ASYMPTOTIC NORMALITY AND OPTIMALITIES IN ESTIMATION OF LARGE GAUSSIAN GRAPHICAL MODEL

By Zhao Ren [*,‡], Tingni Sun[§], Cun-Hui Zhang[†,¶] and Harrison H. Zhou[*,‡]

Yale University[‡], *University of Pennsylvania*[§] *and Rutgers University*[¶]

The Gaussian graphical model, a popular paradigm for studying relationship among variables in a wide range of applications, has attracted great attention in recent years. This paper considers a fundamental question: When is it possible to estimate low-dimensional parameters at parametric square-root rate in a large Gaussian graphical model? A novel regression approach is proposed to obtain asymptotically efficient estimation of each entry of a precision matrix under a sparseness condition relative to the sample size. When the precision matrix is not sufficiently sparse, or equivalently the sample size is not sufficiently large, a lower bound is established to show that it is no longer possible to achieve the parametric rate in the estimation of each entry. This lower bound result, which provides an answer to the delicate sample size question, is established with a novel construction of a subset of sparse precision matrices in an application of Le Cam's Lemma. Moreover, the proposed estimator is proven to have optimal convergence rate when the parametric rate cannot be achieved, under a minimal sample requirement.

The proposed estimator is applied to test the presence of an edge in the Gaussian graphical model or to recover the support of the entire model, to obtain adaptive rate-optimal estimation of the entire precision matrix as measured by the matrix l_q operator norm, and to make inference in latent variables in the graphical model. All these are achieved under a sparsity condition on the precision matrix and a side condition on the range of its spectrum. This significantly relaxes the commonly imposed uniform signal strength condition on the precision matrix, irrepresentable condition on the Hessian tensor operator of the covariance matrix or the ℓ_1 constraint on the precision matrix.

[*] The research of Zhao Ren and Harrison H. Zhou was supported in part by NSF Career Award DMS-0645676 and NSF FRG Grant DMS-0854975.

[†] The research of Cun-Hui Zhang was supported in part by the NSF Grants DMS 0906420, DMS-11-06753 and NSA Grant H98230-11-1-0205.

AMS 2000 subject classifications: Primary 62H12; secondary 62F12, 62G09

Keywords and phrases: asymptotic efficiency, covariance matrix, inference, graphical model, latent graphical model, minimax lower bound, optimal rate of convergence, scaled lasso, precision matrix, sparsity, spectral norm

Numerical results confirm our theoretical findings. The ROC curve of the proposed algorithm, Asymptotic Normal Thresholding (ANT), for support recovery significantly outperforms that of the popular GLasso algorithm.

1. Introduction. Gaussian graphical model, a powerful tool for investigating the relationship among a large number of random variables in a complex system, is used in a wide range of scientific applications. A central question for Gaussian graphical model is to recover the structure of an undirected Gaussian graph. Let $G = (V, E)$ be an undirected graph representing the conditional dependence relationship between components of a random vector $Z = (Z_1, \ldots, Z_p)^T$ as follows. The vertex set $V = \{V_1, \ldots, V_p\}$ represents the components of Z. The edge set E consists of pairs (i, j) indicating the conditional dependence between Z_i and Z_j given all other components. In applications, the following question is fundamental: Is there an edge between V_i and V_j? It is well known that recovering the structure of an undirected Gaussian graph $G = (V, E)$ is equivalent to recovering the support of the population precision matrix of the data in the Gaussian graphical model. Let

$$Z = (Z_1, Z_2, \ldots, Z_p)^T \sim \mathcal{N}(\mu, \Sigma),$$

where $\Sigma = (\sigma_{ij})$ is the population covariance matrix. The precision matrix, denoted by $\Omega = (\omega_{ij})$, is defined as the inverse of covariance matrix, $\Omega = \Sigma^{-1}$. There is an edge between V_i and V_j, i.e., $(i, j) \in E$, if and only if $\omega_{ij} \neq 0$. See, for example, Lauritzen (1996). Consequently, the support recovery of the precision matrix Ω yields the recovery of the structure of the graph G.

Suppose n i.i.d. p-variate random vectors $X^{(1)}, X^{(2)}, \ldots, X^{(n)}$ are observed from the same distribution as Z, i.e. the Gaussian $\mathcal{N}(\mu, \Omega^{-1})$. Assume without loss of generality that $\mu = 0$ hereafter. In this paper, we address the following two fundamental questions: When is it possible to make statistical inference for each individual entry of a precision matrix Ω at the parametric $\sqrt{n}$ rate? When and in what sense is it possible to recover the support of Ω in the presence of some small nonzero $|\omega_{ij}|$?

The problems of estimating a large sparse precision matrix and recovering its support have drawn considerable recent attention. There are mainly two approaches in literature. The first one is a penalized likelihood estimation approach with a lasso-type penalty on entries of the precision matrix. Yuan and Lin (2007) proposed to use the lasso penalty and studied its asymptotic properties when p is fixed. Ravikumar et al. (2011) derived the rate of convergence when the dimension p is high by applying a primal-dual witness construction under an irrepresentability condition on the Hessian tensor operator and a constraint on the matrix l_1 norm of the precision matrix. See also Rothman et al. (2008) and Lam and Fan (2009) for other related results. The other one is the neighborhood-based approach, by running a lasso-type regression or Dantzig selection type of each variable on all the rest of variables to estimate precision matrix column by column. See Meinshausen and Bühlmann (2006), Yuan (2010), Cai, Liu and Luo (2011), Cai, Liu and Zhou (2012) and Sun and Zhang (2012a). The irrepresentability condition is no longer needed in Cai, Liu and Luo (2011) and Cai, Liu and Zhou (2012) for support recovery, but the thresholding level for support recovery depends on the matrix l_1 norm of the precision matrix. The matrix l_1 norm is unknown and large, which makes the support recovery procedures there nonadaptive and thus less practical. In Sun and Zhang (2012a), optimal convergence rate in the spectral norm is achieved without requiring the matrix ℓ_1 norm constraint or the irrepresentability condition. However, support recovery properties of the estimator was not analyzed.

In spite of an extensive literature on the topic, it is still largely unknown the fundamental limit of support recovery in the Gaussian graphical model, let alone an adaptive procedure to achieve the limit.

Statistical inference of low-dimensional parameters at the $\sqrt{n}$ rate has been considered in the closely related linear regression model. Sun and Zhang (2012b) proposed an efficient scaled Lasso estimator of the noise level under the sample size condition $n \gg (s \log p)^2$, where s is the ℓ_0 or capped-ℓ_1 measure of the size of the unknown regression vector. Zhang and Zhang (2011) proposed an asymptotically normal low-dimensional projection estimator for the regression coefficients and their estimator was proven to be asymptotically efficient by van de Geer, Bühlmann and Ritov (2013) in a semiparametric sense under the same sample size condition. The asymptotic efficiency of these estimators can be also understood through the minimum Fisher information in a more general context (Zhang, 2011). Alternative methods for testing and estimation of regression coefficients were proposed in Belloni, Chernozhukov and Hansen (2012), Bühlmann (2012), and Javanmard and Montanari (2013). However, the optimal rate of convergence is unclear from these papers when the sample size condition $n \gg (s \log p)^2$ fails to hold.

This paper makes important advancements in the understanding of statistical inference of low-dimensional parameters in the Gaussian graphical model in the following ways. Let s be the maximum degree of the graph or a certain more relaxed capped-ℓ_1 measure of the complexity of the precision matrix. We prove that the estimation of each ω_{ij} at the parametric $\sqrt{n}$ convergence rate requires the sparsity condition $s \leq O(1)n^{1/2}/\log p$

or equivalently a sample size of order $(s\log p)^2$. We propose an adaptive estimator of individual ω_{ij} and prove its asymptotic normality and efficiency when $n \gg (s\log p)^2$. Moreover, we prove that the proposed estimator achieves the optimal convergence rate when the sparsity condition is relaxed to $s \le c_0 n/\log p$ for a certain positive constant c_0. The efficient estimator of the individual ω_{ij} is then used to construct fully data driven procedures to recover the support of Ω and to make statistical inference about latent variables in the graphical model.

The methodology we are proposing is a novel regression approach briefly described in Sun and Zhang (2012c). In this regression approach, the main task is not to estimate the slope as seen in Meinshausen and Bühlmann (2006), Yuan (2010), Cai, Liu and Luo (2011), Cai, Liu and Zhou (2012) and Sun and Zhang (2012b), but to estimate the noise level. For any index subset A of $\{1, 2, \ldots, p\}$ and a vector Z of length p, we use Z_A to denote a vector of length $|A|$ with elements indexed by A. Similarly for a matrix U and two index subsets A and B of $\{1, 2, \ldots, p\}$ we can define a submatrix $U_{A,B}$ of size $|A| \times |B|$ with rows and columns of U indexed by A and B respectively. Consider $A = \{i, j\}$, for example, $i = 1$ and $j = 2$, then $Z_A = (Z_1, Z_2)^T$ and $\Omega_{A,A} = \begin{pmatrix} \omega_{11} & \omega_{12} \\ \omega_{21} & \omega_{22} \end{pmatrix}$. It is well known that

$$Z_A | Z_{A^c} = \mathcal{N}\left(-\Omega_{A,A}^{-1}\Omega_{A,A^c} Z_{A^c}, \Omega_{A,A}^{-1}\right).$$

This observation motivates us to consider regression with two response variables above. The noise level $\Omega_{A,A}^{-1}$ has only three parameters. When Ω is sufficiently sparse, a penalized regression approach is proposed in Section 2 to obtain an asymptotically efficient estimation of ω_{ij}, i.e., the estimator is asymptotically normal and the variance matches that of the maximum likelihood estimator in the classical setting where the dimension p is a fixed constant. Consider the class of parameter spaces modeling sparse precision matrices with at most $k_{n,p}$ off-diagonal nonzero elements in each column,

$$\mathcal{G}_0(M, k_{n,p}) = \left\{ \begin{array}{c} \Omega = (\omega_{ij})_{1\le i,j\le p} : \max_{1\le j\le p} \sum_{i\ne j} 1\{\omega_{ij} \ne 0\} \le k_{n,p}, \\ \text{and } 1/M \le \lambda_{\min}(\Omega) \le \lambda_{\max}(\Omega) \le M. \end{array} \right\}, \tag{1}$$

where $1\{\cdot\}$ is the indicator function and M is some constant greater than 1. The following theorem shows that a necessary and sufficient condition to obtain a $\sqrt{n}-$consistent estimation of ω_{ij} is $k_{n,p} = O\left(\frac{\sqrt{n}}{\log p}\right)$, and when $k_{n,p} = o\left(\frac{\sqrt{n}}{\log p}\right)$ the procedure to be proposed in Section 2 is asymptotically efficient.

Theorem 1. *Let $X^{(i)} \overset{i.i.d.}{\sim} \mathcal{N}_p(\mu, \Sigma)$, $i = 1, 2, \ldots, n$. Assume that $k_{n,p} \le c_0 n/\log p$ with a sufficiently small constant $c_0 > 0$ and $p \ge k_{n,p}^{\nu}$ with some $\nu > 2$. We have the following* *probablistic results,*

(i). *There exists a constant $\epsilon_0 > 0$ such that*

$$\inf_{i,j} \inf_{\hat{\omega}_{ij}} \sup_{\mathcal{G}_0(M,k_{n,p})} \mathbb{P}\Big\{ |\hat{\omega}_{ij} - \omega_{ij}| \geq \epsilon_0 \max\left\{n^{-1}k_{n,p}\log p, n^{-1/2}\right\}\Big\} \geq \epsilon_0.$$

(ii). *The estimator $\hat{\omega}_{ij}$ defined in (12) is rate optimal in the sense of*

$$\max_{i,j} \sup_{\mathcal{G}_0(M,k_{n,p})} \mathbb{P}\Big\{ |\hat{\omega}_{ij} - \omega_{ij}| \geq M \max\left\{n^{-1}k_{n,p}\log p, n^{-1/2}\right\}\Big\} \to 0,$$

as $(M, n) \to (\infty, \infty)$. Furthermore, the estimator $\hat{\omega}_{ij}$ is asymptotically efficient when $k_{n,p} = o\left(\frac{\sqrt{n}}{\log p}\right)$, i.e., with $F_{ij}^{-1} = \omega_{ii}\omega_{jj} + \omega_{ij}^2$,

$$\sqrt{nF_{ij}}\left(\hat{\omega}_{ij} - \omega_{ij}\right) \xrightarrow{D} \mathcal{N}(0,1)\,. \tag{2}$$

Moreover, the minimax risk of estimating ω_{ij} over the class $\mathcal{G}_0(k, M_{n,p})$ satisfies, provided $n = O\left(p^{\xi}\right)$ with some $\xi > 0$,

$$\inf_{\hat{\omega}_{ij}} \sup_{\mathcal{G}_0(M,k_{n,p})} \mathbb{E}\,|\hat{\omega}_{ij} - \omega_{ij}| \asymp \max\left\{k_{n,p}\frac{\log p}{n}, \sqrt{\frac{1}{n}}\right\}. \tag{3}$$

The lower bound is established through Le Cam's Lemma and a novel construction of a subset of sparse precision matrices. An important implication of the lower bound is that the difficulty of support recovery for sparse precision matrix is different from that for sparse covariance matrix when $k_{n,p} \gg \left(\frac{\sqrt{n}}{\log p}\right)$, and when $k_{n,p} \ll \left(\frac{\sqrt{n}}{\log p}\right)$ the difficulty of support recovery for sparse precision matrix is just the same as that for sparse covariance matrix.

It is worthwhile to point out that the asymptotic efficiency result is obtained without the need to assume the irrepresentable condition or the l_1 constraint of the precision matrix which are commonly required in literature. An immediate consequence of the asymptotic normality result (2) is to test individually whether there is an edge between V_i and V_j in the set E, i.e., the hypotheses $\omega_{ij} = 0$. The result is applied to do adaptive support recovery optimally. In addition, we can strengthen other results in literature under weaker assumptions, and the procedures are adaptive, including adaptive rate-optimal estimation of the precision matrix under various matrix l_q norms, and an extension of our framework for inference and estimation to a class of latent variable graphical models. See Section 3 for details.

Our work on optimal estimation of precision matrix given in the present paper is closely connected to a growing literature on estimation of large covariance matrices. Many regularization methods have been proposed and studied. For example, Bickel and Levina

(2008a,b) proposed banding and thresholding estimators for estimating bandable and sparse covariance matrices respectively and obtained rate of convergence for the two estimators. See also El Karoui (2008) and Lam and Fan (2009). Cai, Zhang and Zhou (2010) established the optimal rates of convergence for estimating bandable covariance matrices. Cai and Zhou (2012) and Cai, Liu and Zhou (2012) obtained the minimax rate of convergence for estimating sparse covariance and precision matrices under a range of losses including the spectral norm loss. In particular, a new general lower bound technique for matrix estimation was developed there. See also Sun and Zhang (2012a).

The proposed estimator was briefly described in Sun and Zhang (2012c) along with a statement of the efficiency of the estimator without proof under the sparsity assumption $k_{n,p} \ll n^{-1/2} \log p$. While we are working on the delicate issue of the necessity of the sparsity condition $k_{n,p} \ll n^{1/2}/\log p$ and the optimality of the method for support recovery and estimation under the general sparsity condition $k_{n,p} \ll n/\log p$, Liu (2013) developed p-values for testing $\omega_{ij} = 0$ and related FDR control methods under the stronger sparsity condition $k_{n,p} \ll n^{1/2}/\log p$. However, his method cannot be converted into confidence intervals, and the optimality of his method is unclear under either sparsity conditions.

The paper is organized as follows. In Section 2, we introduce our methodology and main results for statistical inference. Applications to estimation under the spectral norm, to support recovery and estimation of latent variable graphical model are presented in Section 3. Section 4 discusses extensions of results in Sections 2 and 3. Numerical studies are given in Section 5. Proofs for theorems in Sections 2-3 are given in Sections 6-7. Proofs for main lemmas are given in Section 8. We collect auxiliary results for proving main lemmas in the supplementary material.

Notations. We summarize here some notations to be used throughout the paper. For $1 \leq w \leq \infty$, we use $\|u\|_w$ and $\|A\|_w$ to denote the usual vector l_w norm, given a vector $u \in \mathbb{R}^p$ and a matrix $A = (a_{ij})_{p \times p}$ respectively. In particular, $\|A\|_\infty$ denote the entry-wise maximum $\max_{ij} |a_{ij}|$. We shall write $\|\cdot\|$ without a subscript for the vector l_2 norm. The matrix ℓ_w operator norm of a matrix A is defined by $|||A|||_w = \max_{\|x\|_w=1} \|Ax\|_w$. The commonly used spectral norm $|||\cdot|||$ coincides with the matrix ℓ_2 operator norm $|||\cdot|||_2$.

2. Methodology and Statistical Inference.

In this section we will introduce our methodology for estimating each entry and more generally, a smooth functional of any square submatrix of finite size. Asymptotic efficiency results are stated in Section 2.2 under a sparseness assumption. The lower bound in Section 2.3 shows that the sparseness condition to obtain the asymptotic efficiency in Section 2.2 is sharp.

2.1. *Methodology.* We will first introduce the methodology to estimate each entry ω_{ij}, and discuss its extension to the estimation of functionals of a submatrix of the precision matrix.

The methodology is motivated by the following simple observation with $A = \{i, j\}$,

$$Z_{\{i,j\}} | Z_{\{i,j\}^c} = \mathcal{N}\left(-\Omega_{A,A}^{-1}\Omega_{A,A^c}Z_{\{i,j\}^c}, \Omega_{A,A}^{-1}\right). \tag{4}$$

Equivalently we write

$$(Z_i, Z_j) = Z_{\{i,j\}^c}^T\beta + (\eta_i, \eta_j), \tag{5}$$

where the coefficients and error distributions are

$$\beta = -\Omega_{A^c,A}\Omega_{A,A}^{-1}, \quad (\eta_i, \eta_j)^T \sim \mathcal{N}\left(0, \Omega_{A,A}^{-1}\right). \tag{6}$$

Denote the covariance matrix of $(\eta_i, \eta_j)^T$ by

$$\Theta_{A,A} = \Omega_{A,A}^{-1} = \begin{pmatrix} \theta_{ii} & \theta_{ij} \\ \theta_{ji} & \theta_{jj} \end{pmatrix}.$$

We will estimate $\Theta_{A,A}$ and expect that an efficient estimator of $\Theta_{A,A}$ yields an efficient estimation of entries of $\Omega_{A,A}$ by inverting the estimator of $\Theta_{A,A}$.

Denote the n by p dimensional data matrix by $\mathbf{X}$. The ith row of data matrix is the ith sample $X^{(i)}$. Let $\mathbf{X}_A$ be the columns indexed by $A = \{i, j\}$. Based on the regression interpretation (5), we have the following data version of the multivariate regression model

$$\mathbf{X}_A = \mathbf{X}_{A^c}\beta + \epsilon_A. \tag{7}$$

Here each row of (7) is a sample of the linear model (5). Note that β is a $p-2$ by 2 dimensional coefficient matrix. Denote a sample version of $\Theta_{A,A}$ by

$$\Theta_{A,A}^{ora} = \left(\theta_{kl}^{ora}\right)_{k\in A, l\in A} = \epsilon_A^T\epsilon_A/n \tag{8}$$

which is an oracle MLE of $\Theta_{A,A}$, assuming that we know β, and

$$\Omega_{A,A}^{ora} = \left(\omega_{kl}^{ora}\right)_{k\in A, l\in A} = \left(\Theta_{A,A}^{ora}\right)^{-1}. \tag{9}$$

But of course β is unknown, and we will need to estimate β and plug in its estimator to estimate ϵ_A.

Now we formally introduce the methodology. For each $m \in A = \{i, j\}$, we apply a scaled lasso penalization to the univariate linear regression of $\mathbf{X}_m$ against $\mathbf{X}_{A^c}$ as follows,

$$\left\{\hat{\beta}_m, \hat{\theta}_{mm}^{1/2}\right\} = \arg \min_{b\in\mathbb{R}^{p-2}, \sigma\in\mathbb{R}^+} \left\{ \frac{\|\mathbf{X}_m - \mathbf{X}_{A^c}b\|^2}{2n\sigma} + \frac{\sigma}{2} + \lambda \sum_{k\in A^c} \frac{\|\mathbf{X}_k\|}{\sqrt{n}} |b_k| \right\}, \tag{10}$$

with a weighted ℓ_1 penalty, where the vector b is indexed by A^c. The penalty level will be specified explicitly later. Define the residuals of the scaled lasso regression by

$$\hat{\epsilon}_A = \mathbf{X}_A - \mathbf{X}_{A^c}\hat{\beta},$$

and

$$\hat{\Theta}_{A,A} = \hat{\epsilon}_A^T \hat{\epsilon}_A / n. \tag{11}$$

It can be shown that this definition of $\hat{\theta}_{mm}$ is consistent with the $\hat{\theta}_{mm}$ obtained from the scaled lasso (10) for each $m \in A$. Finally we simply inverse the estimator $\hat{\Theta}_{A,A} = \left(\hat{\theta}_{kl}\right)_{k,l\in A}$ to estimate entries in $\Omega_{A,A}$, i.e.

$$\hat{\Omega}_{A,A} = \hat{\Theta}_{A,A}^{-1}. \tag{12}$$

This methodology can be routinely extended into a more general form. For any subset $B \subset \{1, 2, \ldots, p\}$ with a bounded size, the conditional distribution of Z_B given Z_{B^c} is

$$Z_B | Z_{B^c} = \mathcal{N}\left(-\Omega_{B,B}^{-1}\Omega_{B,B^c}Z_{B^c}, \Omega_{B,B}^{-1}\right), \tag{13}$$

so that the associated multivariate linear regression model is $\mathbf{X}_B = \mathbf{X}_{B^c}\beta_{B,B^c} + \epsilon_B$ with $\beta_{B^c,B} = -\Omega_{B^c,B}\Omega_{B,B}^{-1}$ and $\epsilon_B \sim \mathcal{N}(0, \Omega_{B,B}^{-1})$. Consider a slightly more general problem of estimating a smooth functional of $\Omega_{B,B}^{-1}$, denoted by

$$\zeta := \zeta\left(\Omega_{B,B}^{-1}\right).$$

When $\beta_{B^c,B}$ is known, ϵ_B is sufficient for $\Omega_{B,B}^{-1}$ due to the independence of ϵ_B and $\mathbf{X}_{B^c}$, so that an oracle estimator of ζ can be defined as

$$\zeta^{ora} = \zeta\left(\epsilon_B^T \epsilon_B / n\right).$$

We apply scaled lasso to the univariate linear regression of $\mathbf{X}_m$ against $\mathbf{X}_{B^c}$ for each $m \in B$ as in Equation (10),

$$\left\{\hat{\beta}_m, \hat{\theta}_{mm}^{1/2}\right\} = \arg \min_{b\in\mathbb{R}^{p-|B|}, \sigma\in\mathbb{R}^+} \left\{\frac{\|\mathbf{X}_m - \mathbf{X}_{B^c}b\|^2}{2n\sigma} + \frac{\sigma}{2} + \lambda \sum_{k\in B^c} \frac{\|\mathbf{X}_k\|}{\sqrt{n}} |b_k|\right\}$$

where $|B|$ is the size of subset B. The residual matrix of the model is $\hat{\epsilon}_B = \mathbf{X}_B - \mathbf{X}_{B^c}\hat{\beta}_{B^c,B}$, then the scaled lasso estimator of $\zeta\left(\Omega_{B,B}^{-1}\right)$ is defined by

$$\hat{\zeta} = \zeta\left(\hat{\epsilon}_B^T \hat{\epsilon}_B / n\right). \tag{14}$$

2.2. *Statistical Inference.* For $\lambda > 0$, define capped-ℓ_1 balls as

$$\mathcal{G}^*(M,s,\lambda) = \{\Omega : s_\lambda(\Omega) \le s, 1/M \le \lambda_{\min}(\Omega) \le \lambda_{\max}(\Omega) \le M\}, \tag{15}$$

where $s_\lambda = s_\lambda(\Omega) = \max_j \Sigma_{i \neq j} \min\{1, |\omega_{ij}|/\lambda\}$ for $\Omega = (\omega_{ij})_{1 \le i,j \le p}$. In this paper, λ is of the order $\sqrt{(\log p)/n}$. We omit the subscript λ from s when it is clear from the context. When $|\omega_{ij}|$ is either 0 or larger than λ, s_λ is the maximum node degree of the graph, which is denoted by $k_{n,p}$ in the class of parameter spaces (1). In general, $k_{n,p}$ is an upper bound of the sparseness measurement s_λ. The spectrum of Σ is bounded in the matrix class $\mathcal{G}^*(M,s,\lambda)$ as in the ℓ_0 ball (1). The following theorem gives an error bound for our estimators by comparing them with the oracle MLE (8), and shows that

$$\kappa_{ij}^{ora} = \sqrt{n}\frac{\omega_{ij}^{ora} - \omega_{ij}}{\sqrt{\omega_{ii}\omega_{jj} + \omega_{ij}^2}}$$

is asymptotically standard normal, which implies the oracle MLE (8) is asymptotically normal with mean ω_{ij} and variance $n^{-1}\left(\omega_{ii}\omega_{jj} + \omega_{ij}^2\right)$.

THEOREM 2. *Let $\hat{\Theta}_{A,A}$ and $\hat{\Omega}_{A,A}$ be estimators of $\Theta_{A,A}$ and $\Omega_{A,A}$ defined in (11) and (12) respectively, and $\lambda = (1+\varepsilon)\sqrt{\frac{2\delta \log p}{n}}$ for any $\delta \ge 1$ and $\varepsilon > 0$ in Equation (10).*

(i). *Suppose $s \le c_0 n/\log p$ for a sufficiently small constant $c_0 > 0$. We have*

$$\max_{\Omega \in \mathcal{G}^*(M,s,\lambda)} \max_{A:A=\{i,j\}} \mathbb{P}\left\{\left\|\hat{\Theta}_{A,A} - \Theta_{A,A}^{ora}\right\|_\infty > C_1 s \frac{\log p}{n}\right\} \le o\left(p^{-\delta+1}\right), \tag{16}$$

and

$$\max_{\Omega \in \mathcal{G}^*(M,s,\lambda)} \max_{A:A=\{i,j\}} \mathbb{P}\left\{\left\|\hat{\Omega}_{A,A} - \Omega_{A,A}^{ora}\right\|_\infty > C_1' s \frac{\log p}{n}\right\} \le o\left(p^{-\delta+1}\right), \tag{17}$$

where $\Theta_{A,A}^{ora}$ and $\Omega_{A,A}^{ora}$ are the oracle estimators defined in (8) and (9) respectively and C_1 and C_1' are positive constants depending on $\{\varepsilon, c_0, M\}$ only.

(ii). *There exist constants D_1 and $\vartheta \in (0,\infty)$, and three marginally standard normal random variables Z_{kl}, where $kl = ii, ij, jj$, such that whenever $|Z_{kl}| \le \vartheta\sqrt{n}$ for all kl, we have*

$$\left|\kappa_{ij}^{ora} - Z'\right| \le \frac{D_1}{\sqrt{n}}\left(1 + Z_{ii}^2 + Z_{ij}^2 + Z_{jj}^2\right), \tag{18}$$

where $Z' \sim \mathcal{N}(0,1)$, which can be defined as a linear combination of Z_{kl}, $kl = ii, ij, jj$.

Theorem 2 immediately yields the following results of estimation and inference for ω_{ij}.

Theorem 3. *Let $\hat{\Omega}_{A,A}$ be the estimator of $\Omega_{A,A}$ defined in (12), and $\lambda = (1+\varepsilon)\sqrt{\frac{2\delta\log p}{n}}$ for any $\delta \geq 1$ and $\varepsilon > 0$ in Equation (10). Suppose $s \leq c_0 n/\log p$ for a sufficiently small constant $c_0 > 0$. For any small constant $\epsilon > 0$, there exists a constant $C_2 = C_2(\epsilon, \varepsilon, c_0, M)$ such that*

$$\max_{\Omega\in\mathcal{G}^*(M,s,\lambda)} \max_{1\leq i\leq j\leq p} \mathbb{P}\left\{|\hat{\omega}_{ij} - \omega_{ij}| > C_2 \max\left\{s\frac{\log p}{n}, \sqrt{\frac{1}{n}}\right\}\right\} \leq \epsilon. \tag{19}$$

Moreover, there is constant $C_3 = C_3(\epsilon, \varepsilon, c_0, M)$ such that

$$\max_{\Omega\in\mathcal{G}^*(M,s,\lambda)} \mathbb{P}\left\{\left\|\hat{\Omega} - \Omega\right\|_\infty > C_3 \max\left\{s\frac{\log p}{n}, \sqrt{\frac{\log p}{n}}\right\}\right\} \leq o\left(p^{-\delta+3}\right). \tag{20}$$

Furthermore, $\hat{\omega}_{ij}$ is asymptotically efficient

$$\sqrt{nF_{ij}}\left(\hat{\omega}_{ij} - \omega_{ij}\right) \xrightarrow{D} \mathcal{N}(0,1), \tag{21}$$

when $\Omega \in \mathcal{G}^*(M,s,\lambda)$ and $s = o(\sqrt{n}/\log p)$, where

$$F_{ij}^{-1} = \omega_{ii}\omega_{jj} + \omega_{ij}^2.$$

Remark 1. *The upper bounds $\max\left\{s\frac{\log p}{n}, \sqrt{\frac{1}{n}}\right\}$ and $\max\left\{s\frac{\log p}{n}, \sqrt{\frac{\log p}{n}}\right\}$ in Equations (19) and (20) respectively are shown to be rate-optimal in Section 2.3.*

Remark 2. *The choice of $\lambda = (1+\varepsilon)\sqrt{\frac{2\delta\log p}{n}}$ is common in literature, but can be too big and too conservative, which usually leads to some estimation bias in practice. Let $\mathrm{tq}(\alpha, n)$ denotes the α quantile of the t distribution with n degrees of freedom. In Section 4 we show the value of λ can be reduced to $\lambda_{finite}^{new} = (1+\varepsilon)B/\sqrt{n-1+B^2}$ where $B = \mathrm{tq}\left(1 - \left(\frac{s_{\max}}{p}\right)^\delta/2, n-1\right)$ for every $s_{\max} = o\left(\frac{\sqrt{n}}{\log p}\right)$, which is asymptotically equivalent to $(1+\varepsilon)\sqrt{\frac{2\delta\log(p/s_{\max})}{n}}$. See Section 4 for more details. In simulation studies of Section 5, we use the penalty λ_{finite}^{new} with $\delta = 1$, which gives a good finite sample performance.*

Remark 3. *In Theorems 2 and 3, our goal is to estimate each entry ω_{ij} of the precision matrix Ω. Sometimes it is more natural to consider estimating the partial correlation $r_{ij} = -\omega_{ij}/(\omega_{ii}\omega_{jj})^{1/2}$ between Z_i and Z_j. Let $\hat{\Omega}_{A,A}$ be estimator of $\Omega_{A,A}$ defined in (12). Our estimator of partial correlation r_{ij} is defined as $\hat{r}_{ij} = -\hat{\omega}_{ij}/(\hat{\omega}_{ii}\hat{\omega}_{jj})^{1/2}$. Then the results above can be easily extended to the case of estimating r_{ij}. In particular, under the same assumptions in Theorem 3, the estimator $\hat{r}_{ij}$ is asymptotically efficient: $\sqrt{n(1-r_{ij}^2)^{-2}}(\hat{r}_{ij} - r_{ij})$ converges to $\mathcal{N}(0,1)$ when $s = o(\sqrt{n}/\log p)$. This is stated as Corollary 1 in Sun and Zhang (2012c) without proof.*

The following theorem extends Theorems 2 and 3 to estimation of $\zeta\left(\Omega_{B,B}^{-1}\right)$, a smooth functional of $\Omega_{B,B}^{-1}$ for a bounded size subset B. Assume that $\zeta : \mathbb{R}^{|B|\times|B|} \to \mathbb{R}$ is a unit Lipschitz function in a neighborhood $\left\{G : |||G-\Omega_{B,B}^{-1}||| \le \kappa\right\}$, i.e.,

$$\left|\zeta\left(G\right)-\zeta\left(\Omega_{B,B}^{-1}\right)\right| \le |||G-\Omega_{B,B}^{-1}|||. \tag{22}$$

Theorem 4. *Let $\hat{\zeta}$ be the estimator of ζ defined in (14), and $\lambda=(1+\varepsilon)\sqrt{\frac{2\delta\log p}{n}}$ for any $\delta\ge 1$ and $\varepsilon>0$ in Equation (10). Suppose $s\le c_0 n/\log p$ for a sufficiently small constant $c_0>0$. Then,*

$$\max_{\Omega\in\mathcal{G}^{*}(M,s,\lambda)}\mathbb{P}\left\{\left|\hat{\zeta}-\zeta^{ora}\right|>C_1 s\frac{\log p}{n}\right\}\le o\left(|B|\,p^{-\delta+1}\right), \tag{23}$$

with a constant $C_1=C_1(\varepsilon,c_0,M,|B|)$. Furthermore, $\hat{\zeta}$ is asymptotically efficient

$$\sqrt{nF_\zeta}\left(\hat{\zeta}-\zeta\right)\xrightarrow{D}\mathcal{N}\left(0,1\right), \tag{24}$$

when $\Omega\in\mathcal{G}^{*}(M,s,\lambda)$ and $s=o\left(\sqrt{n}/\log p\right)$, where F_ζ is the Fisher information of estimating ζ for the Gaussian model $\mathcal{N}\left(0,\Omega_{B,B}^{-1}\right)$.

Remark 4. *The results in this section can be easily extended to the weak l_q ball with $0<q<1$ to model the sparsity of the precision matrix. A weak l_q ball of radius c in $\mathbb{R}^p$ is defined as follows,*

$$B_q\left(c\right)=\left\{\xi\in\mathbb{R}^p:\left|\xi_{(j)}^{q}\right|\le cj^{-1},\ \textit{for all } j=1,\dots,p\right\},$$

where $\left|\xi_{(1)}\right|\ge\left|\xi_{(2)}\right|\ge\ldots\ge\left|\xi_{(p)}\right|$. Let

$$\mathcal{G}_q(M,k_{n,p})=\left\{\begin{array}{c}\Omega=\left(\omega_{ij}\right)_{1\le i,j\le p}:\omega_{\cdot j}\in B_q\left(k_{n,p}\right),\\ \textit{and } 1/M\le\lambda_{\min}\left(\Omega\right)\le\lambda_{\max}\left(\Omega\right)\le M.\end{array}\right\}. \tag{25}$$

Since $\xi\in B_q(k)$ implies $\sum_j\min(1,|\xi_j|/\lambda)\le\lfloor k/\lambda^q\rfloor+\{q/(1-q)\}k^{1/q}\lfloor k/\lambda^q\rfloor^{1-1/q}/\lambda$,

$$\mathcal{G}_q(M,k_{n,p})\subseteq\mathcal{G}^{*}(M,s,\lambda),\ 0\le q\le 1, \tag{26}$$

when $k_{n,p}/\lambda^q\le C_q(s\vee 1)$, where $C_q=1+q2^q/(1-q)$ for $0<q<1$ and $C_0=1$. Thus, the conclusions of Theorems 2, 3 and 4 hold with $\mathcal{G}^{*}(M,s,\lambda)$ replaced by $\mathcal{G}_q(M,k_{n,p})$ and s by $k_{n,p}(n/\log p)^{q/2}$, $0\le q<1$.

2.3. *Lower Bound.* In this section, we derive a lower bound for estimating ω_{ij} over the matrix class $\mathcal{G}_0(M, k_{n,p})$ defined in (1). Assume that

$$p \geq k_{n,p}^{\nu} \text{ with } \nu > 2, \tag{27}$$

and

$$k_{n,p} \leq C_0 \frac{n}{\log p} \tag{28}$$

for some $C_0 > 0$. Theorem 5 below implies that the assumption $k_{n,p}\frac{\log p}{n} \to 0$ is necessary for consistent estimation of any single entry of Ω.

We carefully construct a finite collection of distributions $\mathcal{G}_0 \subset \mathcal{G}_0(M, k_{n,p})$ and apply Le Cam's method to show that for any estimator $\hat{\omega}_{ij}$,

$$\sup_{\mathcal{G}_0} \mathbb{P}\left\{ |\hat{\omega}_{ij} - \omega_{ij}| > C_1 k_{n,p} \frac{\log p}{n} \right\} \to 1, \tag{29}$$

for some constant $C_1 > 0$. It is relatively easy to establish the parametric lower bound $\sqrt{\frac{1}{n}}$. These two lower bounds together immediately yield Theorem 5 below.

THEOREM 5. *Suppose we observe independent and identically distributed p-variate Gaussian random variables $X^{(1)}, X^{(2)}, \ldots, X^{(n)}$ with zero mean and precision matrix $\Omega = (\omega_{kl})_{p \times p} \in \mathcal{G}_0(M, k_{n,p})$. Under assumptions (27) and (28), we have the following minimax lower bounds*

$$\inf_{\hat{\omega}_{ij}} \sup_{\mathcal{G}_0(M,k_{n,p})} \mathbb{P}\left\{ |\hat{\omega}_{ij} - \omega_{ij}| > \max\left\{ C_1 \frac{k_{n,p} \log p}{n}, C_2 \sqrt{\frac{1}{n}} \right\} \right\} > c_1 > 0, \tag{30}$$

and

$$\inf_{\hat{\Omega}} \sup_{\mathcal{G}_0(M,k_{n,p})} \mathbb{P}\left\{ \left\| \hat{\Omega} - \Omega \right\|_{\infty} > \max\left\{ C_1' \frac{k_{n,p} \log p}{n}, C_2' \sqrt{\frac{\log p}{n}} \right\} \right\} > c_2 > 0, \tag{31}$$

where C_1, C_2, C_1' and C_2' are positive constants depending on M, ν and C_0 only.

REMARK 5. *The lower bound $\frac{k_{n,p} \log p}{n}$ in Theorem 5 shows that estimation of sparse precision matrix can be very different from estimation of sparse covariance matrix. The sample covariance always gives a parametric rate of estimation of every entry σ_{ij}. But for estimation of sparse precision matrix, when $k_{n,p} \gg \frac{\sqrt{\mathrm{n}}}{\log \mathrm{p}}$, Theorem 5 implies that it is impossible to obtain the parametric rate.*

Remark 6. *Since $\mathcal{G}_0(M, k_{n,p}) \subseteq \mathcal{G}^*(M, k_{n,p}, \lambda)$ by the definitions in (1) and (15), Theorem 5 also provides the lower bound for the larger class. Similarly, Theorem 5 can be easily extended to the weak l_q ball, $0 < q < 1$, defined in (25) and the capped-ℓ_1 ball defined in (15). For these parameter spaces, in the proof of Theorem 5 we only need to define $\mathcal{H}$ as the collection of all $p \times p$ symmetric matrices with exactly $\left(k_{n,p}\left(\frac{n}{\log p}\right)^{q/2} - 1\right)$ rather than $(k_{n,p} - 1)$ elements equal to 1 between the third and the last elements on the first row (column) and the rest all zeros. Then it is easy to check that the sub-parameter space $\mathcal{G}_0$ in (56) is indeed in $\mathcal{G}_q(M, k_{n,p})$. Now under assumptions $p \geq \left(k_{n,p}\left(\frac{n}{\log p}\right)^{q/2}\right)^v$ with $\nu > 2$ and $k_{n,p} \leq C_0 \left(\frac{n}{\log p}\right)^{1-q/2}$,we have the following minimax lower bounds*

$$\inf_{\hat{\omega}_{ij}} \sup_{\mathcal{G}_q(M,k_{n,p})} \mathbb{P}\left\{|\hat{\omega}_{ij} - \omega_{ij}| > \max\left\{C_1 k_{n,p}\left(\frac{\log p}{n}\right)^{1-q/2}, C_2\sqrt{\frac{1}{n}}\right\}\right\} > c_1 > 0,$$

and

$$\inf_{\hat{\Omega}} \sup_{\mathcal{G}_q(M,k_{n,p})} \mathbb{P}\left\{\left\|\hat{\Omega} - \Omega\right\|_\infty > \max\left\{C_1' k_{n,p}\left(\frac{\log p}{n}\right)^{1-q/2}, C_2'\sqrt{\frac{\log p}{n}}\right\}\right\} > c_2 > 0.$$

These lower bounds match the upper bounds for the proposed estimator in Theorem 3 in view of the discussion in Remark 4.

3. Applications.

The asymptotic normality result is applied to obtain rate-optimal estimation of the precision matrix under various matrix l_w norms, to recover the support of Ω adaptively, and to estimate latent graphical models without the need of the irrepresentable condition or the l_1 constraint of the precision matrix commonly required in literature. Our procedure is first obtaining an Asymptotically Normal estimation and then do Thresholding. We thus call it ANT.

3.1. *ANT for Adaptive Support Recovery.*

The support recovery of precision matrix has been studied by several papers. See, for example, Friedman, Hastie and Tibshirani (2008), d'Aspremont, Banerjee and El Ghaoui (2008), Rothman et al. (2008), Ravikumar et al. (2011), Cai, Liu and Luo (2011), and Cai, Liu and Zhou (2012). In these literature, the theoretical properties of the graphical lasso (Glasso), CLIME and ACLIME on the support recovery were obtained. Ravikumar et al. (2011) studied the theoretical properties of Glasso, and showed that Glasso can correctly recover the support under irrepresentable conditions and the condition $\min_{(i,j)\in\mathcal{S}(\Omega)} |\omega_{ij}| \geq c\sqrt{\frac{\log p}{n}}$ for some $c > 0$. Cai, Liu and Luo (2011) does not require irrepresentable conditions, but need to assume that $\min_{(i,j)\in\mathcal{S}(\Omega)} |\omega_{ij}| \geq CM_{n,p}^2\sqrt{\frac{\log p}{n}}$, where $M_{n,p}$ is the matrix l_1 norm of Ω. In Cai, Liu and Zhou (2012), they weakened the condition to $\min_{(i,j)\in\mathcal{S}(\Omega)}|\omega_{ij}| \geq CM_{n,p}\sqrt{\frac{\log p}{n}}$, but the threshold level there is $\frac{C}{2}M_{n,p}\sqrt{\frac{\log p}{n}}$, where C is unknown and $M_{n,p}$ can be very large, which makes the support recovery procedure there impractical.

In this section we introduce an adaptive support recovery procedure based on the variance of the oracle estimator of each entry ω_{ij} to recover the sign of nonzero entries of Ω with high probability. The lower bound condition for $\min_{(i,j)\in\mathcal{S}(\Omega)}|\omega_{ij}|$ is significantly weakened. In particular, we remove the unpleasant matrix l_1 norm $M_{n,p}$. In Theorem 3, when the precision matrix is sparse enough $s = o\left(\frac{\sqrt{n}}{\log p}\right)$, we have the asymptotic normality result for each entry ω_{ij}, $i \neq j$, i.e.,

$$\sqrt{nF_{ij}}\left(\hat{\omega}_{ij} - \omega_{ij}\right) \xrightarrow{D} \mathcal{N}\left(0,1\right),$$

where $F_{ij} = \left(\omega_{ii}\omega_{jj} + \omega_{ij}^2\right)^{-1}$ is the Fisher information of estimating ω_{ij}. The total number of edges is $p\left(p-1\right)/2$. We may apply thresholding to $\hat{\omega}_{ij}$ to correctly distinguish zero and nonzero entries. However, the variance $\omega_{ii}\omega_{jj} + \omega_{ij}^2$ needs to be estimated. We define the adaptive support recovery procedure as follows

$$\hat{\Omega}_{thr} = (\hat{\omega}_{ij}^{thr})_{p\times p}, \tag{32}$$

where $\hat{\omega}_{ii}^{thr} = \hat{\omega}_{ii}$ and $\hat{\omega}_{ij}^{thr} = \hat{\omega}_{ij}1\{|\hat{\omega}_{ij}| \geq \hat{\tau}_{ij}\}$ for $i \neq j$ with

$$\hat{\tau}_{ij} = \sqrt{\frac{2\xi_0\left(\hat{\omega}_{ii}\hat{\omega}_{jj} + \hat{\omega}_{ij}^2\right)\log p}{n}}. \tag{33}$$

Here $\hat{\omega}_{ii}\hat{\omega}_{jj} + \hat{\omega}_{ij}^2$ is the natural estimate of the asymptotic variance of $\hat{\omega}_{ij}$ defined in (12) and ξ_0 is a tuning parameter which can be taken as fixed at any $\xi_0 > 2$. This thresholding estimator is adaptive. The sufficient conditions in the Theorem 6 below for support recovery are much weaker compared with other results in literature.

Define a thresholded population precision matrix as

$$\Omega_{thr} = (\omega_{ij}^{thr})_{p\times p}, \tag{34}$$

where $\omega_{ii}^{thr} = \omega_{ii}$ and $\omega_{ij}^{thr} = \omega_{ij}1\left\{|\omega_{ij}| \geq \sqrt{8\xi(\omega_{ii}\omega_{jj} + \omega_{ij}^2)(\log p)/n}\right\}$, with a certain $\xi > \xi_0$. Recall that $E = E(\Omega) = \{(i,j) : \omega_{ij} \neq 0\}$ is the edge set of the Gauss-Markov graph associated with the precision matrix Ω. Since Ω_{thr} is composed of relatively large components of Ω, $(V, E(\Omega_{thr}))$ can be view as a graph of strong edges. Define

$$\mathcal{S}(\Omega) = \{sgn(\omega_{ij}), \quad 1 \leq i,j \leq p\}.$$

The following theorem shows that with high probability, ANT recovers all the strong edges without false recovery. Moreover, under the uniform signal strength condition

$$|\omega_{ij}| \geq 2\sqrt{\frac{2\xi\left(\omega_{ii}\omega_{jj}+\omega_{ij}^2\right)\log p}{n}},\ \forall\ \omega_{ij}\neq 0. \tag{35}$$

i.e. $\Omega_{thr}=\Omega$, the ANT also recovers the sign matrix $\mathcal{S}(\Omega)$.

Theorem 6. *Let $\lambda=(1+\varepsilon)\sqrt{\frac{2\delta\log p}{n}}$ for any $\delta\geq 3$ and $\varepsilon>0$. Let $\hat{\Omega}_{thr}$ be the ANT estimator defined in (32) with $\xi_0>2$ in the thresholding level (33). Suppose $\Omega\in\mathcal{G}^*(M,s,\lambda)$ with $s=o\left(\sqrt{n/\log p}\right)$. Then,*

$$\lim_{n\to\infty}\mathbb{P}\left(E(\Omega_{thr})\subseteq E(\hat{\Omega}_{thr})\subseteq E(\Omega)\right)=1. \tag{36}$$

where Ω_{thr} is defined in (34) with $\xi>\xi_0$. If in addition (35), then

$$\lim_{n\to\infty}\mathbb{P}\left(\mathcal{S}(\hat{\Omega}_{thr})=\mathcal{S}(\Omega)\right)=1. \tag{37}$$

3.2. *ANT for Adaptive Estimation under the Matrix l_w Norm.* In this section, we consider the rate of convergence under the matrix l_w norm. To control the improbable case for which our estimator $\hat{\Theta}_{A,A}$ is nearly singular, we define our estimator based on the thresholding estimator $\hat{\Omega}_{thr}$ defined in (32),

$$\breve{\Omega}_{thr}=(\hat{\omega}_{ij}^{thr}1\left\{|\hat{\omega}_{ij}|\leq\log p\right\})_{p\times p}. \tag{38}$$

Theorem 7 follows mainly from the convergence rate under element-wise norm and the fact that the upper bound holds for matrix l_1 norm. Then it follows immediately by the inequality $|||M|||_w\leq|||M|||_1$ for any symmetric matrix M and $1\leq w\leq\infty$, which can be proved by applying the Riesz-Thorin interpolation theorem. See, e.g., Thorin (1948). Note that under the assumption $k_{n,p}^2=O\left(n/\log p\right)$, it can be seen from the Equations (17) and (18) in Theorem 2 that with high probability the $\left\|\hat{\Omega}-\Omega\right\|_\infty$ is dominated by $\|\Omega^{ora}-\Omega\|_\infty=O_p\left(\sqrt{\frac{\log p}{n}}\right)$. From there the details of the proof is in nature similar to the Theorem 3 in Cai and Zhou (2012) and thus will be omitted due to the limit of space.

Theorem 7. *Under the assumptions $s^2=O\left(n/\log p\right)$ and $n=O\left(p^\xi\right)$ with some $\xi>0$, the $\breve{\Omega}_{thr}$ defined in (38) with $\lambda=(1+\varepsilon)\sqrt{\frac{2\delta\log p}{n}}$ for sufficiently large $\delta\geq 3+\ 2\xi$ and $\varepsilon>0$ satisfies, for all $1\leq w\leq\infty$ and $k_{n,p}\leq s$*

$$\sup_{\mathcal{G}_0(M,k_{n,p})}\mathbb{E}|||\breve{\Omega}_{thr}-\Omega|||_w^2\leq\sup_{\mathcal{G}^*(M,k_{n,p},\lambda)}\mathbb{E}|||\breve{\Omega}_{thr}-\Omega|||_w^2\leq Cs^2\frac{\log p}{n}. \tag{39}$$

Remark 7. *It follows from Equation (26) that result (39) also holds for the classes of weak ℓ_p balls $\mathcal{G}_q(M, k_{n,p})$ defined in Equation (25), with $s = C_q k_{n,p} \left(\frac{n}{\log p}\right)^{q/2}$,*

$$\sup_{\mathcal{G}_q(M,k_{n,p})} \mathbb{E}|||\breve{\Omega}_{thr} - \Omega|||_w^2 \leq C k_{n,p}^2 \left(\frac{\log p}{n}\right)^{1-q}. \tag{40}$$

Remark 8. *Cai, Liu and Zhou (2012) showed the rates obtained in Equations (39) and (40) are optimal when $p \geq cn^\gamma$ for some $\gamma > 1$ and $k_{n,p} = o\left(n^{1/2} (\log p)^{-3/2}\right)$.*

3.3. *Estimation and Inference for Latent Variable Graphical Model.* Chandrasekaran, Parrilo and Willsky (2012) first proposed a very natural penalized estimation approach and studied its theoretical properties. Their work was discussed and appreciated by several researchers. But it was not clear if the conditions in their paper are necessary and the results there are optimal. Ren and Zhou (2012) observed that the support recovery boundary can be significantly improved from an order of $\sqrt{\frac{p}{n}}$ to $\sqrt{\frac{\log p}{n}}$ under certain conditions including a bounded l_1 norm constraint for the precision matrix. In this section we extend the methodology and results in Section 2 to study latent variable graphical models. The results in Ren and Zhou (2012) are significantly improved under weaker assumptions.

Let O and H be two subsets of $\{1, 2, \ldots, p+h\}$ with $\text{Card}(O) = p$, $\text{Card}(H) = h$ and $O \cup H = \{1, 2, \ldots, p+h\}$. Assume that $\left(X_O^{(i)}, X_H^{(i)}\right)$, $i = 1, \ldots, n$, are i.i.d. $(p+h)$-variate Gaussian random vectors with a positive covariance matrix $\Sigma_{(p+h)\times(p+h)}$. Denote the corresponding precision matrix by $\bar{\Omega}_{(p+h)\times(p+h)} = \Sigma^{-1}_{(p+h)\times(p+h)}$. We only have access to $\left\{X_O^{(1)}, X_O^{(2)}, \ldots, X_O^{(n)}\right\}$, while $\left\{X_H^{(1)}, X_H^{(2)}, \ldots, X_H^{(n)}\right\}$ are hidden and the number of latent components is unknown. Write $\Sigma_{(p+h)\times(p+h)}$ and $\bar{\Omega}_{(p+h)\times(p+h)}$ as follows,

$$\Sigma_{(p+h)\times(p+h)} = \begin{pmatrix} \Sigma_{O,O} & \Sigma_{O,H} \\ \Sigma_{H,O} & \Sigma_{H,H} \end{pmatrix}, \text{ and } \bar{\Omega}_{(p+h)\times(p+h)} = \begin{pmatrix} \bar{\Omega}_{O,O} & \bar{\Omega}_{O,H} \\ \bar{\Omega}_{H,O} & \bar{\Omega}_{H,H} \end{pmatrix},$$

where $\Sigma_{O,O}$ and $\Sigma_{H,H}$ are covariance matrices of $X_O^{(i)}$ and $X_H^{(i)}$ respectively and from the Schur complement we have

$$\Sigma^{-1}_{O,O} = \bar{\Omega}_{O,O} - \bar{\Omega}_{O,H}\bar{\Omega}^{-1}_{H,H}\bar{\Omega}_{H,O}. \tag{41}$$

See, e.g., Horn and Johnson (1990). Define

$$S = \bar{\Omega}_{O,O}, \text{ and } L = \bar{\Omega}_{O,H}\bar{\Omega}^{-1}_{H,H}\bar{\Omega}_{H,O},$$

where $h' = \text{rank}\,(L) = \text{rank}\,(\bar{\Omega}_{O,H}) \leq h$.

We are interested in estimating $\Sigma_{O,O}^{-1}$ as well as S and L. To make the problem identifiable we assume that S is sparse and the effect of latent variables is spread out over all coordinates, i.e.,

$$S = (s_{ij})_{1\le i,j\le p}, \quad \max_{1\le j\le p} \sum_{i\ne j} 1\{s_{ij} \ne 0\} \le k_{n,p}; \tag{42}$$

and

$$L = (l_{ij})_{1\le i,j\le p}, \ |l_{ij}| \le \frac{M_0}{p}. \tag{43}$$

The sparseness of $S = \bar{\Omega}_{O,O}$ can be seen to be inherited from the sparse full precision matrix $\bar{\Omega}_{(p+h)\times(p+h)}$, and it is particularly interesting for us to identify the support of $S = \bar{\Omega}_{O,O}$ and make inference for each entry of S. A sufficient condition for the assumption (43) is that the eigendecomposition of $L = \sum_{i=1}^{h'} \lambda_i u_i u_i^T$ satisfies $\|u_i\|_\infty \le \sqrt{\frac{c_0}{p}}$ for all i and $c_0 \sum_{i=1}^{h'} \lambda_i \le M_0$. See Candès and Recht (2009) for a similar assumption. In addition, we assume that

$$1/M \le \lambda_{\min}(\Sigma_{(p+h)\times(p+h)}) \le \lambda_{\max}(\Sigma_{(p+h)\times(p+h)}) \le M \tag{44}$$

for some universal constant M, and

$$\frac{n}{\log p} = o(p). \tag{45}$$

Equation (44) implies that both the covariance $\Sigma_{O,O}$ of observations $X_O^{(i)}$ and the sparse component $S = \bar{\Omega}_{O,O}$ have bounded spectrum, and $\lambda_{\max}(L) \le M$.

With a slight abuse of notation, denote the precision matrix $\Sigma_{O,O}^{-1}$ of $X_O^{(i)}$ by Ω and its inverse by Θ. We propose to apply the methodology in Section 2 to the observations $X^{(i)}$ which are i.i.d. $\mathcal{N}(0, \Sigma_{O,O})$ with $\Omega = (s_{ij} - l_{ij})_{1\le i,j\le p}$ by considering the following regression

$$\mathbf{X}_A = \mathbf{X}_{O\backslash A}\beta + \epsilon_A \tag{46}$$

for $A = \{i, j\} \subset O$ with $\beta = \Omega_{O\backslash A,A}\Omega_{A,A}^{-1}$ and $\epsilon_A \overset{i.i.d.}{\sim} \mathcal{N}\left(0, \Omega_{A,A}^{-1}\right)$ and the penalized scaled lasso procedure to estimate $\Omega_{A,A}$. When $S = I_p$ and $L = \frac{1}{2}u_0 u_0^T$ with $u_0^T = (1/\sqrt{p}, \ldots, 1/\sqrt{p})$, we see that

$$\max_j \Sigma_{i\ne j} \min\left\{1, \frac{|s_{ij} - l_{ij}|}{\lambda}\right\} = \frac{p-1}{2p\lambda} = O\left(\sqrt{\frac{n}{\log p}}\right).$$

However, to obtain the asymptotic normality result as in Theorem 2, we required

$$\max_j \Sigma_{i\ne j} \min\left\{1, \frac{|s_{ij} - l_{ij}|}{\lambda}\right\} = o\left(\frac{\sqrt{n}}{\log p}\right),$$

which is no longer satisfied for the latent variable graphical model. In Section 7.2 we overcome the difficulty through a new analysis.

THEOREM 8. *Let $\hat{\Omega}_{A,A}$ be the estimator of $\Omega_{A,A}$ defined in (12) with $A = \{i, j\}$ for the regression (46). Let $\lambda = (1+\varepsilon)\sqrt{\frac{2\delta\log p}{n}}$ for any $\delta \geq 1$ and $\varepsilon > 0$. Under the assumptions (42)-(45) and $k_{n,p} = o\left(\frac{\sqrt{n}}{\log p}\right)$ we have*

$$\sqrt{\frac{n}{\omega_{ii}\omega_{jj}+\omega_{ij}^2}}\left(\hat{\omega}_{ij}-\omega_{ij}\right) \overset{D}{\sim} \sqrt{\frac{n}{\omega_{ii}\omega_{jj}+\omega_{ij}^2}}\left(\hat{\omega}_{ij}-s_{ij}\right) \xrightarrow{D} \mathcal{N}(0,1). \tag{47}$$

REMARK 9. *Without condition (45), our estimator may not be asymptotic efficient but still has nice convergence property. We could obtain the following rate of convergence for estimating $\omega_{ij} = s_{ij} - l_{ij}$, provided $k_{n,p} = o\left(\frac{n}{\log p}\right)$, by simply applying Theorem 2 with sparsity* $\max_j \Sigma_{i\neq j} \min\left\{1, \frac{|s_{ij}-l_{ij}|}{\lambda}\right\} = O\left(k_{n,p}+\lambda^{-1}\right)$,

$$\mathbb{P}\left\{|\hat{\omega}_{ij}-\omega_{ij}| > C_3 \max\left\{k_{n,p}\frac{\log p}{n}, \sqrt{\frac{\log p}{n}}\right\}\right\} \leq o\left(p^{-\delta+1}\right),$$

which further implies the rate of convergence for estimating s_{ij}

$$\mathbb{P}\left\{|\hat{\omega}_{ij}-s_{ij}| > C_3 \max\left\{k_{n,p}\frac{\log p}{n}, \sqrt{\frac{\log p}{n}}, \frac{M_0}{p}\right\}\right\} \leq o\left(p^{-\delta+1}\right).$$

Define the adaptive thresholding estimator $\hat{\Omega}_{thr} = (\hat{\omega}_{ij}^{thr})_{p\times p}$ as in (32) and (33). Following the proof of Theorems 6 and 7, we are able to obtain the following results. We shall omit the proof due to the limit of space.

THEOREM 9. *Let $\lambda = (1+\varepsilon)\sqrt{\frac{2\delta\log p}{n}}$ for some $\delta \geq 3$ and $\varepsilon > 0$ in Equation (10). Assume the assumptions (42)-(45) hold. Then*

(i). *Under the assumptions $k_{n,p} = o\left(\sqrt{\frac{n}{\log p}}\right)$ and*

$$|s_{ij}| \geq 2\sqrt{\frac{2\xi_0\left(\omega_{ii}\omega_{jj}+\omega_{ij}^2\right)\log p}{n}}, \ \forall s_{ij} \in \mathcal{S}(S)$$

for some $\xi_0 > 2$, we have

$$\lim_{n\to\infty} \mathbb{P}\left(\mathcal{S}(\hat{\Omega}_{thr}) = \mathcal{S}(S)\right) = 1.$$

(ii). *Under the assumption $k_{n,p}^2 = O\left(n/\log p\right)$ and $n = O\left(p^{\xi}\right)$ with some $\xi > 0$, the $\breve{\Omega}_{thr}$ defined in (38) with sufficiently large $\delta \geq 3+ 2\xi$ satisfies, for all $1 \leq w \leq \infty$,*

$$\mathbb{E}|||\breve{\Omega}_{thr} - S|||_w^2 \leq Ck_{n,p}^2\frac{\log p}{n}.$$

4. Discussion. In the analysis of Theorem 2 and nearly all consequent results in Theorems 3-4, and Theorems 6-9, we have picked the penalty term $\lambda = (1+\varepsilon)\sqrt{\frac{2\delta\log p}{n}}$ for any $\delta \geq 1$ (or $\delta \geq 3$ for support recovery) and $\varepsilon > 0$. This choice of λ can be too conservative and cause some finite sample estimation bias. In this section we show that λ can be chosen smaller.

Let $s_{\max} \leq c_0 \frac{n}{\log p}$ with a sufficiently small constant $c_0 > 0$ and $s_{\max} = O\left(p^t\right)$ for some $t < 1$. Denote the cumulative distribution function of $t_{(n-1)}$ distribution by $F_{t_{(n-1)}}$. Let $\lambda_{finite}^{new} = (1+\varepsilon) B/\sqrt{n-1+B^2}$ where $B = F_{t_{(n-1)}}^{-1}\left(1 - \left(\frac{s_{\max}}{p}\right)^{\delta}/2\right)$. It can be shown that λ_{finite}^{new} is asymptotically equivalent to $\lambda^{new} = (1+\varepsilon)\sqrt{\frac{2\delta\log(p/s_{\max})}{n}}$. We can extend Theorems 2-4 and 6-9 to the new penalties λ^{new} and λ_{finite}^{new}. All the results remain the same except that we need to replace s (or $k_{n,p}$) in those theorems by $s + s_{\max}$ (or $k_{n,p} + s_{\max}$). Since all theorems (except the minimax lower bound Theorem) are derived from Lemma 2, all we need is just an extension of Lemma 2 as follows.

Lemma 1. *Let $\lambda = \lambda^{new}$ or λ_{finite}^{new} for any $\delta \geq 1$ and $\varepsilon > 0$ in Equation (10). Assume $s_{\max} \leq c_0 \frac{n}{\log p}$ with a sufficiently small constant $c_0 > 0$ and $s_{\max} = O\left(p^t\right)$ for some $t < 1$ and define the event E_m as follows,*

$$\begin{aligned}
\left|\hat{\theta}_{mm} - \theta_{mm}^{ora}\right| &\leq C_1'\lambda^2\left(s + s_{\max}\right), \\
\left\|\beta_m - \hat{\beta}_m\right\|_1 &\leq C_2'\lambda\left(s + s_{\max}\right), \\
\left\|\mathbf{X}_{A^c}\left(\beta_m - \hat{\beta}_m\right)\right\|^2/n &\leq C_3'\lambda^2\left(s + s_{\max}\right), \\
\left\|\mathbf{X}_{A^c}^T\epsilon_m/n\right\|_\infty &\leq C_4'\lambda,
\end{aligned}$$

for $m = i$ or j and some constants C_k', $1 \leq k \leq 4$. Under the assumptions of Theorem 2, we have

$$\mathbb{P}\left(E_m^c\right) \leq o\left(p^{-\delta+1}\right).$$

See its proof in the supplementary material for more details. Note that when $s = o\left(\frac{\sqrt{n}}{\log p}\right)$, we have

$$\left\|\mathbf{X}_{A^c}\left(\beta_m - \hat{\beta}_m\right)\right\|^2/n \leq C_3'\lambda^2\left(s + s_{\max}\right) = o\left(1/\sqrt{n}\right)$$

with high probability, for every $s_{\max} = o\left(\frac{\sqrt{n}}{\log p}\right)$. Thus for every choice of $s_{\max} = o\left(\frac{\sqrt{n}}{\log p}\right)$ in the penalty, our procedure leads asymptotically efficient estimation of every entry of the precision matrix as long as $s = o\left(\frac{\sqrt{n}}{\log p}\right)$. In Section 5 we set $\lambda = \lambda_{finite}^{new}$ with $\delta = 1$ for statistical inference and $\delta = 3$ for support recovery.

5. Numerical Studies.

In this section, we present some numerical results for both asymptotic distribution and support recovery. We consider the following 200×200 precision matrix with three blocks. The block sizes are 100×100, 50×50 and 50×50, respectively. Let $(\alpha_1,\alpha_2,\alpha_3)=(1,2,4)$. The diagonal entries are $\alpha_1,\alpha_2,\alpha_3$ in three blocks, respectively. When the entry is in the k-th block, $\omega_{j-1,j}=\omega_{j,j-1}=0.5\alpha_k$, and $\omega_{j-2,j}=\omega_{j,j-2}=0.4\alpha_k$, $k=1,2,3$. The asymptotic variance for estimating each entry can be very different, thus a naive procedure of setting one universal threshold for all entries would likely fail.

We first estimate the entries in the precision matrix, and partial correlations which was discussed in Remark 3, and consider the distributions of the estimators. We generate a random sample of size $n=400$ from a multivariate normal distribution $\mathcal{N}(0,\Sigma)$ with $\Sigma=\Omega^{-1}$. As mentioned in Remark 2, the penalty constant is chosen to be $\lambda_{finite}^{new}=B/\sqrt{n-1+B^2}$, where $B=tq(1-\hat{s}/(2p),n-1)$ with $\hat{s}=\sqrt{n}/\log p$, which is asymptotically equivalent to $\sqrt{(2/n)\log(p/\hat{s})}$.

Table 1 reports the mean and standard error of our estimators for four entries in the precision matrix and the corresponding correlations based on 100 replications. Figure 1 shows the histograms of our estimates with the theoretical normal density super-imposed. We can see that the distributions of our estimates match pretty well to the asymptotic normality in Theorem 3. We have tried other choices of dimensions, e.g. $p=1000$, and obtained similar results.

Table 1
Mean and standard error of the proposed estimators.

	$\omega_{1,2}=0.5$	$\omega_{1,3}=0.4$	$\omega_{1,4}=0$	$\omega_{1,10}=0$
$\widehat{\omega}_{1,j}$	0.469 ± 0.051	0.380 ± 0.054	-0.042 ± 0.043	-0.003± 0.045
	$r_{1,2}=-0.5$	$r_{1,3}=-0.4$	$r_{1,4}=0$	$r_{1,10}=0$
$\widehat{r}_{1,j}$	-0.480 ± 0.037	-0.392 ± 0.043	0.043 ± 0.043	0.003 ± 0.046

Support recovery of a precision matrix is of great interests. We compare our selection results with the GLasso. In addition to the training sample, we generate an independent sample of size 400 from the same distribution for validating the tuning parameter for the GLasso. The GLasso estimators are computed based on training data with a range of penalty levels and we choose a proper penalty level by minimizing likelihood loss $\{\text{trace}(\overline{\Sigma}\widehat{\Omega})-\log\det(\widehat{\Omega})\}$ on the validation sample, where $\overline{\Sigma}$ is the sample covariance matrix. Our ANT estimators are computed based on the training sample only. As stated in Theorem 6, we use a slightly larger penalty constant to allow the selection consistency. Let $\lambda_{finite}^{new}=B/\sqrt{n-1+B^2}$, where $B=tq(1-(\hat{s}/p)^3/2,n-1)$, which is asymptotically equivalent to $\sqrt{(6/n)\log(p/\hat{s})}$. We then apply the thresholding step as in (33) with $\xi_0=2$. Table 2 shows the average selection performances of 10 replications. The true pos-

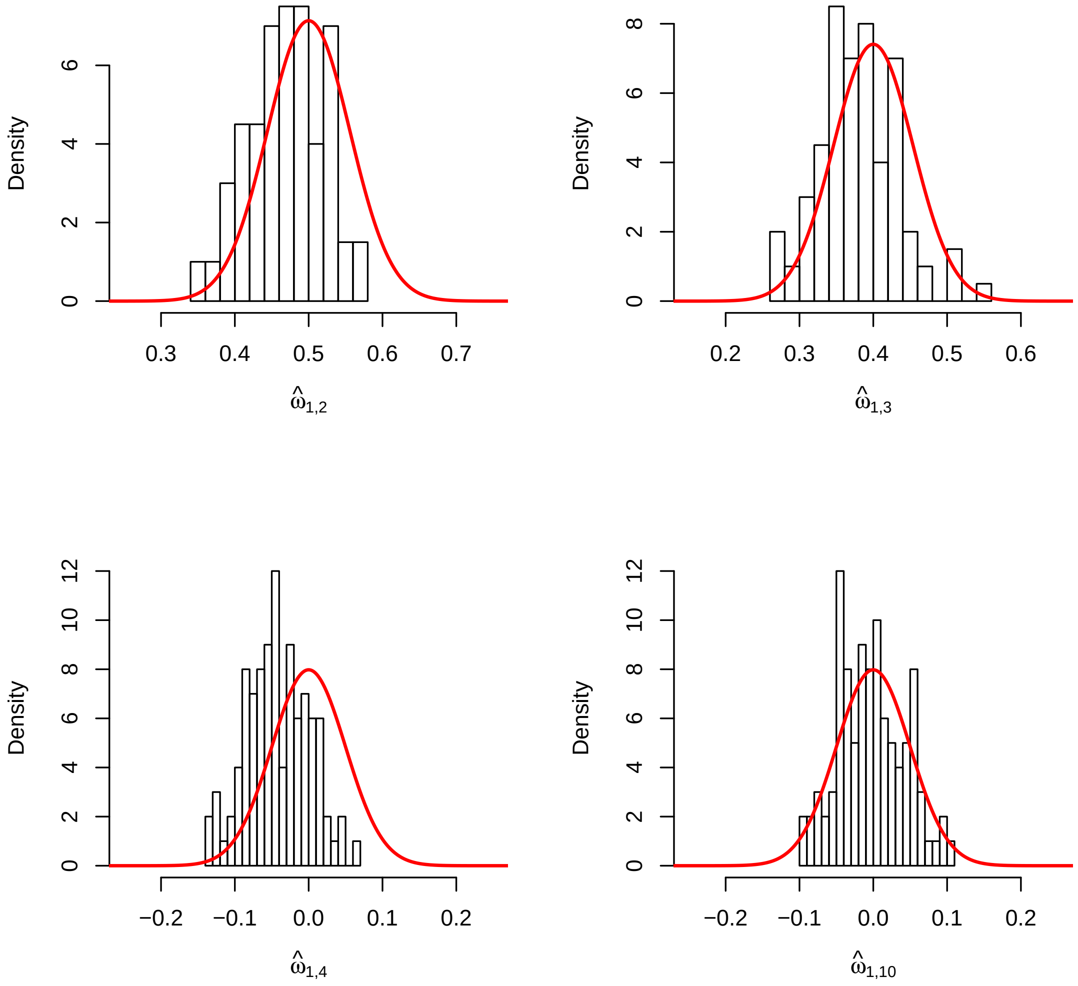

FIG 1. *Histograms of estimated entries. Top: entries $\omega_{1,2}$ and $\omega_{1,3}$ in the precision matrix; bottom: entries $\omega_{1,4}$ and $\omega_{1,10}$ in the precision matrix.*

itive (rate) and false positive (rate) are reported. In addition to the overall performance, the summary statistics are also reported for each block. We can see that while both our ANT method and the graphical Lasso choose all nonzero entries, ANT outperforms the GLasso in the sense of the false positive rate and the false discovery rate.

Table 2
The performance of support recovery

Block	Method	TP	TPR	FP	FPR
Overall	GLasso	391	1	5298.2	0.2716
	ANT	391	1	3.5	0.0004
Block 1	GLasso	197	1	1961	0.4126
	ANT	197	1	1.2	0.0003
Block 2	GLasso	97	1	288.4	0.2557
	ANT	97	1	1.1	0.0010
Block 3	GLasso	97	1	162.1	0.1437
	ANT	97	1	1.1	0.0010

Moreover, we compare our method with the GLasso with various penalty levels. Figure 2 shows the ROC curves for the GLasso with various penalty levels and ANT with various thresholding levels in the follow-up procedure. It is noticed that the GLasso at any penalty level cannot achieve similar performance as ours. In addition, the circle in the plot represents the performance of ANT with the selected threshold level as in (33). The triangle in the plot represents the performance of the graphical Lasso with the penalty level chosen by cross-validation. This again indicates that our method simultaneously achieves a very high true positive rate and a very low false positive rate.

6. Proof of Theorems 1-5.

6.1. *Proof of Theorem 2-4.* We will only prove Theorems 2 and 3. The proof of Theorem 4 is similar to that of Theorems 2 and 3. The following lemma is the key to the proof.

Lemma 2. *Let $\lambda = (1+\varepsilon)\sqrt{\frac{2\delta\log p}{n}}$ for any $\delta \geq 1$ and $\varepsilon > 0$ in Equation (10). Define the event E_m as follows,*

$$\left|\hat{\theta}_{mm} - \theta_{mm}^{ora}\right| \leq C_1'\lambda^2 s, \tag{48}$$

$$\left\|\beta_m - \hat{\beta}_m\right\|_1 \leq C_2'\lambda s, \tag{49}$$

$$\left\|\mathbf{X}_{A^c}\left(\beta_m - \hat{\beta}_m\right)\right\|^2/n \leq C_3'\lambda^2 s, \tag{50}$$

$$\left\|\mathbf{X}_{A^c}^T\epsilon_m/n\right\|_\infty \leq C_4'\lambda, \tag{51}$$

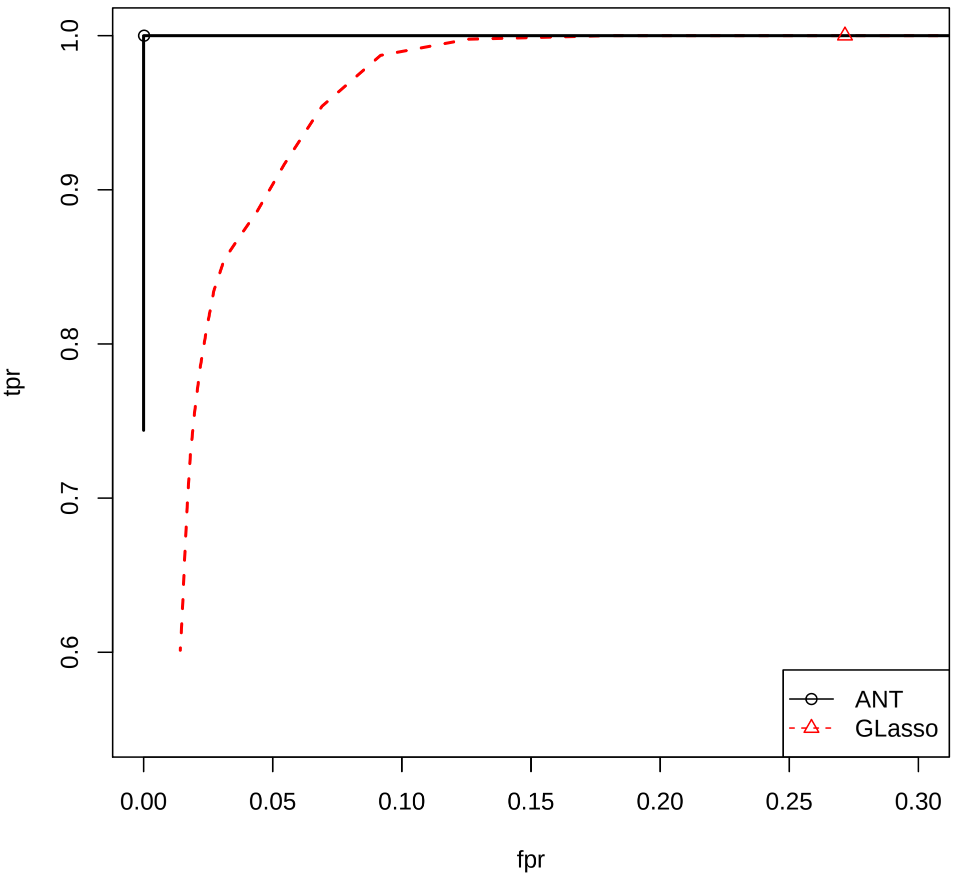

Fig 2. *The ROC curves of the graphical Lasso and ANT. Circle: ANT with our proposed thresholding. Triangle: GLasso with penalty level by CV.*

for $m = i$ or j and some constants C'_k, $1 \leq k \leq 4$. Under the assumptions of Theorem 2, we have

$$\mathbb{P}\left(E_m^c\right) \leq o\left(p^{-\delta+1}\right).$$

6.1.1. *Proof of Theorems 2.* We first prove (i). From Equation (48) of Lemma 2, the large deviation probability in (16) holds for θ_{ii}^{ora} and θ_{jj}^{ora}. We then need only to consider the entry θ_{ij}^{ora}. On the event $E_i \cap E_j$,

$$\begin{aligned}
\left|\hat{\theta}_{ij} - \theta_{ij}^{ora}\right| &= \left|\hat{\epsilon}_i^T \hat{\epsilon}_j / n - \epsilon_i^T \epsilon_j / n\right| \\
&= \left|\left(\epsilon_i + \mathbf{X}_{A^c}\left(\beta_i - \hat{\beta}_i\right)\right)^T \left(\epsilon_j + \mathbf{X}_{A^c}\left(\beta_j - \hat{\beta}_j\right)\right)/n - \epsilon_i^T \epsilon_j / n\right| \\
&\leq \left\|\mathbf{X}_{A^c}^T \epsilon_i / n\right\|_\infty \left\|\beta_j - \hat{\beta}_j\right\|_1 + \left\|\mathbf{X}_{A^c}^T \epsilon_j / n\right\|_\infty \left\|\beta_i - \hat{\beta}_i\right\|_1 \\
&\quad + \left\|\mathbf{X}_{A^c}\left(\beta_i - \hat{\beta}_i\right)\right\| \cdot \left\|\mathbf{X}_{A^c}\left(\beta_j - \hat{\beta}_j\right)\right\| / n \\
&\leq \left(2C'_2 C'_4 + C'_3\right) \lambda^2 s,
\end{aligned}$$

where the last step follows from inequalities (49)-(51) in Lemma 2. Thus we have

$$\mathbb{P}\left\{\left\|\hat{\Theta}_{A,A} - \Theta_{A,A}^{ora}\right\|_\infty > C_1 s \frac{\log p}{n}\right\} \leq o\left(p^{-\delta+1}\right),$$

for some $C_1 > 0$. Since $\Theta_{A,A}$ has a bounded spectrum, the functional $\zeta_{kl}\left(\Theta_{A,A}\right) = \left(\Theta_{A,A}^{-1}\right)_{kl}$ is Lipschitz in a neighborhood of $\Theta_{A,A}$ for $k, l \in A$, then Equation (17) is an immediate consequence of Equation (16).

Now we prove part (ii). Define random vector $\eta^{ora} = \left(\eta_{ii}^{ora}, \eta_{ij}^{ora}, \eta_{jj}^{ora}\right)$, where $\eta_{kl}^{ora} = \sqrt{n}\frac{\theta_{kl}^{ora} - \theta_{kl}}{\sqrt{\theta_{kk}\theta_{ll} + \theta_{kl}^2}}$. The following result is a multidimensional version of KMT quantile inequality: there exist some constants D_0, $\vartheta \in (0, \infty)$ and random normal vector $Z = \left(Z_{ii}, Z_{ij}, Z_{jj}\right) \sim \mathcal{N}\left(0, \breve{\Sigma}\right)$ with $\breve{\Sigma} = Cov(\eta^{ora})$ such that whenever $|Z_{kl}| \leq \vartheta\sqrt{n}$ for all kl, we have

$$\left\|\eta^{ora} - Z\right\|_\infty \leq \frac{D_0}{\sqrt{n}}\left(1 + Z_{ii}^2 + Z_{ij}^2 + Z_{jj}^2\right). \tag{52}$$

See Proposition [KMT] in Mason and Zhou (2012) for one dimensional case and consult Einmahl (1989) for multidimensional case. Note that $\sqrt{n}\eta^{ora}$ can be written as a sum of n i.i.d. random vectors with mean zero and covariance matrix $\breve{\Sigma}$, each of which is sub-exponentially distributed. Hence the assumptions of KMT quantile inequality in literature are satisfied. With a slight abuse of notation, we define $\Theta = (\theta_{ii}, \theta_{ij}, \theta_{jj})$. To prove the desired coupling inequality (18), let's use the Taylor expansion of the function $\omega_{ij}\left(\Theta\right) =$

$-\theta_{ij}/\left(\theta_{ii}\theta_{jj}-\theta_{ij}^2\right)$ to obtain

$$\begin{aligned} & \omega_{ij}^{ora}-\omega_{ij} \\ (53) \qquad = \; & \left\langle \nabla\omega_{ij}\left(\Theta\right),\Theta^{ora}-\Theta\right\rangle + \sum_{|\beta|=2} R_\beta\left(\Theta^{ora}\right)\left(\Theta-\Theta^{ora}\right)^\beta . \end{aligned}$$

The multi-index notation of $\beta=(\beta_1,\beta_2,\beta_3)$ is defined as $|\beta|=\sum_k\beta_k$, $x^\beta=\prod_k x_k^{\beta_k}$ and $D^\beta f\left(x\right)=\frac{\partial^{|\beta|}f}{\partial x_1^{\beta_1}\partial x_2^{\beta_2}\partial x_3^{\beta_3}}$. The derivatives can be easily computed. To save the space, we omit their explicit formulas. The coefficients in the integral form of the remainder with $|\beta|=2$ have a uniform upper bound $\left|R_\beta\left(\Theta_{A,A}^{ora}\right)\right|\leq 2\max_{|\alpha|=2}\max_{\Theta\in B}D^\alpha\omega_{ij}\left(\Theta\right)\leq C_2$, where B is some sufficiently small compact ball with center Θ when Θ^{ora} is in this ball B, which is satisfied by picking a sufficiently small value ϑ in our assumption $\left\|\eta^{ora}\right\|_\infty\leq\vartheta\sqrt{n}$. Recall that κ_{ij}^{ora} and η^{ora} are standardized versions of $\left(\omega_{ij}^{ora}-\omega_{ij}\right)$ and $\left(\Theta-\Theta^{ora}\right)$. Consequently there exist some deterministic constants h_1,h_2,h_3 and D_β with $|\beta|=2$ such that we can rewrite (53) in terms of κ_{ij}^{ora} and η^{ora} as follows,

$$\kappa_{ij}^{ora}=h_1\eta_{ii}^{ora}+h_2\eta_{ij}^{ora}+h_3\eta_{jj}^{ora}+\sum_{|\beta|=2}\frac{D_\beta R_\beta\left(\Theta^{ora}\right)}{\sqrt{n}}\left(\eta^{ora}\right)^\beta ,$$

which, together with Equation (52), completes our proof of Equation (18),

$$\left|\kappa_{ij}^{ora}-Z'\right|\leq\left(\sum_{k=1}^3|h_k|\right)\left\|Z-\eta^{ora}\right\|_\infty+\frac{C_3}{\sqrt{n}}\left\|\eta^{ora}\right\|^2\leq\frac{D_1}{\sqrt{n}}\left(1+Z_{ii}^2+Z_{ij}^2+Z_{jj}^2\right),$$

where constants $C_3,D_1\in(0,\infty)$ and $Z':=h_1Z_1+h_2Z_2+h_3Z_3\sim\mathcal{N}\left(0,1\right)$. The last inequality follows from $\left\|\eta^{ora}\right\|^2\leq C_4\left(Z_{ii}^2+Z_{ij}^2+Z_{jj}^2\right)$ for some large constant C_4, which can be shown using (52) easily.∎

6.1.2. *Proof of Theorems 3.* The triangle inequality gives

$$\begin{aligned} \left|\hat\omega_{ij}-\omega_{ij}\right| &\leq \left|\hat\omega_{ij}-\omega_{ij}^{ora}\right|+\left|\omega_{ij}^{ora}-\omega_{ij}\right|, \\ \left\|\hat\Omega_{A,A}-\Omega_{A,A}\right\|_\infty &\leq \left\|\hat\Omega_{A,A}-\Omega_{A,A}^{ora}\right\|_\infty+\left\|\Omega_{A,A}^{ora}-\Omega_{A,A}\right\|_\infty . \end{aligned}$$

From Equation (17) we have

$$\mathbb{P}\left\{\left\|\hat\Omega_{A,A}-\Omega_{A,A}^{ora}\right\|_\infty>C_1 s\frac{\log p}{n}\right\}\leq o\left(p^{-\delta+1}\right).$$

Now we give a tail bound for $\left|\omega_{ij}^{ora}-\omega_{ij}\right|$ and $\left\|\Omega_{A,A}^{ora}-\Omega_{A,A}\right\|_\infty$ respectively. For the constant $C>0$, we apply Equation (18) to obtain

$$\begin{aligned} \mathbb{P}\left\{\left|\kappa_{ij}^{ora}\right|>C\right\} &\leq \mathbb{P}\left\{\max\left\{|Z_{kl}|\right\}>\vartheta\sqrt{n}\right\}+\bar\Phi\left(\frac{C}{2}\right)+\mathbb{P}\left\{\frac{D_1}{\sqrt{n}}\left(1+Z_{ii}^2+Z_{ij}^2+Z_{jj}^2\right)>\frac{C}{2}\right\} \\ &\leq o(1)+2\exp\left(-C^2/8\right), \end{aligned}$$

according to the inequality $\bar{\Phi}(x) \leq 2\exp(-x^2/2)$ for $x > 0$ and the union bound of three normal tail probabilities. This immediately implies that for large C_4 and large n,

$$\mathbb{P}\left\{\left|\omega_{ij}^{ora} - \omega_{ij}\right| > C_4\sqrt{\frac{1}{n}}\right\} \leq \frac{3}{4}\epsilon,$$

which, together with Equations (17), yields that for $C_2 > C_1 + C_4$,

$$\mathbb{P}\left\{\left|\hat{\omega}_{ij} - \omega_{ij}\right| > C_2 \max\left\{s\frac{\log p}{n}, \sqrt{\frac{1}{n}}\right\}\right\} \leq \epsilon.$$

Similarly, Equation (18) implies

$$\begin{aligned}\mathbb{P}\left\{\left|\kappa_{ij}^{ora}\right| > C\sqrt{\log p}\right\} &\leq \mathbb{P}\left\{\max\left\{|Z_{kl}|\right\} > \vartheta\sqrt{n}\right\} + \bar{\Phi}\left(\frac{C\sqrt{\log p}}{2}\right) \\ &\quad + \mathbb{P}\left\{\frac{D_1}{\sqrt{n}}\left(1 + Z_{ii}^2 + Z_{ij}^2 + Z_{jj}^2\right) > \frac{C\sqrt{\log p}}{2}\right\} \\ &\leq O\left(p^{-C^2/8}\right),\end{aligned}$$

where the first and last components in the first inequality are negligible due to $\log p \leq c_0 n$ with a sufficiently small $c_0 > 0$, which follows from the assumption $s \leq c_0 n/\log p$. That immediately implies that for C_5 large enough,

$$\mathbb{P}\left\{\left\|\Omega_{A,A}^{ora} - \Omega_{A,A}\right\|_\infty > C_5\sqrt{\frac{\log p}{n}}\right\} = o(p^{-\delta}),$$

which, together with Equations (17), yields that for $C_3 > C_1' + C_5$.

$$\mathbb{P}\left\{\left\|\hat{\Omega}_{A,A} - \Omega_{A,A}\right\|_\infty > C_3 \max\left\{s\frac{\log p}{n}, \sqrt{\frac{\log p}{n}}\right\}\right\} \leq o\left(p^{-\delta+1}\right).$$

Thus we have the following union bound over all $\binom{p}{2}$ pairs of (i,j),

$$\mathbb{P}\left\{\left\|\hat{\Omega} - \Omega\right\|_\infty > C_3 \max\left\{s\frac{\log p}{n}, \sqrt{\frac{\log p}{n}}\right\}\right\} \leq p^2/2 \cdot o\left(p^{-\delta+1}\right) = o\left(p^{-\delta+3}\right).$$

Write

$$\sqrt{n}\left(\hat{\Omega}_{A,A} - \Omega_{A,A}\right) = \sqrt{n}\left(\hat{\Omega}_{A,A} - \Omega_{A,A}^{ora}\right) + \sqrt{n}\left(\Omega_{A,A}^{ora} - \Omega_{A,A}\right).$$

Under the assumption $s = o\left(\frac{\sqrt{n}}{\log p}\right)$, we have

$$\sqrt{n}\left\|\hat{\Omega}_{A,A} - \Omega_{A,A}^{ora}\right\|_\infty = o_p(1),$$

which together with Equation (18) further implies

$$\sqrt{n}\left(\hat{\omega}_{ij} - \omega_{ij}\right) \overset{D}{\sim} \sqrt{n}\left(\omega_{ij}^{ora} - \omega_{ij}\right) \overset{D}{\to} \mathcal{N}\left(0, \omega_{ii}\omega_{jj} + \omega_{ij}^2\right). \blacksquare$$

6.2. *Proof of Theorem 5.* In this section we show that the upper bound given in Section 2.2 is indeed rate optimal. We will only establish Equation (30). Equation (31) is an immediate consequence of Equation (30) and the lower bound $\sqrt{\frac{\log p}{n}}$ for estimation of diagonal covariance matrices in Cai, Zhang and Zhou (2010).

The lower bound is established by Le Cam's method. To introduce Le Cam's method we first need to introduce some notation. Consider a finite parameter set $\mathcal{G}_0 = \{\Omega_0, \Omega_1, \ldots, \Omega_{m_*}\} \subset \mathcal{G}_0(M, k_{n,p})$. Let $\mathbb{P}_{\Omega_m}$ denote the joint distribution of independent observations $X^{(1)}$, $X^{(2)}$,..., $X^{(n)}$ with each $X^{(i)} \sim \mathcal{N}\left(0, \Omega_m^{-1}\right)$, $0 \le m \le m_*$, and f_m denote the corresponding joint density, and we define

$$\bar{\mathbb{P}} = \frac{1}{m_*} \sum_{m=1}^{m_*} \mathbb{P}_{\Omega_m}. \tag{54}$$

For two distributions $\mathbb{P}$ and $\mathbb{Q}$ with densities p and q with respect to any common dominating measure μ, we denote the total variation affinity by $\|\mathbb{P} \wedge \mathbb{Q}\| = \int p \wedge q d\mu$. The following lemma is a version of Le Cam's method (cf. Le Cam (1973), Yu (1997)).

Lemma 3. *Let $X^{(i)}$ be i.i.d. $\mathcal{N}(0, \Omega^{-1})$, $i = 1, 2, \ldots, n$, with $\Omega \in \mathcal{G}_0$. Let $\hat{\Omega} = (\hat{\omega}_{kl})_{p \times p}$ be an estimator of $\Omega_m = \left(\omega_{kl}^{(m)}\right)_{p\times p}$, then*

$$\sup_{0 \le m \le m_*} \mathbb{P}\left\{\left|\hat{\omega}_{ij} - \omega_{ij}^{(m)}\right| > \frac{\alpha}{2}\right\} \ge \frac{1}{2}\left\|\mathbb{P}_{\Omega_0} \wedge \bar{\mathbb{P}}\right\|,$$

where $\alpha = \inf_{1 \le m \le m_*}\left|\omega_{ij}^{(m)} - \omega_{ij}^{(0)}\right|$.

Proof of Theorem 5: We shall divide the proof into three steps. Without loss of generality, consider only the cases $(i, j) = (1, 1)$ and $(i, j) = (1, 2)$. For the general case ω_{ii} or ω_{ij} with $i \ne j$, we could always permute the coordinates and rearrange them to the special case ω_{11} or ω_{12}.

Step 1: Constructing the parameter set. We first define Ω_0,

$$\Sigma_0 = \begin{pmatrix} 1 & b & 0 & \ldots & 0 \\ b & 1 & 0 & \ldots & 0 \\ 0 & 0 & 1 & \ldots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}, \text{ and } \Omega_0 = \Sigma_0^{-1} = \begin{pmatrix} \frac{1}{1-b^2} & \frac{-b}{1-b^2} & 0 & \ldots & 0 \\ \frac{-b}{1-b^2} & \frac{1}{1-b^2} & 0 & \ldots & 0 \\ 0 & 0 & 1 & \ldots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}, \tag{55}$$

i.e. $\Sigma_0 = \left(\sigma_{kl}^{(0)}\right)_{p\times p}$ is a matrix with all diagonal entries equal to 1, $\sigma_{12}^{(0)} = \sigma_{21}^{(0)} = b$ and the rest all zeros. Here the constant $0 < b < 1$ is to be determined later. For $\Omega_m, 1 \le m \le m_*$, the construction is as follows. Without loss of generality we assume $k_{n,p} \geq 2$. Denote by $\mathcal{H}$ the collection of all $p \times p$ symmetric matrices with exactly $(k_{n,p} - 1)$ elements equal to 1 between the third and the last elements on the first row (column) and the rest all zeros. Define

$$\mathcal{G}_0 = \left\{ \Omega : \Omega = \Omega_0 \text{ or } \Omega = (\Sigma_0 + aH)^{-1}, \text{ for some } H \in \mathcal{H} \right\}, \tag{56}$$

where $a = \sqrt{\frac{\tau_1 \log p}{n}}$ for some constant τ_1 which is determined later. The cardinality of $\mathcal{G}_0 / \{\Omega_0\}$ is

$$m^* = \text{Card}\,(\mathcal{G}_0) - 1 = \text{Card}\,(\mathcal{H}) = \begin{pmatrix} p-2 \\ k_{n,p} - 1 \end{pmatrix}.$$

We pick the constant $b = \frac{1}{2}(1 - 1/M)$ and $0 < \tau_1 < \min\left\{ \frac{(1-1/M)^2 - b^2}{C_0}, \frac{(1-b^2)^2}{2C_0(1+b^2)}, \frac{(1-b^2)^2}{4\nu(1+b^2)} \right\}$ such that $\mathcal{G}_0 \subset \mathcal{G}_0(M, k_{n,p})$.

First we show that for all Ω_i,

$$1/M \leq \lambda_{\min}(\Omega_i) < \lambda_{\max}(\Omega_i) \leq M. \tag{57}$$

For any matrix Ω_m, $1 \leq m \leq m_*$, some elementary calculations yield that

$$\begin{aligned} \lambda_1\left(\Omega_m^{-1}\right) &= 1 + \sqrt{b^2 + (k_{n,p} - 1)\,a^2}, \lambda_p\left(\Omega_m^{-1}\right) = 1 - \sqrt{b^2 + (k_{n,p} - 1)\,a^2}, \\ \lambda_2\left(\Omega_m^{-1}\right) &= \lambda_3\left(\Omega_m^{-1}\right) = \ldots = \lambda_{p-1}\left(\Omega_m^{-1}\right) = 1. \end{aligned}$$

Since $b = \frac{1}{2}(1 - 1/M)$ and $0 < \tau_1 < \frac{(1-1/M)^2 - b^2}{C_0}$, we can show that

$$\begin{aligned} 1 - \sqrt{b^2 + (k_{n,p} - 1)\,a^2} &\geq 1 - \sqrt{b^2 + \tau_1 C_0} > 1/M, \\ 1 + \sqrt{b^2 + (k_{n,p} - 1)\,a^2} &< 2 - 1/M < M, \end{aligned} \tag{58}$$

which imply

$$1/M \leq \lambda_1^{-1}\left(\Omega_m^{-1}\right) = \lambda_{\min}(\Omega_m) < \lambda_{\max}(\Omega_m) = \lambda_p^{-1}\left(\Omega_m^{-1}\right) \leq M.$$

As for matrix Ω_0, similarly we have

$$\begin{aligned} \lambda_1\left(\Omega_0^{-1}\right) &= 1 + b, \lambda_p\left(\Omega_0^{-1}\right) = 1 - b, \\ \lambda_2\left(\Omega_0^{-1}\right) &= \lambda_3\left(\Omega_0^{-1}\right) = \ldots = \lambda_{p-1}\left(\Omega_0^{-1}\right) = 1, \end{aligned}$$

thus $1/M \leq \lambda_{\min}(\Omega_0) < \lambda_{\max}(\Omega_0) \leq M$ for the choice of $b = \frac{1}{2}(1 - 1/M)$.

Now we show that the number of nonzero off-diagonal elements in Ω_m, $0 \leq m \leq m_*$ is no more than $k_{n,p}$ per row/column. From the construction of Ω_m^{-1}, there exists some permutation matrix P_π such that $P_\pi \Omega_m^{-1} P_\pi^T$ is a two-block diagonal matrix with dimensions $(k_{n,p}+1)$ and $(p-k_{n,p}-1)$, of which the second block is an identity matrix, then $\left(P_\pi \Omega_m^{-1} P_\pi^T\right)^{-1} = P_\pi \Omega_m P_\pi^T$ has the same blocking structure with the first block of dimension $(k_{n,p}+1)$ and the second block being an identity matrix, thus the number of nonzero off-diagonal elements is no more than $k_{n,p}$ per row/column for Ω_m. Therefore, we have $\mathcal{G}_0 \subset \mathcal{G}_0(M, k_{n,p})$ from Equation (57).

Step 2: Bounding α. From the construction of Ω_m^{-1} and the matrix inverse formula, it can be shown that for any precision matrix Ω_m we have

$$\omega_{11}^{(m)} = \frac{1}{1-b^2-(k_{n,p}-1)\,a^2}, \text{ and } \omega_{12}^{(m)} = \frac{-b}{1-b^2-(k_{n,p}-1)\,a^2},$$

for $1 \le m \le m_*$, and for precision matrix Ω_0 we have

$$\omega_{11}^{(0)} = \frac{1}{1-b^2}, \omega_{12}^{(0)} = \frac{-b}{1-b^2}.$$

Since $b^2 + (k_{n,p}-1)\,a^2 < (1-1/M)^2 < 1$ in Equation (58), we have

$$\begin{aligned}
\text{(59)} \qquad \inf_{1\le m\le m_*}\left|\omega_{11}^{(m)} - \omega_{11}^{(0)}\right| &= \frac{(k_{n,p}-1)\,a^2}{(1-b^2)\left(1-b^2-(k_{n,p}-1)\,a^2\right)} \ge C_3 k_{n,p} a^2,\\
\inf_{1\le m\le m_*}\left|\omega_{12}^{(m)} - \omega_{12}^{(0)}\right| &= \frac{b\,(k_{n,p}-1)\,a^2}{(1-b^2)\left(1-b^2-(k_{n,p}-1)\,a^2\right)} \ge C_4 k_{n,p} a^2,
\end{aligned}$$

for some constants $C_3, C_4 > 0$.

Step 3: Bounding the affinity. The following lemma will be proved in Section 8.

LEMMA 4. *Let $\bar{\mathbb{P}}$ be defined in (54). We have*

$$\text{(60)} \qquad \left\|\mathbb{P}_{\Omega_0} \wedge \bar{\mathbb{P}}\right\| \ge C_5$$

for some constant $C_5 > 0$.

Lemma 3, together with Equations (59), (60) and $a = \sqrt{\frac{\tau_1 \log p}{n}}$, imply

$$\begin{aligned}
\sup_{0\le m\le m_*} \mathbb{P}\left\{\left|\hat\omega_{11} - \omega_{11}^{(m)}\right| > \frac{1}{2}\cdot\frac{C_3\tau_1 k_{n,p}\log p}{n}\right\} &\ge C_5/2,\\
\sup_{0\le m\le m_*} \mathbb{P}\left\{\left|\hat\omega_{12} - \omega_{12}^{(m)}\right| > \frac{1}{2}\cdot\frac{C_4\tau_1 k_{n,p}\log p}{n}\right\} &\ge C_5/2,
\end{aligned}$$

which match the lower bound in (30) by setting $C_1 = \min\{C_3\tau_1/2, C_4\tau_1/2\}$ and $c_1 = C_5/2$.

REMARK 10. *Note that $|||\Omega_m|||_1$ is at order of $k_{n,p}\sqrt{\frac{\log p}{n}}$, which implies $\frac{k_{n,p}\log p}{n} = k_{n,p}\sqrt{\frac{\log p}{n}} \cdot \sqrt{\frac{\log p}{n}} \asymp |||\Omega_m|||_1 \sqrt{\frac{\log p}{n}}$. This observation partially explains why in literature* *people need assume bounded matrix l_1 norm of Ω to derive the lower bound rate $\sqrt{\frac{\log p}{n}}$. For the least favorable parameter space, the matrix l_1 norm of Ω cannot be avoided in the upper bound. But the methodology proposed in this paper improves the upper bounds in literature by replacing the matrix l_1 norm for every Ω by only matrix l_1 norm bound of Ω in the least favorable parameter space.*

6.3. *Proof of Theorem 1.* The probabilistic results (i) and (ii) are the immediate consequences of Theorems 2 and 5. We only need to show the minimax rate of convergence result (3). According to the probabilistic lower bound result (30) in Theorem 5, we immediately obtain that

$$\inf_{\hat{\omega}_{ij}} \sup_{\mathcal{G}_0(M,k_{n,p})} \mathbb{E}\left|\hat{\omega}_{ij}-\omega_{ij}\right| \geq c_1 \max\left\{C_1 \frac{k_{n,p}\log p}{n}, C_2\sqrt{\frac{1}{n}}\right\}.$$

Thus it is enough to show there exists some estimator of ω_{ij} such that it attains this upper bound. More precisely, we define the following estimator based on $\hat{\omega}_{ij}$ defined in (12) to control the improbable case for which $\hat{\Theta}_{A,A}$ is nearly singular.

$$\breve{\omega}_{ij} = sgn(\hat{\omega}_{ij}) \cdot \min\left\{\left|\hat{\omega}_{ij}\right|, \log p\right\}.$$

Define the event $G = \left\{\left|\hat{\omega}_{ij} - \omega_{ij}^{ora}\right| \leq C_1 \frac{k_{n,p}\log p}{n}, \left|\omega_{ij}^{ora}\right| \leq 2M\right\}$. Note that the Equations (16) and (18) in Theorem 2 imply $\mathbb{P}\left\{G^c\right\} \leq C\left(p^{-\delta+1} + \exp\left(-cn\right)\right)$ for some constants C and c. Now according to the variance of inverse Wishart distribution, we pick $\delta \geq 2\xi+1$ to complete our proof

$$\begin{aligned}
\mathbb{E}\left|\breve{\omega}_{ij}-\omega_{ij}\right| &\leq \mathbb{E}\left(\left|\breve{\omega}_{ij}-\omega_{ij}^{ora}\right| 1\left\{G\right\}\right) + \mathbb{E}\left(\left|\breve{\omega}_{ij}-\omega_{ij}^{ora}\right| 1\left\{G^c\right\}\right) + \mathbb{E}\left|\omega_{ij}^{ora}-\omega_{ij}\right| \\
&\leq C_1 \frac{k_{n,p}\log p}{n} + \left(\mathbb{P}\left\{G^c\right\}\mathbb{E}\left(\log p + \left|\omega_{ij}^{ora}\right|\right)^2\right)^{1/2} + \left(\mathbb{E}\left(\omega_{ij}^{ora}-\omega_{ij}\right)^2\right)^{1/2} \\
&\leq C_1 \frac{k_{n,p}\log p}{n} + C_2 p^{-\frac{\delta+1}{2}}\log p + C_3\frac{1}{\sqrt{n}} \\
&\leq C' \max\left\{\frac{k_{n,p}\log p}{n}, \sqrt{\frac{1}{n}}\right\},
\end{aligned}$$

where C_2, C_3 and C' are some constants and the last equation follows from the assumption $n = O\left(p^{\xi}\right)$.

7. Proof of Theorems 6-9.

7.1. *Proof of Theorem 6.* When $\delta > 3$, from Theorem 2 it can be shown that the following three results hold:

(i). For any constant $\varepsilon > 0$, we have

$$\mathbb{P}\left\{\sup_{(i,j)}\left|\frac{\hat{\omega}_{ii}\hat{\omega}_{jj}+\hat{\omega}_{ij}^2}{\omega_{ii}\omega_{jj}+\omega_{ij}^2}-1\right|>\varepsilon\right\}\to 0; \tag{61}$$

(ii). There is a constant $C_1 > 0$ such that

$$\mathbb{P}\left\{\sup_{(i,j)}\left|\omega_{ij}^{ora}-\hat{\omega}_{ij}\right|>C_1 s\frac{\log p}{n}\right\}\to 0; \tag{62}$$

(iii). For any constant $2 < \xi_1$, we have

$$\mathbb{P}\left\{\sup_{(i,j)}\frac{\left|\omega_{ij}^{ora}-\omega_{ij}\right|}{\sqrt{\omega_{ii}\omega_{jj}+\omega_{ij}^2}}>\sqrt{\frac{2\xi_1\log p}{n}}\right\}\to 0. \tag{63}$$

In fact, under the assumption $\delta \geq 3$, Equation (17) in Theorem 2 and the union bound over all pair (i, j) imply the second result (62), which further shows the first result (61) because that $\hat{\omega}_{ij}$ and $\hat{\omega}_{ii}$ are consistent estimators and $\omega_{ii}\omega_{jj} + \omega_{ij}^2$ is bounded below and above. For the third result, we apply Equation (18) from Theorem 2 and pick $2 < \xi_2 < \xi_1$ and $a = \sqrt{\xi_1} - \sqrt{\xi_2}$ to show that

$$\begin{aligned}\mathbb{P}\left\{\ \left|\kappa_{ij}^{ora}\right|>\sqrt{2\xi_1\log p}\right\} &\leq \mathbb{P}\left\{\max\left\{|Z_{kl}|\right\}>\vartheta\sqrt{n}\right\}+\bar{\Phi}\left(\sqrt{2\xi_2\log p}\right)\\ &\quad+\mathbb{P}\left\{\frac{D_1}{\sqrt{n}}\left(1+Z_{ii}^2+Z_{ij}^2+Z_{jj}^2\right)>a\sqrt{2\log p}\right\}\\ &\leq O(p^{-\xi_2}\sqrt{\frac{1}{\log p}}),\end{aligned}$$

where the last inequality follows from $\log p = o(n)$. The third result (63) is thus obtained by the union bound with $2 < \xi_2$.

Essentially Equation (36) and Equation (37) are equivalent to each other. Thus we only show that Equation (37) in Theorem 6 is just a simple consequence of results (i), (ii) and (iii). Set $\varepsilon > 0$ sufficiently small and $\xi \in (2, \xi_0)$ sufficiently close to 2 such that

$2\sqrt{2\xi_0}-\sqrt{2\xi_0\left(1+\varepsilon\right)}>\sqrt{2\xi}$ and $\xi_0\left(1-\varepsilon\right)>\xi$, and $2<\xi_1<\xi$. We have

$$\begin{aligned}
&\mathbb{P}\left(\mathcal{S}(\hat{\Omega}_{thr})=\mathcal{S}(\Omega)\right)\\
=\;&\mathbb{P}\;\left(\hat{\omega}_{ij}^{thr}\neq 0\text{ for all }\;(i,j)\in\mathcal{S}(\Omega)\right)+\mathbb{P}\left(\hat{\omega}_{ij}^{thr}=0\text{ for all }\;(i,j)\notin\mathcal{S}(\Omega)\right)\\
=\;&\mathbb{P}\left\{\;|\hat{\omega}_{ij}|>\sqrt{\frac{2\xi_0\left(\hat{\omega}_{ii}\hat{\omega}_{jj}+\hat{\omega}_{ij}^2\right)\log p}{n}}\text{ for all }\;(i,j)\in\mathcal{S}(\Omega)\right\}\\
&+\mathbb{P}\left\{|\hat{\omega}_{ij}|\leq\sqrt{\frac{2\xi_0\left(\hat{\omega}_{ii}\hat{\omega}_{jj}+\hat{\omega}_{ij}^2\right)\log p}{n}}\text{ for all }\;(i,j)\notin\mathcal{S}(\Omega)\right\}\\
\geq\;&\mathbb{P}\left\{\sup_{(i,j)}\frac{|\hat{\omega}_{ij}-\omega_{ij}|}{\sqrt{\omega_{ii}\omega_{jj}+\omega_{ij}^2}}\leq\sqrt{\frac{2\xi\log p}{n}}\right\}-\mathbb{P}\left\{\sup_{(i,j)}\left|\frac{\hat{\omega}_{ii}\hat{\omega}_{jj}+\hat{\omega}_{ij}^2}{\omega_{ii}\omega_{jj}+\omega_{ij}^2}-1\right|>\varepsilon\right\},
\end{aligned}$$

which is bounded below by

$$\mathbb{P}\left\{\sup_{(i,j)}\frac{\left|\omega_{ij}^{ora}-\omega_{ij}\right|}{\sqrt{\omega_{ii}\omega_{jj}+\omega_{ij}^2}}\leq\sqrt{\frac{2\xi_1\log p}{n}}\right\}-\left[\begin{array}{c}\mathbb{P}\left\{\sup_{(i,j)}\left|\omega_{ij}^{ora}-\hat{\omega}_{ij}\right|>C_1 s\frac{\log p}{n}\right\}+\\ \mathbb{P}\left\{\sup_{(i,j)}\left|\frac{\hat{\omega}_{ii}\hat{\omega}_{jj}+\hat{\omega}_{ij}^2}{\omega_{ii}\omega_{jj}+\omega_{ij}^2}-1\right|>\varepsilon\right\}\end{array}\right]=1+o\left(1\right),$$

where $s=o\left(\sqrt{n/\log p}\right)$ implies $s\frac{\log p}{n}=o\left(\sqrt{(\log p)/n}\right)$.∎

7.2. *Proof of Theorem 8.* The proof of this Theorem is very similar to that of Theorem 2. Due to the limit of space, we follow the line of the proof of Theorem 2 and Theorem 3, but only give necessary details when the proof is different from that of Theorem 2. Note that β for the latent variable graphical model is not sparse enough in the sense that

$$\max_j\Sigma_{i\neq j}\min\left\{1,\frac{|s_{ij}-l_{ij}|}{\lambda}\right\}\neq o\left(\frac{\sqrt{n}}{\log p}\right).$$

Our strategy is to decompose it into two parts,

$$\beta=S_{O\backslash A,A}\Omega_{A,A}^{-1}-L_{O\backslash A,A}\Omega_{A,A}^{-1}:=\beta^S-\beta^L,\tag{64}$$

where $\beta^S=S_{O\backslash A,A}\Omega_{A,A}^{-1}$ and $\beta^L=L_{O\backslash A,A}\Omega_{A,A}^{-1}$ correspond the sparse and low-rank components respectively. We expect the penalized estimator $\hat{\beta}$ in (10) is closer to the sparse part β^S than β itself, which motivates us to rewrite the regression model as follows,

$$\mathbf{X}_A=\mathbf{X}_{O\backslash A}\beta^S+\left(\epsilon_A-\mathbf{X}_{O\backslash A}\beta^L\right):=\mathbf{X}_{O\backslash A}\beta^S+\epsilon_A^S,\tag{65}$$

with $\epsilon_A^S=\left(\epsilon_A-\mathbf{X}_{O\backslash A}\beta^L\right)$, and define

$$\Theta_{A,A}^{ora,S}=\left(\epsilon_A^S\right)^T\left(\epsilon_A^S\right)/n.\tag{66}$$

We can show our estimator $\hat{\Theta}_{A,A}$ is within a small ball of the "oracle" $\Theta_{A,A}^{ora,S}$ with radius at the rate of $k_{n,p}\lambda^2$, where $k_{n,p}$ is the sparsity of model (65). More specifically, similar to Lemma 2 for the proof of Theorem 2 we have the following result for the latent variable graphical model. The proof is provided in the supplementary material.

Lemma 5. *Let $\lambda = (1+\varepsilon)\sqrt{\frac{2\delta \log p}{n}}$ for any $\delta \geq 1$ and $\varepsilon > 0$ in Equation (10). Define the event E as follows,*

$$\begin{aligned}
\left|\hat{\theta}_{mm} - \theta_{mm}^{ora,S}\right| &\leq C_1' k_{n,p}\lambda^2, \\
\left\|\beta_m^S - \hat{\beta}_m\right\|_1 &\leq C_2' k_{n,p}\lambda, \\
\left\|\mathbf{X}_{O\backslash A}\left(\beta_m^S - \hat{\beta}_m\right)\right\|^2/n &\leq C_3' k_{n,p}\lambda^2, \\
\left\|\mathbf{X}_{O\backslash A}^T \epsilon_m^S/n\right\|_\infty &\leq C_4'\lambda.
\end{aligned} \tag{67}$$

for $m = i$ and j and some constants C_k', $1 \leq k \leq 4$. Under the assumptions in Theorem 8, we have

$$\mathbb{P}\left\{E^c\right\} \leq o\left(p^{-\delta+1}\right).$$

Similar to the argument in Section 6.1.1, from Lemma 5 we have

$$\left|\hat{\theta}_{ij} - \theta_{ij}^{ora,S}\right| = \left|\left(\epsilon_i^S\right)^T\left(\epsilon_j^S\right)/n - \hat{\epsilon}_i^T\hat{\epsilon}_j/n\right| \leq C k_{n,p}\lambda^2$$

on the event E, thus there is a constant $C_1 > 0$ such that

$$\mathbb{P}\left\{\left\|\hat{\Theta}_{A,A} - \Theta_{A,A}^{ora,S}\right\|_\infty > C_1 k_{n,p}\frac{\log p}{n}\right\} \leq o\left(p^{-\delta+1}\right).$$

Later we will show that there are some constant C_2 and C_3 such that

$$\mathbb{P}\left\{\left\|\Theta_{A,A}^{ora} - \Theta_{A,A}^{ora,S}\right\|_\infty > C_2 k_{n,p}\frac{\log p}{n}\right\} \leq C_3 p^{-2\delta}, \tag{68}$$

which implies

$$\mathbb{P}\left\{\left\|\hat{\Theta}_{A,A} - \Theta_{A,A}^{ora}\right\|_\infty > C_4 k_{n,p}\frac{\log p}{n}\right\} \leq o\left(p^{-\delta+1}\right),$$

for some constant $C_4 > 0$. Then following the proof of Theorem 3 exactly, we establish Theorem 8.

Now we conclude the proof by establishing Equation (68). In fact, we will only show that

$$\mathbb{P}\left\{\left|\theta_{i,i}^{ora} - \theta_{i,i}^{ora,S}\right| > C_2 k_{n,p}\frac{\log p}{n}\right\} \leq C_5 p^{-2\delta},$$

for some $C_5 > 0$. The tail bounds for $\left|\theta_{j,j}^{ora} - \theta_{j,j}^{ora,S}\right|$ and $\left|\theta_{i,j}^{ora} - \theta_{i,j}^{ora,S}\right|$ can be shown similarly. Write

(69)
$$\theta_{i,i}^{ora} - \theta_{i,i}^{ora,S} = \frac{\epsilon_i^T \epsilon_i - \left(\epsilon_i^S\right)^T \left(\epsilon_i^S\right)}{n} = \frac{\epsilon_i^T \epsilon_i}{n} - \left(\epsilon_i - \mathbf{X}_{O\backslash A}\beta_i^L\right)^T \left(\epsilon_i - \mathbf{X}_{O\backslash A}\beta_i^L\right)/n = D_1 + D_2.$$

where

$$\begin{aligned} D_1 &= \frac{2}{n}\epsilon_i^T \mathbf{X}_{O\backslash A}\beta_i^L = \frac{2}{n}\sum_{k=1}^{n} \epsilon_{i,k} \cdot \left(X_{O\backslash A}^{(k)}\beta_i^L\right) \\ D_2 &= \frac{1}{n}\left(\beta_i^L\right)^T \mathbf{X}_{O\backslash A}^T \mathbf{X}_{O\backslash A}\beta_i^L \sim var\left(X_{O\backslash A}^{(k)}\beta_i^L\right) \cdot \chi_{(n)}^2/n. \end{aligned}$$

It is then enough to show that there are constants C_2 and C_5 such that

$$\mathbb{P}\left\{|D_i| > \frac{C_2}{2}k_{n,p}\frac{\log p}{n}\right\} \le \frac{C_5}{2}p^{-2\delta}$$

for each D_i, $i = 1, 2$. We first study D_1, which is an average of n i.i.d. random variables $\epsilon_{i,k} \cdot \left(X_{O\backslash A}^{(k)}\beta_i^L\right)$ with

$$\epsilon_i \sim \mathcal{N}\left(0, \Omega_{A,A}^{-1}\right), \text{ and } X_{O\backslash A}^{(k)}\beta_i^L \sim \mathcal{N}\left(0, \left(\beta_i^L\right)^T \Sigma_{O\backslash A, O\backslash A}\beta_i^L\right),$$

where $\Omega_{A,A}^{-1}$ has bounded spectrum, and $\left(\beta_i^L\right)^T \Sigma_{O\backslash A, O\backslash A}\beta_i^L \le \frac{C_6}{p}$ for some C_6 for the assumption (43) that elements of β^L are at an order of $\frac{C_5}{p}$. From classical large deviations bounds, there exist some uniform constants $c_1, c_2 > 0$ such that

$$\mathbb{P}\left\{\left|\frac{\sum_{k=1}^n \epsilon_{i,k} \cdot \sqrt{p}X_{O\backslash A}^{(k)}\beta_i^L}{n}\right| > t\right\} \le 2\exp\left(-nt^2/c_2\right) \quad \text{for } 0 < t < c_1.$$

See, for example, Theorem 2.8 of Petrov (1995). By setting $t = \sqrt{\frac{2\delta c_2 \log p}{n}} = o\left(1\right)$, we have

$$\mathbb{P}\left\{|D_1| > 2\sqrt{\frac{2\delta c_2 \log p}{np}}\right\} \le 2p^{-2\delta}, \tag{70}$$

where $2\sqrt{\frac{2\delta c_2 \log p}{np}} = o\left(k_{n,p}\frac{\log p}{n}\right)$ from Equation (45). The derivation for the tail bound of D_2 is similar by a large deviation bound for $\chi_{(n)}^2$,

$$\mathbb{P}\left\{\frac{\chi_{(n)}^2}{n} > 1 + t\right\} \le 2\exp\left(-nt^2/c_2\right) \quad \text{for } 0 < t < c_1,$$

which implies

$$\mathbb{P}\left\{|D_2| > var\left(X_{O\backslash A}^{(k)}\beta_i^L\right)\left(1 + \sqrt{\frac{2\delta c_2 \log p}{n}}\right)\right\} \le 2p^{-2\delta}, \tag{71}$$

by setting $t = \sqrt{\frac{2\delta c_2 \log p}{n}} = o(1)$, where $var\left(X^{(k)}_{O\backslash A}\beta^L_i\right)\left(1+\sqrt{\frac{2\delta c_2 \log p}{n}}\right) = O(1/p) = o\left(k_{n,p}\frac{\log p}{n}\right)$ from Equation (45). Equations (70) and (71) imply the desired tail bound (68).

8. Proof of Auxiliary Lemmas.

In this section we prove two key lemmas, Lemmas 2 and 4, for establishing our main results.

8.1. *Proof of Lemma 2.* We first reparameterize Equations (7) and (10) by setting

$$d_k := \frac{\|\mathbf{X}_k\|}{\sqrt{n}} b_k, \quad \text{and} \quad \mathbf{Y} := \mathbf{X}_{A^c} \cdot diag\left(\frac{\sqrt{n}}{\|\mathbf{X}_k\|}\right)_{k\in A^c} \tag{72}$$

to make the analysis cleaner, and then rewrite the regression (7) and the penalized procedure (10) as follows,

$$\mathbf{X}_m = \mathbf{Y} d^{true} + \epsilon_m. \tag{73}$$

and

$$L_\lambda(d,\sigma) \quad : \quad = \frac{\|\mathbf{X}_m - \mathbf{Y}d\|^2}{2n\sigma} + \frac{\sigma}{2} + \lambda \|d\|_1, \tag{74}$$

$$\left\{\hat{d},\hat{\sigma}\right\} \quad : \quad = \arg\min_{d\in\mathbb{R}^{p-2},\sigma\in\mathbb{R}} L_\lambda(d,\sigma), \tag{75}$$

where the true coefficients of the reparameterized regression (74) are

$$d^{true} = \left(d^{true}_k\right)_{k\in A^c}, \quad \text{where } d^{true}_k = \frac{\|\mathbf{X}_k\|}{\sqrt{n}}\beta_{m,k}. \tag{76}$$

Then the oracle estimator of σ can be written as

$$\sigma^{ora} := (\theta^{ora}_{mm})^{1/2} = \frac{\|\mathbf{X}_m - \mathbf{Y}d^{true}\|}{\sqrt{n}} = \frac{\|\epsilon_m\|}{\sqrt{n}}, \tag{77}$$

and we have the following relationship between $\left\{\hat{d},\hat{\sigma}\right\}$ and $\left\{\hat{\beta}_m,\hat{\theta}_{mm}\right\}$,

$$\hat{\beta}_{m,k} = \hat{d}_k \frac{\sqrt{n}}{\|\mathbf{X}_k\|}, \text{ and } \hat{\theta}_{mm} = \hat{\sigma}^2. \tag{78}$$

The proof has two parts. The first part is the algebraic analysis of the solution $\left\{\hat{d},\hat{\sigma}\right\}$ to the regression (74) by which we can define an event E such that Equations (48)-(51) in Lemma 2 hold. The second part of the proof is the probabilistic analysis of the E.

8.1.1. *Algebraic Analysis.* The function $L_\lambda(d,\sigma)$ is jointly convex in (d,σ). For fixed σ, denote the minimizer of $L_\lambda(d,\sigma)$ over all $d \in \mathbb{R}^{p-2}$ by $\hat{d}(\sigma\lambda)$, a function of $\sigma\lambda$, i.e.,

$$\hat{d}(\sigma\lambda) = \arg\min_{d\in\mathbb{R}^{p-2}} L_\lambda(d,\sigma) = \arg\min_{d\in\mathbb{R}^{p-2}} \left\{ \frac{\|\mathbf{X}_m - \mathbf{Y}d\|^2}{2n} + \lambda\sigma \|d\|_1 \right\}. \tag{79}$$

then if we knew $\hat{\sigma}$ in the solution of Equation (75), the solution for the equation is $\left\{\hat{d}(\hat{\sigma}\lambda), \hat{\sigma}\right\}$. Let $\mu = \lambda\sigma$. From the Karush-Kuhn-Tucker condition, $\hat{d}(\mu)$ is the solution of Equation (79) if and only if

$$\begin{aligned} \mathbf{Y}_k^T\left(\mathbf{X}_m - \mathbf{Y}\hat{d}(\mu)\right)/n &= \mu \cdot sgn\left(\hat{d}_k(\mu)\right), \text{ if } \hat{d}_k(\mu) \neq 0, \\ \mathbf{Y}_k^T\left(\mathbf{X}_m - \mathbf{Y}\hat{d}(\mu)\right)/n &\in [-\mu,\mu], \text{ if } \hat{d}_k(\mu) = 0. \end{aligned} \tag{80}$$

To define the event E we need to introduce some notation. Define the l_1 cone invertibility factor (CIF_1) as follows,

$$CIF_1(\alpha, K, \mathbf{Y}) = \inf\left\{ \frac{|K| \left\| \frac{\mathbf{Y}^T\mathbf{Y}}{n} u \right\|_\infty}{\|u_K\|_1} : u \in \mathcal{C}(\alpha, K), u \neq 0 \right\}, \tag{81}$$

where $|K|$ is the cardinality of an index set K, and

$$\mathcal{C}(\alpha, K) = \left\{ u \in \mathbb{R}^{p-2} : \|u_{K^c}\|_1 \leq \alpha \|u_K\|_1 \right\}.$$

Let

$$T = \left\{ k \in A^c, \left| d_k^{true} \right| \geq \lambda \right\} \tag{82}$$

$$\nu = \left\| \mathbf{Y}^T\left(\mathbf{X}_m - \mathbf{Y}d^{true}\right)/n \right\|_\infty = \left\| \mathbf{Y}^T\epsilon_m/n \right\|_\infty, \tag{83}$$

and

$$\tau = \lambda \max\left\{ \frac{4(1+\xi)}{\sigma^{ora}} \left\| \left(d^{true}\right)_{T^c} \right\|_1, \frac{8\lambda |T|}{CIF_1(2\xi+1, T, \mathbf{Y})} \right\} \tag{84}$$

for some $\xi > 1$ be specified later. Define

$$E = \cap_{i=1}^4 I_i \tag{85}$$

where

$$I_1 = \left\{ \nu \leq \sigma^{ora}\lambda \frac{\xi - 1}{\xi + 1}(1-\tau) \right\}, \tag{86}$$

$$I_2 = \left\{ CIF_1(2\xi+1, T, \mathbf{Y}) \geq \frac{C_{cif}}{(\xi+1)^2} > 0 \right\} \text{ with } C_{cif} = 1/\left(10\sqrt{2M^3}\right) \tag{87}$$

$$I_3 = \left\{ \sigma^{ora} \in \left[\sqrt{1/(2M)}, \sqrt{2M} \right] \right\} \tag{88}$$

$$I_4 = \left\{ \frac{\|\mathbf{X}_k\|}{\sqrt{n}} \in \left[\sqrt{1/(2M)}, \sqrt{2M} \right] \text{ for all } k \in A^c \right\}. \tag{89}$$

Now we show Equations (48)-(51) in Lemma 2 hold on the event E defined in Equation (85). The following two results are helpful to establish our result. Their proofs are given in the supplementary material.

Proposition 1. *For any $\xi > 1$, on the event $\left\{\nu \le \mu\frac{\xi-1}{\xi+1}\right\}$, we have*

$$\left\|\hat{d}(\mu) - d^{true}\right\|_1 \le \max\left\{(2+2\xi)\left\|\left(d^{true}\right)_{T^c}\right\|_1, \frac{(\nu+\mu)|T|}{CIF_1(2\xi+1, T, \mathbf{Y})}\right\}, \tag{90}$$

$$\frac{1}{n}\left\|\mathbf{Y}\left(d^{true} - \hat{d}(\mu)\right)\right\|^2 \le (\nu+\mu)\left\|\hat{d}(\mu) - d^{true}\right\|_1. \tag{91}$$

Proposition 2. *Let $\left\{\hat{d}, \hat{\sigma}\right\}$ be the solution of the scaled lasso (75). For any $\xi > 1$, on the event $I_1 = \left\{\nu \le \sigma^{ora}\lambda\frac{\xi-1}{\xi+1}(1-\tau)\right\}$, we have*

$$\left|\frac{\hat{\sigma}}{\sigma^{ora}} - 1\right| \le \tau. \tag{92}$$

Note that $s = \max_j \Sigma_{i\neq j}\min\left\{1, \frac{|\omega_{ij}|}{\lambda}\right\}$ is defined in terms of Ω which has bounded spectrum, then on the event I_4,

$$\max\left\{\left\|\left(d^{true}\right)_{T^c}\right\|_1, \lambda|T|\right\} \le C\lambda s$$

from the definition of T in Equation (82). On their intersection $I_1 \cap I_2 \cap I_3$, Proposition 2 and Proposition 1 with $\mu = \lambda\hat{\sigma}$ imply that there exist some constants c_1, c_2 and c_3 such that

$$\begin{aligned} |\hat{\sigma} - \sigma^{ora}| &\le c_1\lambda\max\left\{\left\|\left(d^{true}\right)_{T^c}\right\|_1, \lambda|T|\right\}, \\ \left\|\hat{d}(\mu) - d^{true}\right\|_1 &\le c_2\max\left\{\left\|\left(d^{true}\right)_{T^c}\right\|_1, \lambda|T|\right\}, \\ \frac{\left\|\mathbf{Y}\left(d^{true} - \hat{d}\right)\right\|^2}{n} &\le c_3\lambda\max\left\{\left\|\left(d^{true}\right)_{T^c}\right\|_1, \lambda|T|\right\}. \end{aligned}$$

then from the definition of β and Equations (6) and (78),

$$\hat{\beta}_{m,k} = \hat{d}_k\frac{\sqrt{n}}{\|\mathbf{X}_k\|},\ \hat{\theta}_{mm} = \hat{\sigma}^2,\ \text{and } \mathbf{X}_{A^c} = \mathbf{Y}\cdot diag\left(\frac{\|\mathbf{X}_k\|}{\sqrt{n}}\right)_{k\in A^c},$$

we immediately have

$$\begin{aligned} \left|\hat{\theta}_{mm} - \hat{\theta}_{mm}^{ora}\right| &= |\hat{\sigma} - \sigma^{ora}|\cdot|\hat{\sigma} + \sigma^{ora}| \le C_1'\lambda^2 s, \\ \left\|\beta_m - \hat{\beta}_m\right\|_1 &\le C_2'\lambda s, \\ \left\|\mathbf{X}_{A^c}\left(\beta_m - \hat{\beta}_m\right)\right\|^2/n &\le C_3'\lambda^2 s, \end{aligned}$$

for some constants C_i', which are exactly Equations (48)-(50). From the definition of events I_1, I_3 and I_4 we obtain Equation (51), i.e., $\left\|\mathbf{X}_{A^c}^T\epsilon_m/n\right\|_\infty \le C_4'\lambda$.

8.1.2. *Probabilistic Analysis.* We will show that

$$\begin{aligned}\mathbb{P}\left\{I_1^c\right\} &\leq O\left(p^{-\delta+1}/\sqrt{\log p}\right), \\ \mathbb{P}\left\{I_i^c\right\} &\leq o(p^{-\delta}) \text{ for } i=2,3 \text{ and } 4,\end{aligned}$$

which implies

$$\mathbb{P}\left\{E\right\} \geq 1-o\left(p^{-\delta+1}\right).$$

We will first consider $\mathbb{P}\left\{I_3^c\right\}$ and $\mathbb{P}\left\{I_4^c\right\}$, then $\mathbb{P}\left\{I_2^c\right\}$, and leave $\mathbb{P}\left\{I_1^c\right\}$ to the last, which relies on the bounds for $\mathbb{P}\left\{I_i^c\right\}$, $2 \leq i \leq 4$.

(1). To study $\mathbb{P}\left\{I_3^c\right\}$ and $\mathbb{P}\left\{I_4^c\right\}$, we need the following tail bound for the chi-squared distribution with n degrees of freedom,

$$\mathbb{P}\left\{\left|\frac{\chi^2_{(n)}}{n}-1\right| \geq t\right\} \leq 2\exp\left(-nt\left(t\wedge 1\right)/8\right), \tag{93}$$

for $t>0$. Since $\sigma^{ora}=\left\|\epsilon_m\right\|/\sqrt{n}$ with $\epsilon_m \sim \mathcal{N}\left(0,\theta_{mm}I_n\right)$, and $\mathbf{X}_k \sim \mathcal{N}\left(0,\sigma_{kk}I_n\right)$ with $\sigma_{kk} \in (1/M, M)$, we have

$$n\left(\sigma^{ora}\right)^2/\theta_{mm} \sim \chi^2_{(n)}, \text{ and } \left\|\mathbf{X}_k\right\|^2/\sigma_{kk} \sim \chi^2_{(n)},$$

then Equation (93) implies

$$\begin{aligned}\mathbb{P}\left\{I_3^c\right\} &= \mathbb{P}\left\{\left(\sigma^{ora}\right)^2 \notin \left[1/\left(2M\right), 2M\right]\right\} \leq \mathbb{P}\left\{\left|\frac{\left(\sigma^{ora}\right)^2}{\theta_{mm}}-1\right| \geq \frac{1}{2}\right\} \\ &\leq 2\exp\left(-n/32\right) \leq o(p^{-\delta}),\end{aligned} \tag{94}$$

and

$$\mathbb{P}\left\{I_4^c\right\} = \mathbb{P}\left\{\frac{\left\|\mathbf{X}_k\right\|^2}{n} \notin \left[1/\left(2M\right), 2M\right] \text{ for some } k \in A^c\right\} \leq 2p\exp\left(-n/32\right) \leq o(p^{-\delta}). \tag{95}$$

(2). To study the term $CIF_1\left(2\xi+1, T, \mathbf{Y}\right)$ of the event I_2, we need to introduce some notation first. Define

$$\pi_a^{\pm}\left(\mathbf{Y}\right) = \max_G\left\{\pm\left(\left\|\mathbf{Y}_A^T\mathbf{Y}_A u/n\right\|-1\right)\right\}, \text{ and } \theta_{a,b}\left(\mathbf{Y}\right) = \max_G v^T\mathbf{Y}_A^T\mathbf{Y}_B u/n.$$

where

$$G = \left\{(A,B,u,v) : \left(|A|, |B|, \|u\|, \|v\|\right) = (a,b,1,1) \text{ with } A \cap B = \emptyset\right\}.$$

then $1+\pi_a^+\left(\mathbf{Y}\right)$ and $1-\pi_a^-\left(\mathbf{Y}\right)$ are the maximal and minimal eigenvalues of all submatrices of $\mathbf{Y}^T\mathbf{Y}/\mathbf{n}$ with dimensions no greater than a respectively, and $\theta_{a,b}\left(\mathbf{Y}\right)$ satisfies that

$$\theta_{a,b}\left(\mathbf{Y}\right) \leq \left(1+\pi_a^+\left(\mathbf{Y}\right)\right)^{1/2}\left(1+\pi_b^+\left(\mathbf{Y}\right)\right)^{1/2} \leq 1+\pi_{a\vee b}^+\left(\mathbf{Y}\right). \tag{96}$$

The following two propositions will be used to establish the probability bound for $\mathbb{P}\left\{I_2^c\right\}$. Proposition 3 follows from Zhang and Huang (2008) Proposition 2(i). The proof of Proposition 4 is given in the supplementary material.

PROPOSITION 3. *Let rows of the data matrix $\mathbf{X}$ be i.i.d. copies of $\mathcal{N}(0,\Sigma)$, and denote the minimal and maximal eigenvalues of Σ by $\lambda_{\min}(\Sigma)$ and $\lambda_{\max}(\Sigma)$ respectively. Assume that $m \leq c\frac{n}{\log p}$ with a sufficiently small constant $c>0$, then for $\varepsilon>0$ we have*

$$\mathbb{P}\left\{(1-h)^2\lambda_{\min}(\Sigma) \leq 1-\pi_m^-(\mathbf{X}) \leq 1+\pi_m^+(\mathbf{X}) \leq (1+h)^2\lambda_{\max}(\Sigma)\right\} \geq 1-\varepsilon, \tag{97}$$

where $h=\sqrt{\frac{m}{n}}+\sqrt{2\frac{m\log p-\log(\varepsilon/2)}{n}}$.

PROPOSITION 4. *For $CIF_1(\alpha,K,\mathbf{Y})$ defined in (81) with $|K|=k$, we have, for any $0<l\leq p-k$,*

$$CIF_1(\alpha,K,\mathbf{Y}) \geq \frac{1}{(1+\alpha)\left((1+\alpha)\wedge\sqrt{1+\frac{l}{k}}\right)}\left(1-\pi_{l+k}^-(\mathbf{Y})-\alpha\sqrt{\frac{k}{4l}}\theta_{4l,k+l}(\mathbf{Y})\right). \tag{98}$$

Note that there exists some constant C such that Cs is an upper bound of the $|T|$ from the definitions of s and set T (82) on I_4. For $l \geq Cs$ Proposition 4 gives

$$\begin{aligned} CIF_1(2\xi+1,T,\mathbf{Y}) &\geq \frac{1}{4(1+\xi)^2}\left(1-\pi_{l+|T|}^-(\mathbf{Y})-(2\xi+1)\sqrt{\frac{|T|}{4l}}\theta_{4l,|T|+l}(\mathbf{Y})\right)\\ &\geq \frac{1}{4(1+\xi)^2}\left(1-\pi_{4l}^-(\mathbf{Y})-(2\xi+1)\sqrt{\frac{Cs}{4l}}\left(1+\pi_{4l}^+(\mathbf{Y})\right)\right), \end{aligned} \tag{99}$$

where the second inequality follows from (96). From the definition $\mathbf{Y}=\mathbf{X}_{A^c}\cdot diag\left(\frac{\sqrt{n}}{\|\mathbf{X}_k\|}\right)_{k\in A^c}$ and the property that $\pi_a^{\pm}(\mathbf{Y})$ are increasing as functions of a, we have

$$\begin{aligned} 1+\pi_{4l}^+(\mathbf{Y}) &\leq \max_{k\in A^c}\left\{\frac{\sqrt{n}}{\|\mathbf{X}_k\|}\right\}\left(1+\pi_{4l}^+(\mathbf{X}_{A^c})\right) \leq \max_{k\in A^c}\left\{\frac{\sqrt{n}}{\|\mathbf{X}_k\|}\right\}\left(1+\pi_{4l}^+(\mathbf{X})\right),\\ \min_{k\in A^c}\left\{\frac{\sqrt{n}}{\|\mathbf{X}_k\|}\right\}\left(1-\pi_{4l}^-(\mathbf{X})\right) &\leq \min_{k\in A^c}\left\{\frac{\sqrt{n}}{\|\mathbf{X}_k\|}\right\}\left(1-\pi_{4l}^-(\mathbf{X}_{A^c})\right) \leq 1-\pi_{4l}^-(\mathbf{Y}), \end{aligned}$$

and by applying Proposition 3 to the data matrix $\mathbf{Y}$ with $m=4l=\left(4(2\xi+1)M^3\right)^2Cs>$

Cs and $\varepsilon = p^{-2\delta}$, we have on I_4

$$\begin{aligned}
& 1 - \pi_{4l}^{-}(\mathbf{Y}) - (2\xi+1)\sqrt{\frac{Cs}{4l}}\left(1+\pi_{4l}^{+}(\mathbf{Y})\right) \\
\geq\ & (1-h)^2 \lambda_{\min}(\Sigma) \min_{k\in A^c}\left\{\frac{\sqrt{n}}{\|\mathbf{X}_k\|}\right\} - (2\xi+1)\sqrt{\frac{Cs}{4l}}(1+h)^2 \lambda_{\max}(\Sigma) \max_{k\in A^c}\left\{\frac{\sqrt{n}}{\|\mathbf{X}_k\|}\right\} \\
\geq\ & (1-h)^2 \frac{1}{\sqrt{2M}M} - (2\xi+1)\sqrt{\frac{Cs}{4l}}(1+h)^2\sqrt{2M}M \\
\geq\ & \frac{1}{\sqrt{2M^3}}(1-h)^2 - \frac{1}{2\sqrt{2M^3}}(1+h)^2 \geq \frac{4}{10\sqrt{2M^3}}
\end{aligned}$$

with probability at least $1-\varepsilon$, where $h = \sqrt{\frac{m}{n}} + \sqrt{2\frac{m\log p - \log\left(p^{-2\delta}/2\right)}{n}} = o(1)$. Note $\mathbb{P}\left\{I_4^c\right\} \leq o(p^{-\delta})$, thus we established that $\mathbb{P}\left\{I_2^c\right\} \leq o(p^{-\delta})$.

(3). Finally we study the probability of event I_1. The following tail probability of t distribution is helpful in the analysis.

Proposition 5. *Let T_n follows a t distribution with n degrees of freedom. Then there exists $\varepsilon_n \to 0$ as $n \to \infty$ such that $\forall t > 0$*

$$\mathbb{P}\left\{T_n^2 > n\left(e^{2t^2/(n-1)} - 1\right)\right\} \leq (1+\varepsilon_n)\, e^{-t^2}/\left(\pi^{1/2}t\right).$$

Please refer to Sun and Zhang (2012b) Lemma 1 for the proof. According to the definition of ν in Equation (83) we have

$$\frac{\nu}{\sigma^{ora}} = \max_{k\in A^c}|h_k|, \text{ with } h_k = \frac{\mathbf{Y}_k^T\epsilon_m}{n\sigma^{ora}} \text{ for } k \in A^c.$$

Note that each column of $\mathbf{Y}$ has norm $\|\mathbf{Y}_k\| = \sqrt{n}$ by the normalization step (72). Given $\mathbf{X}_{A^c}$, equivalently $\mathbf{Y}$, we have

$$\begin{aligned}
\frac{\sqrt{n-1}h_k}{\sqrt{1-h_k^2}} &= \sqrt{\frac{n-1}{n}}\frac{\left(\mathbf{Y}_k^T\epsilon_m/\sqrt{n\theta_{mm}}\right)/\left(\sqrt{\epsilon_m^T\epsilon_m/n\theta_{mm}}\right)}{\left(\sqrt{\left(\epsilon_m^T\epsilon_m - \epsilon_m^T\mathbf{Y}_k\mathbf{Y}_k^T\epsilon_m/n\right)/n\theta_{mm}}\right)/\left(\sqrt{\epsilon_m^T\epsilon_m/n\theta_{mm}}\right)} \\
&= \frac{\left(\mathbf{Y}_k^T\epsilon_m/\sqrt{n\theta_{mm}}\right)}{\left(\sqrt{\left\|P_{\mathbf{Y}_k^c}\epsilon_m\right\|^2/(n-1)\,\theta_{mm}}\right)} \sim t_{(n-1)},
\end{aligned}$$

where $t_{(n-1)}$ is t distribution with $n-1$ degrees of freedom, since the numerator follows a standard normal and the denominator follows an independent $\sqrt{\chi^2_{(n-1)}/(n-1)}$. From

Proposition 5 we have

$$\begin{aligned} & \mathbb{P}\left\{|h_k| > \sqrt{\frac{2t^2}{n}}\right\} \\ = \; & \mathbb{P}\left\{\frac{(n-1)\,h_k^2}{1-h_k^2} > \frac{2\,(n-1)\,t^2/n}{1-2t^2/n}\right\} \le \mathbb{P}\left\{\frac{(n-1)\,h_k^2}{1-h_k^2} > \frac{2\,(n-1)\,t^2/\,(n-2)}{1-t^2/\,(n-2)}\right\} \\ \le \; & \mathbb{P}\left\{\frac{(n-1)\,h_k^2}{1-h_k^2} > (n-1)\left(e^{2t^2/(n-2)}-1\right)\right\} \le (1+\varepsilon_{n-1})\,e^{-t^2}/\left(\pi^{1/2}t\right). \end{aligned}$$

where the first inequality holds when $t^2 \ge 2$, and the second inequality which follows the fact $e^x - 1 \le x/(1-\frac{x}{2})$ for $0 < x < 2$. Now let $t^2 = \delta \log p > 2$, and $\lambda = \left(\sqrt{2\delta}\,(1+\varepsilon)\right)\sqrt{\frac{\log p}{n}}$ with $\xi = 3/\varepsilon + 1$, then we have $\lambda\frac{\xi-1}{\xi+1}(1-\tau) \ge \sqrt{\frac{2\delta \log p}{n}}$ for sufficiently small τ. Clearly, the τ defined in Equation (84) satisfies $\tau = O\left(s\lambda^2\right)$ which is sufficiently small on $I_2 \cap I_3 \cap I_4$. Therefore we have

$$\begin{aligned} \mathbb{P}\left\{\cap_{i=1}^{4} I_i\right\} \; & \ge \; \mathbb{P}\left\{\frac{\nu}{\sigma^{ora}} \le \sqrt{\frac{2\delta \log p}{n}}\right\} - \mathbb{P}\left\{(I_2 \cap I_3 \cap I_4)^c\right\} \\ & \ge \; 1 - p \cdot \mathbb{P}\left\{|h_k| > \sqrt{\frac{2\delta \log p}{n}}\right\} - \mathbb{P}\left\{(I_2 \cap I_3 \cap I_4)^c\right\} \ge 1 - O\left(\frac{p^{-\delta+1}}{\sqrt{\log p}}\right), \qquad (100) \end{aligned}$$

which implies immediately $\mathbb{P}\left\{I_1^c\right\} \le O\left(p^{-\delta+1}/\sqrt{\log p}\right)$.

8.2. *Proof of Lemma 4.* Now we establish the lower bound (60) for the total variation affinity. Since the affinity $\int q_0 \wedge q_1 d\mu = 1 - \frac{1}{2}\int |q_0 - q_1|\, d\mu$ for any two densities q_0 and q_1, Jensen's Inequality implies

$$\left[\int |q_0 - q_1|\, d\mu\right]^2 = \left(\int \left|\frac{q_0 - q_1}{q_0}\right| q_0 d\mu\right)^2 \le \int \frac{(q_0 - q_1)^2}{q_0} d\mu = \int \frac{q_1^2}{q_0} d\mu - 1.$$

Hence $\int q_0 \wedge q_1 d\mu \ge 1 - \frac{1}{2}\left(\int \frac{q_1^2}{q_0} d\mu - 1\right)^{1/2}$. To establish (60), it thus suffices to show that

$$\Delta = \int \frac{\left(\frac{1}{m_*}\sum_{m=1}^{m_*} f_m\right)^2}{f_0} - 1 = \frac{1}{m_*^2}\sum_{m,l}\int\left(\frac{f_m f_l}{f_0} - 1\right) \to 0.$$

The following lemma is used to calculate the term $\int (f_m f_l / f_0 - 1)$ in Δ.

Lemma 6. *Let g_s be the density function of $\mathcal{N}(0, \Sigma_s)$, $s = 0, m$ or l. Then*

$$\int \frac{g_m g_l}{g_0} = \left[\det\left(I - \Sigma_0^{-1}(\Sigma_m - \Sigma_0)\,\Sigma_0^{-1}(\Sigma_l - \Sigma_0)\right)\right]^{-1/2}.$$

Let $\Sigma_m = \Omega_m^{-1}$ for $0 \leq m \leq m_*$. Lemma 6 implies

$$\int \frac{f_m f_l}{f_0} = \left(\int \frac{g_m g_l}{g_0} \right)^n = \left[\det \left(I - \Omega_0 \left(\Sigma_m - \Sigma_0 \right) \Omega_0 \left(\Sigma_l - \Sigma_0 \right) \right) \right]^{-n/2}.$$

Let $J(m,l)$ be the number of overlapping nonzero off-diagonal elements between Σ_m and Σ_l in the first row. Recall the simple structures of Ω_0 (55) and $\Sigma_m - \Sigma_0$ by our construction. Elementary calculations yield that

$$\det \left(I - \Omega_0 \left(\Sigma_m - \Sigma_0 \right) \Omega_0 \left(\Sigma_l - \Sigma_0 \right) \right) = (1 - \frac{1+b^2}{\left(1-b^2\right)^2} J a^2)^2,$$

which is 1 when $J = 0$. Now we set $d := \frac{1+b^2}{(1-b^2)^2} > 1$ to simplify our notation. It is easy to see that the total number of pairs (Σ_m, Σ_l) such that $J(m,l) = j$ is $\binom{p-2}{k_{n,p}-1} \binom{k_{n,p}-1}{j} \binom{p-1-k_{n,p}}{k_{n,p}-1-j}$. Hence,

$$\begin{aligned} \Delta &= \frac{1}{m_*^2} \sum_{0 \leq j \leq k_{n,p}-1} \sum_{J(m,l)=j} \int \left(\frac{f_m f_l}{f_0} - 1 \right) \\ &= \frac{1}{m_*^2} \sum_{0 \leq j \leq k_{n,p}-1} \sum_{J(m,l)=j} \left((1 - dja^2)^{-n} - 1 \right) \\ &\leq \frac{1}{m_*^2} \sum_{1 \leq j \leq k_{n,p}-1} \binom{p-2}{k_{n,p}-1} \binom{k_{n,p}-1}{j} \binom{p-1-k_{n,p}}{k_{n,p}-1-j} (1 - dja^2)^{-n}. \qquad (101) \end{aligned}$$

Note that

$$(1 - dja^2)^{-n} \leq (1 + 2dja^2)^n \leq \exp\left(n 2 dja^2 \right) = p^{2d\tau_1 j}$$

where the first inequality follows from the fact that $dja^2 \leq dk_{n,p} a^2 \leq \frac{1+b^2}{(1-b^2)^2} \tau_1 C_0 < 1/2$. Hence,

$$\begin{aligned} \Delta &\leq \sum_{1 \leq j \leq k_{n,p}-1} \frac{\binom{k_{n,p}-1}{j} \binom{p-1-k_{n,p}}{k_{n,p}-1-j}}{\binom{p-2}{k_{n,p}-1}} p^{2d\tau_1 j} \\ &= \sum_{1 \leq j \leq k_{n,p}-1} \frac{1}{j!} \frac{\left(\frac{(k_{n,p}-1)!}{(k_{n,p}-1-j)!} \right)^2}{\frac{(p-2)!(p-2k_{n,p}+j)!}{[(p-1-k_{n,p})!]^2}} p^{2d\tau_1 j} \\ &\leq \sum_{1 \leq j \leq k_{n,p}-1} \left(\frac{k_{n,p}^2 p^{2d\tau_1}}{p - k_{n,p} - 1} \right)^j, \end{aligned}$$

where the last inequality follows from the facts that $\frac{(k_{n,p}-1)!}{(k_{n,p}-1-j)!}$ is a product of j terms with each term less than $k_{n,p}$ and $\frac{(p-2)!(p-2k_{n,p}+j)!}{[(p-1-k_{n,p})!]^2}$ is bounded below by a product of j terms with each term greater than $(p - k_{n,p} - 1)$. Recall the assumption (27) $p \geq k_{n,p}^v$. So for large enough p, we have $p - k_{n,p} - 1 \geq p/2$ and

$$\begin{aligned} k_{n,p}^2 \frac{p^{2d\tau_1}}{p - k_{n,p} - 1} &\leq 2p^{2/\nu} \cdot \frac{p^{2d\tau_1}}{p} \\ &\leq 2p^{-(v-2)/(2v)}, \end{aligned}$$

where the last step follows from the fact that $\tau_1 \leq (\nu - 2) / (4\nu d)$. Thus

$$\Delta \leq 2 \sum_{1 \leq j \leq k_{n,p}-1} p^{-j(v-2)/(2v)} \to 0,$$

which immediately implies (60). ∎

SUPPLEMENTARY MATERIAL

Supplement to "Asymptotic Normality and Optimalities in Estimation of Large Gaussian Graphical Model"
(doi: 10.1214/00-AOSXXXXSUPP). In this supplement we collect proofs for proving auxiliary Lemma 1 and 5 and Proposition 1, 2 and 4.

References.

Belloni, A., Chernozhukov, V. and Hansen, C. (2012). Inference on treatment effects after selection. *arXiv preprint arXiv:1201.0224.*

Bickel, P. J. and Levina, E. (2008a). Regularized estimation of large covariance matrices. *Ann. Statist.* **36** 199-227.

Bickel, P. J. and Levina, E. (2008b). Covariance regularization by thresholding. *Ann. Statist.* **36** 2577-2604.

Bühlmann, P. (2012). Statistical significance in high-dimensional linear models. *arXiv preprint arXiv:1202.1377.*

Cai, T. T., Liu, W. and Luo, X. (2011). A constrained ℓ_1 minimization approach to sparse precision matrix estimation. *J. Amer. Statist. Assoc.* **106** 594-607.

Cai, T. T., Liu, W. and Zhou, H. H. (2012). Estimating Sparse Precision Matrix: Optimal Rates of Convergence and Adaptive Estimation. *Manuscript.*

Cai, T. T., Zhang, C. H. and Zhou, H. H. (2010). Optimal rates of convergence for covariance matrix estimation. *Ann. Statist.* **38** 2118-2144.

Cai, T. T. and Zhou, H. H. (2012). Optimal rates of convergence for sparse covariance matrix estimation. *Ann. Statist.* **40** 2389–2420.

Candès, E. J. and Recht, B. (2009). Exact matrix completion via convex optimization. *Found. of Comput. Math.* **9** 717-772.

Chandrasekaran, V., Parrilo, P. A. and Willsky, A. S. (2012). Latent variable graphical model selection via convex optimization. *Ann. Statist.* **40** 1935-1967.

d'Aspremont, A., Banerjee, O. and El Ghaoui, L. (2008). First-order methods for sparse covariance selection. *SIAM Journal on Matrix Analysis and its Applications* **30** 56-66.

Einmahl, U. (1989). Extensions of results of Komlós, Major, and Tusnády to the multivariate case. *J. Multivariate Anal.* **28** 20–68.

El Karoui, N. (2008). Operator norm consistent estimation of large dimensional sparse covariance matrices. *Ann. Statist.* **36** 2717-2756.

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

Horn, R. A. and Johnson, C. R. (1990). *Matrix Analysis.* Cambridge University Press.

Javanmard, A. and Montanari, A. (2013). Hypothesis testing in high-dimensional regression under the gaussian random design model: Asymptotic theory. *arXiv preprint arXiv:1301.4240*.

Lam, C. and Fan, J. (2009). Sparsistency and rates of convergence in large covariance matrix estimation. *Ann. Statist.* **37** 4254-4278.

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

Le Cam, L. (1973). Convergence of estimates under dimensionality restrictions. *Ann. Statist.* **1** 38-53.

Liu, W. (2013). Gaussian Graphical Model Estimation with False Discovery Rate Control. *arXiv preprint arXiv:1306.0976*.

Mason, D. and Zhou, H. H. (2012). Quantile Coupling Inequalities and Their Applications. *Probability Surveys* **9** 439-479.

Meinshausen, N. and Bühlmann, P. (2006). High dimensional graphs and variable selection with the Lasso. *Ann. Statist.* **34** 1436-1462.

Petrov, V. V. (1995). *Limit theorems of probability theory.* Oxford Science Publications.

Ravikumar, P., Wainwright, M. J., Raskutti, G. and Yu, B. (2011). High-dimensional covariance estimation by minimizing ℓ_1 penalized log-determinant divergence. *Electron. J. Statist.* **5** 935-980.

Ren, Z. and Zhou, H. H. (2012). Discussion : Latent variable graphical model selection via convex optimization. *Ann. Statist.* **40** 1989-1996.

Rothman, A., Bickel, P., Levian, E. and Zhu, J. (2008). Sparse permutation invariant covariance estimation. *Electron. J. Statist.* **2** 494-515.

Sun, T. and Zhang, C. H. (2012a). Sparse matrix inversion with scaled Lasso. *Manuscript.*

Sun, T. and Zhang, C. H. (2012b). Scaled Sparse Linear Regression. *Biometrika* **99** 879–898.

Sun, T. and Zhang, C. H. (2012c). Comment: Minimax estimation of large covariance matrices under l1 norm. *Statist. Sinica* **22** 1354–1358.

Thorin, G. O. (1948). Convexity theorems generalizing those of M. Riesz and Hadamard with some applications. *Comm. Sem. Math. Univ. Lund.* **9** 1-58.

van de Geer, S., Bühlmann, P. and Ritov, Y. (2013). On asymptotically optimal confidence regions and tests for high-dimensional models. *arXiv preprint arXiv:1303.0518*.

Yu, B. (1997). Assouad, Fano, and Le Cam. *Festschrift for Lucien Le Cam* 423–435.

Yuan, M. (2010). Sparse inverse covariance matrix estimation via linear programming. *J. Mach. Learn. Res.* **11** 2261-2286.

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

Zhang, C.-H. (2011). Statistical inference for high-dimensional data. In *Mathematisches Forschungsinstitut Oberwolfach: Very High Dimensional Semiparametric Models, Report No. 48/2011* 28-31.

Zhang, C. H. and Huang, J. (2008). The sparsity and bias of the Lasso selection in high-dimensional linear regression. *Ann. Statist.* **36** 1567-1594.

Zhang, C.-H. and Zhang, S. S. (2011). Confidence intervals for low-dimensional parameters with high-dimensional data. *arXiv preprint arXiv:1110.2563*.

Department of Statistics
Yale University
New Haven, Connecticut 06511
USA
E-mail: zhao.ren@yale.edu
huibin.zhou@yale.edu

Department of Statistics
The Wharton School
University of Pennsylvania
Philadelphia, Pennsylvania 19104
USA
E-mail: tingni@wharton.upenn.edu

Department of Statistics and Biostatistics
Hill Center, Busch Campus
Rutgers University
Piscataway, New Jersey 08854
USA
E-mail: cunhui@stat.rutgers.edu