provide the following proposition that characterizes the diagonal dominance properties of the precision matrices estimated from (2).

Proposition 4.2. *Under the same conditions as Theorem 3.1, we have that $\hat{K}^\lambda$ is $\frac{1+\alpha+\sqrt{2(1+\alpha)}}{1-\alpha}$-SDD (or that $\hat{\Gamma}^\lambda$ is $\frac{1+\alpha}{1-\alpha}$-SDD) for all $\lambda_{n,p} \leq \lambda < 1$.*

Hence, the $(\sqrt{2}+1)$-SDD assumption on $\hat{K}^\lambda$ (or 1-SDD on $\hat{\Gamma}^\lambda$) is naturally satisfied for all $\lambda \geq \lambda_{n,p}$ whenever Theorem 3.1 holds. In other words, our robustness guarantee in Theorem 4.1 does not impose any new conditions. Intuitively, Theorem 4.1 implies that the set of edges is "monotonically decreasing" by increasing the value of λ. In this case, even if we cannot discover the correct tuning parameter $\lambda_{n,p}$ (this usually happens when the graphical model is too large such that parameter tuning is computationally infeasible), just by choosing any $\lambda \geq \lambda_{n,p}$, (2) is always guaranteed to discover a subset of edges in E^*.

Our robustness guarantee is a remarkable improvement compared to the previous estimators in general high-dimensional distributions [37, 30, 24, 19, 12, 7, 8, 27]. Such property has never been proven by these estimators. One result similar to Theorem 4.1 includes [14, Theorem 13]. More specifically, the authors show that, under certain assumptions with respect to S as discussed below, there exist some particular choices of λ such that the graph estimated by graphical lasso with tuning parameter λ is exactly the same as the naive thresholding estimator $\hat{T}^\lambda := \{(i,j) : i \neq j \text{ and } |S_{ij}| \geq \lambda\}$. Since the naive thresholding estimator $\hat{T}^\lambda$ is monotonically decreasing as we increase λ, such equivalence property implies that the estimate from graphical lasso may also be "monotonically decreasing". Nevertheless, compared with Theorem 4.1, [14] has two problems. First, in order to prove that graphical lasso is monotonically decreasing, we not only need to prove that graphical lasso and thresholding estimator have the same estimate for some particular choices of λ, but also that the two estimators are "uniformly equivalent", i.e., they have the same estimate for all possible choices of $\lambda \geq \lambda_{n,p}$. However, in [14], the author already shows that the two estimators are not equivalent for all the λ's where $\lambda = |S_{ij}|$. In this case, they cannot be uniformly equivalent. Second, as also mentioned by [14], the equivalence conditions are usually difficult to hold when S is positive semidefinite. This has limited [14] to be only applicable to the setting where $p < n$.

Since the 1-SDD assumption with respect to $\hat{K}^\lambda$ is much easier to be satisfied than the $(\sqrt{2}+1)$-SDD assumption provided in Theorem 4.1, it is therefore of interest to understand if the robustness guarantee still holds when $\hat{K}^\lambda$ is only 1-SDD. We end this section with the following remark showing that the estimator (2) is in general not robust if $\hat{K}^\lambda$ is only 1-SDD.

Remark 4.3. We use the following example to show that when $\hat{K}^\lambda$ is only 1-SDD, the support recovery from (2) may not be "monotonically decreasing" with the increment of λ. In this example, we consider the

following sample correlation matrix as input to (2):

$$S = \begin{bmatrix} 1 & 0.21 & 0.2113 & 0.3641 & 0.5648 \\ 0.21 & 1 & 0.2835 & 0.2213 & 0.2265 \\ 0.2113 & 0.2835 & 1 & 0.5072 & 0.526 \\ 0.3641 & 0.2213 & 0.5072 & 1 & 0.8162 \\ 0.5648 & 0.2265 & 0.526 & 0.8162 & 1 \end{bmatrix}$$

and set $\lambda = 0$, the estimated $\hat{K}^\lambda$ is as follows:

$$\hat{K} = \begin{bmatrix} 1.4838 & -0.1284 & 0 & 0 & -0.809 \\ -0.1284 & 1.1097 & -0.2429 & -0.0627 & 0 \\ 0 & -0.2429 & 1.4718 & -0.3171 & -0.4603 \\ 0 & -0.0627 & -0.3171 & 3.0759 & -2.3294 \\ -0.809 & 0 & -0.4603 & -2.3294 & 3.6001 \end{bmatrix}$$

By improving λ up to $\lambda = 0.01$, the estimated $\hat{K}^\lambda$ is changed into the following matrix:

$$\begin{bmatrix} 1.4869 & -0.1241 & 0 & 0 & -0.8061 \\ -0.1241 & 1.1143 & -0.2396 & -0.0602 & -0.0002 \\ 0 & -0.2396 & 1.4739 & -0.3154 & -0.459 \\ 0 & -0.0602 & -0.3154 & 3.0772 & -2.3283 \\ -0.8061 & -0.0002 & -0.459 & -2.3283 & 3.597 \end{bmatrix}$$

By comparing the above two matrices, apparently the support of $\hat{K}^\lambda$ is not "monotonically decreasing" with the increment of λ, since the edge $2 - 4$ is added into the estimated undirected graphical model as we increase λ from 0 to 0.01. Hence, The 1-SDD assumption is not sufficient for the robustness guarantee.

4.2 Characterization of MWSF of inverse M-matrices

In Section 4.1, we have provided the robustness guarantee of our estimator. In this section, we further show that applying such robustness guarantee also gives us the following more general result of inverse M-matrices. For any symmetric matrix A, let MWSF(A) denote the set of maximum weight spanning forests (MWSFs) of A. With a slight abuse of notation, we also denote MWSF(A) as the MWSF of A if the MWSF is unique. Then we have the following result.

Theorem 4.4. *Let M be a $(\sqrt{2}+1)$-SDD M-matrix. If MWSF(M^{-1}) is unique, then MWSF(M^{-1}) is a sub graph of the support of matrix M.*

The proof is given in the appendix. A sufficient condition that MWSF(M^{-1}) is unique is that all entries in M^{-1} have distinct values, which is satisfied with probability 1. Let $\delta_{n,p} := \|S - \Sigma^*\|_\infty$, Theorem 4.4 further gives us the following corollary about recovering the edges in E^* using MWSF(S).

Corollary 4.5. *Suppose K^* is an M-matrix satisfying the same conditions as in Theorem 4.4. Then as $\delta_{n,p} \to 0$, MWSF(S) $\subseteq E^*$.*

For high-dimensional Gaussian distributions, $\delta_{n,p} \asymp \sqrt{\frac{\log p}{n}}$. Hence the condition that $\delta_{n,p} \to 0$ is usually naturally satisfied with high probability whenever $n \gg \log p$. In the high-dimensional setting, estimating the MWSF of S is usually more computationally efficient than applying graphical lasso or other convex optimization based estimators. Thus, Corollary 4.5 gives us a new method to discover a subset of edges in E^* in a much more computationally efficient manner. Furthermore, such new method does not rely on any tuning parameter.

5 Comparison of $\hat{K}^\lambda$ to the thresholding estimator

Motivated by the comparison between graphical lasso and the naive thresholding estimator in [34, 33], in this section, we also compare our estimator with the thresholding estimator $\hat{T}^\lambda := \{(i,j) : i \neq j \ \& \ S_{ij} \geq \lambda\}$. Our main result concerning the equivalence between the two estimators is as follows.

Theorem 5.1. *Assume that $\hat{K}^\lambda$ is γ-SDD for some $\gamma > 2$. Then by choosing the threshold*

$$\lambda' = \frac{1}{\gamma - 1} \max_{u \neq v} \left| \frac{(\hat{K}^\lambda)_{uv}}{(\hat{K}^\lambda)_{uu}(\hat{K}^\lambda)_{vv}} \right| + \lambda,$$

we have that $\hat{T}^{\lambda'} \subseteq \hat{E}^\lambda$. Moreover, $\hat{T}^\lambda \neq \emptyset$ and includes all the edge (i,j) such that

$$\left| \frac{(\hat{K}^\lambda)_{ij}}{(\hat{K}^\lambda)_{ii}(\hat{K}^\lambda)_{jj}} \right| \geq \frac{1}{\gamma - 1} \max_{u \neq v} \left| \frac{(\hat{K}^\lambda)_{uv}}{(\hat{K}^\lambda)_{uu}(\hat{K}^\lambda)_{vv}} \right|.$$

The proof is given in the appendix. By applying Proposition 4.2, we can easily conclude that the γ-SDD condition in Theorem 5.1 is naturally satisfied for all $\lambda \geq \lambda_{n,p}$ whenever Theorem 3.1 holds. In this case, by choosing λ big enough, the thresholding estimator $\hat{T}^\lambda$ is always guaranteed to discover a subset of edges in E^* under the setting where estimator (2) is sparsistent.

Unexpectedly, by applying Theorem 5.1 we can also prove that the naive thresholding estimator is still provably consistent in discovering a subset of edges in E^* even under the settings where the estimator (2) is not sparsistent:

Corollary 5.2. *Consider a MTP$_2$ distribution where the precision matrix K^* is γ-SDD for some $\gamma > 2$. Then there exists some positive constants C and τ such that by choosing the tuning parameter*

$$\lambda := \frac{1}{\gamma - 1} \max_{u \neq v} \left| \frac{K^*_{uv}}{K^*_{uu}K^*_{vv}} \right| + C\sqrt{\frac{\log p}{n}},$$

$\hat{T}^\lambda \subseteq E^*$ with probability $1 - p^{-\tau}$. In addition, for all $(i,j) \in E^*$ where

$$\left| \frac{K^*_{ij}}{K^*_{ii}K^*_{jj}} \right| \geq \frac{1}{\gamma - 1} \max_{u \neq v} \left| \frac{K^*_{uv}}{K^*_{uu}K^*_{vv}} \right| + 2C\sqrt{\frac{\log p}{n}}, \tag{3}$$

$(i,j) \in \hat{T}^\lambda$.

The proof is given in the appendix. Apparently, Corollary 5.2 only requires that K^* is 2-SDD, which is a more relaxed condition than Theorem 3.1. Although the naive thresholding estimator has edge recovery guarantees under a condition that is weaker than our estimator (2), it is only able to discover a subset of edges in E^* that satisfy (3). While our estimator is guaranteed to discover all edges in E^* as $\log p/n \to 0$.

6 Extension to general distributions

So far, we have provided a new estimator to estimate the underlying graphical models in Gaussian and transelliptical MTP$_2$ distributions. In this section, we show that we can further extend our results into

general high-dimensional distributions, i.e., distributions without MTP_2 constraints. Our new estimator for estimating precision matrices of high-dimensional Gaussian distributions is as follows.

$$\tilde{K}^{\lambda} := \underset{K}{\text{argmax}} \ \log\det(K) - \text{trace}(K \cdot S) - \lambda \sum_{i \neq j} |K_{ij}| + \lambda \sum_{i} K_{ii}. \tag{4}$$

Just as graphical lasso, (4) also penalizes off-diagonal entries. Nevertheless, for diagonal entries, the new estimator (4) instead rewards the increment of diagonal entries, which is different from graphical lasso. Our sparsistency and robustness guarantees for the new estimator (4) are shown as follows.

Corollary 6.1. *Consider a Gaussian distribution satisfying either of the following two conditions for some* $0 < \alpha < 1$*:*

- *the inverse Isserlis matrix* Γ^* *is* $\frac{1+\alpha}{1-\alpha}$*-SDD;*
- *the precision matrix* K^* *is* $\frac{1+\alpha+\sqrt{2(1+\alpha)}}{1-\alpha}$*-SDD.*

Then there exists some positive constants C *and* τ *that depend on* α *such that by choosing tuning parameter* $\lambda_{n,p} := C\sqrt{\frac{\log p}{n}}$*, the solution to* (2) *satisfies that*

$$\|\hat{K}^{\lambda_{n,p}} - K^*\|_{\infty} \leq 8{M^*}^2 C\sqrt{\frac{\log p}{n}}$$

with probability $1 - p^{-\tau}$*. Moreover, the support of* $\hat{K}^{\lambda_{n,p}}$ *is a subset of the support of* K^* *and includes all the edges such that* $|K^*_{ij}| > 8{M^*}^2 C\sqrt{\frac{\log p}{n}}$*.*

Corollary 6.2. *Assume for some* $\lambda > 0$*, the solution to* (4) *with tuning parameter* λ *satisfies that* $\tilde{K}^{\lambda}$ *is* $(\sqrt{2}+1)$*-SDD or that* $\tilde{\Gamma}^{\lambda}$ *is SDD. Then for all* $\tilde{\lambda}$ *such that* $\lambda < \tilde{\lambda} < 1$*, the support of* $\tilde{K}^{\tilde{\lambda}}$ *is a subset of the support of* $\tilde{K}^{\lambda}$*. In addition,* $\tilde{\Gamma}^{\tilde{\lambda}}$ *is also* 1*-SDD (or* $\tilde{K}^{\tilde{\lambda}}$ *is also* $(\sqrt{2}+1)$*-SDD).*

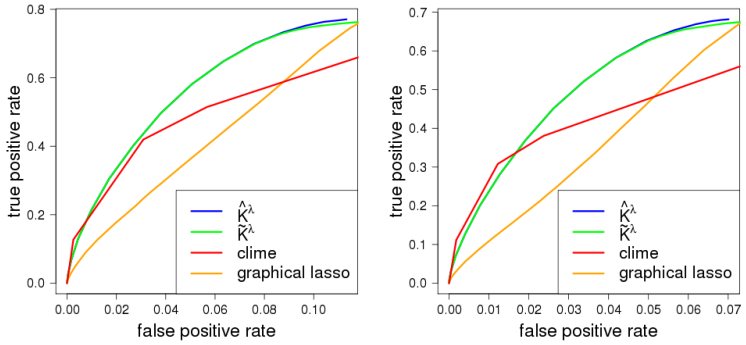
The proofs are given in the appendix. Corollaries 6.1 and 6.2 show that all the theoretical properties for Gaussian MTP_2 distributions still hold for the general Gaussian distributions. This gives us a robust and provably consistent estimator for a broader class of distribution families. Note that in addition to Gaussian distributions, the same result still holds for general transelliptical distributions by taking the rank-based data statistics S^{τ} as a plug-in to (4).

7 Experiments

7.1 Numeric simulations

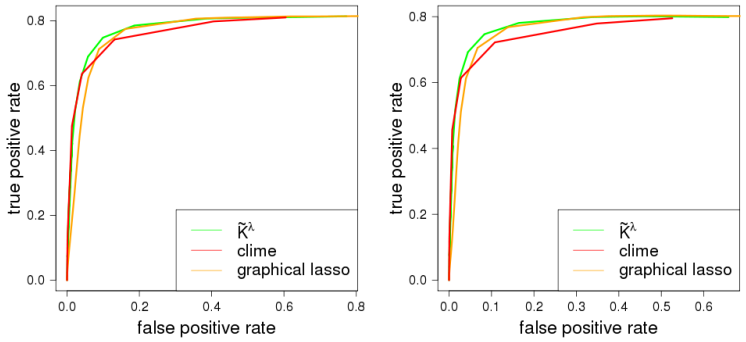
In this section, we perform numerical experiments to compare our estimators with the state of the art methods in both Gaussian MTP_2 distribution setting and general Gaussian distribution setting. For the MTP_2 setting, we consider p-dimensional Gaussian distributions generated according to the following model:

- Rothman model: We generate each Gaussian distribution $\mathcal{N}(\mu_0, \Sigma_0)$ using a model similar to Rothman et al. [31]. In this model, the mean is set to be $\mu_0 = 0$. To generate Σ_0, we let its inverse $K_0 = B + \delta I$, where I is an identity matrix, B is a non-positive matrix where each off-diagonal entry



(a) $p = 100, n = 100$ (b) $p = 200, n = 100$

Figure 2: The averaged ROC curve of support recovery for 100 randomly generated MTP_2 distributions on p nodes with n samples. The expected neighbourhood size of the underlying precision matrices is 5, the condition number is p.



(a) $p = 100, n = 100$ (b) $p = 200, n = 100$

Figure 3: The averaged ROC curve of support recovery for 100 randomly generated general high-dimensional distributions on p nodes with n samples. The expected neighbourhood size of the underlying precision matrices is 5, the condition number is p.

in B is generated independently to be equal to -0.5 with probability $5/(p-1)$ and 0 with probability $1-5/(p-1)$ and each diagonal entry in B is equal to zero. We then choose δ such that the condition number of matrix K_0, i.e., the ratio between the largest eigenvalue and the smallest eigenvalue of K_0, is equal to p. Finally, K_0 is renormalized such that all diagonal entries of K_0 are equal to 1.

The Rothman model is an example of a sparse matrix model with expected neighbourhood size to be equal to 5. Based on this model, we generated 100 realizations of Gaussian distributions with $p = \{100, 200\}$ according to the Rothman model. Then for each distribution, we generated a training sample of size $n = 100$ and applied our estimators $\hat{K}^\lambda$ and $\tilde{K}^\lambda$ and the state of the art graphical lasso [31] and CLIME [7] to estimate the precision matrix using the n samples.

Figure 2 shows the averaged ROC curve in support recovery using different estimators. As expected, $\hat{K}^\lambda$ and $\tilde{K}^\lambda$ significantly outperformed the others. One interesting phenomenon is that the performance of

	$\hat{K}^\lambda$	$\tilde{K}^\lambda$	Graphical lasso
Modularity	0.65	0.65	0.58

Table 1: The modularity score of the graph estimated from different methods. Higher score means better performance in community detection.

$\hat{K}^\lambda$ and $\tilde{K}^\lambda$ are similar, this is also expected since the consistency and the robustness guarantees of the two estimators $\hat{K}^\lambda$ and $\tilde{K}^\lambda$ require similar conditions.

For general distribution setting, we generated simulation data using the same model as the MTP_2 setting except that the edge weights of each off-diagonal entry in B is set to be 0.5. Figure 3 shows the averaged ROC curve in support recovery under the general distribution setting. According to Figure 3, $\tilde{K}^\lambda$ did not significantly outperform the state of the art estimators. This also meets with our theoretical analysis since the γ-SDD assumption for some $\gamma \geq \sqrt{2}+1$ is usually more difficult to hold for general Gaussian distributions. Nevertheless, in the general distribution setting the performance of $\tilde{K}^\lambda$ is still comparable to graphical lasso and CLIME. This shows that $\tilde{K}^\lambda$ is still a competitive estimator for general high-dimensional Gaussian distributions.

7.2 Real data analysis

We also tested our method on a stock price data set from Yahoo! Finance (`finance.yahoo.com`), which has also been used in [23]. This data set consists of the daily closing prices for 452 stocks that are consistently in the S&P 500 index from January 1, 2003 to January 1, 2018. This results in $1,257$ data points and each data point corresponds to the closing prices of the 452 stocks at the end of a single trading day. The 452 stocks are categorized into 10 Global Industry Classification Standard (GICS) sectors, namely Consumer Discretionary (70 stocks), Consumer Staples (35 stocks), Energy (37 stocks), Financials (74 stocks), Health Care (46 stocks), Industrials (59 stocks), Information Technology (64 stocks), Telecommunications Services (6 stocks), Materials (29 stocks) and Utilities (32 stocks).

In this experiment, our goal is to estimate the undirected graphical model of the random vector $X := (X_1, \cdots, X_{452})^T$ where each random variable X_j represents the price daily change of stock j. With $S_j^{(t)}$ denoting the stock closing price of stock j on day t, we let $X_j^{(t)} := \log(S_j^{(t)}/S_j^{(t-1)})$ denote the daily change record of stock j on day t and treat each $X^{(t)} := (X_1^{(t)}, \cdots, X_{452}^{(t)})^T$ as an i.i.d. realization of the random vector X. This gives us 1256 data points to estimate the underlying graphical model.

We run $\hat{K}^\lambda$ and $\tilde{K}^\lambda$ on these data points with different tuning parameters and determine the set of edges in the estimated graphical model using stability selection [28]. For comparison we also used graphical lasso. Here we did not include CLIME since CLIME is not scalable for this data set. To assess performance with regard to graph structure recovery, since the graph structure of the true underlying graphical model is still unknown, we cannot use ROC curve for evaluation. In this experiment, we instead assess each estimated graph based on its performance in grouping stocks from the same sector together. More specifically, we consider the following criterion.

Definition 7.1. [10, Equation (4)] Consider an estimated graph $G := ([p], E)$ with vertex set $[p]$ and edge set E. For each stock j, let c_j denote the sector where stock j belongs to. Let A denote the adjacency matrix of G where $A_{ij} = 1$ whenever stocks i and j are connected in G and k_j denote the number of neighbours of

stock j in graph G. Then the modularity Q is given by

$$Q = \frac{1}{2|E|} \sum_{i,j \in [p]} \Big(A_{ij} - \frac{k_i k_j}{2|E|} \Big) \delta(c_i, c_j)$$

where the δ function $\delta(i, j) = 1$ if $i = j$ and 0 otherwise.

Intuitively, Q represents the difference between the fraction of within-sector edges of the estimated graph and the fraction we would have expected from a random graph. In this case, high Q means that stocks from the same sector tend to be grouped into the same community in the estimated graph while $Q = 0$ means that the community structure has no deviation from a random graph. In particular, [10] claims that $Q \geq 0.3$ indicates a significant community structure in the graph. Table 1 shows the modularity score of the graph estimated from different methods. From the table, we can see that our methods significantly outperform graphical lasso. This shows that our method is a convincing approach in real world applications.

8 Discussion

In this paper, we have developed a new method to estimate the undirected graphical model when the data-generating distribution follows a Gaussian or transelliptical MTP_2 distribution. In particular, we have shown that our estimator is more robust than existing estimators for support recovery in the high-dimensional setting. As a corollary, we also extend such robustness guarantees into estimating precision matrices for general high-dimensional distributions. Since MTP_2 has been used in a wide range of applications and is also implied by many existing models, such as latent tree models in phylogenetics or single factor models in psychology [11], our work provides a viable tool to learn the underlying graphical model in these applications. Real data analysis also shows promise.

One remaining question from this paper is to relax the SDD assumption. In the future, I would also be interested in developing provably consistent algorithms to learn undirected graphical models for high-dimensional MTP_2 distributions without any tuning parameter. Beyond Gaussian distributions, another interesting question is to apply MTP_2 constraint into other distribution families, such as log-concave distributions or discrete distributions.

References

[1] O. Barndorff-Nielsen. *Information and Exponential Families: in Statistical Theory*. John Wiley & Sons, 2014.

[2] R. G. Bartle and D. R. Sherbert. *Introduction to Real Analysis*. Wiley, 2000.

[3] A. Berman and R. J. Plemmons. *Nonnegative Matrices in the Mathematical Sciences*, volume 9. 1994.

[4] E. Bølviken. Probability inequalities for the multivariate normal with non-negative partial correlations. *Scandinavian Journal of Statistics*, pages 49–58, 1982.

[5] S. Boyd and L. Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

[6] L. D. Brown. *Fundamentals of Statistical Exponential Families with Applications in Statistical Decision Theory*, volume 9. Institute of Mathematical Statistics, 1986.

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

[8] T. T. Cai, W. Liu, and H. H. Zhou. Estimating sparse precision matrix: optimal rates of convergence and adaptive estimation. *The Annals of Statistics*, 44(2):455–488, 2016.

[9] D. Carlson and T. L. Markham. Schur complements of diagonally dominant matrices. *Czechoslovak Mathematical Journal*, 29(2):246–251, 1979.

[10] A. Clauset, M. E. Newman, and C. Moore. Finding community structure in very large networks. *Physical review E*, 70(6):066111, 2004.

[11] S. Fallat, S. Lauritzen, K. Sadeghi, C. Uhler, N. Wermuth, and P. Zwiernik. Total positivity in markov structures. *The Annals of Statistics*, 45(3):1152–1184, 2017.

[12] J. Fan, Y. Feng, and Y. Wu. Network exploration via the adaptive LASSO and SCAD penalties. *The Annals of Applied Statistics*, 3(2):521, 2009.

[13] K. Fang, S. Kotz, and K. Ng. *Symmetric multivariate and related distributions*. Chapman & Hall, 1990.

[14] S. Fattahi and S. Sojoudi. Graphical lasso and thresholding: Equivalence and closed-form solutions. *arXiv preprint arXiv:1708.09479*, 2017.

[15] C. M. Fortuin, P. W. Kasteleyn, and J. Ginibre. Correlation inequalities on some partially ordered sets. *Communications in Mathematical Physics*, 22(2):89–103, 1971.

[16] T. P. Hubbard, T. Li, and H. J. Paarsch. Semiparametric estimation in models of first-price, sealed-bid auctions with affiliation. *Journal of Econometrics*, 168(1):4–16, 2012.

[17] H. Joe. Generating random correlation matrices based on partial correlations. *Journal of Multivariate Analysis*, 97(10):2177–2189, 2006.

[18] S. Karlin and Y. Rinott. M-matrices as covariance matrices of multinormal distributions. *Linear algebra and its applications*, 52:419–438, 1983.

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

[20] S. Lauritzen, C. Uhler, and P. Zwiernik. Maximum likelihood estimation in Gaussian models under total positivity. *arXiv preprint arXiv:1702.04031*, 2017.

[21] T. Lei, C. Woo, J. Liu, and F. Zhang. On the Schur complements of diagonally dominant matrices. In *Proceedings of the SIAM Conference on Applied Linear Algebra*, 2003.

[22] H. Liu, F. Han, M. Yuan, J. Lafferty, and L. Wasserman. High-dimensional semiparametric gaussian copula graphical models. *The Annals of Statistics*, 40(4):2293–2326, 2012.

[23] H. Liu, F. Han, and C. Zhang. Transelliptical graphical models. In *Advances in Neural Information Processing Systems*, pages 800–808, 2012.

[24] P. Loh and M. J. Wainwright. Support recovery without incoherence: A case for nonconvex regularization. *The Annals of Statistics*, 45(6):2455–2482, 2017.

[25] D. M. Malioutov, J. K. Johnson, and A. S. Willsky. Walk-sums and belief propagation in gaussian graphical models. *Journal of Machine Learning Research*, 7(Oct):2031–2064, 2006.

[26] B. F. Malle and L. M. Horowitz. The puzzle of negative self-views: An exploration using the schema concept. *Journal of Personality and Social Psychology*, 68(3):470, 1995.

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

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

[29] A. Müller and M. Scarsini. Archimedean copulae and positive dependence. *Journal of Multivariate Analysis*, 93(2):434–445, 2005.

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

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

[32] Martin Slawski and Matthias Hein. Estimation of positive definite M-matrices and structure learning for attractive Gaussian Markov random fields. *Linear Algebra and its Applications*, 473:145–179, 2015.

[33] S. Sojoudi. Equivalence of graphical lasso and thresholding for sparse graphs. *Journal of Machine Learning Research*, 17(115):1–21, 2016.

[34] S. Sojoudi and J. Doyle. Study of the brain functional network using synthetic data. In *Communication, Control, and Computing (Allerton), 2014 52nd Annual Allerton Conference on*, pages 350–357. IEEE, 2014.

[35] C. Uhler. Gaussian graphical models: An algebraic and geometric perspective. *arXiv preprint arXiv:1707.04345*, 2017.

[36] Y. Weiss and W. T. Freeman. Correctness of belief propagation in gaussian graphical models of arbitrary topology. In *Advances in Neural Information Processing Systems*, pages 673–679, 2000.

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

A Theoretical analysis in Section 3

A.1 Proof of Theorem 3.1

Let $\delta_{n,p} := \|S - \Sigma^*\|_\infty$ denote the elementwise maximal difference between the sample correlation matrix S and the true correlation matrix Σ^*. To prove Theorem 3.1, we first need the following theorem.

Theorem A.1. *Consider a Gaussian MTP$_2$ distribution satisfying either of the following two conditions for some $0 < \alpha < 1$:*

- *the inverse Isserlis matrix Γ^* is $\frac{1+\alpha}{1-\alpha}$-SDD;*
- *the precision matrix K^* is $\frac{1+\alpha+\sqrt{2(1+\alpha)}}{1-\alpha}$-SDD.*

Then, if $\delta_{n,p}$ is upper bounded by $\delta_{n,p} \leq \frac{\alpha}{8M^*}$, the solution to (2) with tuning parameter $\lambda = \alpha^{-1}\delta_{n,p}$ satisfies the following ℓ_∞ bound:

$$\|\hat{K}^\lambda - K^*\|_\infty \leq r,$$

where $r := 8M^{*2}\lambda$. Moreover, the support of $\hat{K}^\lambda$ is a subset of the support of K^* and includes all the edges such that $|(K^*)_{ij}| > r$.

A.1.1 Notations

In this section, we introduce a set of notations that will be used to prove Theorem A.1. Given any sample correlation matrix $S \in \mathbb{R}^{p\times p}$ and a tuning parameter $\lambda > 0$, we define $S^\lambda \in \mathbb{R}^{p\times p}$ as $(S^\lambda)_{ij} := S_{ij} - \lambda$, then we can rewrite our estimator (2) as follows:

$$\begin{aligned}\hat{K}^\lambda := \underset{K}{\text{argmax}} \ \ & \log\det(K) - \text{trace}(KS^\lambda) \\ \textit{s.t. } & K_{ij} \leq 0 \quad \forall i \neq j.\end{aligned}$$

Based on the duality of convex optimization, it follows from Proposition 2.2 of [20] that the inverse of $\hat{K}^\lambda$, denoted as $\hat{\Sigma}^\lambda$, can be obtained according to the following objective:

$$\begin{aligned}\hat{\Sigma}^\lambda := \text{argmax} \ \ & \log\det(\Sigma) \\ \textit{s.t. } & \Sigma_{ii} = (S^\lambda)_{ii} \quad \forall\, i, \\ & \Sigma_{ij} \geq (S^\lambda)_{ij} \quad \forall\, i \neq j.\end{aligned}$$

Moreover, $\hat{K}^\lambda$ and $\hat{\Sigma}^\lambda$ has the following complementary slackness property:

$$\begin{aligned}(\hat{\Sigma}^\lambda)_{ij} &= (S^\lambda)_{ij} \ \text{ if } \ (\hat{K}^\lambda)_{ij} \neq 0 \\ (\hat{\Sigma}^\lambda)_{ij} &\geq (S^\lambda)_{ij} \ \text{ if } \ (\hat{K}^\lambda)_{ij} = 0.\end{aligned} \tag{5}$$

Let ν denote a scalar variable such that $\nu \in [0, 1]$, let $S^\nu := (1-\nu)\Sigma^* + \nu S^\lambda$, let $\hat{K}^\nu$ as that

$$\begin{aligned}\hat{K}^\nu := \underset{K}{\text{argmax}} \ \ & \log\det(K) - \text{trace}(KS^\nu) \\ \textit{s.t. } & K_{ij} \leq 0 \quad \forall i \neq j.\end{aligned} \tag{6}$$

It is then a short computation from (6) that $\hat{K}^\nu\big|_{\nu=1} = \hat{K}^\lambda$ and $\hat{K}^\nu\big|_{\nu=0} = K^*$. This allows us to prove Theorem A.1 by taking $\hat{K}^\nu$ as a function of ν and analyzing how $\hat{K}^\nu$ is changed as ν is increased from $\nu = 0$ to $\nu = 1$.

Let $\hat{\Sigma}^\nu := (\hat{K}^\nu)^{-1}$, then $\hat{\Sigma}^\nu$ can be given by

$$\begin{aligned} \hat{\Sigma}^\nu := \operatorname{argmax}\ & \log\det(\Sigma) \\ \textit{s.t.}\ & \Sigma_{ii} = (S^\nu)_{ii} \quad \forall\, i, \\ & \Sigma_{ij} \geq (S^\nu)_{ij} \quad \forall\, i \neq j. \end{aligned} \tag{7}$$

Also, let $\hat{E}^\nu$ denote the support of $\hat{K}^\nu$. Throughout the proof, we denote by $\hat{\Gamma}^\nu := \hat{K}^\nu \otimes \hat{K}^\nu$ and $\hat{H}^\nu := \hat{\Sigma}^\nu \otimes \hat{\Sigma}^\nu$. Also, for any matrix A, let $[A]$ denote the vectorized form of matrix A where each entry in the vector corresponds to an entry in matrix A.

A.1.2 Preliminary lemmas for the proof of Theorem A.1

Lemma A.2. *Suppose a symmetric positive definite matrix A is γ-SDD, then the Kronecker product $A_K := A \otimes A$ is $\frac{\gamma^2}{2\gamma+1}$-SDD.*

Proof. For any diagonal entry in A_K, i.e., $(A_K)_{(i,j),(i,j)}$, we have that

$$\begin{aligned} \sum_{(\ell,m)\neq(i,j)} |(A_K)_{(i,j),(\ell,m)}| &= \sum_{\ell\neq i, m\neq j} |A_{i\ell}A_{jm}| + \sum_{m\neq j} |A_{ii}A_{jm}| + \sum_{\ell\neq i} |A_{i\ell}A_{jj}| \\ &< (2\gamma^{-1} + \gamma^{-2})|A_{ii}A_{jj}| = (2\gamma^{-1} + \gamma^{-2})|(A_K)_{(i,j),(i,j)}|. \end{aligned}$$

□

Lemma A.3. *Consider two symmetric positive definite matrices $K^{(1)}$ and $K^{(2)}$, with inverse denoted by $\Sigma^{(1)}$ and $\Sigma^{(2)}$ respectively. Let $T := \{(i,j) : K^{(1)}_{ij} \neq 0 \text{ or } K^{(2)}_{ij} \neq 0\}$, then*

$$[K^{(2)}]_T - [K^{(1)}]_T = \big(H_{TT}\big)^{-1}\big([\Sigma^{(1)}]_T - [\Sigma^{(2)}]_T\big) + o\big([\Sigma^{(1)}]_T - [\Sigma^{(2)}]_T\big),$$

where $H := \Sigma^{(1)} \otimes \Sigma^{(1)}$.

Proof. It follows from the definition of $K^{(i)}$ and $\Sigma^{(i)}$ that for all $i \in \{1,2\}$,

$$\begin{aligned} K^{(i)} = \operatorname*{argmax}_K\ & \log\det(K) - \operatorname{trace}(K\Sigma^{(i)}) \\ \textit{s.t.}\ & K_{ij} = 0 \quad \text{if } (i,j) \notin T. \end{aligned} \tag{8}$$

In other words, $K^{(i)}$ is the output of the convex program (8) by taking $[\Sigma^{(i)}]_T$ as the data statistics. It then follows from [1, 6] that there exists a homeomorphic function f that maps the space of canonical parameters $[K]_T$ and the space of sufficient statistics $[\Sigma]_T$ such that for all $i \in \{1,2\}$,

$$[\Sigma^{(i)}]_T = f([K^{(i)}]_T) \quad \text{and} \quad [K^{(i)}]_T = f^{-1}([\Sigma^{(i)}]_T).$$

Moreover, [1, 6] also shows that this function f is the gradient of the log-partition function, i.e.

$$f([K]_T) := \nabla \log\det(K),$$

and that f is analytic. Then using standard results from matrix derivatives [5], the gradient of f follows that

$$\nabla f([K]_T)\Big|_{[K]_T=[K^{(1)}]_T} = -H_{TT}.$$

Hence its inverse function satisfies that

$$\nabla f^{-1}([\Sigma]_T)\Big|_{[\Sigma]_T=[\Sigma^{(1)}]_T} = -\big(H_{TT}\big)^{-1},$$

which completes the proof. □

Lemma A.4. *$\hat{\Sigma}^\nu$ is unique and continuous in ν over the domain $[0,1]$.*

Proof. We first prove the existence and uniqueness. Since for all $\nu \in [0,1]$ and all $i \neq j$, $(S^\nu)_{ij} < \sqrt{(S^\nu)_{ii}(S^\nu)_{jj}}$, by applying Theorem 3.5 of [20], we have that $\hat{\Sigma}^\nu$ always exists and is unique.

We next prove continuity, we prove by contradiction. Suppose there exists a ν_0 such that $\hat{\Sigma}^{\nu_0}$ is not continuous, then there exists a sequence $\nu_k \to \nu_0$ such that $\|\hat{\Sigma}^{\nu_k} - \hat{\Sigma}^{\nu_0}\| \geq \epsilon$ for some $\epsilon > 0$ for all k. Apparently all diagonal entries of $\hat{\Sigma}^{\nu_k}$ are bounded between $1-\lambda$ and 1. In addition, using that all $\hat{\Sigma}^{\nu_k}$ are positive definite and inverse M-matrices, we have that for all $i \neq j$ and all k,

$$0 \leq (\hat{\Sigma}^{\nu_k})_{ij} \leq \sqrt{(\hat{\Sigma}^{\nu_k})_{ii}(\hat{\Sigma}^{\nu_k})_{jj}}.$$

Therefore, all off-diagonal entries are also bounded. Since all entries in $\hat{\Sigma}^{\nu_k}$ are bounded, by applying Bolzano-Weierstrass theorem [2], we have that there exists a subsequence $\tilde{\nu}_k \to \nu_0$ that converge to some point $\hat{\Sigma}'$ and that $\|\hat{\Sigma}' - \hat{\Sigma}^{\nu_0}\| \geq \epsilon$ for some $\epsilon > 0$.

Let $\hat{\Sigma}^{\nu_0,\tilde{\nu}_k} := \hat{\Sigma}^{\nu_0} + (S^{\tilde{\nu}_k} - S^{\nu_0})$. Apparently $\hat{\Sigma}^{\nu_0,\tilde{\nu}_k}$ is a feasible solution of (7) if we set the ν in (7) as $\tilde{\nu}_k$. Using that $\hat{\Sigma}^{\tilde{\nu}_k}$ is the global optimum of (7) if we set the ν in (7) as $\tilde{\nu}_k$, we have that $\log\det(\hat{\Sigma}^{\tilde{\nu}_k}) \geq \log\det(\hat{\Sigma}^{\nu_0,\tilde{\nu}_k})$. By taking $k \to \infty$ on both sides and using that $\hat{\Sigma}^{\tilde{\nu}_k} \to \hat{\Sigma}'$ and $\hat{\Sigma}^{\nu_0,\tilde{\nu}_k} \to \hat{\Sigma}^{\nu_0}$ as $k \to \infty$, we further have that

$$\log\det(\hat{\Sigma}') \geq \log\det(\hat{\Sigma}^{\nu_0}). \tag{9}$$

At the same time, for each $i \neq j$, by taking $k \to \infty$ on both sides of the constraint $(\hat{\Sigma}^{\tilde{\nu}_k})_{ij} \geq (S^{\tilde{\nu}_k})_{ij}$ we have that $\hat{\Sigma}'_{ij} \geq (S^{\nu_0})_{ij}$; for each diagonal entry, by taking $k \to \infty$ on both sides of $(\hat{\Sigma}^{\tilde{\nu}_k})_{ii} = (S^{\tilde{\nu}_k})_{ii}$ we have that $\hat{\Sigma}'_{ii} = (S^{\nu_0})_{ii}$. Hence, $\hat{\Sigma}'$ is also a feasible solution of the objective (7) when we set the ν in (7) as ν_0. Based on the uniqueness of the global optimum of (7), we instead conclude that $\log\det(\hat{\Sigma}') < \log\det(\hat{\Sigma}^{\nu_0})$. This contradicts with (9). □

Lemma A.5. *Assume $\hat{\Gamma}^\nu$ is $\frac{1+\alpha}{1-\alpha}$-strictly diagonally dominant, then there exists a $\Delta\nu$ small enough such that for all $\tilde{\nu} > \nu$ while $\tilde{\nu} - \nu \leq \Delta\nu$, $\hat{E}^{\tilde{\nu}} \subseteq \hat{E}^\nu$. Moreover, $(\hat{K}^{\tilde{\nu}})_{ij} > (\hat{K}^\nu)_{ij}$ for all the (i,j)'s where $(\hat{K}^\nu)_{ij} \neq 0$.*

Proof. Apparently the complementary slackness property given in (5) also applies to $\hat{\Sigma}^\nu$ and $\hat{K}^\nu$. Therefore, we can divide the indices of $\hat{K}^\nu$ into the following three subsets:

$$\begin{aligned}
\mathcal{A} :=& \{(i,j) : (\hat{\Sigma}^\nu)_{ij} = S^\nu \text{ and } (\hat{K}^\nu)_{ij} < 0\} \cup \{(i,i) : \forall i \in [p]\},\\
\mathcal{I} :=& \{(i,j) : (\hat{\Sigma}^\nu)_{ij} > S^\nu \text{ and } (\hat{K}^\nu)_{ij} = 0\},\\
\mathcal{B} :=& \{(i,j) : (\hat{\Sigma}^\nu)_{ij} = S^\nu \text{ and } (\hat{K}^\nu)_{ij} = 0\}.
\end{aligned}$$

Similarly, we also divide the indices of $\hat{K}^{\tilde{\nu}}$ into $\tilde{\mathcal{A}}$, $\tilde{\mathcal{I}}$ and $\tilde{\mathcal{B}}$ accordingly. Based on the continuity of $\hat{K}^{\nu}$ as given in Lemma A.4, we have that for $\Delta\nu$ small enough, $\mathcal{A}\cap\tilde{\mathcal{I}}=\emptyset$ and $\mathcal{I}\cap\tilde{\mathcal{A}}=\emptyset$ and therefore

$$\tilde{\mathcal{A}}\subseteq\mathcal{A}\cup\mathcal{B}\quad\text{and}\quad\mathcal{A}\subseteq\tilde{\mathcal{A}}\cup\tilde{\mathcal{B}} \tag{10}$$

for all $\tilde{\nu}\leq\nu+\Delta\nu$.

Based on this definition, it is sufficient to prove $\hat{E}^{\tilde{\nu}}\subseteq\hat{E}^{\nu}$ by proving that $\tilde{\mathcal{A}}\subseteq\mathcal{A}$. Let $T:=\tilde{\mathcal{A}}\cup\mathcal{A}$, by applying Lemma A.3 with $K^{(1)}=\hat{K}^{\nu}$ and $K^{(2)}=\hat{K}^{\tilde{\nu}}$, we have that

$$\begin{aligned}[\hat{K}^{\tilde{\nu}}]_T-[\hat{K}^{\nu}]_T&=\left((\hat{H}^{\nu})_{TT}\right)^{-1}\left([\hat{\Sigma}^{\nu}]_T-[\hat{\Sigma}^{\tilde{\nu}}]_T\right)+o(\tilde{\nu}-\nu)\\&=\left((\hat{H}^{\nu})_{TT}\right)^{-1}\left([S^{\nu}]_T-[S^{\tilde{\nu}}]_T\right)+o(\tilde{\nu}-\nu)\end{aligned} \tag{11}$$

where the second equality uses the fact that for any D satisfying $D\subseteq\mathcal{A}\cup\mathcal{B}$ and $D\subseteq\tilde{\mathcal{A}}\cup\tilde{\mathcal{B}}$,

$$[\hat{\Sigma}^{\nu}]_D-[\hat{\Sigma}^{\tilde{\nu}}]_D=[S^{\nu}]_D-[S^{\tilde{\nu}}]_D.$$

In addition, since $\hat{\Gamma}^{\nu}$ is $\frac{1+\alpha}{1-\alpha}$-SDD and $\left((\hat{H}^{\nu})_{TT}\right)^{-1}$ is a Schur complement of the matrix $\hat{\Gamma}^{\nu}$, by applying Lemma 3.3 we have that $\left((\hat{H}^{\nu})_{TT}\right)^{-1}$ is also $\frac{1+\alpha}{1-\alpha}$-SDD.

Finally, for each entry (i,j) in T, we also have that

$$(1-\alpha)(\tilde{\nu}-\nu)\lambda\leq(S^{\nu})_{ij}-(S^{\tilde{\nu}})_{ij}\leq(1+\alpha)(\tilde{\nu}-\nu)\lambda. \tag{12}$$

By combining (11), (12) and that $\left((\hat{H}^{\nu})_{TT}\right)^{-1}$ is $\frac{1+\alpha}{1-\alpha}$-SDD, it is a short computation to discover that the vector $\left((\hat{H}^{\nu})_{TT}\right)^{-1}\left([\hat{\Sigma}^{\nu}]_T-[\hat{\Sigma}^{\tilde{\nu}}]_T\right)$ is a fully positive vector. Hence, $[\hat{K}^{\tilde{\nu}}]_T-[\hat{K}^{\nu}]_T$ is also a fully positive vector.

In this case, if $\tilde{\mathcal{A}}\setminus\mathcal{A}$ is nonempty, then for all $(i,j)\in\tilde{\mathcal{A}}\setminus\mathcal{A}$, we have that $(\hat{K}^{\tilde{\nu}})_{ij}>(\hat{K}^{\nu})_{ij}$. This contradicts with that $(\hat{K}^{\tilde{\nu}})_{ij}<0$ and $(\hat{K}^{\nu})_{ij}=0$ for all $(i,j)\in\tilde{\mathcal{A}}\setminus\mathcal{A}$. Moreover, since $[\hat{K}^{\tilde{\nu}}]_T-[\hat{K}^{\nu}]_T$ is a positive vector, we also have that for any entry (i,j) where $(\hat{K}^{\nu})_{ij}\neq0$, $(\hat{K}^{\tilde{\nu}})_{ij}>(\hat{K}^{\nu})_{ij}$. □

Lemma A.6. *Assume $\hat{\Gamma}^{\nu}$ is $\frac{1+\alpha}{1-\alpha}$-SDD, then the right-hand and left-hand derivatives satisfy that*

$$\frac{d(\hat{K}^{\nu})_{ij}}{d\nu_+}\leq2\hat{M}^2\lambda\quad\textit{and}\quad\frac{d(\hat{K}^{\nu})_{ij}}{d\nu_-}\leq2\hat{M}^2\lambda,$$

where $\hat{M}$ is the largest diagonal entry of $\hat{K}^{\nu}$.

Proof. It follows from Lemma A.5 that for $\tilde{\nu}-\nu$ sufficiently small, $(\hat{K}^{\tilde{\nu}})_{ij}<0$ for all $(i,j)\in\mathcal{A}$ and hence $\tilde{\mathcal{A}}=\mathcal{A}$. Moreover, Lemma A.5 also tells us that

$$[\hat{K}^{\tilde{\nu}}]_{\mathcal{A}}-[\hat{K}^{\nu}]_{\mathcal{A}}=\left((\hat{H}^{\nu})_{\mathcal{A}\mathcal{A}}\right)^{-1}\left([S^{\nu}]_{\mathcal{A}}-[S^{\tilde{\nu}}]_{\mathcal{A}}\right)+o(\tilde{\nu}-\nu).$$

Since by using Lemma 3.3 and that $\hat{\Gamma}^{\nu}$ is $\frac{1+\alpha}{1-\alpha}$-SDD, we have that

$$\left\|\left((\hat{H}^{\nu})_{\mathcal{A}\mathcal{A}}\right)^{-1}\right\|_{L_1}\leq\|\hat{\Gamma}^{\nu}\|_{L_1}\leq\frac{2}{1+\alpha}\hat{M}^2. \tag{13}$$

By combining this with the fact that $\|S^{\nu} - S^{\tilde{\nu}}\|_{\infty} \leq (1+\alpha)(\tilde{\nu} - \nu)\lambda$, we have that for any entry $(i,j) \in \mathcal{A}$,

$$\left\| \left((\hat{H}^{\nu})_{\mathcal{A}\mathcal{A}} \right)^{-1} \left([S^{\nu}]_{\mathcal{A}} - [S^{\tilde{\nu}}]_{\mathcal{A}} \right) \right\|_{\infty} \leq 2\hat{M}^2 \lambda (\tilde{\nu} - \nu).$$

By applying this onto (13) and taking the limit $\tilde{\nu} \to \nu$, we further have that for any entry (i,j) in $\mathcal{A}$, the right-hand derivative satisfies that

$$\frac{d(\hat{K}^{\nu})_{ij}}{d\nu_{+}} \leq 2\hat{M}^2 \lambda.$$

Similarly, we also obtain the same bound on the left-hand derivative. □

A.1.3 Main proof

We first focus on proving Theorem A.1.

proof of Theorem A.1. We first prove that $\hat{K}^{\lambda}$ is sparsistent whenever Γ^* is $\frac{1+\alpha}{1-\alpha}$-SDD. Our entire proof is splitted into three steps.

First we prove that $\hat{\Gamma}^{\nu}$ is $\frac{1+\alpha}{1-\alpha}$-SDD for all $\nu \in [0,1]$. We prove this by contradiction. Suppose in contradiction there exists some $\nu \in (0,1]$ such that $\hat{\Gamma}^{\nu}$ is not $\frac{1+\alpha}{1-\alpha}$-SDD. Without loss of generality, let $\nu_1 > 0$ denote the smallest ν where $\hat{\Gamma}^{\nu}$ is not $\frac{1+\alpha}{1-\alpha}$-SDD, i.e., $\hat{\Gamma}^{\nu_1}$ is not $\frac{1+\alpha}{1-\alpha}$-SDD while for all $\nu \in [0, \nu_1)$, $\hat{\Gamma}^{\nu}$ is $\frac{1+\alpha}{1-\alpha}$-SDD. Since $\hat{\Gamma}^{\nu}$ is $\frac{1+\alpha}{1-\alpha}$-SDD for $\nu \in [0, \nu_1)$, by applying Lemma A.5, we have that for all $i \neq j$, $|(\hat{K}^{\nu_1})_{ij}| \leq |(\hat{K}^{\nu})_{ij}|$ and for all i, $(\hat{K}^{\nu_1})_{ii} \geq (\hat{K}^{\nu})_{ii}$. In this case, $\hat{K}^{\nu_1}$ is more diagonally dominant than $\hat{K}^{\nu}$ and therefore, $\hat{\Gamma}^{\nu_1}$ is also $\frac{1+\alpha}{1-\alpha}$-SDD. This contradicts with that $\hat{\Gamma}^{\nu_1}$ is not $\frac{1+\alpha}{1-\alpha}$-SDD. Hence, we have that $\hat{\Gamma}^{\nu}$ is $\frac{1+\alpha}{1-\alpha}$-SDD for all $\nu \in [0,1]$.

Next, we prove that the support of $\hat{K}^{\lambda}$ is a subset of the support of K^*. Suppose in contradiction there exists an entry (i,j) such that $(K^*)_{ij} = 0$ while $(\hat{K}^{\lambda})_{ij} \neq 0$. By applying the continuity as given in Lemma A.4, there exists a $\nu_2 < 1$ such that $(\hat{K}^{\nu_2})_{ij} = 0$ while $(\hat{K}^{\nu})_{ij} < 0$ for all $\nu > \nu_2$. However, this contradicts with Lemma A.5 considering that $\hat{\Gamma}^{\nu_2}$ is $\frac{1+\alpha}{1-\alpha}$-SDD.

Finally, we prove that the maximal diagonal entry of $\hat{K}^{\nu}$, denoted as $\hat{M}^{\nu}$, is always smaller than $2M^*$. We prove this by contradiction. Suppose this does not hold, then by continuity there exists a ν_3 such that $\hat{M}^{\nu_3} = 2M^*$. Since all the diagonal entries are monotonically increasing, we have that for all $\nu < \nu_3$, $\hat{M}^{\nu} \leq 2M^*$ and therefore by applying Lemma A.6 on the derivative,

$$\|\hat{K}^{\nu_3} - K^*\|_{\infty} < 8{M^*}^2 \nu_3 \lambda.$$

By combing this with $\|\hat{K}^{\nu_3} - K^*\|_{\infty} \geq M^*$, we have that $\lambda > \frac{1}{8M^*}$. This contradicts with that $\delta_{n,p} \leq \frac{\alpha}{8M^*}$ and $\delta_{n,p} = \alpha\lambda$. In this case, we have that $\hat{M}^{\nu} \leq 2M^*$ for all $\nu \in [0,1]$. By applying Lemma A.6, we have that

$$\|\hat{K}^{\lambda} - K^*\|_{\infty} \leq 8{M^*}^2 \lambda.$$

So far, we have provided the sparsistency guarantee under the condition that Γ^* is $\frac{1+\alpha}{1-\alpha}$-SDD. We next prove that the sparsistency holds under the condition that K^* is $\frac{1+\alpha+\sqrt{2(1+\alpha)}}{1-\alpha}$-SDD. Using Lemma A.2, we have that the $\frac{1+\alpha+\sqrt{2(1+\alpha)}}{1-\alpha}$-SDD condition for K^* is a sufficient condition for Γ^* to be $\frac{1+\alpha}{1-\alpha}$-SDD. Then using the same derivations as mentioned above, the sparsistency guarantees still hold. □

Then we are ready to prove Theorem 3.1.

proof of Theorem 3.1. By applying Theorem A.1, it remains to prove that there exists some positive constants C and τ such that $\delta_{n,p} \leq C\sqrt{\frac{\log p}{n}}$ with probability $1 - p^{-\tau}$. This is a direct consequence by applying [30, Lemma 1]. □

A.2 Proof of Lemma 3.3

For any matrix $A \in \mathbb{R}^{p\times p}$, let $|A| \in \mathbb{R}^{p\times p}$ denote the matrix where each (i,j)-th entry in $|A|$ is equal to the absolute value of A_{ij}. To prove Lemma 3.3, we first need the following lemma:

Lemma A.7. *For any symmetric matrix A, let B as that*

$$B_{ij} = \begin{cases} |A_{ij}| & i = j \\ -|A_{ij}| & i \neq j, \end{cases}$$

Then if A is SDD, we have that B^{-1} always exists. In addition, $B^{-1} \geq |A^{-1}|$, i.e., B^{-1} is entrywise bigger than $|A^{-1}|$.

Proof. Apparently B is also strictly diagonally dominant. Hence B is positive definite and its inverse exists. To prove that $B^{-1} \geq |A^{-1}|$, we decompose A as that $A = D_A + O_A$ where D_A is a diagonal matrix and all diagonal elements of O_A are zero. We also decompose B as $B = D_B + O_B$. Since the operator norm of $D_B^{-1}O_B$ is less than 1, one can rewrite

$$B^{-1} = D_B^{-1} + \sum_{k=1}^{\infty}(-1)^k(D_B^{-1}O_B)^k D_B^{-1}.$$

Since $D_B = |D_A|$ and $-O_B = |O_A|$, one can further rewrite that

$$B^{-1} = |D_A^{-1}| + \sum_{k=1}^{\infty}(|D_A^{-1}||O_A|)^k|D_A^{-1}|. \tag{14}$$

At the same time one can rewrite A^{-1} as that

$$A^{-1} = D_A^{-1} + \sum_{k=1}^{\infty}(-1)^k(D_A^{-1}O_A)^k D_A^{-1}. \tag{15}$$

Since for all $k \geq 0$, $|(-1)^k(D_A^{-1}O_A)^k D_A^{-1}| \leq (|D_A^{-1}||O_A|)^k|D_A^{-1}|$, it then follows from the decomposition provided in (14) and (15) that $B^{-1} \geq |A^{-1}|$. □

Then we our main proof of Lemma 3.3 is as follows

Proof. Using Schur complement, it is a short computation to show that for any $i \in [k]$,

$$\begin{aligned} |A'_{ii}| - \sum_{j\neq i \& j\in[k]} |A'_{ij}| &= |A_{ii} - A_{iS'}A_{S'S'}^{-1}A_{S'i}| - \sum_{j\neq i \& j\in[k]} |A_{ij} - A_{iS'}A_{S'S'}^{-1}A_{S'j}| \\ &\geq |A_{ii}| - \sum_{j\neq i \& j\in[k]} |A_{ij}| - \sum_{j\in[k]} |A_{iS'}A_{S'S'}^{-1}A_{S'j}| \\ &\geq |A_{ii}| - \sum_{j\neq i \& j\in[k]} |A_{ij}| - \sum_{j\in[k]} |A_{iS'}||A_{S'S'}^{-1}||A_{S'j}| \end{aligned} \tag{16}$$

Let $B \in \mathbb{R}^{p-k,p-k}$ denote the matrix constructed from $A_{S'S'}$ such that

$$B_{ij} = \begin{cases} |A_{i+k,j+k}| & i = j \\ -|A_{i+k,j+k}| & i \neq j. \end{cases}$$

Apparently, it follows from Lemma A.7 that $B^{-1} \geq |A^{-1}_{S'S'}|$. In this way, we can further get that

$$|A_{iS'}||A^{-1}_{S'S'}| \sum_{j\in[k]} |A_{S'j}| \leq |A_{iS'}||B^{-1}| \sum_{j\in[k]} |A_{S'j}|.$$

Based on the strict diagonal dominance property of A, we have that for all $\ell \in [p] \setminus [k]$,

$$|A_{\ell\ell}| - \sum_{m\neq\ell \& m\in[p]\setminus[k]} |A_{\ell m}| > \sum_{j\in[k]} |A_{\ell j}|.$$

Then using that

$$|A_{\ell\ell}| - \sum_{m\neq\ell \& m\in[p]\setminus[k]} |A_{\ell m}| = \sum_{1\leq t\leq p-k} B_{\ell-k,t},$$

we further have that $\sum_{j\in[k]} |A_{S'j}| < B\mathbf{1}$ and therefore

$$|A_{iS'}||A^{-1}_{S'S'}| \sum_{j\in[k]} |A_{S'j}| < |A_{iS'}|B^{-1}B\mathbf{1} < |A_{iS'}|\mathbf{1}. \tag{17}$$

By taking this back into (16) we have that

$$|A'_{ii}| - \sum_{j\neq i \& j\in[k]} |A'_{ij}| > |A_{ii}| - \sum_{j\neq i \& j\in[k]} |A_{ij}| - |A_{iS'}|\mathbf{1} > (1 - \frac{1}{\gamma})|A_{ii}|.$$

In addition, we also have that

$$\begin{aligned} \sum_{j\neq i \& j\in[k]} |A'_{ij}| &\leq \sum_{j\neq i \& j\in[k]} |A_{ij}| + \sum_{j\in[k]} |A_{iS'}||A^{-1}_{S'S'}||A_{S'j}| \\ &\leq \sum_{j\neq i \& j\in[k]} |A_{ij}| + |A_{iS'}|\mathbf{1} < \frac{1}{\gamma}|A_{ii}|. \end{aligned}$$

By combining them together, we finally have that

$$|A'_{ii}| - \sum_{j\neq i \& j\in[k]} |A'_{ij}| > (\gamma - 1) \sum_{j\neq i \& j\in[k]} |A'_{ij}|,$$

which completes the closure property of γ-SDD. For the bound on matrix L_1 norm, observe that for all i, it follows from (17) that

$$\begin{aligned} \sum_{j\in[k]} |A'_{ij}| &= \sum_{j\in[k]} |A_{ij} - A_{iS'}A^{-1}_{S'S'}A_{S'j}| \leq \sum_{j\in[k]} |A_{ij}| + \sum_{j\in[k]} |A_{iS'}A^{-1}_{S'S'}A_{S'j}| \\ &\leq \sum_{j\in[k]} |A_{ij}| + \sum_{j\in[k]} |A_{iS'}||A^{-1}_{S'S'}||A_{S'j}| \leq \sum_{j\in[k]} |A_{ij}| + \sum_{j\in[p]\setminus[k]} |A_{ij}|. \end{aligned}$$

This gives us the bound on the matrix L_1 norm. □

A.3 Proof of Proposition 3.4

We first prove that Proposition 3.4 holds when X follows an elliptical MTP$_2$ distribution $X \sim \mathrm{EC}(\mu^*, \Sigma^*, g)$. Without loss of generality it is sufficient to assume that $\mu^* = 0$. To prove that the K^* is an M-matrix, it is sufficient to prove that $K^*_{ij} \leq 0$ for all $i \neq j$. We next prove that $K^*_{ij} \leq 0$ for an arbitrary choice of off-diagonal entry (i, j).

Let p_X denote the probability density function of the random vector X, to prove that K^* is an M-matrix, we consider two p-dimensional vectors $x^{(1)}, x^{(2)}$ such that $x^{(1)}$ satisfies that $(x^{(1)})_i = 1$, $(x^{(1)})_j = -1$ and all other entries are zero; and $x^{(2)}$ satisfies that $(x^{(2)})_i = -1$, $(x^{(2)})_j = 1$ and all other entries are zero. In this case, we can easily have that

$$p_X(x^{(1)})p_X(x^{(2)}) = \det(\Sigma^*)g^2(K^*_{ii} + K^*_{jj} - 2K^*_{ij})$$

and that

$$p_X(x^{(1)} \vee x^{(2)})p_X(x^{(1)} \wedge x^{(2)}) = \det(\Sigma^*)g^2(K^*_{ii} + K^*_{jj} + 2K^*_{ij}).$$

Then using that the function g is monotonically decreasing and that

$$p_X(x^{(1)})p_X(x^{(2)}) \leq p_X(x^{(1)} \vee x^{(2)})p_X(x^{(1)} \wedge x^{(2)})$$

since it is MTP$_2$, one can easily have that $K^*_{ii} + K^*_{jj} - 2K^*_{ij} \geq K^*_{ii} + K^*_{jj} + 2K^*_{ij}$. Therefore, $K^*_{ij} \leq 0$, which completes the proof for the elliptical distributions.

For transelliptical distributions $X \sim \mathrm{TE}(\Sigma^*, g; f_1, \cdots, f_p)$, let random vector $Z := (f_1(X_1), \cdots, f_p(X_p))$. Then we easily have that $Z \sim \mathrm{EC}(0, \Sigma^*, g)$. Moreover, the probability density function of Z, denoted as p_Z, satisfies that

$$p_Z(z) = \frac{p_X(f_1^{-1}(z_1), \cdots, f_p^{-1}(z_p))}{f_1'(x_1) \cdots f_p'(x_p)},$$

Since for every $x^{(1)}, x^{(2)}$, $p_X(x^{(1)})p_X(x^{(2)}) \leq p_X(x^{(1)} \vee x^{(2)})p_X(x^{(1)} \wedge x^{(2)})$. By combining this with the above equality, one also has that for every $z^{(1)}$ and $z^{(2)}$,

$$p_Z(z^{(1)})p_Z(z^{(2)}) \leq p_Z(z^{(1)} \vee z^{(2)})p_Z(z^{(1)} \wedge z^{(2)}).$$

Hence, the random vector Z is also MTP$_2$. Then it immediately follows from the proof of elliptical setting that the latent generalized precision matrix K^* is an M-matrix.

A.4 Proof of Proposition 3.5

Proof. It follows from [23, Theorem 5.1] that there exists some positive constants c_1 and c_2 such that the following concentration equality holds for all $i, j \in [p]$.

$$\mathbb{P}\Big(S^\tau_{ij} - \Sigma^*_{ij} \geq t\Big) \leq c_1 \exp(-c_2 n t^2).$$

Then by choosing $t = \sqrt{(\tau + 2/c_2)\frac{\log p}{n}}$ for some positive constant $\tau > 0$ we have that

$$\begin{aligned}
&\mathbb{P}\Big(\|S^\tau - \Sigma^*\|_\infty \geq \sqrt{(\tau + 2/c_2)\frac{\log p}{n}}\Big) \\
&\leq \sum_{i \geq j} \mathbb{P}\Big(S^\tau_{ij} - \Sigma^*_{ij} \geq \sqrt{(\tau + 2/c_2)\frac{\log p}{n}}\Big) \\
&\leq \sum_{i \geq j} c_1 \exp(-c_2(\tau + 2/c_2)\log p) \leq c_1 \exp(-c_2 \tau \log p),
\end{aligned}$$

where the first inequality is based on the union bound. This completes the proof. □

B Theoretical analysis in Section 4

B.1 Proof of Theorem 4.1

For any $\tilde{\lambda} > \lambda$, by redefining S^ν as

$$(S^\nu)_{ij} := (1 - \nu)(S^\lambda)_{ij} + \nu(S^{\tilde{\lambda}})_{ij},$$

it is a short computation that $\hat{K}^\nu\big|_{\nu=0} = \hat{K}^\lambda$ and $\hat{K}^\nu\big|_{\nu=1} = \hat{K}^{\tilde{\lambda}}$. Then Theorem 4.1 immediately follows from the proof of Theorem A.1 by choosing $\alpha = 0$.

B.2 Proof of Proposition 4.2

It directly follows from Lemma A.5 that $(\hat{K}^{\lambda_{n,p}})_{i,i} \geq K^*_{i,i}$ for all $i \in [p]$ and $|(\hat{K}^{\lambda_{n,p}})_{i,j}| \leq |K^*_{i,j}|$ for all $i \neq j$. Hence, $\hat{K}^{\lambda_{n,p}}$ is more diagonally dominant than K^*. By combining this with the fact that K^* is $\frac{1+\alpha+\sqrt{2(1+\alpha)}}{1-\alpha}$-SDD (or that Γ^* is $\frac{1+\alpha}{1-\alpha}$-SDD), we can further conclude that $\hat{K}^{\lambda_{n,p}}$ is $\frac{1+\alpha+\sqrt{2(1+\alpha)}}{1-\alpha}$-SDD (or that $\hat{\Gamma}^{\lambda_{n,p}}$ is $\frac{1+\alpha}{1-\alpha}$-SDD).

B.3 Proof of Theorem 4.4

In order to prove Theorem 4.4, we need the following two lemmas:

Lemma B.1. *Suppose the M-matrix M is γ-SDD for some $\gamma > 1$, then we have that*

$$\begin{aligned}
(M^{-1})_{ij} &< \Big|\frac{M_{ij}}{M_{ii}M_{jj}}\Big| + \frac{1}{\gamma - 1}\max_{u \neq j}\Big|\frac{M_{uj}}{M_{uu}M_{jj}}\Big| \quad \text{if } i \neq j, \\
(M^{-1})_{ij} &\geq \Big|\frac{M_{ij}}{M_{ii}M_{jj}}\Big| \qquad\qquad\qquad\qquad\qquad\quad \forall i, j.
\end{aligned}$$

Proof. We decompose M as $M = D - O$, where D corresponds to the diagonal part of M and O corresponds to the absolute of the off-diagonal part of M. In this case, using the decomposition that

$$M^{-1} = D^{-1} + \sum_{k=1}^{\infty}(-1)^k(D^{-1}(-O))^k D^{-1},$$

we have that for any off-diagonal entry (i,j) in M^{-1},

$$(M^{-1})_{ij} = \left(D^{-1}OD^{-1}\right)_{ij} + \sum_{k=1}^{\infty} \left((D^{-1}O)^k D^{-1}OD^{-1}\right)_{ij}. \tag{18}$$

Then using the fact that for any two matrices $A, B \in \mathbb{R}^{p\times p}$, $(AB)_{ij} \leq \|A\|_{L_1} \max_{u\neq j} |B_{uj}|$, one can rewrite that

$$\begin{aligned}\left((D^{-1}O)^k D^{-1}OD^{-1}\right)_{ij} &\leq \|D^{-1}O\|_{L_1} \max_{u\neq j} \left((D^{-1}O)^{k-1} D^{-1}OD^{-1}\right)_{uj} \\ &< \frac{1}{\gamma} \max_{u\neq j} \left((D^{-1}O)^{k-1} D^{-1}OD^{-1}\right)_{uj},\end{aligned}$$

where the second inequality comes from the fact that M is γ-SDD. We further have that for all $k \geq 1$,

$$\left((D^{-1}O)^k D^{-1}OD^{-1}\right)_{ij} < \frac{1}{\gamma^k} \max_{u\neq j} \left(D^{-1}OD^{-1}\right)_{uj}.$$

By summing all the k's together, one can give an upper bound of (18) as that

$$\begin{aligned}(M^{-1})_{ij} &< \left(D^{-1}OD^{-1}\right)_{ij} + \sum_{k=1}^{\infty} \frac{1}{\gamma^k} \max_{u\neq j} \left(D^{-1}OD^{-1}\right)_{uj} \\ &= \left(D^{-1}OD^{-1}\right)_{ij} + \frac{1}{\gamma - 1} \max_{u\neq j} \left(D^{-1}OD^{-1}\right)_{uj},\end{aligned}$$

which completes the first inequality in the statement. For the second inequality, using again the decomposition as given in (18), and the fact that the $\left((D^{-1}O)^k D^{-1}OD^{-1}\right)_{ij}$'s are always positive, we have that for all i, j,

$$(M^{-1})_{ij} \geq \left(D^{-1}OD^{-1}\right)_{ij}.$$

□

The next lemma concerns the estimate from (2) when we instead take a submatrix of S as the sample correlation matrix in (2). Recall that $\hat{K}^\lambda$ denotes the solution of (2) with S as the input sample correlation matrix and λ as the tuning parameter, we have the following lemma.

Lemma B.2. *Let $S_{[p]\setminus\{u\},[p]\setminus\{u\}} \in \mathbb{R}^{p-1\times p-1}$ denote the submatrix of S that removes the u-th row and the u-th column, let $\hat{K}' \in \mathbb{R}^{p-1\times p-1}$ be as that*

$$\begin{aligned}\hat{K}^\lambda := \underset{K}{\textit{argmax}} \ &\log\det(K) - \textit{trace}(KS_{[p]\setminus\{u\},[p]\setminus\{u\}}) + \lambda \sum_{i,j} K_{ij} \\ &\textit{s.t. } K_{ij} \leq 0 \quad \forall i \neq j\end{aligned}$$

for some $\lambda \geq 0$. Then $\hat{K}'$ satisfies that

$$\hat{K}' = (\hat{K}^\lambda)_{[p]\setminus\{u\},[p]\setminus\{u\}} - (\hat{K}^\lambda)_{[p]\setminus\{u\},u} \left((\hat{K}^\lambda)_{uu}\right)^{-1} (\hat{K}^\lambda)_{u,[p]\setminus\{u\}}.$$

Proof. This follows from the convergence of Algorithm 1 in [20]. □

Then we get the main proof of Theorem 4.4 as follows:

proof of Theorem 4.4. It is sufficient to prove the theorem by proving that for each edge (i,j) in MWSF(M^{-1}), the edge (i,j) is also in the support of M. To prove that each edge $(i,j) \in$ MWSF(M^{-1}) is also in the support of M, we consider two cases.

Case 1: $(M^{-1})_{ij} \leq \min_u (M^{-1})_{uu}$.

For any $\lambda \geq 0$, let $\bar{K}^\lambda$ be as that

$$\bar{K}^\lambda := \underset{K}{\text{argmax}} \ \log\det(K) - \text{trace}(KM^{-1}) + \lambda \sum_{i,j} K_{ij}$$
$$\textit{s.t. } K_{ij} \leq 0 \quad \forall i \neq j,$$

i.e., $\bar{K}^\lambda$ is the solution to (2) by taking M^{-1} as the input matrix with tuning parameter λ. Also, for any λ, let $\bar{E}^\lambda$ denote the estimated graphical model corresponding to $\bar{K}^\lambda$ and let $\bar{\Sigma}^\lambda := (\bar{K}^\lambda)^{-1}$. Apparently by choosing $\lambda_1 := 0$, the solution is $\bar{K}^{\lambda_1} = M$. In this case, if we can prove that there exists a $\lambda_0 > \lambda_1$ such that the edge $(i,j) \in \bar{E}^{\lambda_0}$, by applying Theorem 4.1, we can easily have that $(i,j) \in \bar{E}^{\lambda_1}$, or equivalent, the edge (i,j) is in the support of M.

Let C_1 and C_2 denote two disjoint sets of nodes that are complement with each other ($C_1 \cup C_2 = [p]$) and are separated in MWSF(M^{-1}) after removing the edge (i,j). In this case, it follows from the uniqueness of MWSF(M^{-1}) that M_{ij} is the bigger than any other $M_{\ell m}$'s where $\ell \in C_1$ and $m \in C_2$. Then by choosing the tuning parameter to be some $\lambda_2 > \lambda_1$ such that

$$\lambda_2 < (M^{-1})_{ij} \quad \text{and} \quad \lambda_2 > \max_{\ell \in C_1, m \in C_2 \ \& \ (\ell,m) \neq (i,j)} (M^{-1})_{\ell m},$$

it follows from the complementary slackness property as given in (5) that in the estimated support $\bar{E}^{\lambda_2}$, C_1 and C_2 would be either disconnected in $\bar{E}^{\lambda_2}$ or connected by only one edge (i,j). We can therefore prove that the edge $(i,j) \in \bar{E}^{\lambda_2}$ by proving that the two disjoint sets C_1 and C_2 are connected in $\bar{E}^{\lambda_2}$. We prove this by contradiction. If in contradiction C_1 and C_2 are disconnected, it is a short computation that $(\bar{\Sigma}^{\lambda_2})_{ij} = 0$. Since $(\bar{\Sigma}^{\lambda_2})_{ij} \geq (M^{-1})_{ij} - \lambda_2$, we can further conclude that $(M^{-1})_{ij} \leq \lambda_2$, which contradicts with the fact that $(M^{-1})_{ij} > \lambda_2$. Hence C_1 and C_2 are connected and furthermore, $(i,j) \in \bar{E}^{\lambda_2}$, which completes the proof for the case 1.

Case 2: There exist some $u \in [p]$ such that $(M^{-1})_{uu} < (M^{-1})_{ij}$.

Without loss of generality, we only consider the case where there is only one u such that $(M^{-1})_{uu} < (M^{-1})_{ij}$. Using that M is $(\sqrt{2}+1)$-SDD, we have that for any $v \neq u$, $|\frac{M_{vu}}{M_{vv}M_{uu}}| < \frac{1}{\sqrt{2}+1}\frac{1}{M_{uu}}$. By applying Lemma B.1, we have that $(M^{-1})_{vu} < (M^{-1})_{uu}$. Then by choosing the tuning parameter to be some $\lambda_3 > \lambda_1$ such that

$$\lambda_3 < (M^{-1})_{uu} \quad \text{and} \quad \lambda_3 > \max_{v \neq u} (M^{-1})_{vu},$$

it follows from the complementary slackness property as given in (5) that $(\hat{K}^{\lambda_3})_{uv} = 0$ for all $v \neq u$. In this case, by choosing the S, $S_{[p]\setminus\{u\},[p]\setminus\{u\}}$ and λ in Lemma B.2 to be M^{-1}, $(M^{-1})_{[p]\setminus\{u\},[p]\setminus\{u\}}$ and λ_3 respectively, we have that the $\hat{K}'$ as defined in Lemma B.2 satisfies that

$$\begin{aligned}\hat{K}' &= (\bar{K}^{\lambda_3})_{[p]\setminus\{u\},[p]\setminus\{u\}} - (\bar{K}^{\lambda_3})_{[p]\setminus\{u\},u}\Big((\bar{K}^{\lambda_3})_{uu}\Big)^{-1}(\bar{K}^{\lambda})_{u,[p]\setminus\{u\}} \\ &= (\bar{K}^{\lambda_3})_{[p]\setminus\{u\},[p]\setminus\{u\}},\end{aligned}$$

where the second equality is based on the fact that $(\bar{K}^{\lambda_3})_{[p]\setminus\{u\},[p]\setminus\{u\}} = 0$.

Since the edge $(i,j) \in \text{MWSF}((M^{-1})_{[p]\setminus\{u\},[p]\setminus\{u\}})$, it immediately follows from the proof of Case 1 that the edge (i,j) is also in the support of $\hat{K}'$, or equivalently the support of $(\bar{K}^{\lambda_3})_{[p]\setminus\{u\},[p]\setminus\{u\}}$. Hence, the edge (i,j) also exists in $\bar{E}^{\lambda_3}$. By again applying Theorem 4.1 we have that $(i,j) \in \bar{E}^{\lambda_1}$, or equivalently, in the support of M. □

C Theoretical analysis in Section 5

C.1 Proof of Theorem 5.1

It is a short computation from Lemma B.1 that for all entries (i,j) where $(\hat{K}^{\lambda})_{ij} = 0$,

$$(\hat{\Sigma}^{\lambda})_{ij} < \frac{1}{\gamma - 1} \max_{u \neq v} \Big| \frac{(\hat{K}^{\lambda})_{uv}}{(\hat{K}^{\lambda})_{uu}(\hat{K}^{\lambda})_{vv}} \Big|.$$

In this case, if we choose

$$\lambda' = \frac{1}{\gamma - 1} \max_{u \neq v} \Big| \frac{(\hat{K}^{\lambda})_{uv}}{(\hat{K}^{\lambda})_{uu}(\hat{K}^{\lambda})_{vv}} \Big| + \lambda,$$

we have that for any $(i,j) \notin \hat{E}^{\lambda}$, $S_{ij} < (\hat{\Sigma}^{\lambda})_{ij} + \lambda \leq \lambda'$. Therefore, $\hat{T}^{\lambda'} \subseteq \hat{E}^{\lambda}$. In addition, observe that for any $(r,\ell) \in \hat{E}^{\lambda}$ such that

$$\Big| \frac{(\hat{K}^{\lambda})_{r\ell}}{(\hat{K}^{\lambda})_{rr}(\hat{K}^{\lambda})_{\ell\ell}} \Big| \geq \frac{1}{\gamma - 1} \max_{u \neq v} \Big| \frac{(\hat{K}^{\lambda})_{uv}}{(\hat{K}^{\lambda})_{uu}(\hat{K}^{\lambda})_{vv}} \Big|, \tag{19}$$

we also have that

$$\begin{aligned} S_{r\ell} &\overset{(a)}{=} (\hat{\Sigma}^{\lambda})_{r\ell} + \lambda \overset{(b)}{\geq} \Big| \frac{(\hat{K}^{\lambda})_{r\ell}}{(\hat{K}^{\lambda})_{rr}(\hat{K}^{\lambda})_{\ell\ell}} \Big| + \lambda \\ &\geq \frac{1}{\gamma - 1} \max_{u \neq v} \Big| \frac{(\hat{K}^{\lambda})_{uv}}{(\hat{K}^{\lambda})_{uu}(\hat{K}^{\lambda})_{vv}} \Big| + \lambda = \lambda', \end{aligned}$$

where the equality (a) is based on the complementary slackness property (5) and inequality (b) is based on (19). Hence, we also have that $(r,\ell) \in \hat{T}^{\lambda'}$, which proves the second statement in Theorem 5.1.

C.2 Proof of Corollary 5.2

Proof. By replacing the $\hat{K}^{\lambda}$ with K^*, it directly follows from the proof of Theorem 5.1 that for all the off-diagonal entries (i,j) where

$$\Sigma^*_{ij} \geq \frac{1}{\gamma - 1} \max_{u \neq v} \Big| \frac{K^*_{uv}}{K^*_{uu}K^*_{vv}} \Big|,$$

we also have that $(i,j) \in E^*$. By combining this with the fact that there exist some positive constants C, τ such that with probability $1 - p^{-\tau}$, $\|S - \Sigma^*\|_{\infty} \leq C\sqrt{\frac{\log p}{n}}$, it is a short computation that $\hat{T}^{\lambda} \subseteq E^*$ with probability $1 - p^{-\tau}$. In addition, for all the (r,ℓ)'s where

$$\Big| \frac{K^*_{r\ell}}{K^*_{rr}K^*_{\ell\ell}} \Big| \geq \frac{1}{\gamma - 1} \max_{u \neq v} \Big| \frac{K^*_{uv}}{K^*_{uu}K^*_{vv}} \Big| + 2C\sqrt{\frac{\log p}{n}},$$

it is a short computation from Lemma B.1 that

$$\Sigma^*_{r\ell} \geq \frac{1}{\gamma - 1} \max_{u \neq v} \left| \frac{K^*_{uv}}{K^*_{uu} K^*_{vv}} \right| + 2C \sqrt{\frac{\log p}{n}}.$$

Using that $\|S - \Sigma^*\|_\infty \leq C \sqrt{\frac{\log p}{n}}$, we further have that

$$S_{r\ell} \geq \frac{1}{\gamma - 1} \max_{u \neq v} \left| \frac{K^*_{uv}}{K^*_{uu} K^*_{vv}} \right| + C \sqrt{\frac{\log p}{n}},$$

which completes the proof. □

D Theoretical analysis in Section 6

D.1 Proof of Corollary 6.1

Just as in proving Theorem 3.1, to prove Corollary 6.1, we first consider proving the following corollary.

Corollary D.1. *Consider a Gaussian distribution satisfying either of the following two conditions for some* $0 < \alpha < 1$*:*

- *the inverse Isserlis matrix* Γ^* *is* $\frac{1+\alpha}{1-\alpha}$*-SDD;*
- *the precision matrix* K^* *is* $\frac{1+\alpha+\sqrt{2(1+\alpha)}}{1-\alpha}$*-SDD.*

Then, if $\delta_{n,p}$ *is upper bounded by* $\delta_{n,p} \leq \frac{\alpha}{8M^*}$*, the solution to* (2) *with tuning parameter* $\lambda = \alpha^{-1}\delta_{n,p}$ *satisfies the following* ℓ_∞ *bound:*

$$\|\hat{K}^\lambda - K^*\|_\infty \leq r,$$

where $r := 8M^{*2}\lambda$*. Moreover, the support of* $\hat{K}^\lambda$ *is a subset of the support of* K^* *and includes all the edges such that* $|(K^*)_{ij}| > r$*.*

With a slight abuse of notation, we redefine $S^\nu \in \mathbb{R}^{p \times p}$ as that $S^\nu := (1 - \nu)\Sigma^* + \nu S$. To prove Corollary D.1, we propose a new estimator $\tilde{K}^\nu$ as that

$$\tilde{K}^\nu := \operatorname*{argmax}_K \log\det(K) - \operatorname{trace}(K \cdot S^\nu) - \lambda \sum_{i \neq j} \nu |K_{ij}| + \lambda\nu \sum_i K_{ii}. \tag{20}$$

It is a short computation from (20) that $\tilde{K}^\nu\big|_{\nu=0} = K^*$ and $\tilde{K}^\nu\big|_{\nu=1} = \tilde{K}^\lambda$. In this case, again we can understand the consistency guarantees of $\tilde{K}^\lambda$ by understanding how $\tilde{K}^\nu$ is changed as ν is increased from $\nu = 0$ to $\nu = 1$, just as in proving Theorem A.1. Let $\tilde{\Gamma}^\nu := \tilde{K}^\nu \otimes \tilde{K}^\nu$ and $\tilde{H}^\nu := \tilde{\Sigma}^\nu \otimes \tilde{\Sigma}^\nu$. In order to prove Corollary D.1, we still need to prove the following lemmas:

Lemma D.2. *$\tilde{K}^\nu$ is unique and continuous in ν over the domain $[0, 1]$.*

Proof. We first consider uniqueness. One can rewrite (20) as that

$$\underset{K}{\text{maximize}} \ \log\det(K) - \text{trace}\Big(K \cdot (S^\nu - \lambda\nu I)\Big) - \lambda\nu \sum_{i\neq j} |K_{ij}|$$

where $I \in \mathbb{R}^{p\times p}$ is the identity matrix. This allows us to think of $\tilde{K}^\nu$ as the global optimum of graphical lasso by choosing $\lambda\nu$ as the penalization parameter and choosing $S^\nu - \lambda\nu I$ as the input sample correlation matrix. It then follows from Lemma 3 of [30] that $\tilde{K}^\nu$ is unique as long as $\lambda\nu < 1$.

For continuity, the continuity of (20) follows immediately from the continuity of graphical lasso as given in [14]. □

Lemma D.3. *Assume $\tilde{\Gamma}^\nu$ is $\frac{1+\alpha}{1-\alpha}$-SDD, then there exists a $\Delta\nu$ small enough such that for all $\tilde{\nu} > \nu$ while $\tilde{\nu} - \nu \le \Delta\nu$, $\tilde{E}^{\tilde{\nu}} \subseteq \tilde{E}^\nu$. Moreover, $|(\tilde{K}^{\tilde{\nu}})_{ij}| < |(\tilde{K}^\nu)_{ij}|$ for all $(i,j) \in \tilde{E}^\nu$; $(\tilde{K}^{\tilde{\nu}})_{ii} > (\tilde{K}^\nu)_{ii}$ for all $i \in [p]$.*

Proof. It follows from standard results in graphical lasso [30] that the global optimum $\tilde{K}^\nu$ and its inverse $\tilde{\Sigma}^\nu$ follows the following sub-differential condition:

$$\begin{aligned}
(\tilde{\Sigma}^\nu)_{ii} &= (S^\nu)_{ii} - \lambda\nu \\
(\tilde{\Sigma}^\nu)_{ij} &= (S^\nu)_{ij} + \lambda\nu \text{sign}((\tilde{K}^\nu)_{ij}) \quad \text{if } i \neq j \ \&\ (\tilde{K}^\nu)_{ij} \neq 0 \\
(S^\nu)_{ij} - \lambda\nu \le (\tilde{\Sigma}^\nu)_{ij} &\le (S^\nu)_{ij} + \lambda\nu \quad \text{if } i \neq j \ \&\ (\tilde{K}^\nu)_{ij} = 0.
\end{aligned}$$

We therefore divide all the indices into the following five subsets:

$$\begin{aligned}
\mathcal{A}_+ :=& \{(i,j) : (\tilde{\Sigma}^\nu)_{ij} = S^\nu + \lambda\nu \text{ and } (\tilde{K}^\nu)_{ij} > 0\}, \\
\mathcal{A}_- :=& \{(i,j) : (\tilde{\Sigma}^\nu)_{ij} = S^\nu - \lambda\nu \text{ and } (\tilde{K}^\nu)_{ij} < 0\}, \\
\mathcal{I} :=& \{(i,j) : S^\nu - \lambda\nu < (\tilde{\Sigma}^\nu)_{ij} < S^\nu + \lambda\nu \text{ and } (\tilde{K}^\nu)_{ij} = 0\}, \\
\mathcal{B}_+ :=& \{(i,j) : (\tilde{\Sigma}^\nu)_{ij} = S^\nu + \lambda\nu \text{ and } (\tilde{K}^\nu)_{ij} = 0\}, \\
\mathcal{B}_- :=& \{(i,j) : (\tilde{\Sigma}^\nu)_{ij} = S^\nu - \lambda\nu \text{ and } (\tilde{K}^\nu)_{ij} = 0\}.
\end{aligned}$$

Similarly, we also divide indices of $\tilde{K}^{\tilde{\nu}}$ into $\tilde{\mathcal{A}}_+$, $\tilde{\mathcal{A}}_-$, $\tilde{\mathcal{I}}$, $\tilde{\mathcal{B}}_+$ and $\tilde{\mathcal{B}}_-$ accordingly. In this case, let $\mathcal{A} := \mathcal{A}_+ \cup \mathcal{A}_- \cup \{(i,i) : i \in [p]\}$ and $\tilde{\mathcal{A}} := \tilde{\mathcal{A}}_+ \cup \tilde{\mathcal{A}}_- \cup \{(i,i) : i \in [p]\}$, it immediately follows from the proof of Lemma A.5 that with $T := \mathcal{A} \cup \tilde{\mathcal{A}}$,

$$[\tilde{K}^{\tilde{\nu}} - \tilde{K}^\nu]_T = \Big((\tilde{H}^\nu)_{TT}\Big)^{-1} [\tilde{\Sigma}^\nu - \tilde{\Sigma}^{\tilde{\nu}}]_T + o(\tilde{\nu} - \nu),$$

where $\Big((\tilde{H}^\nu)_{TT}\Big)^{-1}$ is $\frac{1+\alpha}{1-\alpha}$-SDD, and that

$$\begin{aligned}
(1-\alpha)(\tilde{\nu}-\nu)\lambda \le (\tilde{\Sigma}^\nu - \tilde{\Sigma}^{\tilde{\nu}})_{ij} &\le (1+\alpha)(\tilde{\nu}-\nu)\lambda \qquad \text{if } i = j \text{ or } (i,j) \in \mathcal{A}_- \cup \tilde{\mathcal{A}}_- \\
-(1+\alpha)(\tilde{\nu}-\nu)\lambda \le (\tilde{\Sigma}^\nu - \tilde{\Sigma}^{\tilde{\nu}})_{ij} &\le -(1-\alpha)(\tilde{\nu}-\nu)\lambda \qquad \text{if } (i,j) \in \mathcal{A}_+ \cup \tilde{\mathcal{A}}_+.
\end{aligned}$$

By combining them together, it is a short computation to discover that $\Big((\tilde{H}^\nu)_{TT}\Big)^{-1} [\tilde{\Sigma}^\nu - \tilde{\Sigma}^{\tilde{\nu}}]_T$ is a fully positive vector. Hence, $[\tilde{K}^{\tilde{\nu}} - \tilde{K}^\nu]_T$ is also a fully positive vector. This gives us that $(\tilde{K}^{\tilde{\nu}})_{ij} > (\tilde{K}^\nu)_{ij}$ for all $i = j$ and for all $(i,j) \in \mathcal{A}_- \cup \tilde{\mathcal{A}}_-$. In addition, we also have that $(\tilde{K}^{\tilde{\nu}})_{ij} < (\tilde{K}^\nu)_{ij}$ for all $(i,j) \in \mathcal{A}_+ \cup \tilde{\mathcal{A}}_+$. In this case, following from the proof of Lemma A.5, we further have that $\tilde{\mathcal{A}}_- \subseteq \mathcal{A}_-$ and $\tilde{\mathcal{A}}_+ \subseteq \mathcal{A}_+$ and hence $\tilde{E}^{\tilde{\nu}} \subseteq \tilde{E}^\nu$. □

Lemma D.4. *Assume $\tilde{\Gamma}^\nu$ is $\frac{1+\alpha}{1-\alpha}$-SDD, then the right-hand and right-hand derivatives satisfy that*

$$\frac{d|(\hat{K}^\nu)_{ij}|}{d\nu_+} \geq -2\hat{M}^2\lambda \text{ for } i \neq j \quad \text{and} \quad \frac{d(\hat{K}^\nu)_{ii}}{d\nu_+} \leq 2\hat{M}^2\lambda;$$

and that

$$\frac{d|(\hat{K}^\nu)_{ij}|}{d\nu_-} \geq -2\hat{M}^2\lambda \text{ for } i \neq j \quad \text{and} \quad \frac{d(\hat{K}^\nu)_{ii}}{d\nu_-} \leq 2\hat{M}^2\lambda.$$

Proof. It immediately follows from the proof of Lemmas D.3 and A.6. □

Then we have the formal proof of Corollary D.1 as follows

Proof. Given Lemmas D.2, D.3 and D.4, our proof of Corollary D.1 follows from the proof of Theorem 3.1. □

D.2 Proof of Corollary 6.2

For any $\tilde{\lambda} > \lambda$, we redefine $\tilde{K}^\nu$ as that

$$\tilde{K}^\nu := \operatorname*{argmax}_K \log\det(K) - \text{trace}(K \cdot S) - \big(\lambda + (\tilde{\lambda} - \lambda)\nu\big)\sum_{i \neq j}|K_{ij}| + \big(\lambda + (\tilde{\lambda} - \lambda)\nu\big)\sum_i K_{ii}.$$

It is then a short computation that $\tilde{K}^\nu\big|_{\nu=0} = \tilde{K}^\lambda$ and $\tilde{K}^\nu\big|_{\nu=1} = \tilde{K}^{\tilde{\lambda}}$. This allows us to prove the robustness guarantee by understanding how $\tilde{K}^\nu$ is changed as ν is increased from 0 to 1. Then the robustness guarantee immediately follows from the proof of Corollary D.1.