The Annals of Statistics
2015, Vol. 43, No. 3, 991–1026
DOI: 10.1214/14-AOS1286

ASYMPTOTIC NORMALITY AND OPTIMALITIES IN ESTIMATION OF LARGE GAUSSIAN GRAPHICAL MODELS

BY ZHAO REN*, TINGNI SUN†, CUN-HUI ZHANG[1,‡] AND HARRISON H. ZHOU[2,§]

*University of Pittsburgh**, *University of Maryland*†, *Rutgers University*‡ *and Yale 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 of this is 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, irrepresentability condition on the Hessian tensor operator of the covariance matrix or the ℓ_1 constraint on the precision matrix. 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.

Received August 2013; revised October 2014.

[1]Supported in part by the NSF Grants DMS-11-06753 and DMS-12-09014 and NSA Grant H98230-11-1-0205.

[2]Supported in part by NSF Career Award DMS-06-45676 and NSF FRG Grant DMS-08-54975.

MSC2010 subject classifications. Primary 62H12; secondary 62F12, 62G09.

Key words 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.

1. Introduction. The 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 models is how 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, that is, $(i, j) \in E$, 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, that is, 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 $n^{-1/2}$ 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 the literature. The first approach 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 selection consistency and related error bounds 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) for Frobenius-based error bounds and Lam and Fan (2009) for concave penalized likelihood estimation without the irrepresentability condition. The second approach, proposed earlier by Meinshausen and Bühlmann (2006), is neighborhood-based. It estimates the precision matrix column by column by running the lasso or Dantzig selector for each variable against all the rest of variables; see Yuan (2010), Cai, Liu and Luo (2011), Cai, Liu and Zhou (2012) and Sun and Zhang (2013). 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 (2013), 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 were not analyzed.

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

Statistical inference of low-dimensional parameters at the $n^{-1/2}$ rate has been considered in the closely related linear regression model. Sun and Zhang (2012a) 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 coefficient vector. Zhang and Zhang (2014) proposed an asymptotically normal low-dimensional projection estimator (LDPE) for the regression coefficients under the same sample size condition. Both estimators converge at the $n^{-1/2}$ rate, and their asymptotic efficiency can be understood from the minimum Fisher information in a more general context [Zhang (2011)]. A proof of the asymptotic efficiency of the LDPE was given in van de Geer et al. (2014) where the generalized linear model was also considered. Alternative methods for testing and estimation of regression coefficients were proposed in Belloni, Chernozhukov and Hansen (2014), Bühlmann (2013), Javanmard and Montanari (2014) and Liu (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. Please see Section 5.3 for more details of their connection with this paper.

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 $n^{-1/2}$ convergence rate requires the sparsity condition $s = O(n^{1/2}/\log p)$ 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 $s \leq c_0 n/\log p$ 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 (2012b). 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 (2012a),

but to estimate the noise level. For a vector Z of length p and any index subset A of $\{1, 2, \ldots, p\}$, we denote by Z_A the sub-vector of Z with elements indexed by A. Similarly for a matrix U and two index subsets A and B of $\{1, 2, \ldots, p\}$, we denote by $U_{A,B}$ the $|A| \times |B|$ sub-matrix of U with elements in rows in A and columns in B. Consider $A = \{i, j\}$ with $i \neq j$, so that $Z_A = (Z_i, Z_j)^T$ and $\Omega_{A,A} = \left(\begin{smallmatrix} \omega_{ii} & \omega_{ij} \\ \omega_{ji} & \omega_{jj} \end{smallmatrix}\right)$. It is well known that

$$Z_A | Z_{A^c} \sim \mathcal{N}\big(-\Omega_{A,A}^{-1} \Omega_{A,A^c} Z_{A^c}, \Omega_{A,A}^{-1}\big).$$

This observation motivates us to consider the estimation of individual entries of Ω, ω_{ii} and ω_{ij}, by estimating the noise level in the regression of the two response variables in A against the variables in A^c. 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} in the following sense: The estimator is asymptotically normal, and its asymptotic 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 $k_{n,p}$ nonzero elements in each column,

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

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 $n^{-1/2}$-consistent estimation of ω_{ij} is $k_{n,p} = O(\frac{\sqrt{n}}{\log p})$, and when $k_{n,p} = o(\frac{\sqrt{n}}{\log p})$ 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 $3 \le k_{n,p} \le c_0 n / \log p$ with a sufficiently small constant $c_0 > 0$ and $p \ge k_{n,p}^{\nu}$ with some $\nu > 2$.*

(i) *There exists a constant $\varepsilon_0 > 0$ such that*

$$\inf_{i,j} \inf_{\hat{\omega}_{ij}} \sup_{\mathcal{G}_0(M,k_{n,p})} \mathbb{P}\big\{|\hat{\omega}_{ij} - \omega_{ij}| \ge \varepsilon_0 \max\{n^{-1} k_{n,p} \log p, n^{-1/2}\}\big\} \ge \varepsilon_0.$$

Moreover, the minimax risk of estimating ω_{ij} over the class $\mathcal{G}_0(M, k_{n,p})$ satisfies

$$\inf_{\hat{\omega}_{ij}} \sup_{\mathcal{G}_0(M,k_{n,p})} \mathbb{E}|\hat{\omega}_{ij} - \omega_{ij}| \asymp \max\{n^{-1} k_{n,p} \log p, n^{-1/2}\} \tag{2}$$

uniformly in (i, j), provided that $n = O(p^{\xi})$ with some $\xi > 0$.

(ii) *The estimator $\hat{\omega}_{ij}$ defined in (10) in Section 2 is rate optimal in the sense of*

$$\lim_{(C,n) \to (\infty,\infty)} \max_{i,j} \sup_{\mathcal{G}_0(M,k_{n,p})} \mathbb{P}\big\{|\hat{\omega}_{ij} - \omega_{ij}| \ge C \max\{n^{-1} k_{n,p} \log p, n^{-1/2}\}\big\} = 0.$$

Furthermore, the estimator $\hat{\omega}_{ij}$ *is asymptotically efficient when* $k_{n,p} = o(\frac{\sqrt{n}}{\log p})$, *that is, with* $F_{ij} = (\omega_{ii}\omega_{jj} + \omega_{ij}^2)^{-1}$ *being the Fisher information for estimating* ω_{ij} *and* $\hat{F}_{ij} = (\hat{\omega}_{ii}\hat{\omega}_{jj} + \hat{\omega}_{ij}^2)^{-1}$ *its estimate*,

$$\sqrt{n\hat{F}_{ij}}(\hat{\omega}_{ij} - \omega_{ij}) \xrightarrow{D} \mathcal{N}(0, 1), \qquad \hat{F}_{ij}/F_{ij} \to 1. \tag{3}$$

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 matrices is different from that for sparse covariance matrices when $k_{n,p} \gg (\frac{\sqrt{n}}{\log p})$, and when $k_{n,p} = o(\frac{\sqrt{n}}{\log p})$ the difficulty of support recovery for sparse precision matrices is just the same as that for sparse covariance matrices.

It is worthwhile to point out that the asymptotic efficiency result is obtained without the need to assume the irrepresentability condition or the ℓ_1 constraint of the precision matrix which are commonly required in the literature. For preconceived (i, j), two immediate consequences of (3) are efficient interval estimation of ω_{ij} and efficient test for the existence of an edge between V_i and V_j in the graphical model, that is, the hypotheses $\omega_{ij} = 0$. However, the impact of Theorem 1 is much broader. We derive fully adaptive thresholded versions of the estimator and prove that the thresholded estimators achieve rate optimality in support recovery without assuming the irrepresentability condition and in various matrix norms for the estimation of the entire precision matrix Ω under weaker assumptions than the requirements of existing results in the literature. In addition, we extend our inference and estimation framework to a class of latent variable graphical models. See Section 3 for details.

Our work on optimal estimation of precision matrices given in the present paper is closely connected to a growing literature on the estimation of large covariance matrices. Many regularization methods have been proposed and studied. For example, Bickel and Levina (2008a, 2008b) 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. More recently, Sun and Zhang (2013) proposed to apply a scaled lasso to estimate Ω and proved its rate optimality in spectrum norm without imposing an ℓ_1 norm assumption on Ω.

The proposed estimator was briefly described in Sun and Zhang (2012b) along with a statement of the efficiency of the estimator without proof under the spar-

sity assumption $k_{n,p} = o(n^{-1/2} \log p)$. While we are working on the delicate issue of the necessity of the sparsity condition $k_{n,p} = o(n^{1/2}/\log p)$ and the optimality of the method for support recovery and estimation under the general sparsity condition $k_{n,p} = o(n/\log p)$, Liu (2013) developed p-values for testing $\omega_{ij} = 0$ and related FDR control methods under the stronger sparsity condition $k_{n,p} = o(n^{1/2}/\log p)$. However, his method cannot be directly 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 the estimation under the spectral norm, support recovery and the estimation of latent variable graphical models are presented in Section 3. Results on linear regression are presented in Section 4 to support the main theory. Section 5 discusses possible extensions of our results and the connection between our and existing results. Numerical studies are presented in Section 6. The proof for the novel lower bound result is given in Section 7. Additional proofs are provided in Ren et al. (2015).

Notation. We summarize here some notation to be used throughout the paper. For $1 \le w \le \infty$, 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 introduce our methodology for estimating each entry and more generally, a smooth functional of any square submatrix of fixed size. Asymptotic efficiency results are stated in Section 2.3 under a sparseness assumption. The lower bound in Section 2.4 shows that the sparseness condition is sharp for the asymptotic efficiency proved in Section 2.3.

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 $A = \{i, j\}$:

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

Equivalently we write a bivariate linear model

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

where the coefficients and error distributions are

$$\boldsymbol{\beta} = \boldsymbol{\beta}_{A^c,A} = -\Omega_{A^c,A}\Omega_{A,A}^{-1}, \qquad (\eta_i, \eta_j)^T \sim \mathcal{N}\big(0, \Omega_{A,A}^{-1}\big). \tag{6}$$

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 the 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 the data matrix is the ith sample $X^{(i)}$. Let $\mathbf{X}_A$ be the sub-matrix of $\mathbf{X}$ composed of columns indexed by A. Based on the regression interpretation (5), we have the following data version of the multivariate regression model

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

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

$$\Theta_{A,A}^{\mathrm{ora}} = (\theta_{ij}^{\mathrm{ora}})_{i\in A, j\in A} = \boldsymbol{\varepsilon}_A^T\boldsymbol{\varepsilon}_A/n, \tag{8}$$

which is an oracle MLE of $\Theta_{A,A}$ based on the extra knowledge of $\boldsymbol{\beta}$. The oracle MLE of $\Omega_{A,A}$ is

$$\Omega_{A,A}^{\mathrm{ora}} = (\omega_{ij}^{\mathrm{ora}})_{i\in A, j\in A} = (\Theta_{A,A}^{\mathrm{ora}})^{-1}. \tag{9}$$

Of course $\boldsymbol{\beta}$ is unknown, and we will need to estimate $\boldsymbol{\beta}$ and plug in its estimator to estimate $\boldsymbol{\varepsilon}_A$. This general scheme can be formally written as

$$\hat{\Omega}_{A,A} = (\hat{\omega}_{ij})_{i\in A, j\in A} = \hat{\Theta}_{A,A}^{-1}, \qquad \hat{\Theta}_{A,A} = (\hat{\theta}_{ij})_{i,j\in A} = \hat{\boldsymbol{\varepsilon}}_A^T\hat{\boldsymbol{\varepsilon}}_A/n, \tag{10}$$

where $\hat{\boldsymbol{\varepsilon}}_A$ is the estimated residual corresponding to a suitable estimator of $\boldsymbol{\beta}_{A^c,A}$, that is,

$$\hat{\boldsymbol{\varepsilon}}_A = \mathbf{X}_A - \mathbf{X}_{A^c}\hat{\boldsymbol{\beta}}_{A^c,A}. \tag{11}$$

Now we introduce specific estimators of $\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}_{A^c,A} = (\hat{\boldsymbol{\beta}}_i, \hat{\boldsymbol{\beta}}_j)$. For each $m \in A = \{i, j\}$, we apply a scaled lasso estimator to the univariate linear regression of $\mathbf{X}_m$ against $\mathbf{X}_{A^c}$ as follows:

$$\{\hat{\boldsymbol{\beta}}_m, \hat{\theta}_{mm}^{1/2}\} = \mathop{\arg\min}_{b\in\mathbb{R}^{p-2}, \sigma\in\mathbb{R}^+} \left\{ \frac{\|\mathbf{X}_m - \mathbf{X}_{A^c}b\|^2}{2n\sigma} + \frac{\sigma}{2} + \lambda \sum_{k\in A^c} \frac{\|\mathbf{X}_k\|}{\sqrt{n}}|b_k| \right\}, \tag{12}$$

with a weighted ℓ_1 penalty, where the vector b is indexed by A^c. This is equivalent to standardizing the design vector to length $\sqrt{n}$ and then applying the ℓ_1 penalty to the new coefficients $(\|\mathbf{X}_k\|/\sqrt{n})b_k$. The penalty level λ will be specified explicitly later. It can be shown that the definition of $\hat{\theta}_{mm}$ in (10) is consistent with the $\hat{\theta}_{mm}$ obtained from the scaled lasso (12) for each $m \in A$ and each A. We also consider the following least squares estimator (LSE) in the model $\hat{S}_{mm}$ selected in (12):

$$\{\hat{\boldsymbol{\beta}}_m, \hat{\theta}_{mm}^{1/2}\} = \mathop{\arg\min}_{b\in\mathbb{R}^{p-2}, \sigma\in\mathbb{R}^+} \left\{ \frac{\|\mathbf{X}_m - \mathbf{X}_{A^c}b\|^2}{2n\sigma} + \frac{\sigma}{2} : \mathrm{supp}(b) \subseteq \hat{S}_{mm} \right\}, \tag{13}$$

where $\operatorname{supp}(b)$ denotes the support of vector b.

Different versions of scaled lasso, in the sense of scale-free simultaneous estimation of the regression coefficients and noise level, have been considered in Städler, Bühlmann and van de Geer (2010), Antoniadis (2010) and Sun and Zhang (2010, 2012a) among others. The $\hat{\boldsymbol{\beta}}_m$ in (12) is equivalent to the square-root lasso in Belloni, Chernozhukov and Wang (2011). Theoretical properties of the LSE after model selection, given in (13), were studied in Sun and Zhang (2012a, 2013).

Our 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

$$Z_B | Z_{B^c} = \mathcal{N}\big(-\Omega_{B,B}^{-1}\Omega_{B,B^c}Z_{B^c}, \Omega_{B,B}^{-1}\big), \tag{14}$$

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

$$\zeta = \zeta\big(\Omega_{B,B}^{-1}\big).$$

When $\boldsymbol{\beta}_{B^c,B}$ 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 maximum likelihood estimator of ζ can be defined as

$$\zeta^{\mathrm{ora}} = \zeta\big(\boldsymbol{\varepsilon}_B^T\boldsymbol{\varepsilon}_B/n\big).$$

We apply an adaptive regularized estimator $\hat{\boldsymbol{\beta}}_{B^c,B}$ by regressing $\mathbf{X}_B$ against $\mathbf{X}_{B^c}$, for example, a penalized LSE or the LSE after model selection. We estimate the residual matrix by $\hat{\boldsymbol{\varepsilon}}_B = \mathbf{X}_B - \mathbf{X}_{B^c}\hat{\boldsymbol{\beta}}_{B^c,B}$, and $\zeta(\Omega_{B,B}^{-1})$ by

$$\hat{\zeta} = \zeta\big(\hat{\boldsymbol{\varepsilon}}_B^T\hat{\boldsymbol{\varepsilon}}_B/n\big). \tag{15}$$

2.2. *Computational complexity.* For statistical inference about a single entry ω_{ij} of the precision matrix Ω with preconceived i and j, the computational cost of the estimator (10) is of the same order as that of a single run of the scaled lasso (12).

For the estimation of the entire precision matrix Ω, the definition of (10) requires the computation of $\hat{\omega}_{ij}$ for $\binom{p}{2}$ different $A = \{i, j\}$, $i < j$. However, the computational cost for these $\binom{p}{2}$ different $\hat{\omega}_{ij}$ is no greater than that of $(1 + \bar{s})p$ runs of (12) where $\bar{s}$ is the average size of the selected model for regressing a single $\mathbf{X}_j$ against the other $p - 1$ variables. This can be seen as follows. Define the "one-versus-rest" estimator as

$$\Big\{\hat{\boldsymbol{\beta}}_{-i,i}^{(1)}, \sqrt{\hat{\theta}_{ii}^{(1)}}\Big\} = \underset{b\in\mathbb{R}^{p-1},\sigma\in\mathbb{R}^+}{\arg\min}\left\{\frac{\|\mathbf{X}_i - \mathbf{X}_{\{i\}^c}b\|^2}{2n\sigma} + \frac{\sigma}{2} + \lambda\sum_{k\neq i}\frac{\|\mathbf{X}_k\|}{\sqrt{n}}|b_k|\right\}$$

and $\hat{S}_i^{(1)} = \operatorname{supp}(\hat{\boldsymbol{\beta}}_{-i,i}^{(1)})$. For $j \notin \{i\} \cup \hat{S}_i^{(1)}$, the "two-versus-rest" estimator (12) satisfies $\{\hat{\boldsymbol{\beta}}_m, \hat{\theta}_{mm}^{1/2}\} = \{\hat{\boldsymbol{\beta}}_{\{i,j\}^c,i}^{(1)}, \sqrt{\hat{\theta}_{ii}^{(1)}}\}$ when $m = i$ and $A = \{i, j\}$. Thus we only need to carry out $1 + |\hat{S}_i^{(1)}|$ runs of (12) to compute the two-versus-rest estimator $\{\hat{\boldsymbol{\beta}}_m, \hat{\theta}_{mm}^{1/2}\}$ for all $m = i$ and $A = \{i, j\}$, $j \neq i$, where $|\hat{S}_i^{(1)}|$ denotes the cardinality of the set $\hat{S}_i^{(1)}$. Consequently, the total required runs of the scaled lasso (12) is $\sum_{i=1}^p (1 + |\hat{S}_i^{(1)}|) = (1 + \bar{s})p$. It follows from Theorem 11 below that $(1 + \bar{s})p$ is of the order $\#\{(i, j) : \omega_{ij} \neq 0\}$. Thus for the computation of the estimator (10) for the entire precision matrix Ω, the order of the total number of runs of (12) is the total number of edges of the graphical model corresponding to Ω.

2.3. *Statistical inference.* Our analysis can be outlined as follows. We prove that estimators in the form of (10) possess the asymptotic normality and efficiency properties claimed in Theorem 1 when the following conditions hold for certain fixed constant C_0, $\varepsilon_\Omega \to 0$ and all $\delta \geq 1$:

$$\max_{A:A=\{i,j\}} \mathbb{P}\{\|\mathbf{X}_{A^c}(\hat{\boldsymbol{\beta}}_{A^c,A} - \beta_{A^c,A})\|^2 \geq C_0 s\delta \log p\} \leq p^{-\delta+1}\varepsilon_\Omega, \tag{16}$$

$$\max_{A:A=\{i,j\}} \mathbb{P}\{\|\overline{\mathbf{D}}_{A^c}^{1/2}(\hat{\boldsymbol{\beta}}_{A^c,A} - \beta_{A^c,A})\|_1 \geq C_0 s\sqrt{\delta(\log p)/n}\} \leq p^{-\delta+1}\varepsilon_\Omega, \tag{17}$$

with $\overline{\mathbf{D}} = \operatorname{diag}(\mathbf{X}^T\mathbf{X}/n)$, and for $\theta_{ii}^{\mathrm{ora}} = \|\mathbf{X}_i - \mathbf{X}_{A^c}\boldsymbol{\beta}_{A^c,i}\|^2/n$,

$$\max_{A:A=\{i,j\}} \mathbb{P}\left\{\left|\frac{\hat{\theta}_{ii}}{\theta_{ii}^{\mathrm{ora}}} - 1\right| \geq C_0 s\delta(\log p)/n\right\} \leq p^{-\delta+1}\varepsilon_\Omega, \tag{18}$$

with a certain complexity measure s of the precision matrix Ω, provided that the spectrum of Ω is bounded, and the sample size n is no smaller than $(s \log p)^2/c_0$ for a sufficiently small $c_0 > 0$. This is carried out by comparing the estimator in (10) with the oracle MLE in (8) and (9) and proving

$$\kappa_{ij}^{\mathrm{ora}} = \sqrt{n}\frac{\omega_{ij}^{\mathrm{ora}} - \omega_{ij}}{\sqrt{\omega_{ii}\omega_{jj} + \omega_{ij}^2}} \to \mathcal{N}(0, 1),$$

or equivalently the asymptotic normality of the oracle MLE in (9) with mean ω_{ij} and variance $n^{-1}(\omega_{ii}\omega_{jj} + \omega_{ij}^2)$. We then prove (16), (17) and (18) for both the scaled lasso estimator (12) and the LSE after the scaled lasso selection (13). Moreover, we prove that certain thresholded versions of the proposed estimator possesses global optimality properties, as discussed below Theorem 1, under the same boundedness condition on the spectrum of Ω and a more relaxed condition on the sample size.

For the ℓ_0 class $\mathcal{G}_0(M, k_{n,p})$ in (1), the complexity measure for the precision matrix Ω is the maximum degree $s = k_{n,p}$ of the corresponding graph. The ℓ_0

complexity measure can be relaxed to a capped-ℓ_1 measure as follows. For $\lambda > 0$, define capped-ℓ_1 balls as

$$\mathcal{G}^*(M, s, \lambda) = \{\Omega : s_\lambda(\Omega) \le s, 1/M \le \lambda_{\min}(\Omega) \le \lambda_{\max}(\Omega) \le M\}, \tag{19}$$

where $s_\lambda = s_\lambda(\Omega) = \max_j \sum_{i=1}^p \min\{1, |\omega_{ij}|/\lambda\}$ 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. In general, maximum node degree is an upper bound of the capped-ℓ_1 measure 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 bounds the difference between our estimator and the oracle estimators and the difference between the standardized oracle estimator and a standard normal variable.

THEOREM 2. *Let $\Theta^{\text{ora}}_{A,A}$ and $\Omega^{\text{ora}}_{A,A}$ be the oracle MLE defined in (8) and (9), respectively, and $\hat\Theta_{A,A}$ and $\hat\Omega_{A,A}$ be estimators of $\Theta_{A,A}$ and $\Omega_{A,A}$ defined in (10). Let $\delta \ge 1$. Suppose $s \le c_0 n/\log p$ for a sufficiently small constant $c_0 > 0$.*

(i) *Suppose that conditions (16), (17) and (18) hold with C_0 and ε_Ω. Then*

$$\max_{A:A=\{i,j\}} \mathbb{P}\left\{\|\hat\Theta_{A,A} - \Theta^{\text{ora}}_{A,A}\|_\infty > C_1 s\frac{\delta\log p}{n}\right\} \le 6\varepsilon_\Omega p^{-\delta+1} + \frac{4p^{-\delta+1}}{\sqrt{2\log p}} \tag{20}$$

with a positive constant C_1 depending on $\{C_0, \max_{m\in A=\{i,j\}}\theta_{mm}\}$ only, and

$$\max_{A:A=\{i,j\}} \mathbb{P}\left\{\|\hat\Omega_{A,A} - \Omega^{\text{ora}}_{A,A}\|_\infty > C_1' s\frac{\delta\log p}{n}\right\} \le 6\varepsilon_\Omega p^{-\delta+1} + \frac{4p^{-\delta+1}}{\sqrt{2\log p}} \tag{21}$$

with a constant $C_1' > 0$ depending on $\{c_0C_1, \max_{m\in A=\{i,j\}}\{\omega_{mm}, \theta_{mm}\}\}$ only.

(ii) *Let $\lambda = (1+\varepsilon)\sqrt{\frac{2\delta\log p}{n}}$ with $\varepsilon > 0$ in (12), $\hat\beta_{A^c,A}$ be the scaled lasso estimator (12) or the LSE after the scaled lasso selection (13). Then (16), (17) and (18), and thus (20) and (21), hold for all $\Omega \in \mathcal{G}^*(M, s, \lambda)$ with a certain constant C_0 depending on $\{\varepsilon, c_0, M\}$ only and*

$$\max_{\Omega\in\mathcal{G}^*(M,s,\lambda)} \varepsilon_\Omega = o(1). \tag{22}$$

(iii) *Let $\kappa^{\text{ora}}_{ij} = \sqrt{n}(\omega^{\text{ora}}_{ij} - \omega_{ij})/\sqrt{\omega_{ii}\omega_{jj} + \omega_{ij}^2}$. There exist constants D_1 and $\vartheta \in (0, \infty)$, and four marginally standard normal random variables Z', Z_{kl}, where $kl = ii, ij, jj$, such that whenever $|Z_{kl}| \le \vartheta\sqrt{n}$ for all kl, we have*

$$|\kappa^{\text{ora}}_{ij} - Z'| \le \frac{D_1}{\sqrt{n}}(1 + Z_{ii}^2 + Z_{ij}^2 + Z_{jj}^2). \tag{23}$$

Moreover, Z' 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 $\Omega_{A,A}$ in (10) with the components of $\hat{\boldsymbol{\varepsilon}}_A$ being the estimated residuals (11) of (12) or (13). Set $\lambda = (1+\varepsilon)\sqrt{\frac{2\delta \log p}{n}}$ in (12) with certain $\delta \geq 1$ and $\varepsilon > 0$. Suppose $s \leq c_0 n/\log p$ for a sufficiently small constant $c_0 > 0$. For any small constant $\varepsilon_0 > 0$, there exists a constant $C_2 = C_2(\varepsilon_0, \varepsilon, c_0, M)$ such that*

$$\max_{\Omega \in \mathcal{G}^*(M,s,\lambda)} \max_{1 \leq i \leq j \leq p} \mathbb{P}\left\{ |\hat{\omega}_{ij} - \omega_{ij}| > C_2 \max\left\{ s\frac{\log p}{n}, \sqrt{\frac{1}{n}} \right\} \right\} \leq \varepsilon_0. \tag{24}$$

Moreover, there exists a constant $C_3 = C_3(\delta, \varepsilon, c_0, M)$ such that

$$\max_{\Omega \in \mathcal{G}^*(M,s,\lambda)} \mathbb{P}\left\{ \|\hat{\Omega} - \Omega\|_\infty > C_3 \max\left\{ s\frac{\log p}{n}, \sqrt{\frac{\log p}{n}} \right\} \right\} = o(p^{-\delta+3}). \tag{25}$$

Furthermore, $\hat{\omega}_{ij}$ is asymptotically efficient with a consistent variance estimate

$$\sqrt{nF_{ij}}(\hat{\omega}_{ij} - \omega_{ij}) \xrightarrow{D} \mathcal{N}(0,1), \qquad \hat{F}_{ij}/F_{ij} \to 1, \tag{26}$$

uniformly for all i, j and $\Omega \in \mathcal{G}^*(M, s, \lambda)$, provided that $s = o(\sqrt{n}/\log p)$, where

$$F_{ij} = \left(\omega_{ii}\omega_{jj} + \omega_{ij}^2\right)^{-1}, \qquad \hat{F}_{ij} = \left(\hat{\omega}_{ii}\hat{\omega}_{jj} + \hat{\omega}_{ij}^2\right)^{-1}.$$

REMARK 1. The upper bounds $\max\{s\frac{\log p}{n}, \sqrt{\frac{1}{n}}\}$ and $\max\{s\frac{\log p}{n}, \sqrt{\frac{\log p}{n}}\}$ in equations (24) and (25), respectively, are shown to be rate-optimal in Section 2.4.

REMARK 2. The choice of $\lambda = (1+\varepsilon)\sqrt{\frac{2\delta \log p}{n}}$ is common in the literature, but can be too big and too conservative, which usually leads to some estimation bias in practice. Let $L_n(t)$ be the negative quantile function of $\mathcal{N}(0, 1/n)$, which satisfies $L_n(t) \approx \sqrt{(2/n)\log p}$. In Sections 4 and 5.1 we show the value of λ can be reduced to $(1+\varepsilon)L_n(k/p)$ when $\delta \vee k = o(\sqrt{n}/\log p)$.

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 (10). 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 assumptions of 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 asymptotic normality result was stated as Corollary 1 in Sun and Zhang (2012b) without proof.

The following theorem extends Theorems 2 and 3 to the estimation of $\zeta(\Omega_{B,B}^{-1})$, a smooth functional of $\Omega_{B,B}^{-1}$ for a fixed size subset B. Assume that $\zeta : \mathbb{R}^{|B|\times|B|} \to$

$\mathbb{R}$ is a unit Lipschitz function in a neighborhood $\{G : |||G - \Omega_{B,B}^{-1}||| \le \kappa\}$, that is,

$$|\zeta(G) - \zeta(\Omega_{B,B}^{-1})| \le |||G - \Omega_{B,B}^{-1}|||. \tag{27}$$

THEOREM 4. *Let $\hat{\zeta}$ be the estimator of ζ defined in* (15) *with the components of $\hat{\boldsymbol{\varepsilon}}_B$ being the estimated residuals* (11) *of the estimators* (12) *or* (13). *Set the penalty level $\lambda = (1+\varepsilon)\sqrt{\frac{2\delta \log p}{n}}$ in* (12) *with certain $\delta \ge 1$ and $\varepsilon > 0$. Suppose $s \le c_0 n/\log p$ for a sufficiently small constant $c_0 > 0$. Then*

$$\max_{\Omega \in \mathcal{G}^*(M,s,\lambda)} \mathbb{P}\left\{|\hat{\zeta} - \zeta^{\mathrm{ora}}| > C_1 s \frac{\log p}{n}\right\} = o\big(|B| p^{-\delta+1}\big), \tag{28}$$

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

$$\sqrt{n F_\zeta}(\hat{\zeta} - \zeta) \xrightarrow{D} \mathcal{N}(0,1), \tag{29}$$

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

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:

$$B_q(c) = \{\xi \in \mathbb{R}^p : |\xi_{(j)}|^q \le c j^{-1}, \text{ for all } j = 1, \ldots, p\},$$

where $|\xi_{(1)}| \ge |\xi_{(2)}| \ge \cdots \ge |\xi_{(p)}|$. Let

$$\mathcal{G}_q(M, k_{n,p}) = \left\{ \begin{array}{c} \Omega = (\omega_{ij})_{1 \le i,j \le p} : \omega_{\cdot j} \in B_q(k_{n,p}), \\ \text{and } 1/M \le \lambda_{\min}(\Omega) \le \lambda_{\max}(\Omega) \le M \end{array} \right\}. \tag{30}$$

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

$$\mathcal{G}_q(M, k_{n,p}) \subseteq \mathcal{G}^*(M, s, \lambda), \qquad 0 \le q < 1, \tag{31}$$

when $C_q k_{n,p}/\lambda^q \le s$, where $C_q = 1 + q 2^{1/q-1}/(1-q)$ for $0 < q < 1$ and $C_0 = 1$. We state the extension in the following corollary.

COROLLARY 1. *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 $k_{n,p}(n/\log p)^{q/2}$, $0 \le q < 1$.*

2.4. *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

$$p \ge k_{n,p}^{\nu} \qquad \text{with } \nu > 2 \tag{32}$$

and

$$3 \le k_{n,p} \le C_0 \frac{n}{\log p} \tag{33}$$

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 $\hat{\omega}_{ij}$,

$$\sup_{\mathcal{G}_0} \mathbb{P}\left\{|\hat{\omega}_{ij} - \omega_{ij}| > C_1 k_{n,p} \frac{\log p}{n}\right\} \to 1, \tag{34}$$

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 (32) and (33), we have the following minimax lower bounds*:

$$\inf_{\hat{\omega}_{ij}} \sup_{\mathcal{G}_0(M,k_{n,p})} \mathbb{P}\left\{|\hat{\omega}_{ij} - \omega_{ij}| > \max\left\{C_1 \frac{k_{n,p}\log p}{n}, C_2\sqrt{\frac{1}{n}}\right\}\right\} > c_1 > 0 \tag{35}$$

and

$$\inf_{\hat{\Omega}} \sup_{\mathcal{G}_0(M,k_{n,p})} \mathbb{P}\left\{\|\hat{\Omega} - \Omega\|_\infty > \max\left\{C_1' \frac{k_{n,p}\log p}{n}, C_2'\sqrt{\frac{\log p}{n}}\right\}\right\} > c_2 > 0, \tag{36}$$

where c_1, c_2, C_1, C_2, C_1' and C_2' are positive constants depending on M, ν and C_0 only.

REMARK 4. 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 for every entry σ_{ij}. But for estimation of sparse precision matrix, when $k_{n,p} \gg \frac{\sqrt{n}}{\log p}$, Theorem 5 implies that it is impossible to obtain the parametric rate.

REMARK 5. Since $\mathcal{G}_0(M, k_{n,p}) \subseteq \mathcal{G}^*(M, k_{n,p}, \lambda)$ by the definitions in (1) and (19), 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 (30) and the capped-ℓ_1 ball defined in (19). 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 $(k_{n,p}(\frac{n}{\log p})^{q/2} - 1)$ 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 (77) is indeed in $\mathcal{G}_q(M, k_{n,p})$. Now under assumptions $p \geq (k_{n,p}(\frac{n}{\log p})^{q/2})^{\nu}$ with $\nu > 2$ and $k_{n,p} \leq C_0(\frac{n}{\log p})^{1-q/2}$, 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

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

These lower bounds match the upper bounds in Corollary 1 for the proposed estimator.

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 irrepresentability condition or the ℓ_1 constraint of the precision matrix commonly required in literature. In our procedure, we first obtain an asymptotically normal estimation and then 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 works, 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 a strong irrepresentability condition and a uniform signal strength condition $\min_{(i,j):\omega_{ij}\neq 0}|\omega_{ij}| \geq c\sqrt{\frac{\log p}{n}}$ for some $c > 0$. Cai, Liu and Luo (2011) do not require irrepresentability conditions, but need to assume that $\min_{(i,j):\omega_{ij}\neq 0}|\omega_{ij}| \geq CM_{n,p}^2\sqrt{\frac{\log p}{n}}$, where $M_{n,p}$ is the matrix ℓ_1 norm of Ω. In Cai, Liu and Zhou (2012), they weakened the condition to $\min_{(i,j):\omega_{ij}\neq 0}|\omega_{ij}| \geq CM_{n,p}\sqrt{\frac{\log p}{n}}$, 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 $\min_{(i,j):\omega_{ij}\neq 0}|\omega_{ij}|$ 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 $s = o(\frac{\sqrt{n}}{\log p})$, we have the asymptotic normality result for each entry ω_{ij}, $i \neq j$, that is,

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

where $F_{ij} = (\omega_{ii}\omega_{jj} + \omega_{ij}^2)^{-1}$ is the Fisher information of estimating ω_{ij}. The total number of edges is $p(p-1)/2$. 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:

$$\hat{\Omega}_{\mathrm{thr}} = \big(\hat{\omega}_{ij}^{\mathrm{thr}}\big)_{p\times p}, \tag{37}$$

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

$$\hat{\tau}_{ij} = \sqrt{\frac{2\xi_0(\hat{\omega}_{ii}\hat{\omega}_{jj} + \hat{\omega}_{ij}^2)\log p}{n}}. \tag{38}$$

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 (10), 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 Theorem 6 below for support recovery are much weaker than other results in literature.

Define a thresholded population precision matrix as

$$\Omega_{\mathrm{thr}} = \big(\omega_{ij}^{\mathrm{thr}}\big)_{p\times p}, \tag{39}$$

where $\omega_{ii}^{\mathrm{thr}} = \omega_{ii}$ and $\omega_{ij}^{\mathrm{thr}} = \omega_{ij} 1\{|\omega_{ij}| \geq \sqrt{8\xi(\omega_{ii}\omega_{jj} + \omega_{ij}^2)(\log p)/n}\}$, 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, $(V, E(\Omega_{\mathrm{thr}}))$ can be viewed as a graph of strong edges. Define

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

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

$$|\omega_{ij}| \geq 2\sqrt{\frac{2\xi(\omega_{ii}\omega_{jj} + \omega_{ij}^2)\log p}{n}} \qquad \forall \omega_{ij} \neq 0; \tag{40}$$

that is, $\Omega_{\mathrm{thr}} = \Omega$, and 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* $\hat{\Omega}_{\mathrm{thr}}$ *be the ANT estimator defined in* (37) *with* $\xi_0 > 2$ *in the thresholding level* (38). *Suppose* $\Omega \in \mathcal{G}^*(M, s, \lambda)$ *with* $s = o(\sqrt{n/\log p})$. *Then*

$$\lim_{n\to\infty} \mathbb{P}\big(E(\Omega_{\mathrm{thr}}) \subseteq E(\hat{\Omega}_{\mathrm{thr}}) \subseteq E(\Omega)\big) = 1, \tag{41}$$

where Ω_{thr} *is defined in* (39) *with* $\xi > \xi_0$. *If in addition* (40) *holds, then*

$$\lim_{n\to\infty} \mathbb{P}\big(\mathcal{S}(\hat{\Omega}_{\mathrm{thr}}) = \mathcal{S}(\Omega)\big) = 1. \tag{42}$$

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 impact of extremely small tail probability of near singularity of the low-dimensional estimator $\hat{\Theta}_{A,A}$, we define a truncated version of the estimator $\hat{\Omega}_{\mathrm{thr}}$ defined in (37),

$$\breve{\Omega}_{\mathrm{thr}} = \left(\hat{\omega}_{ij}^{\mathrm{thr}} \min\left\{1, \frac{\log p}{|\hat{\omega}_{ij}|}\right\}\right)_{p\times p}. \tag{43}$$

Theorem 7 below follows mainly from the convergence rate under element-wise norm and the fact that the upper bound holds for the matrix ℓ_1 norm. This argument uses the inequality $|||M|||_w \le |||M|||_1$ for symmetric matrices M and $1 \le w \le \infty$, which follows from the Riesz–Thorin interpolation theorem; see, for example, Thorin (1948). Note that under the assumption $s^2 = O(n/\log p)$, it can be seen from the equations (21) and (23) in Theorem 2 that with high probability the $\|\hat{\Omega} - \Omega\|_\infty$ is dominated by $\|\Omega^{\mathrm{ora}} - \Omega\|_\infty = O_p(\sqrt{\frac{\log p}{n}})$. From there the details of the proof is similar in nature to those of Theorem 3 in Cai and Zhou (2012) and thus will be omitted due to the limit of space.

THEOREM 7. *Under the assumptions* $s^2 = O(n/\log p)$ *and* $n = O(p^\xi)$ *with some* $\xi > 0$, *the* $\breve{\Omega}_{\mathrm{thr}}$ *defined in* (43) *with* $\lambda = (1+\varepsilon)\sqrt{\frac{2\delta \log p}{n}}$ *for sufficiently large* $\delta \ge 3 + 2\xi$ *and* $\varepsilon > 0$ *satisfies, for all* $1 \le w \le \infty$ *and* $k_{n,p} \le s$,

$$\sup_{\mathcal{G}_0(M,k_{n,p})} \mathbb{E}|||\breve{\Omega}_{\mathrm{thr}} - \Omega|||_w^2 \le \sup_{\mathcal{G}^*(M,k_{n,p},\lambda)} \mathbb{E}|||\breve{\Omega}_{\mathrm{thr}} - \Omega|||_w^2 \le Cs^2 \frac{\log p}{n}. \tag{44}$$

REMARK 6. It follows from equation (31) that result (44) also holds for the classes of weak ℓ_p balls $\mathcal{G}_q(M, k_{n,p})$ defined in equation (30), with $s = C_q k_{n,p} (\frac{n}{\log p})^{q/2}$,

$$\sup_{\mathcal{G}_q(M,k_{n,p})} \mathbb{E}|||\breve{\Omega}_{\mathrm{thr}} - \Omega|||_w^2 \le Ck_{n,p}^2 \left(\frac{\log p}{n}\right)^{1-q}. \tag{45}$$

REMARK 7. Cai, Liu and Zhou (2012) showed that the rates obtained in equations (44) and (45) are optimal when $p \ge cn^{\alpha_0}$ for some $\alpha_0 > 1$ and $k_{n,p} = o(n^{1/2}(\log p)^{-3/2})$.

REMARK 8. Although the estimator $\breve{\Omega}_{\mathrm{thr}}$ is symmetric, it is not guaranteed to be positive definite. It follows from Theorem 7 that $\breve{\Omega}_{\mathrm{thr}}$ is positive definite with high probability. When it is not positive definite, we can always pick the smallest $c_a \ge 0$ such that $c_a I + \breve{\Omega}_{\mathrm{thr}}$ is positive semidefinite. It is trivial to see that $(c_a + 1/n)I + \breve{\Omega}_{\mathrm{thr}}$ is positive definite, sparse and enjoys the same rate of convergence as $\breve{\Omega}_{\mathrm{thr}}$ for the loss functions considered in this paper.

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 has been discussed and appreciated by several researchers, but it has never been clear if the conditions in their paper are necessary and the results 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 $\mathrm{Card}(O) = p$, $\mathrm{Card}(H) = h$ and $O \cup H = \{1, 2, \ldots, p + h\}$. Assume that $(X_O^{(i)}, X_H^{(i)})$, $i = 1, \ldots, n$, 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 $\{X_O^{(1)}, X_O^{(2)}, \ldots, X_O^{(n)}\}$, while $\{X_H^{(1)}, X_H^{(2)}, \ldots, X_H^{(n)}\}$ 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:

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

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

$$\Sigma^{-1}_{O,O} = \bar{\Omega}_{O,O} - \bar{\Omega}_{O,H}\bar{\Omega}^{-1}_{H,H}\bar{\Omega}_{H,O}; \tag{46}$$

see, for example, Horn and Johnson (1990). Define

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

and $h' = \mathrm{rank}(L)$. We note that $h' = \mathrm{rank}(\bar{\Omega}_{O,H}) \le h$.

We focus on the estimation of $\Sigma^{-1}_{O,O}$ and S, as the estimation of L can be naturally carried out based on our results as in Chandrasekaran, Parrilo and Willsky (2012) and Ren and Zhou (2012). To make the problem identifiable we assume that S is sparse, and the observed and latent variables are weakly correlated in the following sense:

$$S = (s_{ij})_{1\le i,j\le p}, \qquad \max_{1\le j\le p}\sum_{i=1}^{p} 1\{s_{ij} \ne 0\} \le k_{n,p}, \tag{47}$$

and that for certain $a_n \to 0$

$$L = (l_{ij})_{1\le i,j\le p}, \qquad \max_j \sum_i l_{ij}^2 \le (a_n/n)\log p. \tag{48}$$

The sparseness of $S = \bar{\Omega}_{O,O}$ can be seen as inherited from that of the full precision matrix $\bar{\Omega}_{(p+h)\times(p+h)}$. It is particularly interesting for us to identify the support of $S = \bar{\Omega}_{O,O}$ and make inference for each entry of S. The ℓ_2 condition (48) on L is of a weaker form than the ℓ_1 and ℓ_∞ conditions imposed in Ren and Zhou (2012). In addition, we assume that for some universal constant M,

$$1/M \le \lambda_{\min}(\Sigma_{(p+h)\times(p+h)}) \le \lambda_{\max}(\Sigma_{(p+h)\times(p+h)}) \le M, \tag{49}$$

which 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.

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

$$\mathbf{X}_A = \mathbf{X}_{O\setminus A}\boldsymbol{\beta} + \boldsymbol{\varepsilon}_A \tag{50}$$

for $A = \{i, j\} \subset O$ with $\boldsymbol{\beta} = \Omega_{O\setminus A,A}\Omega_{A,A}^{-1}$ and $\boldsymbol{\varepsilon}_A \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \Omega_{A,A}^{-1})$.

To obtain the asymptotic normality result, condition (19) of Theorem 2 requires

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

with $\lambda \asymp \sqrt{(\log p)/n}$. However, when L is coherent [Candès and Recht (2009)] in the sense of $\{\max_j \sum_i |l_{ij}|\}^2 \asymp p \max_i \sum_j l_{ij}^2 \asymp p(a_n/n)\log p$,

$$\begin{aligned}\max_j \sum_{i=1}^p \min\left\{1, \frac{|s_{ij} - l_{ij}|}{\lambda}\right\} &\ge \max_j \sum_{s_{ij}=0} \frac{|l_{ij}|}{\lambda} \\ &\asymp \frac{\sqrt{p} - \sqrt{k_{n,p}}}{\lambda}\left(\frac{a_n \log p}{n}\right)^{1/2} \asymp \sqrt{a_n p}.\end{aligned}$$

Thus the conditions of Theorem 2 are not satisfied for the latent variable graphical model when $a_n p(\log p)^2 \ge n$. We overcome the difficulty through a new analysis.

THEOREM 8. *Let $\hat{\Omega}_{A,A}$ be the estimator of $\Omega_{A,A}$ defined in (10) with $A = \{i, j\}$ for the regression (50), where the components of $\hat{\boldsymbol{\varepsilon}}_A$ are the estimated residuals of (12) or (13). Let $\lambda = (1+\varepsilon)\sqrt{\frac{2\delta\log p}{n}}$ for certain $\delta \ge 1$ and $\varepsilon > 0$. Under assumptions (47)–(49) and $k_{n,p} \le c_0 n/\log p$ with a small c_0, we have*

$$\mathbb{P}\{|\hat{\omega}_{ij} - \omega_{ij}| > C_3 \max\{k_{n,p} n^{-1}\log p, n^{-1/2}\}\} = o(p^{-\delta+1})$$

for a certain constant C_3, and

$$\mathbb{P}\{|\hat{\omega}_{ij} - s_{ij}| > C_3 \max\{k_{n,p} n^{-1}\log p, n^{-1/2}, \sqrt{(a_n/n)\log p}\}\} = o(p^{-\delta+1}).$$

If the condition on $k_{n,p}$ is strengthened to $k_{n,p} = o(\frac{\sqrt{n}}{\log p})$, then

$$\sqrt{\frac{n}{\omega_{ii}\omega_{jj} + \omega_{ij}^2}}(\hat{\omega}_{ij} - \omega_{ij}) \xrightarrow{D} \mathcal{N}(0, 1). \tag{51}$$

REMARK 9. If, in addition, $\ell_{ij} = o(n^{-1/2})$, then (51) implies

$$\sqrt{\frac{n}{\omega_{ii}\omega_{jj} + \omega_{ij}^2}}(\hat{\omega}_{ij} - s_{ij}) \xrightarrow{D} \mathcal{N}(0, 1). \tag{52}$$

Define the adaptive thresholded estimator $\hat{\Omega}_{\text{thr}} = (\hat{\omega}_{ij}^{\text{thr}})_{p\times p}$ as in (37) and (38). 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 (12). Assume assumptions (47)–(49) hold. Then*:

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

$$|s_{ij}| \geq 2\sqrt{\frac{2\xi_0(\omega_{ii}\omega_{jj} + \omega_{ij}^2)\log p}{n}} \qquad \forall s_{ij} \neq 0$$

for some $\xi_0 > 2$, we have

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

(ii) *Under the assumption $k_{n,p}^2 = O(n/\log p)$ and $n = O(p^{\xi})$ with some $\xi > 0$, the $\breve{\Omega}_{\text{thr}}$ defined in (43) with sufficiently large $\delta \geq 3 + 2\xi$ satisfies, for all $1 \leq w \leq \infty$,*

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

4. Regression revisited. The key element of our analysis is to establish (16), (17) and (18) for the scaled lasso estimator (12) and the LSE after the scaled lasso selection (13). The existing literature has provided theorems and arguments to carry out this task. However, several issues still require extension of existing results or explanation and modification of existing proofs. For example, the LSE after model selection is not as well understood as the lasso, and biased regression models are typically studied inexplicably, if at all. Another issue is that the penalty level used in theorems in previous sections could be too large for good numerical performance, especially for $\delta \geq 3$ in (25) of Theorems 3 and Theorems 6, 7 and 9. These issues were addressed in previous versions of this paper (arXiv:1309.6024)

in separate lemmas. In this section, we provide a streamlined presentation of these regression results required in our analysis.

Let $\widetilde{\mathbf{X}} = (\widetilde{\mathbf{X}}_1, \ldots, \widetilde{\mathbf{X}}_{\widetilde{p}})$ be an $n \times \widetilde{p}$ standardized design matrix with $\|\widetilde{\mathbf{X}}_k\|^2 = n$ for all $k = 1, \ldots, \widetilde{p}$, and $\widetilde{\mathbf{Y}}$ be a response vector satisfying

$$\widetilde{\mathbf{Y}}|\widetilde{\mathbf{X}} \sim \mathcal{N}(\widetilde{\mathbf{X}}\boldsymbol{\gamma}, \sigma^2 \mathbf{I}_{n\times n}). \tag{53}$$

For the scaled lasso $\{\hat{\boldsymbol{\beta}}_m, \hat{\theta}_{mm}^{1/2}\}$ in (12), $\{\overline{\mathbf{D}}_{A^c}^{1/2}\hat{\boldsymbol{\beta}}_m, \hat{\theta}_{mm}^{1/2}\}$ can be written as

$$\{\hat{\boldsymbol{\gamma}}, \hat{\sigma}\} = \arg\min_{\{\boldsymbol{\gamma},\sigma\}}\left\{\frac{\|\widetilde{\mathbf{Y}} - \widetilde{\mathbf{X}}\boldsymbol{\gamma}\|^2}{2n\sigma} + \frac{\sigma}{2} + \lambda_0\|\boldsymbol{\gamma}\|_1\right\}, \tag{54}$$

with $m \in A = \{i, j\}$, $\widetilde{\mathbf{X}} = \mathbf{X}_{A^c}\overline{\mathbf{D}}_{A^c}^{-1/2}$, $\overline{\mathbf{D}} = \operatorname{diag}(\mathbf{X}^T\mathbf{X}/n)$, $\widetilde{\mathbf{Y}} = \mathbf{X}_m$ and $\boldsymbol{\gamma} = \overline{\mathbf{D}}_{A^c}^{1/2}\boldsymbol{\beta}_m$. For the LSE after model selection in (13), $\{\overline{\mathbf{D}}_{A^c}^{1/2}\hat{\boldsymbol{\beta}}_m, \hat{\theta}_{mm}^{1/2}\}$ can be written as

$$\{\hat{\boldsymbol{\gamma}}^{\mathrm{lse}}, \hat{\sigma}^{\mathrm{lse}}\} = \arg\min_{\{\boldsymbol{\gamma},\sigma\}}\left\{\frac{\|\widetilde{\mathbf{Y}} - \widetilde{\mathbf{X}}\boldsymbol{\gamma}\|^2}{2n\sigma} + \frac{\sigma}{2} : \operatorname{supp}(\boldsymbol{\gamma}) \subseteq \operatorname{supp}(\hat{\boldsymbol{\gamma}})\right\}. \tag{55}$$

Moreover, for both estimators, conditions (16), (17) and (18) are consequences of

$$\mathbb{P}\{\|\widetilde{\mathbf{X}}(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}^{\mathrm{target}})\|^2 \le C_0 s(\sigma^{\mathrm{ora}})^2\delta\log\widetilde{p}\} \ge 1 - \widetilde{p}^{1-\delta}\widetilde{\varepsilon}_0, \tag{56}$$

$$\mathbb{P}\{\|\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}^{\mathrm{target}}\|_1 \le C_0 s\sigma^{\mathrm{ora}}\sqrt{\delta(\log\widetilde{p})/n}\} \ge 1 - \widetilde{p}^{1-\delta}\widetilde{\varepsilon}_0 \tag{57}$$

and

$$\mathbb{P}\left\{\left|\frac{\hat{\sigma}}{\sigma^{\mathrm{ora}}} - 1\right| \le C_0 s\delta(\log\widetilde{p})/n \le 1/2\right\} \ge 1 - \widetilde{p}^{1-\delta}\widetilde{\varepsilon}_0, \tag{58}$$

with $\sigma^{\mathrm{ora}} = \|\widetilde{\mathbf{Y}} - \widetilde{\mathbf{X}}\boldsymbol{\gamma}^{\mathrm{target}}\|/\sqrt{n}$, $\boldsymbol{\gamma}^{\mathrm{target}} = \boldsymbol{\gamma}$ and $\delta \ge 1$, provided that C_0 is fixed and $\widetilde{\varepsilon}_0 \to 0$ uniformly in $m \in A = \{i, j\}$ and Ω in the class in (19); see Proposition 1. In the latent variable graphical model, Theorems 8 and 9 require (56), (57) and (58) for a certain sparse $\boldsymbol{\gamma}^{\mathrm{target}}$ in a biased linear model when (53) does not provide a sufficiently sparse $\boldsymbol{\gamma}$. We note that both $\boldsymbol{\gamma}$ and $\boldsymbol{\gamma}^{\mathrm{target}}$ are allowed to be random variables here.

To carry out an analysis of the lasso, one has to make a choice among different ways of controlling the correlations between the design and noise vectors in (53),

$$\widetilde{\mathbf{Z}} = (\widetilde{Z}_1, \ldots, \widetilde{Z}_{\widetilde{p}}) = \frac{\widetilde{\mathbf{X}}^T(\widetilde{\mathbf{Y}} - \widetilde{\mathbf{X}}\boldsymbol{\gamma})}{\sqrt{n}\|\widetilde{\mathbf{Y}} - \widetilde{\mathbf{X}}\boldsymbol{\gamma}\|}. \tag{59}$$

A popular choice is to bound $\widetilde{\mathbf{Z}}$ with the ℓ_∞ norm as it is the dual of the ℓ_1 penalty. This has led to the sparse Riesz [Zhang and Huang (2008)], restricted eigenvalue [Bickel, Ritov and Tsybakov (2009), Koltchinskii (2009)], compatibility [van de Geer and Bühlmann (2009)], cone invertibility [Ye and Zhang (2010), Zhang and Zhang (2012)] and other similar conditions on the design matrix. Sun and Zhang

(2012a) took this approach to analyze (54) and (55) with the compatibility and cone invertibility factors. Another approach is to control the sparse ℓ_2 norm of $\widetilde{\mathbf{Z}}$ to allow smaller penalty levels in the analysis; See Zhang (2009) and Ye and Zhang (2010) for analyses of the lasso and Sun and Zhang (2013) for an analysis of the scaled estimators (54) and (55).

Here we take a different approach by using two threshold levels, a smaller one to bound an overwhelming majority of the components of $\widetilde{\mathbf{Z}}$ and a larger one to bound its ℓ_∞ norm. This allows us to use both a small penalty level associated with the smaller threshold level and the compatibility condition.

For $\alpha \geq 0$ and index sets K, the compatibility constant is defined as

$$\phi_{\mathrm{comp}}(\alpha, K, \widetilde{\mathbf{X}}) = \inf\left\{ \frac{|K|^{1/2}\|\widetilde{\mathbf{X}}u\|}{n^{1/2}\|u_K\|_1} : u \in \mathcal{C}(\alpha, K), u \neq 0 \right\},$$

where $|K|$ is the cardinality of K and $\mathcal{C}(\alpha, K) = \{u \in \mathbb{R}^{\widetilde{p}} : \|u_{K^c}\|_1 \leq \alpha\|u_K\|_1\}$. We may want to control the size of selected models with the upper sparse eigenvalue, defined as

$$\kappa^*(m, \widetilde{\mathbf{X}}) = \max_{\|u\|=1, \|u\|_0 \leq m} \|\widetilde{\mathbf{X}}u\|^2/n.$$

We impose the following conditions on the target coefficient vector and the design:

$$\mathbb{P}\{\mathrm{Cond}_1\} \geq 1 - \widetilde{\varepsilon}_1, \qquad \mathrm{Cond}_1 = \left\{ |K| + \sum_{k \notin K} \frac{|\boldsymbol{\gamma}_k^{\mathrm{target}}/\sigma^{\mathrm{ora}}|}{\sqrt{(2/n)\log \widetilde{p}}} \leq s_1 \right\} \tag{60}$$

for a certain index set K, and

$$\mathbb{P}\{\mathrm{Cond}_2\} \geq 1 - \widetilde{\varepsilon}_1, \qquad \mathrm{Cond}_2 = \left\{ \max_{|J \setminus K| \leq s_2} \phi_{\mathrm{comp}}^{-2}(\alpha, J, \widetilde{\mathbf{X}}) \leq C_2 \right\}. \tag{61}$$

For small penalty levels and the LSE after model selection, we also need

$$\mathbb{P}\{\mathrm{Cond}_3\} \geq 1 - \widetilde{\varepsilon}_1, \qquad \mathrm{Cond}_3 = \{\kappa^*(s_3, \widetilde{\mathbf{X}}) \leq C_3\}. \tag{62}$$

Finally, for $\boldsymbol{\gamma}^{\mathrm{target}} \neq \boldsymbol{\gamma}$, we need the condition

$$\begin{aligned} &\mathbb{P}\{\mathrm{Cond}_4\} \geq 1 - \widetilde{\varepsilon}_1, \\ &\mathrm{Cond}_4 = \{C_4\|\widetilde{\mathbf{X}}(\boldsymbol{\gamma}^{\mathrm{target}} - \boldsymbol{\gamma})\| \leq \sigma^{\mathrm{ora}}\sqrt{\log(\widetilde{p}/\widetilde{\varepsilon}_1)}\}. \end{aligned} \tag{63}$$

In (60), (61), (62) and (63), s_j are allowed to change with $\{n, \widetilde{p}\}$, while α and C_j are fixed constants. These conditions also make sense for deterministic designs with $\widetilde{\varepsilon}_1 = 0$ for deterministic conditions.

Let k and ε be positive real numbers and λ_0 be a penalty level satisfying

$$\lambda_0 \geq (1 + \varepsilon)L_{n-3/2}(k/\widetilde{p}), \tag{64}$$

where $L_n(t) = n^{-1/2}\Phi^{-1}(1-t)$ is the $\mathcal{N}(0, 1/n)$ negative quantile function. Let

$$\varepsilon_1 \geq \frac{e^{1/(4n-6)^2}4k/s_2}{L_1^4(k/\widetilde{p}) + 2L_1^2(k/\widetilde{p})} + \left(\frac{L_1(\widetilde{\varepsilon}_1/\widetilde{p})}{L_1(k/\widetilde{p})} + \frac{e^{1/(4n-6)^2}/\sqrt{2\pi}}{L_1(k/\widetilde{p})}\right)\sqrt{\frac{C_3}{s_2}}. \tag{65}$$

We note that $L_n(t) = n^{-1/2}L_1(t) \leq \sqrt{(2/n)\log(1/t)}$ for $t \leq 1/2$, so that the right-hand side of (65) is of the order $k/\{s_2(\log p)^2\} + \sqrt{\delta/s_2}$. Thus condition (65) is easily satisfied even when ε_1 is a small positive number and k is a moderately large number. Moreover, λ depends on δ only through $\sqrt{\delta/s_2}$ in (65).

THEOREM 10. *Let $\{\hat{\boldsymbol{\gamma}}, \hat{\sigma}\}$ be as in (54) with data in (53) and a penalty level in (64). Let $\widetilde{\varepsilon}_1 < 1$ and $\lambda^* = L_{n-3/2}(\widetilde{\varepsilon}_1/\widetilde{p})$. Suppose $\lambda^* \leq 1$ and $\delta s(\log \widetilde{p})/n \leq c_0$.*

(i) *Let $\boldsymbol{\gamma}^{\text{target}} = \boldsymbol{\gamma}$, $s \geq s_1$, $s_2 = 0$, $k \leq \widetilde{\varepsilon}_1$ in (64), $\alpha = 1 + 2/\varepsilon$ and $4\widetilde{\varepsilon}_1 \leq \widetilde{p}^{1-\delta}\widetilde{\varepsilon}_0$. Then there exists a constant C_0 depending on $\{\alpha, C_2\}$ only such that when $C_0c_0 \leq 1/2$, (60) and (61) imply (56), (57) and (58).*

(ii) *Let $\boldsymbol{\gamma}^{\text{target}} = \boldsymbol{\gamma}$, $s \geq s_1 + s_2$, $1 \leq s_2 \leq s_3$, $k \geq 1$ and $\varepsilon_1 < \varepsilon$ in (64) and (65), $\alpha \geq \sqrt{2}(\varepsilon - \varepsilon_1)_+^{-1}\{1 + \varepsilon + L_1(\widetilde{\varepsilon}_1/\widetilde{p})/L_1(k/\widetilde{p})\}$ and $(5 + e^{1/(4n-6)^2})\widetilde{\varepsilon}_1 \leq \widetilde{p}^{1-\delta}\widetilde{\varepsilon}_0$. Then there exists a constant C_0 depending on $\{\alpha, \varepsilon, \varepsilon_1, C_2\}$ only such that when $C_0c_0 \leq 1/2$, (60), (61) and (62) imply (56), (57) and (58).*

(iii) *Let $s \geq s_1 + s_2$, $1 \leq s_2 \leq s_3$, $k \geq 1$ and $\varepsilon_1 < \varepsilon$ in (64) and (65), $\varepsilon_1 < \varepsilon_2 < \varepsilon$, $\alpha \geq 2(\varepsilon - \varepsilon_2)_+^{-1}\{1 + \varepsilon + L_1(\widetilde{\varepsilon}_1/\widetilde{p})/L_1(k/\widetilde{p})\}$, $(6 + e^{1/(4n-6)^2})\widetilde{\varepsilon}_1 \leq \widetilde{p}^{1-\delta}\widetilde{\varepsilon}_0$ and $C_4 \geq \sqrt{(4/L_1^2(k/\widetilde{p}))\log(\widetilde{p}/\widetilde{\varepsilon}_1)}/\min(\sqrt{2}-1, \varepsilon_2 - \varepsilon_1)$. Then there exists a constant C_0 depending on $\{\alpha, \varepsilon, \varepsilon_2, C_2\}$ only such that when $C_0c_0 \leq 1/2$, (60), (61), (62) and (63) imply (56), (57) and (58).*

In Theorem 10, s_1 in (60) represents the complexity or the size of the coefficient vector, and s_2 represents the number of false positives we are willing to accept with the penalty level in (64). Thus s is an upper bound for the total number of estimated coefficients, true or false. We summarize parallel results for the LSE after model selection as follows.

THEOREM 11. *Let $\{\hat{\boldsymbol{\gamma}}, \hat{\sigma}\}$ be as in (54) and $\{\hat{\boldsymbol{\gamma}}^{\text{lse}}, \hat{\sigma}^{\text{lse}}\}$ as in (55).*

(i) *The following bounds always hold*:

$$\hat{\sigma}^2 - (\hat{\sigma}^{\text{lse}})^2 = \|\widetilde{\mathbf{X}}(\hat{\boldsymbol{\gamma}}^{\text{lse}} - \hat{\boldsymbol{\gamma}})\|^2/n \leq \frac{(\hat{\sigma}\lambda_0)^2|\hat{S}|}{\phi^2_{\text{comp}}(0, \hat{S}, \widetilde{\mathbf{X}})} \tag{66}$$

with $\hat{S} = \text{supp}(\hat{\boldsymbol{\gamma}})$ and

$$\|\hat{\boldsymbol{\gamma}}^{\text{lse}} - \hat{\boldsymbol{\gamma}}\|_1 \leq \frac{\hat{\sigma}\lambda_0|\hat{S}|}{\phi^2_{\text{comp}}(0, \hat{S}, \widetilde{\mathbf{X}})}. \tag{67}$$

(ii) *Let λ_0 be a penalty level satisfying* (64) *and $\varepsilon_1 < \varepsilon_2 < \varepsilon_3 < \varepsilon$. Suppose the conditions of Theorem* 10 *hold and that the constant factor C_0 in Theorem* 10 *satisfies*

$$C_0 s\delta(\log \widetilde{p})/n \le \frac{\varepsilon - \varepsilon_3}{1+\varepsilon}, \qquad \frac{C_0 s\delta(\log \widetilde{p})}{(\varepsilon_3 - \varepsilon_2)^2 L_1^2(k/\widetilde{p})} \le \frac{s_3}{C_3}.$$

Then, for the parameters defined in the respective parts of Theorem 10,

$$\mathbb{P}\big\{|\hat{S}| < s_3 + s\big\} \ge 1 - \widetilde{p}^{1-\delta}\widetilde{\varepsilon}_0. \tag{68}$$

If in addition, condition (61) *is strengthened to*

$$\mathbb{P}\Big\{\Big(\max_{|J\setminus K|\le s_2} \phi_{\mathrm{comp}}^{-2}(\alpha, J, \widetilde{\mathbf{X}})\Big) \vee \Big(\max_{|J|\le s_3+s} \phi_{\mathrm{comp}}^{-2}(0, J, \widetilde{\mathbf{X}})\Big) > C_2\Big\} \le \widetilde{\varepsilon}_1, \tag{69}$$

then the conclusions of Theorem 10 *hold with $\{\hat{\boldsymbol{\gamma}}, \hat{\sigma}\}$ replaced by $\{\hat{\boldsymbol{\gamma}}^{\mathrm{lse}}, \hat{\sigma}^{\mathrm{lse}}\}$.*

We collect some probability bounds for the regularity conditions in the following proposition. Consider deterministic coefficient vectors $\boldsymbol{\beta}^{\mathrm{target}}$ satisfying

$$|K| + \sum_{j\notin K} \frac{C_1|\beta_j^{\mathrm{target}}|}{\sqrt{(2/n)\log \widetilde{p}}} \le s_1. \tag{70}$$

PROPOSITION 1. *Let $\mathbf{X}$ be a $n \times p$ matrix with i.i.d. $\mathcal{N}(0, \boldsymbol{\Sigma})$ rows, $A^c \subset \{1, \ldots, p\}$ with $|A^c| = \widetilde{p}$, $\overline{\mathbf{D}} = \mathrm{diag}(\mathbf{X}^T\mathbf{X}/n)$, $\widetilde{\mathbf{X}} = \mathbf{X}_{A^c}\overline{\mathbf{D}}_{A^c}^{-1/2}$, $\boldsymbol{\gamma} = \overline{\mathbf{D}}_{A^c}^{1/2}\boldsymbol{\beta}_{A^c}$ and $\boldsymbol{\gamma}^{\mathrm{target}} = \overline{\mathbf{D}}_{A^c}^{1/2}\boldsymbol{\beta}_{A^c}^{\mathrm{target}}$. Suppose $1/M \le \lambda_{\min}(\boldsymbol{\Sigma}) \le \lambda_{\max}(\boldsymbol{\Sigma}) \le M$ with a fixed M. Let $\lambda_1 = \sqrt{(2/n)\log(\widetilde{p}/\widetilde{\varepsilon}_1)}$. Then, for a certain constant C_* depending on M only,*

$$(70) \Rightarrow (60) \qquad \text{when } C_1 \ge \sqrt{M(1+\lambda_1)}, \tag{71}$$

$$C_2 \ge C_*\big\{1 + \max\big\{|K| + s_2, s + s_3\big\}\lambda_1^2\big\} \Rightarrow (69) \Rightarrow (61), \tag{72}$$

$$C_3 \ge C_*\big\{1 + s_3\lambda_1^2\big\} \Rightarrow (62), \tag{73}$$

and for $\lambda_2 = \sqrt{(2/n)\log(1/\widetilde{\varepsilon}_1)}$ and any coefficient vectors $\boldsymbol{\beta}^{\mathrm{target}}$ and $\boldsymbol{\beta}$,

$$C_4C_*(1+\lambda_2)\big\|\boldsymbol{\beta}^{\mathrm{target}} - \boldsymbol{\beta}\big\| \le \lambda_1 \Rightarrow (63). \tag{74}$$

Moreover, when $\lambda_2 \le 1/2$, σ^{ora} can be replaced by $\sqrt{\mathbb{E}(\sigma^{\mathrm{ora}})^2}$ or C_* in (56) *and* (57).

5. Discussion.

5.1. *Alternative choice of penalty level for finite sample performance.* In Theorem 2 and nearly all consequent results in Theorems 3–4 and 6–9, we have picked the penalty level $\lambda = (1+\varepsilon)\sqrt{(2\delta/n)\log p}$ for $\delta \geq 1$ ($\delta \geq 3$ for support recovery) and $\varepsilon > 0$. This choice of λ can be too conservative and may cause some finite sample estimation bias. However, in view of Theorem 10(ii) and (iii), the results in these theorems in Sections 2 and 3 still hold for penalty levels no smaller than $\lambda = (1+\varepsilon)L_n(k/p) \approx (1+\varepsilon)\sqrt{(2/n)\log(p/k)}$, which weakly depends on δ through (65) and the requirement of $\varepsilon > \varepsilon_1$.

Condition (65), with $\varepsilon < \varepsilon_1$, $\widetilde{\varepsilon}_1 = \widetilde{p}^{1-\delta}$ and $\widetilde{p} = p-2$ for the estimation of precision matrix, is the key for the choice of the smaller penalty level $\lambda = (1+\varepsilon)L_n(k/p)$. It provides theoretical justifications for the choice of $k \in [1, n]$ or even up to $k \asymp n\log p$ for the theory to work. Let $s_{\max} = c_0 n/\log p$ with a sufficiently small constant $c_0 > 0$, which can be viewed as the largest possible $s \geq s_1 + s_2$ in our theory. Suppose $n \leq p^{t_0}$ for some fixed $t_0 < 1$ and the bound C_3 for the upper sparse eigenvalue can be treated as fixed in (62) for $s_2 \leq s_{\max}$. For $\lambda = (1+\varepsilon)\sqrt{(2/n)\log(p/k)}$ with $k \leq n\log p$ and $s_2 \leq s_{\max}$, condition (65) can be written as

$$\varepsilon > \varepsilon_1 \geq \frac{(1+o(1))(k/n)s_{\max}/(c_0 s_2)}{(1-t_0)(1+(1-t_0)\log p)} + \big(\sqrt{\delta}+o(1)\big)\sqrt{C_3/s_2},$$

which holds for sufficiently small $ks_{\max}/(ns_2\log p)$. This allows $k \asymp n\log p$ for $s_2 = s_{\max}$. For the asymptotic normality, we need $s_2 = o(\sqrt{n}/\log p)$, so that $k = o(n^{1/2}\log p)$ is sufficient.

5.2. *Statistical inference under unbounded condition number.* The main results in this paper assume that the spectrum of the precision matrix Ω is bounded below and above by a universal constant M in (1). Then the dependency of the key result (20) on M is hidden in the constant C_1 in front of the rate $s\frac{\log p}{n}$ for bounding $\|\hat{\theta}_{A,A} - \theta^{\mathrm{ora}}_{A,A}\|_\infty$ in Theorem 2. The inference result in (26) follows as long as this bound $C_1 s\frac{\log p}{n}$ is dominated by the parametric square-root rate of $\theta^{\mathrm{ora}}_{A,A}$, or equivalently $s = o(\sqrt{n}/\log p)$. In fact, following the proof of Theorem 2, the requirement can be somewhat weakened to $\lambda_{\max}(\Omega) \leq M$ and $\max_j \sigma_{jj} \leq M$.

It would be interesting to consider a slightly more general case, where we assume $\max_j \sigma_{jj} \leq C$ and $\lambda_{\max}(\Omega) \leq M_{n,p}$ with absolute constant C and possibly a large constant $M_{p,n} \to \infty$ as $(n, p) \to \infty$. In this setting, the condition number of Ω may not be bounded. Suppose we would like to make inference for ω_{12} and assume $\max\{\omega_{11}, \omega_{22}\} \leq C$ to make its inverse Fisher information bounded. We are able to show that (26) holds as long as $s = o(\sqrt{n}/(M_{n,p}\log p))$ under this setting. In fact, the regression model (4) is still valid with bounded noise level $\theta_{mm} \leq \sigma_{mm} \leq C$ for $m \in A = \{1, 2\}$. However, the compatibility condition (61)

may not hold with absolute constant because the smallest eigenvalue of the population Gram matrix $\lambda_{\min}(\Sigma_{A^cA^c})$ is possibly as small as $M_{n,p}^{-1}$. Taking this possible compatibility constant $M_{n,p}^{-1}$ into account, we can obtain $\|\hat{\theta}_{A,A} - \theta_{A,A}^{\mathrm{ora}}\|_\infty = O_p(sM_{n,p}\frac{\log p}{n})$ while the sufficient statistics $\theta_{A,A}^{\mathrm{ora}}$ still has square-root rate. As a consequence, the inference result in (26) holds as long as $s = o(\sqrt{n}/(M_{n,p}\log p))$. We would like to point out that to guarantee compatibility condition (61) indeed holds at the level $C_2 \asymp M_{n,p}$, an extra condition $\sqrt{s(\log p)/n} = o(M_{n,p}^{-1})$ is required; see Corollary 1 in Raskutti, Wainwright and Yu (2010). However, when $s = o(\sqrt{n}/(M_{n,p}\log p))$, this condition is automatically satisfied. The argument above can be made rigorous.

5.3. *Related works*. Our methodology in this paper is related to Zhang and Zhang (2014) who proposed a LDPE approach for making inference in a high-dimensional linear model. Since $\hat{\boldsymbol{\varepsilon}}_A$ can be viewed as an approximate projection of $\mathbf{X}_A$ to the direction of $\boldsymbol{\varepsilon}_A$ in (10), the estimator in (10) can be viewed as an LDPE as Zhang and Zhang (2014) discussed in the regression context. See also van de Geer et al. (2014) and Javanmard and Montanari (2014). When appropriately applying their approach to our setting, their result is asymptotically equivalent to ours and also obtains the asymptotic normality. In this section, we briefly discuss their approach in the large graphical model setting.

Consider $A = \{1, 2\}$. While our method regresses two nodes $\mathbf{X}_A$ against all other nodes $\mathbf{X}_{A^c}$ and focuses on the estimation of the two by two dimensional covariance matrix $\Omega_{A,A}^{-1}$ of the noise, their approach consists of the following two steps. First, one node $\mathbf{X}_1$ is regressed against all other nodes $\mathbf{X}_{1^c}$ using scaled lasso with coefficient $\hat{\beta}^{(\mathrm{init})}$. As equation (4) suggests, the noise level is ω_{11}^{-1}, and the coefficient for the column $\mathbf{X}_2$ is $\beta_2 = -\omega_{12}\omega_{11}^{-1}$. Then in the second step, to correct the bias of the initial estimator $\hat{\beta}_2^{(\mathrm{init})}$ obtained in the first step for the coefficient vector β_2, a score vector $\mathbf{z}$ is picked and applied to obtain the final estimator of β_2 as follows:

$$\hat{\beta}_2 = \hat{\beta}_2^{(\mathrm{init})} + \mathbf{z}^T\big(\mathbf{X}_1 - \mathbf{X}_{1^c}\hat{\beta}_2^{(\mathrm{init})}\big)/\mathbf{z}^T\mathbf{X}_2,$$

where $\mathbf{z}$ is the residue after regressing $\mathbf{X}_2$ against all remaining columns in step one $\mathbf{X}_{A^c}$ using scaled lasso again. To obtain the final estimator of ω_{12}, the estimator $\hat{\beta}_2$ of $-\omega_{12}\omega_{11}^{-1}$ should be scaled by an accurate estimator of ω_{11}^{-1}, which uses the variance component of the scaled lasso estimator in the first step. It seems that two approaches are quite different. However, both approaches do the same thing: they try to estimate the partial correlation of node Z_1 and Z_2 and hence are asymptotically equivalent. Compared with their approach, our method enjoys simper form and clearer interpretation. It is worthwhile to point out that the main contribution of this paper is understanding the fundamental limit of the Gaussian graphical model in making statistical inference, which is not covered by other works.

5.4. *Unknown mean* μ. In the Introduction, we assume $Z \sim \mathcal{N}(\mu, \Sigma)$ and $\mu = 0$ without loss of generality. This can be seen as follows. Suppose we observe an $n \times p$ data matrix $\mathbf{X}$ with i.i.d. rows from $\mathcal{N}(\mu, \Sigma)$. Let $u^{(i)}$, $i = 1, \ldots, n$, be n-dimensional orthonormal row vectors with $u^{(n)} = (1, \ldots, 1)/\sqrt{n}$. Then $u^{(i)}\mathbf{X}$ are i.i.d. p-dimensional row vectors from $\mathcal{N}(0, \Sigma)$. Thus we can simply apply our methods and theory to the sample $\{u^{(i)}\mathbf{X}, i = 1, \ldots, n-1\}$.

6. Numerical studies. In this section, we present some numerical results for both asymptotic distribution and support recovery. We generate the data from $p \times p$ precision matrices with three blocks. Two cases are considered: $p = 200, 800$. The ratio of block sizes is $2:1:1$; that is, for a 200×200 matrix, the block sizes are 100×100, 50×50 and 50×50, respectively. The diagonal entries are $\alpha_1, \alpha_2, \alpha_3$ in three blocks, respectively, where $(\alpha_1, \alpha_2, \alpha_3) = (1, 2, 4)$. When the entry is in the kth 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 simple procedure with a single threshold level for all entries is not likely to perform well.

We first estimate the entries in the precision matrix and partial correlations as discussed in Remark 3, and consider the distributions of these estimators. We generate a random sample of size $n = 400$ from a multivariate Gaussian distribution $\mathcal{N}(0, \Sigma)$ with $\Sigma = \Omega^{-1}$. For the proposed estimators defined through (10) and (11) with the scaled lasso (12) or the LSE after model selection (13), we pick $\lambda = n^{-1/2}L_n(1/p) \approx \sqrt{(2/n)\log p}$; that is, $k = 1$ in (64) with small adjustment in n and p ignored. This is justified by our theoretical results as discussed in Section 5.1.

Table 1 reports the mean and standard error of our estimators for four entries in the precision matrices and the corresponding correlations. In addition, we report the point estimates by the GLasso [Friedman, Hastie and Tibshirani (2008)] and CLIME [Cai, Liu and Luo (2011)] for comparison. For $p = 800$, the results for the GLasso are based on 10 replications, while all other entries in the table are based on 100 replications. The GLasso is computed by the R package "glasso" with penalized diagonal (default option), while the CLIME estimators are computed by the R package "fastclime" [Pang, Liu and Vanderbei (2014)]. As the GLasso and CLIME are designed for estimating precision matrices as high-dimensional objects, it is not surprising that the proposed estimator outperforms them in estimation accuracy for individual entries. Figures 1 and 2 show the histograms of the proposed estimates with the theoretical Gaussian density in Theorem 3 super-imposed. They demonstrated that the histograms match pretty well to the asymptotic distribution, especially for the LSE after model selection. The asymptotic normality leads to the following $(1-\alpha)$ confidence intervals for ω_{ij} and r_{ij}:

$$\Big(\hat{\omega}_{ij} - z_{\alpha/2}\sqrt{(\hat{\omega}_{i,i}\hat{\omega}_{j,j} + \hat{\omega}_{i,j}^2)/n}, \omega_{ij} + z_{\alpha/2}\sqrt{(\hat{\omega}_{i,i}\hat{\omega}_{j,j} + \hat{\omega}_{i,j}^2)/n}\Big),$$

$$(\hat{r}_{i,j} - z_{\alpha/2}(1-\hat{r}_{i,j}^2)/\sqrt{n}, \hat{r}_{i,j} + z_{\alpha/2}(1-\hat{r}_{i,j}^2)/\sqrt{n}),$$

TABLE 1
Mean and standard error of GLasso, CLIME and proposed estimators

p		$\omega_{1,2} = 0.5$	$\omega_{1,3} = 0.4$	$\omega_{1,4} = 0$	$\omega_{1,10} = 0$
200	GLasso	0.368 ± 0.039	0.282 ± 0.038	-0.056 ± 0.03	-0.001 ± 0.01
	CLIME	0.776 ± 0.479	0.789 ± 0.556	0.482 ± 1.181	0.002 ± 0.017
	$\hat{\omega}_{i,j}$	0.459 ± 0.05	0.372 ± 0.052	-0.049 ± 0.041	-0.003 ± 0.044
	$\hat{\omega}_{i,j}^{\mathrm{LSE}}$	0.503 ± 0.059	0.401 ± 0.061	-0.006 ± 0.049	-0.002 ± 0.052
800	GLasso	0.801 ± 0.039	0.258 ± 0.031	0.19 ± 0.014	-0.063 ± 0.028
	CLIME	1.006 ± 0.255	0.046 ± 0.140	0.022 ± 0.071	0.018 ± 0.099
	$\hat{\omega}_{i,j}$	0.436 ± 0.049	0.361 ± 0.047	-0.057 ± 0.044	0.001 ± 0.044
	$\hat{\omega}_{i,j}^{\mathrm{LSE}}$	0.491 ± 0.059	0.396 ± 0.058	0 ± 0.052	-0.003 ± 0.05
p		$\mathbf{r}_{1,2} = -0.5$	$\mathbf{r}_{1,3} = -0.4$	$\mathbf{r}_{1,4} = 0$	$\mathbf{r}_{1,10} = 0$
200	$\hat{r}_{i,j}$	-0.477 ± 0.037	-0.391 ± 0.043	0.051 ± 0.043	0.003 ± 0.046
	$\hat{r}_{i,j}^{\mathrm{LSE}}$	-0.485 ± 0.04	-0.386 ± 0.046	0.006 ± 0.047	0.002 ± 0.049
800	$\hat{r}_{i,j}$	-0.468 ± 0.039	-0.392 ± 0.041	0.06 ± 0.045	-0.001 ± 0.048
	$\hat{r}_{i,j}^{\mathrm{LSE}}$	-0.475 ± 0.041	-0.382 ± 0.044	0 ± 0.049	0.002 ± 0.048

where $z_{\alpha/2}$ is the z-score such that $P(\mathcal{N}(0,1) > z_{\alpha/2}) = \alpha/2$. Table 2 reports the empirical coverage probabilities for 95% confidence intervals, which matches well to the assigned confidence level.

Support recovery of a precision matrix is of great interest. We compare our selection results with the GLasso and CLIME. 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 and CLIME. These estimators are computed based on the entire training sample with a range of penalty levels and a proper penalty level is chosen by minimizing the negative likelihood $\{\mathrm{trace}(\overline{\Sigma}\hat{\Omega}) - \log\det(\hat{\Omega})\}$ on the validation sample, where $\overline{\Sigma}$ is the sample covariance matrix. The proposed ANT estimators are computed based on the training sample only with $\xi_0 = 2$ in the thresholding step as in (38). Tables 3 and 4 present the average selection performances as measured in the true positive, false positive and the corresponding rates. In addition to the overall performance, the summary statistics are reported for each block. The results demonstrate the selection consistency property of both ANT methods and substantial false positive for the GLasso and CLIME. It should be pointed out that the ANT takes the advantage of an additional thresholding step, while the GLasso and CLIME do not. A possible explanation of the false positive for the GLasso is a tendency for the likelihood criterion with the validation sample to pick a small penalty level. However, such an explanation seems not to hold for the CLIME, which demonstrated much lower false positive than the GLasso, as the true positive rate of the CLIME is consistently maintained at about 95% for $p = 200$ and 85% for $p = 800$.

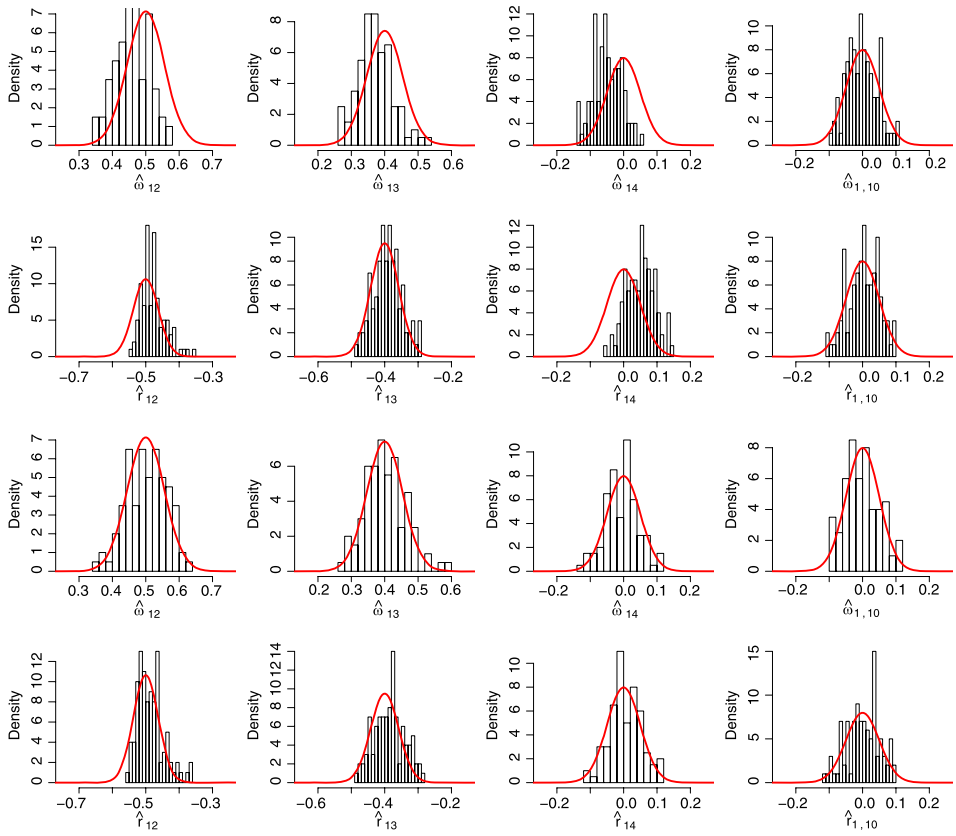

FIG. 1. *Histograms of estimated entries for* $p = 200$. *The first row*: *scaled lasso for entries* $\omega_{1,2}$ *and* $\omega_{1,3}$ *in the precision matrix*; *the second row*: *scaled lasso for entries* $\omega_{1,4}$ *and* $\omega_{1,10}$; *the third and fourth rows*: *LSE after scaled lasso selection*.

Moreover, we compare the ANT with the GLasso and CLIME in a range of penalty levels. Figure 3 plots the ROC curves for the GLasso and CLIME with various penalty levels and the ANT with various thresholding levels in the follow-up procedure. It demonstrates that the CLIME outperforms the GLasso, but the two methods perform significantly more poorly than the ANT in the experiment. In addition, the circle in the plot represents the performance of the ANT with the selected threshold level as in (38). The triangle and diamond in the plot represents the performance of the GLasso and CLIME with the penalty level chosen by cross-validation, respectively. This again indicates that our method simultaneously achieves a very high true positive rate and a very low false positive rate.

7. Proof of Theorem 5. In this section we show that the upper bound given in Section 2.3 is indeed rate optimal. We will only establish equation (35). Equa-

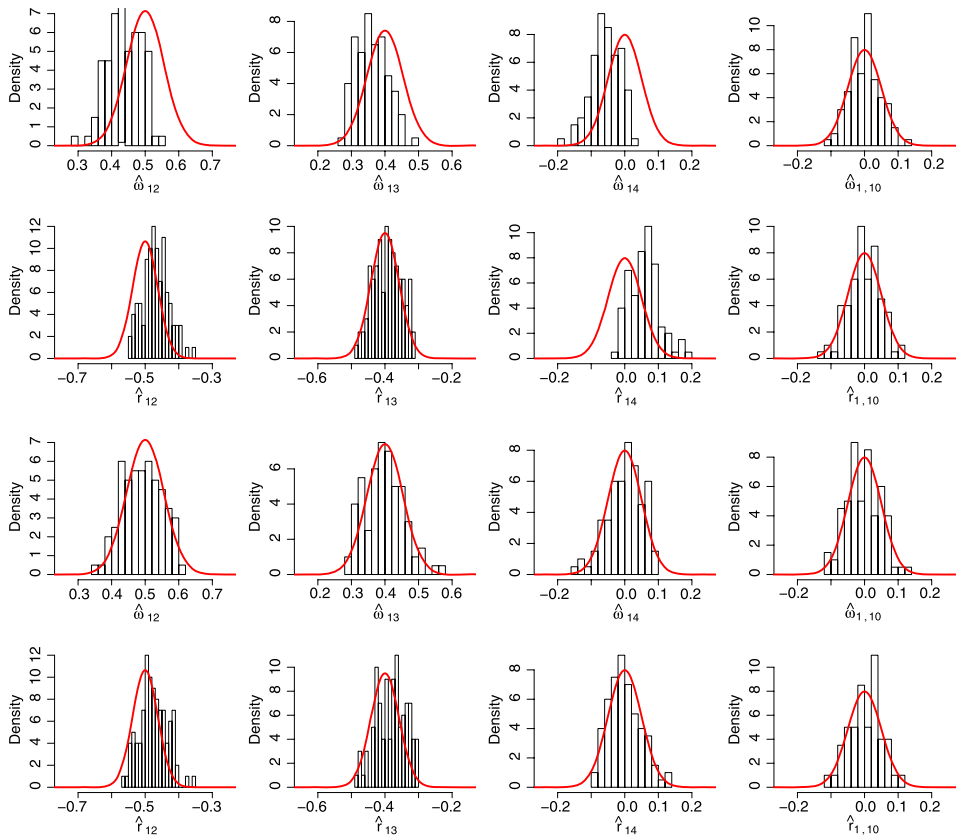

FIG. 2. *Histograms of estimated entries for* $p = 800$. *The first row*: *scaled lasso for entries* $\omega_{1,2}$ *and* $\omega_{1,3}$ *in the precision matrix*; *the second row*: *scaled lasso for entries* $\omega_{1,4}$ *and* $\omega_{1,10}$; *the third and fourth rows*: *LSE after scaled lasso selection*.

tion (36) is an immediate consequence of equation (35) 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 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 $X^{(1)}, X^{(2)}, \ldots, X^{(n)}$ with each $X^{(i)} \sim \mathcal{N}(0, \Omega_m^{-1})$, $0 \le m \le m_*$ and f_m denoting the corresponding joint density, and we define

$$\bar{\mathbb{P}} = \frac{1}{m_*} \sum_{m=1}^{m_*} \mathbb{P}_{\Omega_m}. \tag{75}$$

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).

TABLE 2
Empirical coverage probabilities of the 95% confidence intervals

p	**(i, j)**	**(1, 2)**	**(1, 3)**	**(1, 4)**	**(1, 10)**
200	$\hat{\omega}_{i,j}$	0.87	0.89	0.87	0.98
	$\hat{\omega}_{i,j}^{\mathrm{LSE}}$	0.96	0.91	0.94	0.98
	$\hat{r}_{i,j}$	0.94	0.94	0.87	0.98
	$\hat{r}_{i,j}^{\mathrm{LSE}}$	0.93	0.94	0.94	0.97
800	$\hat{\omega}_{i,j}$	0.74	0.88	0.84	0.95
	$\hat{\omega}_{i,j}^{\mathrm{LSE}}$	0.93	0.93	0.96	0.96
	$\hat{r}_{i,j}$	0.89	0.98	0.83	0.95
	$\hat{r}_{i,j}^{\mathrm{LSE}}$	0.90	0.94	0.96	0.96

LEMMA 1. *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 $\Omega_m = (\omega_{kl}^{(m)})_{p\times p}$, then*

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

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

TABLE 3
The performance of support recovery ($p = 200$, 100 replications)

Block	**Method**	**TP**	**TPR**	**FP**	**FPR**
Overall	GLasso	391	1	5322.24	0.2728
	CLIME	372.61	0.953	588.34	0.0302
	ANT	391	1	0.04	0
	ANT-LSE	390.97	0.9999	0.01	0
Block 1	GLasso	197	1	1981.1	0.4168
	CLIME	188.47	0.9567	205.58	0.0433
	ANT	197	1	0	0
	ANT-LSE	196.98	0.9999	0	0
Block 2	GLasso	97	1	293.93	0.2606
	CLIME	92.29	0.9514	72.89	0.0646
	ANT	97	1	0	0
	ANT-LSE	96.99	0.9999	0	0
Block 3	GLasso	97	1	160.93	0.1427
	CLIME	91.85	0.9469	72.94	0.0647
	ANT	97	1	0	0
	ANT-LSE	97	1	0	0

TABLE 4
The performance of support recovery ($p = 800$, 10 *replications*)

Block	Method	TP	TPR	FP	FPR
Overall	GLasso	1590.7	0.9998	44785.6	0.1408
	CLIME	1365.9	0.8585	134.6	4e–04
	ANT	1589	0.9987	0	0
	ANT-LSE	1586.2	0.997	0	0
Block 1	GLasso	797	1	19694.5	0.2493
	CLIME	687.5	0.8626	71.4	9e–04
	ANT	795.8	0.9985	0	0
	ANT-LSE	794.8	0.9972	0	0
Block 2	GLasso	397	1	2133.4	0.1094
	CLIME	339.4	0.8549	29.6	0.0015
	ANT	396.7	0.9992	0	0
	ANT-LSE	395.8	0.997	0	0
Block 3	GLasso	396.7	0.9992	664.7	0.0341
	CLIME	339	0.8539	32.6	0.0017
	ANT	396.5	0.9987	0	0
	ANT-LSE	395.6	0.9965	0	0

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 $i \neq j$, we could always permute the coordinates and

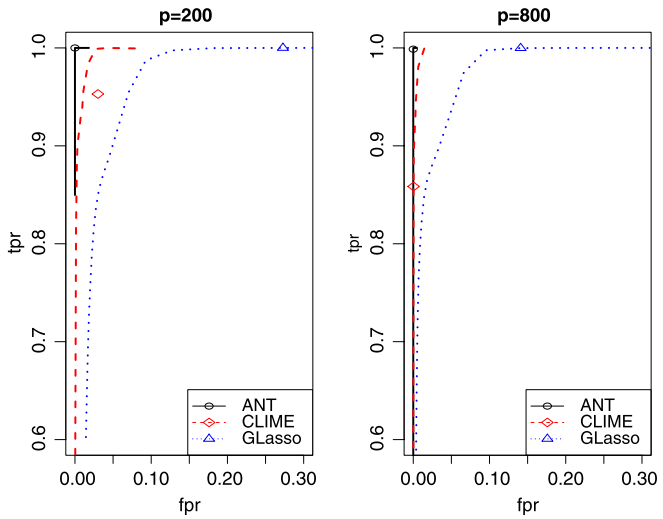

FIG. 3. *The ROC curves. Circle*: *ANT with the proposed thresholding. Triangle*: *GLasso with penalty level by CV. Diamond*: *CLIME with penalty level by CV.*

rearrange them to the special case ω_{11} or ω_{12}.

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

$$\Sigma_0 = \begin{pmatrix} 1 & b & 0 & \cdots & 0 \\ b & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix} \quad \text{and}$$

$$\Omega_0 = \Sigma_0^{-1} = \begin{pmatrix} \dfrac{1}{1-b^2} & \dfrac{-b}{1-b^2} & 0 & \cdots & 0 \\ \dfrac{-b}{1-b^2} & \dfrac{1}{1-b^2} & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}; \tag{76}$$

that is, $\Sigma_0 = (\sigma_{kl}^{(0)})_{p\times p}$ 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 Ω_m, $1 \le m \le m_*$, the construction is as follows. Without loss of generality we assume $k_{n,p} \ge 3$. Denote by $\mathcal{H}$ the collection of all $p \times p$ symmetric matrices with exactly $(k_{n,p} - 2)$ elements equal to 1 between the third and the last elements on the first row (column) and the rest all zeros. Define

$$\mathcal{G}_0 = \{\Omega : \Omega = \Omega_0 \text{ or } \Omega = (\Sigma_0 + aH)^{-1}, \text{ for some } H \in \mathcal{H}\}, \tag{77}$$

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

$$m^* = \operatorname{Card}(\mathcal{G}_0) - 1 = \operatorname{Card}(\mathcal{H}) = \binom{p-2}{k_{n,p}-2}.$$

We pick the constant $b = \frac{1}{2}(1 - 1/M)$ and

$$0 < \tau_1 < \min\left\{\frac{(1-1/M)^2 - b^2}{C_0}, \frac{(1-b^2)^2}{2C_0(1+b^2)}, \frac{(1-b^2)^2}{4\nu(1+b^2)}\right\},$$

and prove that $\mathcal{G}_0 \subset \mathcal{G}_0(M, k_{n,p})$.

First we show that for all Ω_i,

$$1/M \le \lambda_{\min}(\Omega_i) < \lambda_{\max}(\Omega_i) \le M. \tag{78}$$

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

$$\lambda_1(\Omega_m^{-1}) = 1 + \sqrt{b^2 + (k_{n,p}-2)a^2}, \qquad \lambda_p(\Omega_m^{-1}) = 1 - \sqrt{b^2 + (k_{n,p}-2)a^2},$$
$$\lambda_2(\Omega_m^{-1}) = \lambda_3(\Omega_m^{-1}) = \cdots = \lambda_{p-1}(\Omega_m^{-1}) = 1.$$

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

$$\begin{aligned} &1 - \sqrt{b^2 + (k_{n,p} - 2)a^2} \geq 1 - \sqrt{b^2 + \tau_1 C_0} > 1/M, \\ &1 + \sqrt{b^2 + (k_{n,p} - 2)a^2} < 2 - 1/M < M, \end{aligned} \tag{79}$$

which imply

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

As for matrix Ω_0, similarly we have

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

and 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 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 $k_{n,p}$ and $(p - k_{n,p})$, of which the second block is an identity matrix. Then $(P_\pi \Omega_m^{-1} P_\pi^T)^{-1} = P_\pi \Omega_m P_\pi^T$ has the same blocking structure with the first block of dimension $k_{n,p}$ and the second block being an identity matrix. Thus the number of nonzero 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 (78).

Step 2: *Bounding* $\boldsymbol{\alpha}$. From the construction of Ω_m^{-1} and the matrix inverse formula, we have that for any precision matrix Ω_m,

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

for $1 \leq m \leq m_*$, and for the precision matrix Ω_0,

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

Since $b^2 + (k_{n,p} - 2)a^2 < (1 - 1/M)^2 < 1$ in equation (79), we have

$$\begin{aligned} \inf_{1 \leq m \leq m_*} |\omega_{11}^{(m)} - \omega_{11}^{(0)}| &= \frac{(k_{n,p} - 2)a^2}{(1 - b^2)(1 - b^2 - (k_{n,p} - 2)a^2)} \geq C_3 k_{n,p} a^2, \\ \inf_{1 \leq m \leq m_*} |\omega_{12}^{(m)} - \omega_{12}^{(0)}| &= \frac{b(k_{n,p} - 2)a^2}{(1 - b^2)(1 - b^2 - (k_{n,p} - 2)a^2)} \geq C_4 k_{n,p} a^2, \end{aligned} \tag{80}$$

for some constants $C_3, C_4 > 0$.

Step 3: *Bounding the affinity*. The following lemma is proved in Ren et al. (2015).

LEMMA 2. *Let* $\bar{\mathbb{P}}$ *be defined in* (75). *We have*

$$\|\mathbb{P}_{\Omega_0} \wedge \bar{\mathbb{P}}\| \geq C_5 \tag{81}$$

for some constant $C_5 > 0$.

Lemma 1, together with equations (80), (81) and $a = \sqrt{\frac{\tau_1 \log p}{n}}$, imply

$$\sup_{0\leq m\leq m_*} \mathbb{P}\left\{|\hat{\omega}_{11} - \omega_{11}^{(m)}| > \frac{1}{2} \cdot \frac{C_3\tau_1 k_{n,p} \log p}{n}\right\} \geq C_5/2,$$

$$\sup_{0\leq m\leq m_*} \mathbb{P}\left\{|\hat{\omega}_{12} - \omega_{12}^{(m)}| > \frac{1}{2} \cdot \frac{C_4\tau_1 k_{n,p} \log p}{n}\right\} \geq C_5/2,$$

which match the lower bound in (35) 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 of the order $k_{n,p}\sqrt{\frac{\log p}{n}}$, 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 the literature we need to assume the 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. However, the methodology proposed in this paper improves the upper bounds in the 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.

SUPPLEMENTARY MATERIAL

Supplement to "Asymptotic normality and optimalities in estimation of large Gaussian graphical model" (DOI: 10.1214/14-AOS1286SUPP; .pdf). In this supplement we collect proofs of Theorems 1–3 in Section 2, proofs of Theorems 6, 8 in Section 3 and proofs of Theorems 10–11 as well as Proposition 1 in Section 4.

REFERENCES

ANTONIADIS, A. (2010). Comment: ℓ_1-penalization for mixture regression models [MR2677722]. *TEST* **19** 257–258. MR2677723

BELLONI, A., CHERNOZHUKOV, V. and HANSEN, C. (2014). Inference on treatment effects after selection among high-dimensional controls. *Rev. Econ. Stud.* **81** 608–650. MR3207983

BELLONI, A., CHERNOZHUKOV, V. and WANG, L. (2011). Square-root lasso: Pivotal recovery of sparse signals via conic programming. *Biometrika* **98** 791–806. MR2860324

BICKEL, P. J. and LEVINA, E. (2008a). Regularized estimation of large covariance matrices. *Ann. Statist.* **36** 199–227. MR2387969

BICKEL, P. J. and LEVINA, E. (2008b). Covariance regularization by thresholding. *Ann. Statist.* **36** 2577–2604. MR2485008

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

BÜHLMANN, P. (2013). Statistical significance in high-dimensional linear models. *Bernoulli* **19** 1212–1242. MR3102549

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

CAI, T. T., LIU, W. and ZHOU, H. H. (2012). Estimating sparse precision matrix: Optimal rates of convergence and adaptive estimation. Preprint. Available at arXiv:1212.2882.

CAI, T. T., ZHANG, C.-H. and ZHOU, H. H. (2010). Optimal rates of convergence for covariance matrix estimation. *Ann. Statist.* **38** 2118–2144. MR2676885

CAI, T. T. and ZHOU, H. H. (2012). Optimal rates of convergence for sparse covariance matrix estimation. *Ann. Statist.* **40** 2389–2420. MR3097607

CANDÈS, E. J. and RECHT, B. (2009). Exact matrix completion via convex optimization. *Found. Comput. Math.* **9** 717–772. MR2565240

CHANDRASEKARAN, V., PARRILO, P. A. and WILLSKY, A. S. (2012). Latent variable graphical model selection via convex optimization. *Ann. Statist.* **40** 1935–1967. MR3059067

D'ASPREMONT, A., BANERJEE, O. and EL GHAOUI, L. (2008). First-order methods for sparse covariance selection. *SIAM J. Matrix Anal. Appl.* **30** 56–66. MR2399568

EL KAROUI, N. (2008). Operator norm consistent estimation of large-dimensional sparse covariance matrices. *Ann. Statist.* **36** 2717–2756. MR2485011

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 Univ. Press, Cambridge. MR1084815

JAVANMARD, A. and MONTANARI, A. (2014). Hypothesis testing in high-dimensional regression under the Gaussian random design model: Asymptotic theory. *IEEE Trans. Inform. Theory* **60** 6522–6554. MR3265038

KOLTCHINSKII, V. (2009). The Dantzig selector and sparsity oracle inequalities. *Bernoulli* **15** 799–828. MR2555200

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

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

LE CAM, L. (1973). Convergence of estimates under dimensionality restrictions. *Ann. Statist.* **1** 38–53. MR0334381

LIU, W. (2013). Gaussian graphical model estimation with false discovery rate control. *Ann. Statist.* **41** 2948–2978. MR3161453

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

PANG, H., LIU, H. and VANDERBEI, R. (2014). The FASTCLIME package for linear programming and large-scale precision matrix estimation in R. *J. Mach. Learn. Res.* **15** 489–493.

RASKUTTI, G., WAINWRIGHT, M. J. and YU, B. (2010). Restricted eigenvalue properties for correlated Gaussian designs. *J. Mach. Learn. Res.* **11** 2241–2259. MR2719855

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. Stat.* **5** 935–980. MR2836766

REN, Z. and ZHOU, H. H. (2012). Discussion: Latent variable graphical model selection via convex optimization [MR3059067]. *Ann. Statist.* **40** 1989–1996. MR3059072

REN, Z., SUN, T., ZHANG, C.-H. and ZHOU, H. H. (2015). Supplement to "Asymptotic normality and optimalities in estimation of large Gaussian graphical models." DOI:10.1214/14-AOS1286SUPP.

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

STÄDLER, N., BÜHLMANN, P. and VAN DE GEER, S. (2010). ℓ_1-penalization for mixture regression models. *TEST* **19** 209–256. MR2677722

SUN, T. and ZHANG, C.-H. (2010). Comment: ℓ_1-penalization for mixture regression models [MR2677722]. *TEST* **19** 270–275. MR2677726

SUN, T. and ZHANG, C.-H. (2012a). Scaled sparse linear regression. *Biometrika* **99** 879–898. MR2999166

SUN, T. and ZHANG, C.-H. (2012b). Comment: "Minimax estimation of large covariance matrices under ℓ_1-norm" [MR3027084]. *Statist. Sinica* **22** 1354–1358. MR3027086

SUN, T. and ZHANG, C.-H. (2013). Sparse matrix inversion with scaled lasso. *J. Mach. Learn. Res.* **14** 3385–3418. MR3144466

THORIN, G. O. (1948). Convexity theorems generalizing those of M. Riesz and Hadamard with some applications. *Comm. Sem. Math. Univ. Lund* [*Medd. Lunds Univ. Mat. Sem.*] **9** 1–58. MR0025529

VAN DE GEER, S. A. and BÜHLMANN, P. (2009). On the conditions used to prove oracle results for the Lasso. *Electron. J. Stat.* **3** 1360–1392. MR2576316

VAN DE GEER, S., BÜHLMANN, P., RITOV, Y. and DEZEURE, R. (2014). On asymptotically optimal confidence regions and tests for high-dimensional models. *Ann. Statist.* **42** 1166–1202. MR3224285

YE, F. and ZHANG, C.-H. (2010). Rate minimaxity of the Lasso and Dantzig selector for the ℓ_q loss in ℓ_r balls. *J. Mach. Learn. Res.* **11** 3519–3540. MR2756192

YU, B. (1997). Assouad, Fano, and Le Cam. In *Festschrift for Lucien Le Cam* 423–435. Springer, New York.

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

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

ZHANG, T. (2009). Some sharp performance bounds for least squares regression with L_1 regularization. *Ann. Statist.* **37** 2109–2144. MR2543687

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. MR2435448

ZHANG, C.-H. and ZHANG, T. (2012). A general theory of concave regularization for high-dimensional sparse estimation problems. *Statist. Sci.* **27** 576–593. MR3025135

ZHANG, C.-H. and ZHANG, S. S. (2014). Confidence intervals for low dimensional parameters in high dimensional linear models. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **76** 217–242. MR3153940

Z. REN
DEPARTMENT OF STATISTICS
UNIVERSITY OF PITTSBURGH
PITTSBURGH, PENNSYLVANIA 15260
USA
E-MAIL: zren@pitt.edu

T. SUN
DEPARTMENT OF MATHEMATICS
UNIVERSITY OF MARYLAND
COLLEGE PARK, MARYLAND 20742
USA
E-MAIL: tingni@umd.edu

C.-H. ZHANG
DEPARTMENT OF STATISTICS AND BIOSTATISTICS
HILL CENTER, BUSCH CAMPUS
RUTGERS UNIVERSITY
PISCATAWAY, NEW JERSEY 08854
USA
E-MAIL: cunhui@stat.rutgers.edu

H. H. ZHOU
DEPARTMENT OF STATISTICS
YALE UNIVERSITY
NEW HAVEN, CONNECTICUT 06511
USA
E-MAIL: huibin.zhou@yale.edu