The Annals of Statistics
2016, Vol. 44, No. 2, 455–488
DOI: 10.1214/13-AOS1171

ESTIMATING SPARSE PRECISION MATRIX: OPTIMAL RATES OF CONVERGENCE AND ADAPTIVE ESTIMATION

BY T. TONY CAI[1], WEIDONG LIU[2] AND HARRISON H. ZHOU[3]

University of Pennsylvania, Shanghai Jiao Tong University and Yale University

Precision matrix is of significant importance in a wide range of applications in multivariate analysis. This paper considers adaptive minimax estimation of sparse precision matrices in the high dimensional setting. Optimal rates of convergence are established for a range of matrix norm losses. A fully data driven estimator based on adaptive constrained ℓ_1 minimization is proposed and its rate of convergence is obtained over a collection of parameter spaces. The estimator, called ACLIME, is easy to implement and performs well numerically.

A major step in establishing the minimax rate of convergence is the derivation of a rate-sharp lower bound. A "two-directional" lower bound technique is applied to obtain the minimax lower bound. The upper and lower bounds together yield the optimal rates of convergence for sparse precision matrix estimation and show that the ACLIME estimator is adaptively minimax rate optimal for a collection of parameter spaces and a range of matrix norm losses simultaneously.

1. Introduction. Precision matrix plays a fundamental role in many high-dimensional inference problems. For example, knowledge of the precision matrix is crucial for classification and discriminant analyses. Furthermore, precision matrix is critically useful for a broad range of applications such as portfolio optimization, speech recognition and genomics. See, for example, Lauritzen (1996), Yuan and Lin (2007) and Saon and Chien (2011). Precision matrix is also closely connected to the graphical models which are a powerful tool to model the relationships among a large number of random variables in a complex system and are used in a wide array of scientific applications. It is well known that recovering the structure of an undirected Gaussian graph is equivalent to the recovery of the support of the precision matrix. See, for example, Lauritzen (1996), Meinshausen

Received January 2013; revised June 2013.

[1]Supported in part by NSF FRG Grant DMS-08-54973 and NSF Grant DMS-12-08982.

[2]Supported by NSFC Grants 11201298, 11322107 and 11431006, the Program for Professor of Special Appointment (Eastern Scholar) at Shanghai Institutions of Higher Learning, Shanghai Pujiang Program, Shanghai Shuguang Program, Foundation for the Author of National Excellent Doctoral Dissertation of PR China and 973 Program (2015CB856004).

[3]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. Constrained ℓ_1-minimization, covariance matrix, graphical model, minimax lower bound, optimal rate of convergence, precision matrix, sparsity, spectral norm.

and Bühlmann (2006) and Cai, Liu and Luo (2011). Liu, Lafferty and Wasserman (2009) extended the result to a more general class of distributions called nonparanormal distributions.

The problem of estimating a large precision matrix and recovering its support has drawn considerable recent attention and a number of methods have been introduced. Meinshausen and Bühlmann (2006) proposed a neighborhood selection method for recovering the support of a precision matrix. Penalized likelihood methods have also been introduced for estimating sparse precision matrices. Yuan and Lin (2007) proposed an ℓ_1 penalized normal likelihood estimator and studied its theoretical properties. See also Friedman, Hastie and Tibshirani (2008), d'Aspremont, Banerjee and El Ghaoui (2008), Rothman et al. (2008), Lam and Fan (2009) and Ravikumar et al. (2011). Yuan (2010) applied the Dantzig selector method to estimate the precision matrix and gave the convergence rates for the estimator under the matrix ℓ_1 norm and spectral norm. Cai, Liu and Luo (2011) introduced an estimator called CLIME using a constrained ℓ_1 minimization approach and obtained the rates of convergence for estimating the precision matrix under the spectral norm and Frobenius norm.

Although many methods have been proposed and various rates of convergence have been obtained, it is unclear which estimator is optimal for estimating a sparse precision matrix in terms of convergence rate. This is due to the fact that the minimax rates of convergence, which can serve as a fundamental benchmark for the evaluation of the performance of different procedures, is still unknown. The goals of the present paper are to establish the optimal minimax rates of convergence for estimating a sparse precision matrix under a class of matrix norm losses and to introduce a fully data driven adaptive estimator that is simultaneously rate optimal over a collection of parameter spaces for each loss in this class.

Let $X_1, \ldots, X_n$ be a random sample from a p-variate distribution with a covariance matrix $\Sigma = (\sigma_{ij})_{1 \le i,j \le p}$. The goal is to estimate the inverse of Σ, the precision matrix $\Omega = (\omega_{ij})_{1 \le i,j \le p}$. It is well known that in the high-dimensional setting structural assumptions are needed in order to consistently estimate the precision matrix. The class of sparse precision matrices, where most of the entries in each row/column are zero or negligible, is of particular importance as it is related to sparse graphs in the Gaussian case. For a matrix A and a number $1 \le w \le \infty$, the matrix ℓ_w norm is defined as $\|A\|_w = \sup_{|x|_w \le 1} |Ax|_w$. The sparsity of a precision matrix can be modeled by the ℓ_q balls with $0 \le q < 1$. More specifically, we define the parameter space $\mathcal{G}_q(c_{n,p}, M_{n,p})$ by

$$\mathcal{G}_q(c_{n,p}, M_{n,p}) = \left\{ \begin{array}{c} \Omega = (\omega_{ij})_{1 \le i,j \le p} : \max_j \sum_{i=1}^{p} |\omega_{ij}|^q \le c_{n,p}, \\ \|\Omega\|_1 \le M_{n,p}, \lambda_{\max}(\Omega)/\lambda_{\min}(\Omega) \le M_1, \Omega \succ 0 \end{array} \right\}, \tag{1.1}$$

where $0 \le q < 1$, $M_{n,p}$ and $c_{n,p}$ are positive and bounded away from 0 and allowed to grow as n and p grow, $M_1 > 0$ is a given constant, $\lambda_{\max}(\Omega)$ and $\lambda_{\min}(\Omega)$ are the largest and smallest eigenvalues of Ω, respectively, and $c_1 n^{\beta} \le p \le \exp(\gamma n)$ for some constants $\beta > 1$, $c_1 > 0$ and $\gamma > 0$. The notation $A \succ 0$ means that A is symmetric and positive definite. In the special case of $q = 0$, a matrix in $\mathcal{G}_0(c_{n,p}, M_{n,p})$ has at most $c_{n,p}$ nonzero elements on each row/column.

Our analysis establishes the minimax rates of convergence for estimating the precision matrices over the parameter space $\mathcal{G}_q(c_{n,p}, M_{n,p})$ under the matrix ℓ_w norm losses for $1 \le w \le \infty$. We shall first introduce a new method using an adaptive constrained ℓ_1 minimization approach for estimating the sparse precision matrices. The estimator, called ACLIME, is fully data-driven and easy to implement. The properties of the ACLIME are then studied in detail under the matrix ℓ_w norm losses. In particular, we establish the rates of convergence for the ACLIME estimator which provide upper bounds for the minimax risks.

A major step in establishing the minimax rates of convergence is the derivation of rate sharp lower bounds. As in the case of estimating sparse covariance matrices, conventional lower bound techniques, which are designed and well suited for problems with parameters that are scalar or vector-valued, fail to yield good results for estimating sparse precision matrices under the spectral norm. In the present paper, we apply the "two-directional" lower bound technique first developed in Cai and Zhou (2012) for estimating sparse covariance matrices. This lower bound method can be viewed as a simultaneous application of Assouad's lemma along the row direction and Le Cam's method along the column direction. The lower bounds match the rates in the upper bounds for the ACLIME estimator, and thus yield the minimax rates.

By combining the minimax lower and upper bounds developed in later sections, the main results on the optimal rates of convergence for estimating a sparse precision matrix under various norms can be summarized in the following theorem. We focus here on the exact sparse case of $q = 0$; the optimal rates for the general case of $0 \le q < 1$ are given in the end of Section 4. Here, for two sequences of positive numbers a_n and b_n, $a_n \asymp b_n$ means that there exist positive constants c and C independent of n such that $c \le a_n/b_n \le C$.

THEOREM 1.1. *Let $X_i \overset{i.i.d.}{\sim} N_p(\mu, \Sigma)$, $i = 1, 2, \ldots, n$, and let $1 \le c_{n,p} = o(n^{1/2}(\log p)^{-3/2})$. The minimax risk of estimating the precision matrix $\Omega = \Sigma^{-1}$ over the class $\mathcal{G}_0(c_{n,p}, M_{n,p})$ based on the random sample $\{X_1, \ldots, X_n\}$ satisfies*

$$\inf_{\hat{\Omega}} \sup_{\mathcal{G}_0(k, M_{n,p})} \mathbb{E}\|\hat{\Omega} - \Omega\|_w^2 \asymp M_{n,p}^2 c_{n,p}^2 \frac{\log p}{n} \tag{1.2}$$

for all $1 \le w \le \infty$.

In view of Theorem 1.1, the ACLIME estimator to be introduced in Section 2, which is fully data driven, attains the optimal rates of convergence simultaneously for all k-sparse precision matrices in the parameter spaces $\mathcal{G}_0(k, M_{n,p})$ with

$k \ll n^{1/2}(\log p)^{-3/2}$ under the matrix ℓ_w norm for all $1 \le w \le \infty$. The commonly used spectral norm coincides with the matrix ℓ_2 norm. For a symmetric matrix A, it is known that the spectral norm $\|A\|_2$ is equal to the largest magnitude of eigenvalues of A. When $w = 1$, the matrix ℓ_1 norm is simply the maximum absolute column sum of the matrix. As will be seen in Section 4, the adaptivity holds for the general ℓ_q balls $\mathcal{G}_q(c_{n,p}, M_{n,p})$ with $0 \le q < 1$. The ACLIME procedure is thus rate optimally adaptive to both the sparsity patterns and the loss functions.

In addition to its theoretical optimality, the ACLIME estimator is computationally easy to implement for high dimensional data. It can be computed column by column via linear programming and the algorithm is easily scalable. A simulation study is carried out to investigate the numerical performance of the ACLIME estimator. The results show that the procedure performs favorably in comparison to CLIME.

Our work on optimal estimation of precision matrix given in the present paper is closely connected to a growing literature on estimation of large covariance matrices. Many regularization methods have been proposed and studied. For example, Bickel and Levina (2008a, 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 Yuan (2012) introduced an adaptive block thresholding estimator which is simultaneously rate optimal over large collections of bandable covariance matrices. Cai and Zhou (2012) obtained the minimax rate of convergence for estimating sparse covariance matrices under a range of losses including the spectral norm loss. In particular, a new general lower bound technique was developed. Cai and Liu (2011) introduced an adaptive thresholding procedure for estimating sparse covariance matrices that automatically adjusts to the variability of individual entries.

The rest of the paper is organized as follows. The ACLIME estimator is introduced in detail in Section 2 and its theoretical properties are studied in Section 3. In particular, a minimax upper bound for estimating sparse precision matrices is obtained. Section 4 establishes a minimax lower bound which matches the minimax upper bound derived in Section 2 in terms of the convergence rate. The upper and lower bounds together yield the optimal minimax rate of convergence. A simulation study is carried out in Section 5 to compare the performance of the ACLIME with that of the CLIME estimator. Section 6 gives the optimal rate of convergence for estimating sparse precision matrices under the Frobenius norm and discusses connections and differences of our work with other related problems. The proofs are given in Section 7.

2. Methodology. In this section, we introduce an adaptive constrained ℓ_1 minimization procedure, called ACLIME, for estimating a precision matrix Ω. The properties of the estimator are then studied in Section 3 under the matrix ℓ_w norm losses for $1 \le w \le \infty$ and a minimax upper bound is established. The upper bound together with the lower bound given in Section 4 will show that the ACLIME estimator is adaptively rate optimal.

We begin with basic notation and definitions. For a vector $a = (a_1, \dots, a_p)^T \in \mathbb{R}^p$, define $|a|_1 = \sum_{j=1}^p |a_j|$ and $|a|_2 = \sqrt{\sum_{j=1}^p a_j^2}$. For a matrix $A = (a_{ij}) \in \mathbb{R}^{p\times q}$, we define the elementwise ℓ_w norm by $|A|_w = (\sum_{i,j} |a_{ij}|^w)^{1/w}$. The Frobenius norm of A is the elementwise ℓ_2 norm. I denotes a $p \times p$ identity matrix. For any two index sets T and T' and matrix A, we use $A_{TT'}$ to denote the $|T| \times |T'|$ matrix with rows and columns of A indexed by T and T', respectively.

For an i.i.d. random sample $\{X_1, \dots, X_n\}$ of p-variate observations drawn from a population X, let the sample mean $\bar{X} = \frac{1}{n}\sum_{k=1}^n X_k$ and the sample covariance matrix

$$\Sigma^* = (\sigma_{ij}^*)_{1\le i,j\le p} = \frac{1}{n-1}\sum_{l=1}^n (X_l - \bar{X})(X_l - \bar{X})^T, \tag{2.1}$$

which is an unbiased estimate of the covariance matrix $\Sigma = (\sigma_{ij})_{1\le i,j\le p}$.

It is well known that in the high dimensional setting, the inverse of the sample covariance matrix either does not exist or is not a good estimator of Ω. As mentioned in the Introduction, a number of methods for estimating Ω have been introduced in the literature. In particular, Cai, Liu and Luo (2011) proposed an estimator called CLIME by solving the following optimization problem:

$$\min |\Omega|_1 \quad \text{subject to:} \quad |\Sigma^*\Omega - I|_\infty \le \tau_n, \qquad \Omega \in \mathbb{R}^{p\times p}, \tag{2.2}$$

where $\tau_n = CM_{n,p}\sqrt{\log p/n}$ for some constant C. The convex program (2.2) can be further decomposed into p vector-minimization problems. Let e_i be a standard unit vector in $\mathbb{R}^p$ with 1 in the ith coordinate and 0 in all other coordinates. For $1 \le i \le p$, let $\hat{\omega}_i$ be the solution of the following convex optimization problem:

$$\min |\omega|_1 \quad \text{subject to} \quad |\Sigma_n\omega - e_i|_\infty \le \tau_n, \tag{2.3}$$

where ω is a vector in $\mathbb{R}^p$. The final CLIME estimator of Ω is obtained by putting the columns $\hat{\omega}_i$ together and applying an additional symmetrization step. This estimator is easy to implement and possesses a number of desirable properties as shown in Cai, Liu and Luo (2011).

The CLIME estimator has, however, two drawbacks. One is that the estimator is not rate optimal, as will be shown later. Another drawback is that the procedure is not adaptive in the sense that the tuning parameter τ_n is not fully specified and needs to be chosen through an empirical method such as cross-validation.

To overcome these drawbacks of CLIME, we now introduce an adaptive constrained ℓ_1-minimization for inverse matrix estimation (ACLIME). The estimator is fully data-driven and adaptive to the variability of individual entries. A key technical result which provides the motivation for the new procedure is the following fact.

LEMMA 2.1. *Let $X_1, \ldots, X_n \overset{i.i.d.}{\sim} N_p(\mu, \Sigma)$ with $\log p = O(n^{1/3})$. Set $S^* = (s^*_{ij})_{1\le i,j\le p} = \Sigma^*\Omega - I_{p\times p}$, where Σ^* is the sample covariance matrix defined in (2.1). Then*

$$\operatorname{Var}(s^*_{ij}) = \begin{cases} n^{-1}(1+\sigma_{ii}\omega_{ii}), & \text{for } i = j, \\ n^{-1}\sigma_{ii}\omega_{jj}, & \text{for } i \ne j \end{cases}$$

and for all $\delta \ge 2$,

$$\begin{aligned} &\mathbb{P}\left\{\left|(\Sigma^*\Omega - I_{p\times p})_{ij}\right| \le \delta\sqrt{\frac{\sigma_{ii}\omega_{jj}\log p}{n}}, \forall 1 \le i, j \le p\right\} \\ &\quad \ge 1 - O\big((\log p)^{-1/2} p^{-\delta^2/4+1}\big). \end{aligned} \tag{2.4}$$

REMARK 2.1. The condition $\log p = O(n^{1/3})$ in Lemma 2.1 can be relaxed to $\log p = o(n)$. Under the condition $\log p = o(n)$ Theorem 1 in Chapter VIII of Petrov (1975) implies that equation (2.4) still holds by replacing the probability bound $1 - O((\log p)^{-1/2} p^{-\delta^2/4+1})$ with $1 - O((\log p)^{-1/2} p^{-\delta^2/4+1+o(1)})$. We then need the constant $\delta > 2$ so that $(\log p)^{-1/2} p^{-\delta^2/4+1+o(1)} = o(1)$.

A major step in the construction of the adaptive data-driven procedure is to make the constraint in (2.2) and (2.3) adaptive to the variability of individual entries based on Lemma 2.1, instead of using a single upper bound λ_n for all the entries. In order to apply Lemma 2.1, we need to estimate the diagonal elements of Σ and Ω, σ_{ii} and w_{jj}, $i, j = 1, \ldots, p$. Note that σ_{ii} can be easily estimated by the sample variances σ^*_{ii}, but ω_{jj} are harder to estimate. Hereafter, $(A)_{ij}$ denotes the (i, j)th entry of the matrix A, $(a)_j$ denotes the jth element of the vector a. Denote $b_j = (b_{1j}, \ldots, b_{pj})'$.

The ACLIME procedure has two steps: The first step is to estimate ω_{jj} and the second step is to apply a modified version of the CLIME procedure to take into account of the variability of individual entries.

Step 1: *Estimating* ω_{jj}. Note that $\sigma_{ii}\omega_{jj} \le (\sigma_{ii} \vee \sigma_{jj})\omega_{jj}$ and $(\sigma_{ii} \vee \sigma_{jj})\omega_{jj} \ge 1$, which implies $\sqrt{\sigma_{ii}\omega_{jj}} \le (\sigma_{ii} \vee \sigma_{jj})\omega_{jj}$, that is, $2\sqrt{\frac{\sigma_{ii}\omega_{jj}\log p}{n}} \le 2(\sigma_{ii} \vee \sigma_{jj})\omega_{jj}\sqrt{\frac{\log p}{n}}$. From equation (2.4), we consider

$$\left|(\Sigma^*\Omega - I_{p\times p})_{ij}\right| \le 2(\sigma_{ii} \vee \sigma_{jj})\omega_{jj}\sqrt{\frac{\log p}{n}}, \qquad 1 \le i, j \le p. \tag{2.5}$$

Let $\hat\Omega_1 := (\hat\omega^1_{ij}) = (\hat\omega^1_{\cdot 1}, \ldots, \hat\omega^1_{\cdot p})$ be a solution to the following optimization problem:

$$\hat\omega^1_{\cdot j} = \mathop{\arg\min}_{b_j \in R^p}\big\{|b_j|_1 : |\hat\Sigma b_j - e_j|_\infty \le \lambda_n\big(\sigma^*_{ii} \vee \sigma^*_{jj}\big) \times b_{jj}, b_{jj} > 0\big\}, \tag{2.6}$$

where $b_j = (b_{1j}, \ldots, b_{pj})'$, $1 \le j \le p$, $\hat{\Sigma} = \Sigma^* + n^{-1} I_{p\times p}$ and

$$\lambda_n = \delta \sqrt{\frac{\log p}{n}}. \tag{2.7}$$

Here, δ is a constant which can be taken as 2. The estimator $\hat{\Omega}_1$ yields estimates of the conditional variance ω_{jj}, $1 \le j \le p$. More specifically, we define the estimates of ω_{jj} by

$$\breve{\omega}_{jj} = \hat{\omega}^1_{jj} I\left\{\sigma^*_{jj} \le \sqrt{\frac{n}{\log p}}\right\} + \sqrt{\frac{\log p}{n}} I\left\{\sigma^*_{jj} > \sqrt{\frac{n}{\log p}}\right\}.$$

Step 2: *Adaptive estimation*. Given the estimates $\breve{\omega}_{jj}$, the final estimator $\hat{\Omega}$ of Ω is constructed as follows. First, we obtain $\tilde{\Omega}^1 =: (\tilde{\omega}^1_{ij})$ by solving p optimization problems: for $1 \le j \le p$

$$\tilde{\omega}^1_{\cdot j} = \mathop{\arg\min}_{b \in R^p} \left\{ |b|_1 : \left|(\hat{\Sigma} b - e_j)_i\right| \le \lambda_n \sqrt{\sigma^*_{ii} \breve{\omega}_{jj}}, 1 \le i \le p \right\}, \tag{2.8}$$

where λ_n is given in (2.7). We then obtain the estimator $\hat{\Omega}$ by symmetrizing $\tilde{\Omega}^1$,

$$\begin{gathered} \hat{\Omega} = (\hat{\omega}_{ij}) \\ \text{where } \hat{\omega}_{ij} = \hat{\omega}_{ji} = \tilde{\omega}^1_{ij} I\{|\tilde{\omega}^1_{ij}| \le |\tilde{\omega}^1_{ji}|\} + \tilde{\omega}^1_{ji} I\{|\tilde{\omega}^1_{ij}| > |\tilde{\omega}^1_{ji}|\}. \end{gathered} \tag{2.9}$$

We shall call the estimator $\hat{\Omega}$ adaptive CLIME, or ACLIME. The estimator adapts to the variability of individual entries by using an entry-dependent threshold for each individual ω_{ij}. Note that the optimization problem (2.6) is convex and can be cast as a linear program. The constant δ in (2.7) can be taken as 2 and the resulting estimator will be shown to be adaptively minimax rate optimal for estimating sparse precision matrices.

REMARK 2.2. Note that $\delta = 2$ used in the constraint sets is tight, it can not be further reduced in general. If one chooses the constant $\delta < 2$, then with probability tending to 1, the true precision matrix will no longer belong to the feasible sets. To see this, consider $\Sigma = \Omega = I_{p\times p}$ for simplicity. It follows from Liu, Lin and Shao (2008) and Cai and Jiang (2011) that

$$\sqrt{\frac{n}{\log p}} \max_{1 \le i < j \le p} |\hat{\sigma}_{ij}| \to 2$$

in probability. Thus, $P(|\hat{\Sigma}\Omega - I_{p\times p}|_\infty > \lambda_n) \to 1$, which means that if $\delta < 2$, the true Ω lies outside of the feasible set with high probability and solving the corresponding minimization problem cannot lead to a good estimator of Ω.

REMARK 2.3. The CLIME estimator uses a universal tuning parameter $\lambda_n = CM_{n,p}\sqrt{\log p/n}$ which does not take into account the variations in the variances σ_{ii} and the conditional variances ω_{jj}. It will be shown that the convergence rate of CLIME obtained by Cai, Liu and Luo (2011) is not optimal. The quantity $M_{n,p}$ is the upper bound of the matrix ℓ_1 norm which is unknown in practice. The cross validation method can be used to choose the tuning parameter in CLIME. However, the estimator obtained through CV can be variable and its theoretical properties are unclear. In contrast, the ACLIME procedure proposed in the present paper does not depend on any unknown parameters and it will be shown that the estimator is minimax rate optimal.

3. Properties of ACLIME and minimax upper bounds. We now study the properties of the ACLIME estimator $\hat{\Omega}$ proposed in Section 2. We shall begin with the Gaussian case where $X \sim N(\mu, \Sigma)$. Extensions to non-Gaussian distributions will be discussed later. The following result shows that the ACLIME estimator adaptively attains the convergence rate of

$$M_{n,p}^{1-q} c_{n,p} \left(\frac{\log p}{n}\right)^{(1-q)/2}$$

over the class of sparse precision matrices $\mathcal{G}_q(c_{n,p}, M_{n,p})$ defined in (1.1) under the matrix ℓ_w norm losses for all $1 \le w \le \infty$. The lower bound given in Section 4 shows that this rate is indeed optimal and thus ACLIME adapts to both sparsity patterns and this class of loss functions.

THEOREM 3.1. *Suppose we observe a random sample* $X_1, \dots, X_n \overset{i.i.d.}{\sim} N_p(\mu, \Sigma)$. *Let* $\Omega = \Sigma^{-1}$ *be the precision matrix. Let* $\delta \ge 2$, $\log p = O(n^{1/3})$ *and*

$$c_{n,p} = O\bigl(n^{1/2-q/2}/(\log p)^{3/2-q/2}\bigr). \tag{3.1}$$

Then for some constant $C > 0$

$$\inf_{\Omega \in \mathcal{G}_q(c_{n,p}, M_{n,p})} \mathbb{P}\left(\|\hat{\Omega} - \Omega\|_w \le C M_{n,p}^{1-q} c_{n,p} \left(\frac{\log p}{n}\right)^{(1-q)/2}\right) \ge 1 - O\bigl((\log p)^{-1/2} p^{-\delta^2/4+1}\bigr)$$

for all $1 \le w \le \infty$.

For $q = 0$, a sufficient condition for estimating Ω consistently under the spectral norm is

$$M_{n,p} c_{n,p} \sqrt{\frac{n}{\log p}} = o(1), \quad \text{i.e.,} \quad M_{n,p} c_{n,p} = o\left(\sqrt{\frac{n}{\log p}}\right).$$

This implies that the total number of nonzero elements on each column needs be $\ll \sqrt{n}$ in order for the precision matrix to be estimated consistently over

$\mathcal{G}_0(c_{n,p}, M_{n,p})$. In Theorem 4.1 we show that the upper bound $M_{n,p}c_{n,p}\sqrt{\frac{\log p}{n}}$ is indeed rate optimal over $\mathcal{G}_0(c_{n,p}, M_{n,p})$.

REMARK 3.1. Following Remark 2.1, the condition $\log p = O(n^{1/3})$ in Theorem 3.1 can be relaxed to $\log p = o(n)$. In Theorem 3.1, the constant δ then needs to be strictly larger than 2, and the probability bound $1 - O((\log p)^{-1/2} p^{-\delta^2/4+1})$ is replaced by $1 - O((\log p)^{-1/2} p^{-\delta^2/4+1+o(1)})$. By a similar argument, in the following Theorems 3.2 and 3.3, we need only to assume $\log p = o(n)$.

We now consider the rate of convergence under the expectation. For technical reasons, we require the constant $\delta \geq 3$ in this case.

THEOREM 3.2. *Suppose we observe a random sample* $X_1, \dots, X_n \overset{i.i.d.}{\sim} N_p(\mu, \Sigma)$. *Let* $\Omega = \Sigma^{-1}$ *be the precision matrix. Let* $\log p = o(n)$ *and* $\delta \geq 3$. *Suppose that* $p \geq n^{13/(\delta^2-8)}$ *and*

$$c_{n,p} = o\big((n/\log p)^{1/2-q/2}\big).$$

The ACLIME estimator $\hat{\Omega}$ *satisfies, for all* $1 \leq w \leq \infty$ *and* $0 \leq q < 1$,

$$\sup_{\mathcal{G}_q(c_{n,p},M_{n,p})} \mathbb{E}\|\hat{\Omega} - \Omega\|_w^2 \leq C M_{n,p}^{2-2q} c_{n,p}^2 \left(\frac{\log p}{n}\right)^{1-q},$$

for some constant $C > 0$.

Theorem 3.2 can be extended to non-Gaussian distributions.Let $Z = (Z_1, Z_2, \dots, Z_n)'$ be a p-variate random variable with mean μ and covariance matrix $\Sigma = (\sigma_{ij})_{1 \leq i,j \leq p}$. Let $\Omega = (\omega_{ij})_{1 \leq i,j \leq p}$ be the precision matrix. Define $Y_i = (Z_i - \mu_i)/\sigma_{ii}^{1/2}$, $1 \leq i \leq p$ and $(W_1, \dots, W_p)' := \Omega(Z - \mu)$. Assume that there exist some positive constants η and M such that for all $1 \leq i \leq p$,

$$\mathbb{E}\exp\big(\eta Y_i^2\big) \leq M, \qquad \mathbb{E}\exp\big(\eta W_i^2/\omega_{ii}\big) \leq M. \tag{3.2}$$

Then we have the following result.

THEOREM 3.3. *Suppose we observe an i.i.d. sample* $X_1, \dots, X_n$ *with the precision matrix* Ω *satisfying condition* (3.2). *Let* $\log p = o(n)$ *and* $p \geq n^\gamma$ *for some* $\gamma > 0$. *Suppose that*

$$c_{n,p} = o\big((n/\log p)^{1/2-q/2}\big).$$

Then there is a δ *depending only on* η, M *and* γ *such that the ACLIME estimator* $\hat{\Omega}$ *satisfies, for all* $1 \leq w \leq \infty$ *and* $0 \leq q < 1$,

$$\sup_{\mathcal{G}_q(c_{n,p},M_{n,p})} \mathbb{E}\|\hat{\Omega} - \Omega\|_w^2 \leq C M_{n,p}^{2-2q} c_{n,p}^2 \left(\frac{\log p}{n}\right)^{1-q},$$

for some constant $C > 0$.

REMARK 3.2. Under condition (3.2), it can be shown that an analogous result to Lemma 2.1 in Section 2 holds with some δ depending only on η and M. Thus, it can be proved that, under condition (3.2), Theorem 3.3 holds. The proof is similar to that of Theorem 3.2. A practical way to choose δ is using cross validation.

REMARK 3.3. Theorems 3.1, 3.2 and 3.3 follow mainly from the convergence rate under the element-wise ℓ_∞ norm and the inequality $\|M\|_w \le \|M\|_1$ for any symmetric matrix M from Lemma 7.2. The convergence rate under element-wise norm plays an important role in graphical model selection and in establishing the convergence rate under other matrix norms, such as the Frobenius norm $\|\cdot\|_F$. Indeed, from the proof, Theorems 3.1, 3.2 and 3.3 hold under the matrix ℓ_1 norm. More specifically, under the conditions of Theorems 3.2 and 3.3 we have

$$\sup_{\mathcal{G}_q(c_{n,p},M_{n,p})} \mathbb{E}|\hat{\Omega}-\Omega|_\infty^2 \le CM_{n,p}^2 \frac{\log p}{n},$$

$$\sup_{\mathcal{G}_q(c_{n,p},M_{n,p})} \mathbb{E}\|\hat{\Omega}-\Omega\|_1^2 \le CM_{n,p}^{2-2q} c_{n,p}^2 \left(\frac{\log p}{n}\right)^{1-q},$$

$$\sup_{\mathcal{G}_q(c_{n,p},M_{n,p})} \frac{1}{p}\mathbb{E}\|\hat{\Omega}-\Omega\|_F^2 \le CM_{n,p}^{2-q} c_{n,p} \left(\frac{\log p}{n}\right)^{1-q/2}.$$

REMARK 3.4. The results in this section can be easily extended to the weak ℓ_q ball with $0 \le q < 1$ to model the sparsity of the precision matrix Ω. A weak ℓ_q ball of radius c in R^p is defined as follows:

$$B_q(c) = \{\xi \in \mathbb{R}^p : |\xi|_{(k)}^q \le ck^{-1}, \text{for all } k = 1, \dots, p\},$$

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

$$\mathcal{G}_q^*(c_{n,p}, M_{n,p}) = \left\{ \begin{array}{c} c\Omega = (\omega_{ij})_{1\le i,j\le p} : \omega_{\cdot,j} \in B_q(c_{n,p}), \\ \|\Omega\|_1 \le M_{n,p}, \lambda_{\max}(\Omega)/\lambda_{\min}(\Omega) \le M_1, \Omega \succ 0 \end{array} \right\}. \tag{3.3}$$

Theorems 3.1, 3.2 and 3.3 hold with the parameter space $G_q(c_{n,p}, M_{n,p})$ replaced by $G_q^*(c_{n,p}, M_{n,p})$ by a slight extension of Lemma 7.1 for the ℓ_q ball to for the weak ℓ_q ball similar to equation (51) in Cai and Zhou (2012).

4. Minimax lower bounds. Theorem 3.2 shows that the ACLIME estimator adaptively attains the rate of convergence

$$M_{n,p}^{2-2q} c_{n,p}^2 \left(\frac{\log p}{n}\right)^{1-q} \tag{4.1}$$

under the squared matrix ℓ_w norm loss for $1 \le w \le \infty$ over the collection of the parameter spaces $\mathcal{G}_q(c_{n,p}, M_{n,p})$. In this section, we shall show that the rate of convergence given in (4.1) cannot be improved by any other estimator and thus is indeed optimal among all estimators by establishing minimax lower bounds for estimating sparse precision matrices under the squared matrix ℓ_w norm.

THEOREM 4.1. *Let* $X_1, \dots, X_n \overset{i.i.d.}{\sim} N_p(\mu, \Sigma)$ *with* $p > c_1 n^{\beta}$ *for some constants* $\beta > 1$ *and* $c_1 > 0$. *Assume that*

$$cM_{n,p}^q\left(\frac{\log p}{n}\right)^{q/2} \le c_{n,p} = o\bigl(M_{n,p}^q n^{(1-q)/2}(\log p)^{-(3-q)/2}\bigr) \tag{4.2}$$

for some constant $c > 0$. *The minimax risk for estimating the precision matrix* $\Omega = \Sigma^{-1}$ *over the parameter space* $\mathcal{G}_q(c_{n,p}, M_{n,p})$ *under the condition* (4.2) *satisfies*

$$\inf_{\hat{\Omega}} \sup_{\mathcal{G}_q(c_{n,p}, M_{n,p})} \mathbb{E}\|\hat{\Omega} - \Omega\|_w^2 \ge C M_{n,p}^{2-2q} c_{n,p}^2 \left(\frac{\log p}{n}\right)^{1-q}$$

for some constant $C > 0$ *and for all* $1 \le w \le \infty$.

The proof of Theorem 4.1 is involved. We shall discuss the key technical tools and outline the important steps in the proof of Theorem 4.1 in this section. The detailed proof is given in Section 7.

4.1. *A general technical tool.* We use a lower bound technique introduced in Cai and Zhou (2012), which is particularly well suited for treating "two-directional" problems such as matrix estimation. The technique can be viewed as a generalization of both Le Cam's method and Assouad's lemma, two classical lower bound arguments. Let X be an observation from a distribution $\mathbb{P}_\theta$ where θ belongs to a parameter set Θ which has a special tensor structure. For a given positive integer r and a finite set $B \subset \mathbb{R}^p/\{0_{1\times p}\}$, let $\Gamma = \{0, 1\}^r$ and $\Lambda \subseteq B^r$. Define

$$\Theta = \Gamma \otimes \Lambda = \bigl\{(\gamma, \lambda) : \gamma \in \Gamma \text{ and } \lambda \in \Lambda\bigr\}. \tag{4.3}$$

In comparison, the standard lower bound arguments work with either Γ or Λ alone. For example, Assouad's lemma considers only the parameter set Γ and Le Cam's method typically applies to a parameter set like Λ with $r = 1$. Cai and Zhou (2012) gives a lower bound for the maximum risk over the parameter set Θ to the problem of estimating a functional $\psi(\theta)$, belonging to a metric space with metric d.

We need to introduce a few notation before formally stating the lower bound. For two distributions $\mathbb{P}$ and $\mathbb{Q}$ with densities p and q with respect to any common dominating measure μ, the total variation affinity is given by $\|\mathbb{P} \wedge \mathbb{Q}\| = \int p \wedge q\, d\mu$. For a parameter $\gamma = (\gamma_1, \dots, \gamma_r) \in \Gamma$ where $\gamma_i \in \{0, 1\}$, define

$$H(\gamma, \gamma') = \sum_{i=1}^r |\gamma_i - \gamma_i'| \tag{4.4}$$

be the Hamming distance on $\{0, 1\}^r$.

Let $D_\Lambda = \mathrm{Card}(\Lambda)$. For a given $a \in \{0, 1\}$ and $1 \le i \le r$, we define the mixture distribution $\bar{\mathbb{P}}_{a,i}$ by

$$\bar{\mathbb{P}}_{a,i} = \frac{1}{2^{r-1} D_\Lambda} \sum_\theta \{\mathbb{P}_\theta : \gamma_i(\theta) = a\}. \tag{4.5}$$

So $\bar{\mathbb{P}}_{a,i}$ is the mixture distribution over all P_θ with $\gamma_i(\theta)$ fixed to be a while all other components of θ vary over all possible values. In our construction of the parameter set for establishing the minimax lower bound, r is the number of possibly nonzero rows in the upper triangle of the covariance matrix and Λ is the set of matrices with r rows to determine the upper triangle matrix.

LEMMA 4.1. *For any estimator T of $\psi(\theta)$ based on an observation from the experiment $\{\mathbb{P}_\theta, \theta \in \Theta\}$, and any $s > 0$*

$$\max_{\Theta} 2^s \mathbb{E}_\theta d^s\big(T, \psi(\theta)\big) \geq \alpha \frac{r}{2} \min_{1\leq i\leq r} \|\bar{\mathbb{P}}_{0,i} \wedge \bar{\mathbb{P}}_{1,i}\|, \tag{4.6}$$

where $\bar{\mathbb{P}}_{a,i}$ is defined in equation (4.5) and α is given by

$$\alpha = \min_{\{(\theta,\theta'):H(\gamma(\theta),\gamma(\theta'))\geq 1\}} \frac{d^s(\psi(\theta), \psi(\theta'))}{H(\gamma(\theta), \gamma(\theta'))}. \tag{4.7}$$

We introduce some new notation to study the affinity $\|\bar{\mathbb{P}}_{0,i} \wedge \bar{\mathbb{P}}_{1,i}\|$ in equation (4.6). Denote the projection of $\theta \in \Theta$ to Γ by $\gamma(\theta) = (\gamma_i(\theta))_{1\leq i\leq r}$ and to Λ by $\lambda(\theta) = (\lambda_i(\theta))_{1\leq i\leq r}$. More generally we define $\gamma_A(\theta) = (\gamma_i(\theta))_{i\in A}$ for a subset $A \subseteq \{1, 2, \ldots, r\}$, a projection of θ to a subset of Γ. A particularly useful example of set A is

$$\{-i\} = \{1, \ldots, i-1, i+1, \ldots, r\},$$

for which $\gamma_{-i}(\theta) = (\gamma_1(\theta), \ldots, \gamma_{i-1}(\theta), \gamma_{i+1}(\theta), \gamma_r(\theta))$. $\lambda_A(\theta)$ and $\lambda_{-i}(\theta)$ are defined similarly. We denote the set $\{\lambda_A(\theta) : \theta \in \Theta\}$ by Λ_A. For $a \in \{0, 1\}$, $b \in \{0, 1\}^{r-1}$, and $c \in \Lambda_{-i} \subseteq B^{r-1}$, let

$$D_{\Lambda_i(a,b,c)} = \operatorname{Card}\big\{\gamma \in \Lambda : \gamma_i(\theta) = a, \gamma_{-i}(\theta) = b \text{ and } \lambda_{-i}(\theta) = c\big\}$$

and define

$$\bar{\mathbb{P}}_{(a,i,b,c)} = \frac{1}{D_{\Lambda_i(b,c)}} \sum_{\theta} \big\{\mathbb{P}_\theta : \gamma_i(\theta) = a, \gamma_{-i}(\theta) = b \text{ and } \lambda_{-i}(\theta) = c\big\}. \tag{4.8}$$

In other words, $\bar{\mathbb{P}}_{(a,i,b,c)}$ is the mixture distribution over all $\mathbb{P}_\theta$ with $\lambda_i(\theta)$ varying over all possible values while all other components of θ remain fixed.

The following lemma gives a lower bound for the affinity in equation (4.6). See Section 2 of Cai and Zhou (2012) for more details.

LEMMA 4.2. *Let $\bar{\mathbb{P}}_{a,i}$ and $\bar{\mathbb{P}}_{(a,i,b,c)}$ be defined in equations (4.5) and (4.8), respectively, then*

$$\|\bar{\mathbb{P}}_{0,i} \wedge \bar{\mathbb{P}}_{1,i}\| \geq \operatorname*{Average}_{\gamma_{-i},\lambda_{-i}} \big\|\bar{\mathbb{P}}_{(0,i,\gamma_{-i},\lambda_{-i})} \wedge (\bar{\mathbb{P}}_{(1,i,\gamma_{-i},\lambda_{-i})})\big\|,$$

where the average over γ_{-i} and λ_{-i} is induced by the uniform distribution over Θ.

4.2. *Lower bound for estimating sparse precision matrix.* We now apply the lower bound technique developed in Section 4.1 to establish rate sharp results under the matrix ℓ_w norm. Let $X_1, \dots, X_n \overset{\text{i.i.d.}}{\sim} N_p(\mu, \Omega^{-1})$ with $p > c_1 n^\beta$ for some $\beta > 1$ and $c_1 > 0$, where $\Omega \in \mathcal{G}_q(c_{n,p}, M_{n,p})$. The proof of Theorem 4.1 contains four major steps. We first reduce the minimax lower bound under the general matrix ℓ_w norm, $1 \le w \le \infty$, to under the spectral norm. In the second step, we construct in detail a subset $\mathcal{F}_*$ of the parameter space $\mathcal{G}_q(c_{n,p}, M_{n,p})$ such that the difficulty of estimation over $\mathcal{F}_*$ is essentially the same as that of estimation over $\mathcal{G}_q(c_{n,p}, M_{n,p})$, the third step is the application of Lemma 4.1 to the carefully constructed parameter set, and finally in the fourth step we calculate the factors α defined in (4.7) and the total variation affinity between two multivariate normal mixtures. We outline the main ideas of the proof here and leave detailed proof of some technical results to Section 7.

PROOF OF THEOREM 4.1. We shall divide the proof into four major steps.

Step 1: *Reducing the general problem to the lower bound under the spectral norm*. The following lemma implies that the minimax lower bound under the spectral norm yields a lower bound under the general matrix ℓ_w norm up to a constant factor 4.

LEMMA 4.3. *Let $X_1, \dots, X_n \overset{i.i.d.}{\sim} N(\mu, \Omega^{-1})$, and $\mathcal{F}$ be any parameter space of precision matrices. The minimax risk for estimating the precision matrix Ω over $\mathcal{F}$ satisfies*

$$\inf_{\hat\Omega} \sup_{\mathcal{F}} \mathbb{E}\|\hat\Omega - \Omega\|_w^2 \ge \frac{1}{4} \inf_{\hat\Omega} \sup_{\mathcal{F}} \mathbb{E}\|\hat\Omega - \Omega\|_2^2 \tag{4.9}$$

for all $1 \le w \le \infty$.

Step 2: *Constructing the parameter set*. Let $r = \lceil p/2 \rceil$ and let B be the collection of all vectors $(b_j)_{1 \le j \le p}$ such that $b_j = 0$ for $1 \le j \le p - r$ and $b_j = 0$ or 1 for $p - r + 1 \le j \le p$ under the constraint $\|b\|_0 = k$ (to be defined later). For each $b \in B$ and each $1 \le m \le r$, define a $p \times p$ matrix $\lambda_m(b)$ by making the mth row of $\lambda_m(b)$ equal to b and the rest of the entries 0. It is clear that $\mathrm{Card}(B) = \binom{r}{k}$. Set $\Gamma = \{0, 1\}^r$. Note that each component b_i of $\lambda = (b_1, \dots, b_r) \in \Lambda$ can be uniquely associated with a $p \times p$ matrix $\lambda_i(b_i)$. Λ is the set of all matrices λ with the every column sum less than or equal to $2k$. Define $\Theta = \Gamma \otimes \Lambda$ and let $\varepsilon_{n,p} \in \mathbb{R}$ be fixed. (The exact value of $\varepsilon_{n,p}$ will be chosen later.) For each $\theta = (\gamma, \lambda) \in \Theta$ with $\gamma = (\gamma_1, \dots, \gamma_r)$ and $\lambda = (b_1, \dots, b_r)$, we associate θ with a precision matrix $\Omega(\theta)$ by

$$\Omega(\theta) = \frac{M_{n,p}}{2}\left[I_p + \varepsilon_{n,p} \sum_{m=1}^{r} \gamma_m \lambda_m(b_m)\right].$$

Finally, we define a collection $\mathcal{F}_*$ of precision matrices as

$$\mathcal{F}_* = \left\{\Omega(\theta) : \Omega(\theta) = \frac{M_{n,p}}{2}\left[I_p + \varepsilon_{n,p}\sum_{m=1}^{r}\gamma_m\lambda_m(b_m)\right], \theta = (\gamma, \lambda) \in \Theta\right\}.$$

We now specify the values of $\varepsilon_{n,p}$ and k. Set

$$\varepsilon_{n,p} = \upsilon\sqrt{\frac{\log p}{n}} \qquad \text{for some } 0 < \upsilon < \min\left\{\left(\frac{c}{2}\right)^{1/q}, \frac{\beta - 1}{8\beta}\right\}, \tag{4.10}$$

and

$$k = \lceil 2^{-1}c_{n,p}(M_{n,p}\varepsilon_{n,p})^{-q}\rceil - 1, \tag{4.11}$$

which is at least 1 from equation (4.10). Now we show $\mathcal{F}_*$ is a subset of the parameter space $\mathcal{G}_q(c_{n,p}, M_{n,p})$. From the definition of k in (4.11) note that

$$\max_{j\le p}\sum_{i\ne j}|\omega_{ij}|^q \le 2\cdot 2^{-1}\rho_{n,p}(M_{n,p}\varepsilon_{n,p})^{-q}\cdot\left(\frac{M_{n,p}}{2}\varepsilon_{n,p}\right)^q \le c_{n,p}. \tag{4.12}$$

From equation (4.2), we have $c_{n,p} = o(M_{n,p}^q n^{(1-q)/2}(\log p)^{-(3-q)/2})$, which implies

$$2k\varepsilon_{n,p} \le c_{n,p}\varepsilon_{n,p}^{1-q}M_{n,p}^{-q} = o(1/\log p), \tag{4.13}$$

then

$$\max_i\sum_j|\omega_{ij}| \le \frac{M_{n,p}}{2}(1 + 2k\varepsilon_{n,p}) \le M_{n,p}. \tag{4.14}$$

Since $\|A\|_2 \le \|A\|_1$, we have

$$\left\|\varepsilon_{n,p}\sum_{m=1}^{r}\gamma_m\lambda_m(b_m)\right\|_2 \le \left\|\varepsilon_{n,p}\sum_{m=1}^{r}\gamma_m\lambda_m(b_m)\right\|_1 \le 2k\varepsilon_{n,p} = o(1),$$

which implies that every $\Omega(\theta)$ is diagonally dominant and positive definite, and

$$\lambda_{\max}(\Omega) \le \frac{M_{n,p}}{2}(1 + 2k\varepsilon_{n,p}) \quad \text{and} \quad \lambda_{\min}(\Omega) \ge \frac{M_{n,p}}{2}(1 - 2k\varepsilon_{n,p}) \tag{4.15}$$

which immediately implies

$$\frac{\lambda_{\max}(\Omega)}{\lambda_{\min}(\Omega)} < M_1. \tag{4.16}$$

Equations (4.12), (4.14), (4.15) and (4.16) all together imply $\mathcal{F}_* \subset \mathcal{G}_q(c_{n,p}, M_{n,p})$.

Step 3: *Applying the general lower bound argument*. Let $X_1, \dots, X_n \overset{\text{i.i.d.}}{\sim} N_p(0, (\Omega(\theta))^{-1})$ with $\theta \in \Theta$ and denote the joint distribution by P_θ. Applying Lemmas 4.1 and 4.2 to the parameter space Θ, we have

$$
\begin{aligned}
&\inf_{\hat{\Omega}} \max_{\theta\in\Theta} 2^2 E_\theta \|\hat{\Omega} - \Omega(\theta)\|_2^2 \\
&\quad \geq \alpha \cdot \frac{p}{4} \cdot \min_i \operatorname*{Average}_{\gamma_{-i},\lambda_{-i}} \|\bar{\mathbb{P}}_{(0,i,\gamma_{-i},\lambda_{-i})} \wedge \bar{\mathbb{P}}_{(1,i,\gamma_{-i},\lambda_{-i})}\|,
\end{aligned} \tag{4.17}
$$

where

$$
\alpha = \min_{\{(\theta,\theta'):H(\gamma(\theta),\gamma(\theta'))\geq 1\}} \frac{\|\Omega(\theta) - \Omega(\theta')\|_2^2}{H(\gamma(\theta), \gamma(\theta'))} \tag{4.18}
$$

and $\bar{\mathbb{P}}_{0,i}$ and $\bar{\mathbb{P}}_{1,i}$ are defined as in (4.5).

Step 4: *Bounding the per comparison loss* α *defined in* (4.18) *and the affinity* $\min_i \operatorname{Average}_{\gamma_{-i},\lambda_{-i}} \|\bar{\mathbb{P}}_{(0,i,\gamma_{-i},\lambda_{-i})} \wedge \bar{\mathbb{P}}_{(1,i,\gamma_{-i},\lambda_{-i})}\|$ in (4.17). This is done separately in the next two lemmas which are proved in detailed in Section 7.

LEMMA 4.4. *The per comparison loss* α *defined in* (4.18) *satisfies*

$$
\alpha \geq \frac{(M_{n,p}k\varepsilon_{n,p})^2}{4p}.
$$

LEMMA 4.5. *Let* $X_1, \dots, X_n \overset{i.i.d.}{\sim} N(0, (\Omega(\theta))^{-1})$ *with* $\theta \in \Theta$ *and denote the joint distribution by* $\mathbb{P}_\theta$. *For* $a \in \{0, 1\}$ *and* $1 \leq i \leq r$, *define* $\bar{\mathbb{P}}_{(a,i,b,c)}$ *as in* (4.8). *Then there exists a constant* $c_1 > 0$ *such that*

$$
\min_i \operatorname*{Average}_{\gamma_{-i},\lambda_{-i}} \|\bar{\mathbb{P}}_{(0,i,\gamma_{-i},\lambda_{-i})} \wedge \bar{\mathbb{P}}_{(1,i,\gamma_{-i},\lambda_{-i})}\| \geq c_1.
$$

Finally, the minimax lower bound for estimating a sparse precision matrix over the collection $\mathcal{G}_q(c_{n,p}, M_{n,p})$ is obtained by putting together (4.17) and Lemmas 4.4 and 4.5,

$$
\begin{aligned}
&\inf_{\hat{\Omega}} \sup_{\mathcal{G}_q(c_{n,p},M_{n,p})} \mathbb{E}\|\hat{\Omega} - \Omega(\theta)\|_2^2 \\
&\quad \geq \max_{\Omega(\theta)\in\mathcal{F}_*} E_\theta \|\hat{\Omega} - \Omega(\theta)\|_2^2 \geq \frac{(M_{n,p}k\varepsilon_{n,p})^2}{4p} \cdot \frac{p}{16} \cdot c_1 \\
&\quad \geq \frac{c_1}{64}(M_{n,p}k\varepsilon_{n,p})^2 = c_2 M_{n,p}^{2-2q} c_{n,p}^2 \left(\frac{\log p}{n}\right)^{1-q},
\end{aligned}
$$

for some constant $c_2 > 0$. $\square$

Putting together the minimax upper and lower bounds in Theorems 3.2 and 4.1 as well as Remark 3.4 yields the optimal rates of convergence for estimating Ω over the collection of the ℓ_q balls $\mathcal{G}_q(c_{n,p}, M_{n,p})$ defined in (1.1) as well as the collection of the weak ℓ_q balls $\mathcal{G}_q^*(c_{n,p}, M_{n,p})$ defined in (3.3).

THEOREM 4.2. *Suppose we observe a random sample* $X_i \overset{i.i.d.}{\sim} N_p(\mu, \Sigma)$, $i = 1, 2, \dots, n$. *Let* $\Omega = \Sigma^{-1}$ *be the precision matrix. Assume that* $\log p = O(n^{1/3})$ *and*

$$cM_{n,p}^q \left(\frac{\log p}{n}\right)^{q/2} \le c_{n,p} = o\bigl(M_{n,p}^q n^{(1-q)/2} (\log p)^{-(3-q)/2}\bigr) \tag{4.19}$$

for some constant $c > 0$. *Then*

$$\inf_{\hat{\Omega}} \sup_{\Omega \in \mathcal{G}} \mathbb{E}\|\hat{\Omega} - \Omega\|_w^2 \asymp M_{n,p}^{2-2q} c_{n,p}^2 \left(\frac{\log p}{n}\right)^{1-q} \tag{4.20}$$

for all $1 \le w \le \infty$, *where* $\mathcal{G} = \mathcal{G}_q(c_{n,p}, M_{n,p})$ *or* $\mathcal{G}_q^*(c_{n,p}, M_{n,p})$.

5. Numerical results. In this section, we consider the numerical performance of ACLIME. In particular, we shall compare the performance of ACLIME with that of CLIME. The following three graphical models are considered. Let $D = \mathrm{diag}(U_1, \dots, U_p)$, where U_i, $1 \le i \le p$, are i.i.d. uniform random variables on the interval $(1, 5)$. Let $\Sigma = \Omega^{-1} = D^{1/2}\Omega_1^{-1}D^{1/2}$. The matrix D makes the diagonal entries in Σ and Ω different.

- *Band graph*. Let $\Omega_1 = (\omega_{ij})$, where $\omega_{ii} = 1$, $\omega_{i,i+1} = \omega_{i+1,i} = 0.6$, $\omega_{i,i+2} = \omega_{i+2,i} = 0.3$, $\omega_{ij} = 0$ for $|i - j| \ge 3$.
- AR(1) *model*. Let $\Omega_1 = (\omega_{ij})$, where $\omega_{ij} = (0.6)^{|j-i|}$.
- *Erdős–Rényi random graph*. Let $\Omega_2 = (\omega_{ij})$, where $\omega_{ij} = u_{ij} * \delta_{ij}$, δ_{ij} is the Bernoulli random variable with success probability 0.05 and u_{ij} is uniform random variable with distribution $U(0.4, 0.8)$. We let $\Omega_1 = \Omega_2 + (|\min(\lambda_{\min})| + 0.05)I_p$. It is easy to check that the matrix Ω_1 is symmetric and positive definite.

We generate $n = 200$ random training samples from $N_p(0, \Sigma)$ distribution for $p = 50, 100, 200, 300, 400$. For ACLIME, we set $\delta = 2$ in Step 1 and choose δ in Step 2 by a cross validation method. To this end, we generate an additional 200 testing samples. The tuning parameter in CLIME is selected by cross validation. Note that ACLIME chooses different tuning parameters for different columns and CLIME chooses a universal tuning parameter. The log-likehood loss

$$L(\hat{\Sigma}_1, \Omega) = \log\bigl(\det(\Omega)\bigr) - \langle \hat{\Sigma}_1, \Omega \rangle,$$

where $\hat{\Sigma}_1$ is the sample covariance matrix of the testing samples, is used in the cross validation method. For δ in (2.7), we let $\delta = \delta_j = j/50$, $1 \le j \le 100$. For each δ_j, ACLIME $\hat{\Omega}(\delta_j)$ is obtained and the tuning parameter δ in (2.7) is selected by minimizing the following log-likehood loss:

$$\hat{\delta} = \hat{j}/50 \qquad \text{where } \hat{j} = \mathop{\arg\min}_{1 \le j \le 100} L\big(\hat{\Sigma}_1, \hat{\Omega}(\delta_j)\big).$$

The tuning parameter λ_n in CLIME is also selected by cross validation. The detailed steps can be found in Cai, Liu and Luo (2011).

The empirical errors of ACLIME and CLIME estimators under various settings are summarized in Table 1 below. Three losses under the spectral norm, matrix ℓ_1 norm and Frobenius norm are given to compare the performance between ACLIME and CLIME. As can be seen from Table 1, ACLIME, which is tuning-free, outperforms CLIME in most of the cases for each of the three graphs.

6. Discussions. We established in this paper the optimal rates of convergence and introduced an adaptive method for estimating sparse precision matrices under the matrix ℓ_w norm losses for $1 \le w \le \infty$. The minimax rate of convergence under the Frobenius norm loss can also be easily established. As seen in the proof of Theorems 3.1 and 3.2, with probability tending to one,

$$|\hat{\Omega} - \Omega|_\infty \le C M_{n,p} \sqrt{\frac{\log p}{n}}, \tag{6.1}$$

for some constant $C > 0$. From equation (6.1) one can immediately obtain the following risk upper bound under the Frobenius norm, which can be shown to be rate optimal using a similar proof to that of Theorem 4.1.

THEOREM 6.1. *Suppose we observe a random sample $X_i \overset{i.i.d.}{\sim} N_p(\mu, \Sigma)$, $i = 1, 2, \dots, n$. Let $\Omega = \Sigma^{-1}$ be the precision matrix. Under the assumption (4.19), the minimax risk of estimating the precision matrix Ω over the class $\mathcal{G}_q(c_{n,p}, M_{n,p})$ defined in (1.1) satisfies*

$$\inf_{\hat{\Omega}} \sup_{\mathcal{G}_q(c_{n,p}, M_{n,p})} \mathbb{E}\frac{1}{p}\|\hat{\Omega} - \Omega\|_F^2 \asymp M_{n,p}^{2-q} c_{n,p} \left(\frac{\log p}{n}\right)^{1-q/2}.$$

As shown in Theorem 4.2, the optimal rate of convergence for estimating sparse precision matrices under the squared ℓ_w norm loss is $M_{n,p}^{2-2q} c_{n,p}^2 (\frac{\log p}{n})^{1-q}$. It is interesting to compare this with the minimax rate of convergence for estimating sparse covariance matrices under the same loss which is $c_{n,p}^2 (\frac{\log p}{n})^{1-q}$ [cf. Theorem 1 in Cai and Zhou (2012)]. These two convergence rates are similar, but have an important distinction. The difficulty of estimating a sparse covariance matrix does not depend on the ℓ_1 norm bound $M_{n,p}$, while the difficulty of estimating a sparse precision matrix does.

TABLE 1

Comparisons of ACLIME and CLIME for the three graphical models under three matrix norm losses. Inside the parentheses are the standard deviations of the empirical errors over 100 *replications*

	ACLIME					**CLIME**				
p	**50**	**100**	**200**	**300**	**400**	**50**	**100**	**200**	**300**	**400**
					Spectral norm					
Band	0.30 (0.01)	0.45 (0.01)	0.65 (0.01)	0.85 (0.01)	0.91 (0.01)	0.32 (0.01)	0.50 (0.01)	0.72 (0.01)	1.06 (0.01)	1.08 (0.01)
AR(1)	0.75 (0.01)	1.04 (0.01)	1.25 (0.01)	1.35 (0.01)	1.37 (0.01)	0.73 (0.01)	1.05 (0.01)	1.30 (0.01)	1.41 (0.01)	1.51 (0.01)
E-R	0.65 (0.03)	0.95 (0.02)	2.62 (0.02)	4.28 (0.02)	5.69 (0.02)	0.72 (0.03)	1.21 (0.04)	2.28 (0.02)	3.99 (0.03)	6.00 (0.01)
					Matrix ℓ_1 norm					
Band	0.62 (0.02)	0.79 (0.01)	0.94 (0.01)	1.14 (0.01)	1.21 (0.01)	0.65 (0.02)	0.86 (0.02)	0.99 (0.01)	1.35 (0.01)	1.37 (0.01)
AR(1)	1.19 (0.02)	1.62 (0.02)	1.93 (0.01)	1.99 (0.01)	2.04 (0.01)	1.17 (0.01)	1.59 (0.01)	1.89 (0.01)	2.03 (0.01)	2.16 (0.01)
E-R	1.47 (0.08)	2.15 (0.06)	5.47 (0.05)	10.58 (0.09)	11.38 (0.16)	1.53 (0.06)	2.34 (0.06)	5.20 (0.04)	9.85 (0.10)	11.77 (0.03)
					Frobenius norm					
Band	0.80 (0.01)	1.61 (0.02)	3.11 (0.02)	5.10 (0.03)	6.28 (0.01)	0.83 (0.01)	1.73 (0.02)	3.29 (0.03)	5.90 (0.01)	6.93 (0.01)
AR(1)	1.47 (0.02)	2.73 (0.01)	4.72 (0.01)	6.23 (0.01)	7.44 (0.01)	1.47 (0.02)	2.82 (0.02)	4.97 (0.01)	6.64 (0.01)	8.30 (0.01)
E-R	1.53 (0.05)	3.15 (0.03)	9.89 (0.07)	18.53 (0.12)	26.10 (0.26)	1.62 (0.04)	3.61 (0.05)	8.86 (0.04)	17.14 (0.11)	27.67 (0.02)

As mentioned in the Introduction, an important related problem to the estimation of precision matrix is the recovery of a Gaussian graph which is equivalent to the estimation of the support of Ω. Let $G = (V, E)$ be an undirected graph representing the conditional independence relations between the components of a random vector X. The vertex set V contains the components of X, $V = X = \{V_1, \ldots, V_p\}$. The edge set E consists of ordered pairs (i, j), indicating conditional dependence between the components V_i and V_j. An edge between V_i and V_j is in the set E, that is, $(i, j) \in E$, if and only $\omega_{ij} = 0$. The adaptive CLIME estimator, with an additional thresholding step, can recover the support of Ω. Define the estimator of the support of Ω by

$$\widehat{\mathrm{SUPP}(\Omega)} = \{(i, j) : |\hat{\omega}_{ij}| \geq \tau_{ij}\},$$

where the choice of τ_{ij} depends on the bound $|\hat{\omega}_{ij} - \omega_{ij}|$. Equation (6.1) implies that the right threshold levels are $\tau_{ij} = CM_{n,p}\sqrt{\log p/n}$. If the magnitudes of the nonzero entries exceed $2CM_{n,p}\sqrt{\log p/n}$, then $\widehat{\mathrm{SUPP}(\Omega)}$ recovers the support of Ω exactly. In the context of covariance matrix estimation, Cai and Liu (2011) introduced an adaptive entry-dependent thresholding procedure to recover the support of Σ. That method is based on the sharp bound

$$\max_{1 \leq i \leq j \leq p} |\hat{\sigma}_{ij} - \sigma_{ij}| \leq 2\sqrt{\hat{\theta}_{ij} \log p/n},$$

where $\hat{\theta}_{ij}$ is an estimator of $\mathrm{Var}((X_i - \mu_i)(X_j - \mu_j))$. It is natural to ask whether one can use data and entry-dependent threshold levels τ_{ij} to recover the support of Ω. It is clearly that the optimal choice of τ_{ij} depends on the sharp bounds for $|\hat{\omega}_{ij} - \omega_{ij}|$ which are much more difficult to establish than in the covariance matrix case.

Several recent papers considered the estimation of nonparanormal graphical models where the population distribution is non-Gaussian; see Xue and Zou (2012) and Liu et al. (2012). The nonparanormal model assumes that the variables follow a joint normal distribution after a set of unknown marginal monotone transformations. Xue and Zou (2012) estimated the nonparanormal model by applying CLIME (and graphical lasso, neighborhood Dantzig selector) to the adjusted Spearman's rank correlations. ACLIME can also be used in such a setting. It would be interesting to investigate the properties of the resulting estimator under the nonparanormal model. Detailed analysis is involved and we leave this as future work.

7. Proofs. In this section, we prove the main results, Theorems 3.1 and 3.2, and the key technical results, Lemmas 4.3, 4.4 and 4.5, used in the proof of Theorem 4.1. The proof of Lemma 4.5 is involved. We begin by proving Lemma 2.1 stated in Section 2 and collecting a few additional technical lemmas that will be used in the proofs of the main results.

7.1. *Proof of Lemma 2.1 and additional technical lemmas.*

PROOF OF LEMMA 2.1. Let $\tilde{\Sigma} = (\tilde{\sigma}_{ij}) = n^{-1} \sum_{k=1}^{n-1} X_k X_k'$. Note that Σ^* has the same distribution as that of $\tilde{\Sigma}$ with $X_k \sim N(0, \Sigma)$. So we can replace Σ^* in Section 2 by $\tilde{\Sigma}_n = \tilde{\Sigma} + n^{-1} I_{p\times p}$ and assume $X_k \sim N(0, \Sigma)$. Let $A_n = 1 - O((\log p)^{-1/2} p^{-\delta^2/4+1})$ and set $\lambda_n = \delta\sqrt{\log p/n} + O((n\log p)^{-1/2})$. It suffices to prove that with probability greater than A_n,

$$
\begin{aligned}
\left|\sum_{k=1}^{n-1} X_{ki} X_k' \omega_{\cdot j}\right| &\le n\lambda_n \sqrt{\sigma_{ii}\omega_{jj}} \qquad \text{for } i \ne j,\\
\left|\sum_{k=1}^{n-1} X_{kj} X_k' \omega_{\cdot j} - n\right| &\le n\lambda_n \sqrt{\sigma_{jj}\omega_{jj}} - 1 \qquad \text{for } i = j.
\end{aligned}
\tag{7.1}
$$

Note that $\operatorname{Cov}(X_k'\Omega) = \Omega$, $\operatorname{Var}(X_k'\omega_{\cdot j}) = \omega_{jj}$ and $\operatorname{Cov}(X_{ki} X_k'\omega_{\cdot j}) = \sum_{k=1}^{p} \sigma_{ik} \times \omega_{kj} = 0$ for $i \ne j$. So X_{ki} and $X_k'\omega_{\cdot j}$ are independent. Hence, $\mathbb{E}(X_{ki} X_k'\omega_{\cdot j})^3 = 0$. By Theorem 5.23 and (5.77) in Petrov (1995), we have

$$
\begin{aligned}
&\mathbb{P}\left(\left|\sum_{k=1}^{n-1} X_{ki} X_k'\omega_{\cdot j}\right| \ge n\lambda_n\sqrt{\sigma_{ii}\omega_{jj}}\right)\\
&\quad = \big(1+o(1)\big)\mathbb{P}\big(|N(0,1)| \ge \delta\sqrt{\log p}\big) \le C(\log p)^{-1/2} p^{-\delta^2/2}.
\end{aligned}
\tag{7.2}
$$

We next prove the second inequality in (7.1). We have $\mathbb{E}(X_{kj} X_k'\omega_{\cdot j}) = 1$ and $\operatorname{Var}(X_{kj} X_k'\omega_{\cdot j}) = \sigma_{jj}\omega_{jj} + 1$. Since $1 = \mathbb{E}(X_{kj} X_k'\omega_{\cdot j}) \le \mathbb{E}^{1/2}(X_{kj} X_k'\omega_{\cdot j})^2 \le \sigma_{jj}^{1/2}\omega_{jj}^{1/2}$, we have $\sigma_{jj}\omega_{jj} \ge 1$. Note that $\mathbb{E}\exp(t_0(X_{kj} X_k'\omega_{\cdot j})^2/(1+\sigma_{jj}\omega_{jj}) \le c_0$ for some absolute constants t_0 and c_0. By Theorem 5.23 in Petrov (1995),

$$
\begin{aligned}
&\mathbb{P}\left(\left|\sum_{k=1}^{n-1} X_{kj} X_k'\omega_{\cdot j} - n + 1\right| \ge \big(\delta + O\big((\log p)^{-1}\big)\big)\sqrt{\sigma_{jj}\omega_{jj}\log p}\right)\\
&\quad \le \mathbb{P}\left(\left|\sum_{k=1}^{n-1} X_{kj} X_k'\omega_{\cdot j} - n + 1\right| \ge \frac{(\delta + O((\log p)^{-1}))}{\sqrt{2}}\sqrt{(\sigma_{jj}\omega_{jj}+1)\log p}\right)\\
&\quad \le C(\log p)^{-1/2} p^{-\delta^2/4}.
\end{aligned}
\tag{7.3}
$$

This, together with (7.2), yields (7.1). □

LEMMA 7.1. *Let $\hat{\Omega}$ be any estimator of Ω and set $t_n = |\hat{\Omega} - \Omega|_\infty$. Then on the event*

$$\{|\hat{\omega}_{\cdot j}|_1 \le |\omega_{\cdot j}|, \textit{ for } 1 \le j \le p\},$$

we have

$$\|\hat{\Omega} - \Omega\|_1 \le 12 c_{n,p} t_n^{1-q}. \tag{7.4}$$

PROOF. Define

$$h_j = \hat{\omega}_{\cdot j} - \omega_{\cdot j},\ h_j^1 = \left(\hat{\omega}_{ij} I\{|\hat{\omega}_{ij}| \geq 2t_n\}; 1 \leq i \leq p\right)^T - \omega_j, \qquad h_j^2 = h_j - h_j^1.$$

Then

$$|\omega_{\cdot j}|_1 - |h_j^1|_1 + |h_j^2|_1 \leq |\omega_{\cdot j} + h_j^1|_1 + |h_j^2|_1 = |\hat{\omega}_{\cdot j}|_1 \leq |\omega_{\cdot j}|_1,$$

which implies that $|h_j^2|_1 \leq |h_j^1|_1$. This follows that $|h_j|_1 \leq 2|h_j^1|_1$. So we only need to upper bound $|h_j^1|_1$. We have

$$\begin{aligned} |h_j^1|_1 &\leq \sum_{i=1}^p |\hat{\omega}_{ij} - \omega_{ij}| I\{|\hat{\omega}_{ij}| \geq 2t_n\} + \sum_{i=1}^p |\omega_{ij}| I\{|\hat{\omega}_{ij}| < 2t_n\} \\ &\leq \sum_{i=1}^p t_n I\{|\omega_{ij}| \geq t_n\} + \sum_{i=1}^p |\omega_{ij}| I\{|\omega_{ij}| < 3t_n\} \leq 4c_{n,p} t_n^{1-q}. \end{aligned}$$

So (7.4) holds. □

The following lemma is a classical result. It implies that, if we only consider estimators of symmetric matrices, an upper bound under the matrix ℓ_1 norm is an upper bound for the general matrix ℓ_w norm for all $1 \leq w \leq \infty$, and a lower bound under the matrix ℓ_2 norm is also a lower bound for the general matrix ℓ_w norm. We give a proof to this lemma to be self-contained.

LEMMA 7.2. *Let A be a symmetric matrix, then*

$$\|A\|_2 \leq \|A\|_w \leq \|A\|_1$$

for all $1 \leq w \leq \infty$.

PROOF. The Riesz–Thorin interpolation theorem [see, e.g., Thorin (1948)] implies

$$\|A\|_w \leq \max\{\|A\|_{w_1}, \|A\|_{w_2}\} \qquad \text{for all } 1 \leq w_1 \leq w \leq w_2 \leq \infty. \tag{7.5}$$

Set $w_1 = 1$ and $w_2 = \infty$, then equation (7.5) yields $\|A\|_w \leq \max\{\|A\|_1, \|A\|_\infty\}$ for all $1 \leq w \leq \infty$. When A is symmetric, we know $\|A\|_1 = \|A\|_\infty$, then immediately we have $\|A\|_w \leq \|A\|_1$. Since 2 is sandwiched between w and $\frac{w}{w-1}$, and $\|A\|_w = \|A\|_{w/w-1}$ by duality, from equation (7.5) we have $\|A\|_2 \leq \|A\|_w$ for all $1 \leq w \leq \infty$ when A symmetric. □

7.2. *Proof of Theorems* 3.1 *and* 3.2. We first prove Theorem 3.1. From Lemma 7.2, it is enough to consider the $w=1$ case. Because $\lambda_{\max}(\Omega)/\lambda_{\min}(\Omega) \le M_1$, we have $\max_j \sigma_{jj} \max_j \omega_{jj} \le M_1$. By Lemma 2.1, we have with probability greater than A_n,

$$\begin{aligned}
|\hat{\Omega}_1 - \Omega|_\infty &= \bigl|(\Omega\hat{\Sigma} - I_{p\times p})\hat{\Omega}_1 + \Omega(I_{p\times p} - \hat{\Sigma}\hat{\Omega}_1)\bigr|_\infty \\
&\le C\|\hat{\Omega}_1\|_1\sqrt{\frac{\log p}{n}} + 2\|\Omega\|_1 \max_j \sigma_{jj} \max_j \hat{\omega}^1_{jj}\sqrt{\frac{\log p}{n}} \\
&\le C\|\hat{\Omega}_1\|_1\sqrt{\frac{\log p}{n}} + 2\|\Omega\|_1 \max_j \sigma_{jj} \max_j \omega_{jj} \max_j \frac{\hat{\omega}^1_{jj}}{\omega_{jj}}\sqrt{\frac{\log p}{n}} \\
&\le C\|\hat{\Omega}_1\|_1\sqrt{\frac{\log p}{n}} + 2M_1\|\Omega\|_1 \max_j \frac{\hat{\omega}^1_{jj}}{\omega_{jj}}\sqrt{\frac{\log p}{n}}.
\end{aligned} \tag{7.6}$$

We first assume that $\max_i \omega_{ii} > 0.5\sqrt{\log p/n}$. Then we have $\max_i \omega_{ii}^{-q} \le 2(n/\log p)^{q/2}$. Note that $\|\hat{\Omega}_1\|_1 \le \|\Omega\|_1 \le c_{n,q} \max_i \omega_{ii}^{1-q}$ and $\max_i \omega_{ii}/\min_i \omega_{ii} \le \lambda_{\max}(\Omega)/\lambda_{\min}(\Omega) \le M_1$. By the above inequality,

$$\max_i\left|\frac{\hat{\omega}^1_{ii}}{\omega_{ii}} - 1\right| \le Cc_{n,p}\max_i \omega_{ii}^{-q}\sqrt{\frac{\log p}{n}} + 3Mc_{n,p}\max_i \omega_{ii}^{-q}\max_j \frac{\hat{\omega}^1_{jj}}{\omega_{jj}}\sqrt{\frac{\log p}{n}}$$

with probability greater than A_n. Thus, by the conditions in Theorems 3.2 and 3.1, we have

$$\max_i\left|\frac{\hat{\omega}^1_{ii}}{\omega_{ii}} - 1\right| = \begin{cases} o(1), & \text{under conditions of Theorem 3.2,} \\ O\bigl(1/(\log p)\bigr), & \text{under conditions of Theorem 3.1,} \end{cases}$$

with probability greater than A_n. By (7.1), we can see that, under conditions of Theorem 3.1, Ω belongs to the feasible set in (2.8) with probability greater than A_n. Under conditions of Theorem 3.2, Ω belongs to the feasible set in (2.8) with probability greater than $1 - O((\log p)^{-1/2}p^{-\delta^2/4+1+o(1)})$. So by a similar argument as in (7.6), we can get $|\tilde{\Omega}^1 - \Omega|_\infty \le CM_{n,p}\sqrt{\frac{\log p}{n}}$ and

$$|\hat{\Omega} - \Omega|_\infty \le CM_{n,p}\sqrt{\frac{\log p}{n}}.$$

By Lemma 7.1, we see that

$$\|\hat{\Omega} - \Omega\|_1 \le CM_{n,p}^{1-q}c_{n,p}(\log p/n)^{(1-q)/2}.$$

We consider the case $\max_i \omega_{ii} \le 0.5\sqrt{\log p/n}$. Under this setting, we have $\min_{1\le i\le j}\sigma^*_{ii} > \sqrt{n/\log p}$ with probability greater than A_n. Hence, $\check{\omega}_{ii} = \sqrt{\log p/n} \ge \omega_{ii}$ and Ω belongs to the feasible set in (2.8) with probability greater than A_n. So $\|\hat{\Omega}\|_1 \le \|\Omega\|_1 \le Cc_{n,p}(\log p/n)^{(1-q)/2}$. This proves Theorem 3.1.

To prove Theorem 3.2, note that $\|\hat{\Omega}\|_1 \le \|\tilde{\Omega}^1\|_1 \le \|\hat{\Sigma}^{-1}\|_1 \le np^{1/2}$. We have

$$\begin{aligned}\mathbb{E}\|\hat{\Omega}-\Omega\|^2 &\le CM_{n,p}^{2-2q}c_{n,p}^2\left(\frac{\log p}{n}\right)^{1-q}\\ &\quad + C\left(n^2p + M_{n,p}^{2-2q}c_{n,p}^2\right)p^{-\delta^2/4+1+o(1)}(\log p)^{-1/2}\\ &\le CM_{n,p}^{2-2q}c_{n,p}^2\left(\frac{\log p}{n}\right)^{1-q}.\end{aligned}$$

This proves Theorem 3.2.

7.3. *Proof of Lemma* 4.3. We first show that the minimax lower bound over all possible estimators is at the same order of the minimax lower over only estimators of symmetric matrices under each matrix ℓ_w norm. For each estimator $\hat{\Omega}$, we define a projection of $\hat{\Omega}$ to the parameter space $\mathcal{F}$,

$$\hat{\Omega}_{\text{project}} = \arg\min_{\Omega\in\mathcal{F}} \|\hat{\Omega}-\Omega\|_w,$$

which is symmetric, then

$$\begin{aligned}\sup_{\mathcal{F}} \mathbb{E}\|\hat{\Omega}_{\text{project}}-\Omega\|_w^2 &\le \sup_{\mathcal{F}} \mathbb{E}\left[\|\hat{\Omega}-\hat{\Omega}_{\text{project}}\|_w + \|\hat{\Omega}-\Omega\|_w\right]^2\\ &\le \sup_{\mathcal{F}} \mathbb{E}\left[\|\hat{\Omega}-\Omega\|_w + \|\hat{\Omega}-\Omega\|_w\right]^2 \qquad (7.7)\\ &= 4\sup_{\mathcal{F}} \mathbb{E}\|\hat{\Omega}-\Omega\|_w^2,\end{aligned}$$

where the first inequality follows from the triangle inequality and the second one follows from the definition of $\hat{\Omega}_{\text{project}}$. Since equation (7.7) holds for every $\hat{\Omega}$, we have

$$\inf_{\hat{\Omega},\text{symmetric}} \sup_{\mathcal{F}} \mathbb{E}\|\hat{\Omega}-\Omega\|_w^2 \le 4\inf_{\hat{\Omega}}\sup_{\mathcal{F}} \mathbb{E}\|\hat{\Omega}-\Omega\|_w^2.$$

From Lemma 7.2, we have

$$\inf_{\hat{\Omega},\text{symmetric}} \sup_{\mathcal{F}} \mathbb{E}\|\hat{\Omega}-\Omega\|_w^2 \ge \inf_{\hat{\Omega},\text{symmetric}} \sup_{\mathcal{F}} \mathbb{E}\|\hat{\Omega}-\Omega\|_2^2 \ge \inf_{\hat{\Omega}}\sup_{\mathcal{F}} \mathbb{E}\|\hat{\Omega}-\Omega\|_2^2,$$

which, together with equation (7.7), establishes Lemma 4.3.

7.4. *Proof of Lemma* 4.4. Let $v = (v_i)$ be a column vector with length p, and

$$v_i = \begin{cases} 1, & p - \lceil p/2\rceil + 1 \le i \le p,\\ 0, & \text{otherwise}\end{cases}$$

that is, $v = (1\{p - \lceil p/2\rceil + 1 \le i \le p\})_{p\times 1}$. Set

$$w = (w_i) = \left[\Omega(\theta) - \Omega(\theta')\right]v.$$

Note that for each i, if $|\gamma_i(\theta) - \gamma_i(\theta')| = 1$, we have $|w_i| = \frac{M_{n,p}}{2} k\varepsilon_{n,p}$. Then there are at least $H(\gamma(\theta), \gamma(\theta'))$ number of elements w_i with $|w_i| = \frac{M_{n,p}}{2} k\varepsilon_{n,p}$, which implies

$$\|[\Sigma(\theta) - \Sigma(\theta')]v\|_2^2 \geq H(\gamma(\theta), \gamma(\theta')) \cdot \left(\frac{M_{n,p}}{2} k\varepsilon_{n,p}\right)^2.$$

Since $\|v\|^2 = \lceil p/2 \rceil \leq p$, the equation above yields

$$\|\Omega(\theta) - \Omega(\theta')\|^2 \geq \frac{\|[\Omega(\theta) - \Omega(\theta')]v\|_2^2}{\|v\|^2} \geq \frac{H(\gamma(\theta), \gamma(\theta')) \cdot ((M_{n,p}/2)k\varepsilon_{n,p})^2}{p},$$

that is,

$$\frac{\|\Omega(\theta) - \Omega(\theta')\|^2}{H(\gamma(\theta), \gamma(\theta'))} \geq \frac{(M_{n,p}k\varepsilon_{n,p})^2}{4p}$$

when $H(\gamma(\theta), \gamma(\theta')) \geq 1$.

7.5. *Proof of Lemma* 4.5. Without loss of generality, we assume that $M_{n,p}$ is a constant, since the total variance affinity is scale invariant. The proof of the bound for the affinity given in Lemma 4.5 is involved. We break the proof into a few major technical lemmas Without loss of generality, we consider only the case $i = 1$ and prove that there exists a constant $c_2 > 0$ such that $\|\bar{\mathbb{P}}_{1,0} \wedge \bar{\mathbb{P}}_{1,1}\| \geq c_2$. The following lemma turns the problem of bounding the total variation affinity into a chi-square distance calculation on Gaussian mixtures. Define

$$\Theta_{-1} = \{(b, c) : \text{there exists a } \theta \in \Theta \text{ such that } \gamma_{-1}(\theta) = b \text{ and } \lambda_{-1}(\theta) = c\},$$

which is the set of all values of the upper triangular matrix $\Omega(\theta)$ could possibly take, with the first row leaving out.

LEMMA 7.3. *If there is a constant $c_2 < 1$ such that*

$$\operatorname*{Average}_{(\gamma_{-1},\lambda_{-1})\in\Theta_{-1}} \left\{\int \left(\frac{d\bar{\mathbb{P}}_{(1,1,\gamma_{-1},\lambda_{-1})}}{d\bar{\mathbb{P}}_{(1,0,\gamma_{-1},\lambda_{-1})}}\right)^2 d\bar{\mathbb{P}}_{(1,0,\gamma_{-1},\lambda_{-1})} - 1\right\} \leq c_2^2, \tag{7.8}$$

then $\|\bar{\mathbb{P}}_{1,0} \wedge \bar{\mathbb{P}}_{1,1}\| \geq 1 - c_2 > 0$.

From the definition of $\bar{\mathbb{P}}_{(1,0,\gamma_{-1},\lambda_{-1})}$ in equation (4.8) and θ in equation (4.3), $\gamma_1 = 0$ implies $\bar{\mathbb{P}}_{(1,0,\gamma_{-1},\lambda_{-1})}$ is a single multivariate normal distribution with a precision matrix,

$$\Omega_0 = \begin{pmatrix} 1 & \mathbf{0}_{1\times(p-1)} \\ \mathbf{0}_{(p-1)\times 1} & \mathbf{S}_{(p-1)\times(p-1)} \end{pmatrix}, \tag{7.9}$$

where $\mathbf{S}_{(p-1)\times(p-1)} = (s_{ij})_{2\le i,j\le p}$ is uniquely determined by $(\gamma_{-1}, \lambda_{-1}) = ((\gamma_2, \dots, \gamma_r), (\lambda_2, \dots, \lambda_r))$ with

$$s_{ij} = \begin{cases} 1, & i = j, \\ \varepsilon_{n,p}, & \gamma_i = \lambda_i(j) = 1, \\ 0, & \text{otherwise.} \end{cases}$$

Let

$$\Lambda_1(c) = \{a : \text{there exists a } \theta \in \Theta \text{ such that } \lambda_1(\theta) = a \text{ and } \lambda_{-1}(\theta) = c\},$$

which gives the set of all possible values of the first row with rest of rows given, that is, $\lambda_{-1}(\theta) = c$, and define $p_{\lambda_{-1}} = \operatorname{Card}(\Lambda_1(\lambda_{-1}))$, the cardinality of all possible λ_1 such that $(\lambda_1, \lambda_{-1}) \in \Lambda$ for the given λ_{-1}. Then from definitions in equations (4.8) and (4.3) $\bar{\mathbb{P}}_{(1,1,\gamma_{-1},\lambda_{-1})}$ is an average of $\binom{p_{\lambda_{-1}}}{k}$ multivariate normal distributions with precision matrices of the following form:

$$\begin{pmatrix} 1 & \mathbf{r}_{1\times(p-1)} \\ \mathbf{r}_{(p-1)\times 1} & \mathbf{S}_{(p-1)\times(p-1)} \end{pmatrix}, \tag{7.10}$$

where $\|\mathbf{r}\|_0 = k$ with nonzero elements of r equal $\varepsilon_{n,p}$ and the submatrix $\mathbf{S}_{(p-1)\times(p-1)}$ is the same as the one for Σ_0 given in (7.9). It is helpful to observe that $p_{\lambda_{-1}} \ge p/4$. Let $n_{\lambda_{-1}}$ be the number of columns of λ_{-1} with column sum equal to $2k$ for which the first row has no choice but to take value 0 in this column. Then we have $p_{\lambda_{-1}} = \lceil p/2 \rceil - n_{\lambda_{-1}}$. Since $n_{\lambda_{-1}} \cdot 2k \le \lceil p/2 \rceil \cdot k$, the total number of 1's in the upper triangular matrix by the construction of the parameter set, we thus have $n_{\lambda_{-1}} \le \lceil p/2 \rceil / 2$, which immediately implies $p_{\lambda_{-1}} = \lceil p/2 \rceil - n_{\lambda_{-1}} \ge \lceil p/2 \rceil / 2 \ge p/4$.

With Lemma 7.3 in place, it remains to establish equation (7.8) in order to prove Lemma 4.5. The following lemma is useful for calculating the cross product terms in the chi-square distance between Gaussian mixtures. The proof of the lemma is straightforward and is thus omitted.

LEMMA 7.4. *Let g_i be the density function of $N(0, \Omega_i^{-1})$ for $i = 0, 1$ and 2. Then*

$$\int \frac{g_1 g_2}{g_0} = \frac{\det(I)}{[\det(I - \Omega_1^{-1}(\Omega_2 - \Omega_0)\Omega_2^{-1}(\Omega_1 - \Omega_0))]^{1/2}}.$$

Let Ω_i, $i = 1$ or 2, be two precision matrices of the form (7.10). Note that Ω_i, $i = 0, 1$ or 2, differs from each other only in the first row/column. Then $\Omega_i - \Omega_0$, $i = 1$ or 2, has a very simple structure. The nonzero elements only appear in the first row/column, and in total there are $2k$ nonzero elements. This property immediately implies the following lemma which makes the problem of studying the determinant in Lemma 7.4 relatively easy.

LEMMA 7.5. *Let Ω_i, $i = 1$ and 2, be the precision matrices of the form (7.10). Define J to be the number of overlapping $\varepsilon_{n,p}$'s between Ω_1 and Ω_2 on the first row, and*

$$Q \stackrel{\triangle}{=} (q_{ij})_{1\le i,j\le p} = (\Omega_2 - \Omega_0)(\Omega_1 - \Omega_0).$$

There are index subsets I_r and I_c in $\{2,\dots,p\}$ with $\operatorname{Card}(I_r) = \operatorname{Card}(I_c) = k$ and $\operatorname{Card}(I_r \cap I_c) = J$ such that

$$q_{ij} = \begin{cases} J\varepsilon_{n,p}^2, & i = j = 1, \\ \varepsilon_{n,p}^2, & i \in I_r \text{ and } j \in I_c, \\ 0, & \text{otherwise} \end{cases}$$

and the matrix $(\Omega_2 - \Omega_0)(\Omega_1 - \Omega_0)$ has rank 2 with two identical nonzero eigenvalues $J\varepsilon_{n,p}^2$.

Let

$$R_{\lambda_1,\lambda_1'}^{\gamma_{-1},\lambda_{-1}} = -\log\det\bigl(I - \Omega_1^{-1}(\Omega_2 - \Omega_0)\Omega_2^{-1}(\Omega_1 - \Omega_0)\bigr), \tag{7.11}$$

where Ω_0 is defined in (7.9) and determined by $(\gamma_{-1}, \lambda_{-1})$, and Ω_1 and Ω_2 have the first row λ_1 and λ_1' respectively. We drop the indices λ_1, λ_1' and $(\gamma_{-1}, \lambda_{-1})$ from Ω_i to simplify the notation. Define

$$\begin{aligned} &\Theta_{-1}(a_1, a_2) \\ &\quad = \{0,1\}^{r-1} \otimes \bigl\{c \in \Lambda_{-1} : \text{there exist } \theta_i \in \Theta, i = 1 \text{ and } 2, \\ &\qquad \text{such that } \lambda_1(\theta_i) = a_i, \lambda_{-1}(\theta_i) = c\bigr\}. \end{aligned}$$

It is a subset of Θ_{-1} in which the element can pick both a_1 and a_2 as the first row to form parameters in Θ. From Lemma 7.4 the left-hand side of equation (7.8) can be written as

$$\begin{aligned} &\operatorname*{Average}_{(\gamma_{-1},\lambda_{-1})\in\Theta_{-1}} \Bigl\{ \operatorname*{Average}_{\lambda_1,\lambda_1'\in\Lambda_1(\lambda_{-1})} \Bigl[\exp\Bigl(\frac{n}{2}\cdot R_{\lambda_1,\lambda_1'}^{\gamma_{-1},\lambda_{-1}}\Bigr) - 1\Bigr]\Bigr\} \\ &\quad = \operatorname*{Average}_{\lambda_1,\lambda_1'\in B}\Bigl\{ \operatorname*{Average}_{(\gamma_{-1},\lambda_{-1})\in\Theta_{-1}(\lambda_1,\lambda_1')} \Bigl[\exp\Bigl(\frac{n}{2}\cdot R_{\lambda_1,\lambda_1'}^{\gamma_{-1},\lambda_{-1}}\Bigr) - 1\Bigr]\Bigr\}. \end{aligned}$$

The following result shows that $R_{\lambda_1,\lambda_1'}^{\gamma_{-1},\lambda_{-1}}$ is approximately $-\log\det(I - (\Omega_2 - \Omega_0)(\Omega_1 - \Omega_0))$ which is equal to $-2\log(1 - J\varepsilon_{n,p}^2)$ from Lemma 7.5. Define

$$\begin{aligned} \Lambda_{1,J} = \bigl\{&(\lambda_1, \lambda_1') \in \Lambda_1 \otimes \Lambda_1 : \\ &\text{the number of overlapping } \varepsilon_{n,p}\text{'s between } \lambda_1 \text{ and } \lambda_1' \text{ is } J\bigr\}. \end{aligned}$$

LEMMA 7.6. *For $R_{\lambda_1,\lambda_1'}$ defined in equation* (7.11), *we have*

$$R^{\gamma_{-1},\lambda_{-1}}_{\lambda_1,\lambda_1'} = -2\log(1 - J\varepsilon^2_{n,p}) + R^{\gamma_{-1},\lambda_{-1}}_{1,\lambda_1,\lambda_1'} \tag{7.12}$$

where $R^{\gamma_{-1},\lambda_{-1}}_{1,\lambda_1,\lambda_1'}$ satisfies

$$\operatorname*{Average}_{(\lambda_1,\lambda_1')\in\Lambda_{1,J}}\left[\operatorname*{Average}_{(\gamma_{-1},\lambda_{-1})\in\Theta_{-1}(\lambda_1,\lambda_1')} \exp\left(\frac{n}{2}R^{\gamma_{-1},\lambda_{-1}}_{1,\lambda_1,\lambda_1'}\right)\right] = 1 + o(1), \tag{7.13}$$

where J is defined in Lemma 7.5.

7.5.1. *Proof of equation* (7.8). We are now ready to establish equation (7.8) which is the key step in proving Lemma 4.5. It follows from equation (7.12) in Lemma 7.6 that

$$\begin{aligned}
&\operatorname*{Average}_{\lambda_1,\lambda_1'\in B}\left\{\operatorname*{Average}_{(\gamma_{-1},\lambda_{-1})\in\Theta_{-1}(\lambda_1,\lambda_1')}\left[\exp\left(\frac{n}{2}R^{\gamma_{-1},\lambda_{-1}}_{\lambda_1,\lambda_1'}\right) - 1\right]\right\} \\
&\quad = \operatorname*{Average}_{J}\left\{\exp[-n\log(1 - J\varepsilon^2_{n,p})]\right. \\
&\qquad \left.\times \operatorname*{Average}_{(\lambda_1,\lambda_1')\in\Lambda_{1,J}}\left[\operatorname*{Average}_{(\gamma_{-1},\lambda_{-1})\in\Theta_{-1}(\lambda_1,\lambda_1')} \exp\left(\frac{n}{2}R^{\gamma_{-1},\lambda_{-1}}_{1,\lambda_1,\lambda_1'}\right)\right] - 1\right\}.
\end{aligned}$$

Recall that J is the number of overlapping $\varepsilon_{n,p}$'s between Σ_1 and Σ_2 on the first row. It can be shown that J has the hypergeometric distribution with

$$\begin{aligned}
&\mathbb{P}(\text{number of overlapping } \varepsilon_{n,p}\text{'s} = j) \\
&\quad = \binom{k}{j}\binom{p_{\lambda_{-1}} - k}{k - j}\Big/\binom{p_{\lambda_{-1}}}{k} \le \left(\frac{k^2}{p_{\lambda_{-1}} - k}\right)^j.
\end{aligned} \tag{7.14}$$

Equation (7.14) and Lemma 7.6, together with equation (4.10), imply

$$\begin{aligned}
&\operatorname*{Average}_{(\gamma_{-1},\lambda_{-1})\in\Theta_{-1}}\left\{\int\left(\frac{d\bar{\mathbb{P}}_{(1,1,\gamma_{-1},\lambda_{-1})}}{d\bar{\mathbb{P}}_{(1,0,\gamma_{-1},\lambda_{-1})}}\right)^2 d\bar{\mathbb{P}}_{(1,0,\gamma_{-1},\lambda_{-1})} - 1\right\} \\
&\quad \le \sum_{j\ge 0}\left(\frac{k^2}{p/4 - 1 - k}\right)^j \{\exp[-n\log(1 - j\varepsilon^2_{n,p})]\cdot(1 + o(1)) - 1\} \\
&\quad \le (1 + o(1))\sum_{j\ge 1}(p^{(\beta-1)/\beta})^{-j}\exp[2j(\upsilon^2\log p)] + o(1) \\
&\quad \le C\sum_{j\ge 1}(p^{(\beta-1)/\beta - 2\upsilon^2})^{-j} + o(1) < c_2^2,
\end{aligned} \tag{7.15}$$

where the last step follows from $\upsilon^2 < \frac{\beta-1}{8\beta}$, and $k^2 \le [c_{n,p}(M_{n,p}\varepsilon_{n,p})^{-q}]^2 = O(\frac{n}{\log^3 p}) = o(\frac{p^{1/\beta}}{\log p})$ from equations (4.11) and (4.2) and the condition $p > c_1 n^\beta$ for some $\beta > 1$, and c_2 is a positive constant.

7.6. *Proof of Lemma* 7.6. Let

$$A = I - \bigl[I - \Omega_1^{-1}(\Omega_2 - \Omega_0)\Omega_2^{-1}(\Omega_1 - \Omega_0)\bigr] \times \bigl[I - (\Omega_2 - \Omega_0)(\Omega_1 - \Omega_0)\bigr]^{-1}. \tag{7.16}$$

Since $\|\Omega_i - \Omega_0\| \le \|\Omega_i - \Omega_0\|_1 \le 2k\varepsilon_{n,p} = o(1/\log p)$ from equation (4.13), it is easy to see that

$$\|A\| = O(k\varepsilon_{n,p}) = o(1). \tag{7.17}$$

Define

$$R_{1,\lambda_1,\lambda_1'}^{\gamma_{-1},\lambda_{-1}} = -\log\det(I - A).$$

Then we can rewrite $R_{\lambda_1,\lambda_1'}^{\gamma_{-1},\lambda_{-1}}$ as follows:

$$\begin{aligned} R_{\lambda_1,\lambda_1'}^{\gamma_{-1},\lambda_{-1}} &= -\log\det\bigl(I - \Omega_1^{-1}(\Omega_2 - \Omega_0)\Omega_2^{-1}(\Omega_1 - \Omega_0)\bigr) \\ &= -\log\det\bigl([I - A]\cdot\bigl[I - (\Omega_2 - \Omega_0)(\Omega_1 - \Omega_0)\bigr]\bigr) \\ &= -\log\det\bigl[I - (\Omega_2 - \Omega_0)(\Omega_1 - \Omega_0)\bigr] - \log\det(I - A) \\ &= -2\log\bigl(1 - J\varepsilon_{n,p}^2\bigr) + R_{1,\lambda_1,\lambda_1'}^{\gamma_{-1},\lambda_{-1}}, \end{aligned} \tag{7.18}$$

where the last equation follows from Lemma 7.5. To establish Lemma 7.6, it is enough to establish equation (7.13).

Define

$$\begin{aligned} A_1 &= (a_{1,ij}) \\ &= I - \bigl[I - (\Omega_1 - \Omega_0 + I)^{-1}(\Omega_2 - \Omega_0)(\Omega_2 - \Omega_0 + I)^{-1}(\Omega_1 - \Omega_0)\bigr] \\ &\quad \times \bigl[I - (\Omega_2 - \Omega_0)(\Omega_1 - \Omega_0)\bigr]^{-1}. \end{aligned}$$

Similar to equation (7.17), we have $\|A_1\| = o(1)$, and write

$$R_{1,\lambda_1,\lambda_1'}^{\gamma_{-1},\lambda_{-1}} = -\log\det(I - A_1) - \log\det\bigl[(I - A_1)^{-1}(I - A)\bigr].$$

To establish equation (7.13), it is enough to show that

$$\exp\Bigl[-\frac{n}{2}\log\det(I - A_1)\Bigr] = 1 + o(1), \tag{7.19}$$

and that

$$\text{(7.20)} \qquad \underset{(\gamma_{-1},\lambda_{-1})\in\Theta_{-1}(\lambda_1,\lambda_1')}{\text{Average}} \exp\left(-\frac{n}{2}\log\det\left[(I-A_1)^{-1}(I-A)\right]\right) = 1+o(1).$$

The proof for equation (7.19) is as follows. Write

$$\Omega_1 - \Omega_0 = \begin{pmatrix} 0 & \mathbf{v}_{1\times(p-1)} \\ (\mathbf{v}_{1\times(p-1)})^T & \mathbf{0}_{(p-1)\times(p-1)} \end{pmatrix} \quad \text{and}$$
$$\Omega_2 - \Omega_0 = \begin{pmatrix} 0 & \mathbf{v}^*_{1\times(p-1)} \\ (\mathbf{v}^*_{1\times(p-1)})^T & \mathbf{0}_{(p-1)\times(p-1)} \end{pmatrix},$$

where $\mathbf{v}_{1\times(p-1)} = (v_j)_{2\le j\le p}$ satisfies $v_j = 0$ for $2 \le j \le p-r$ and $v_j = 0$ or 1 for $p-r+1 \le j \le p$ with $\|\mathbf{v}\|_0 = k$, and $\mathbf{v}^*_{1\times(p-1)} = (v_j^*)_{2\le j\le p}$ satisfies a similar property. Without loss of generality, we consider only a special case with

$$v_j = \begin{cases} 1, & p-r+1 \le j \le p-r+k, \\ 0, & \text{otherwise} \end{cases} \quad \text{and}$$
$$v_j^* = \begin{cases} 1, & p-r+k-J \le j \le p-r+2k-J, \\ 0, & \text{otherwise.} \end{cases}$$

Note that we may write A_1 as

$$A_1 = B_1 B_2,$$

where B_1 can be written as a polynomial of $\Omega_1 - \Omega_0$ and $\Omega_2 - \Omega_0$, and B_2 can be written as a polynomial of $(\Omega_2 - \Omega_0)(\Omega_1 - \Omega_0)$ by Taylor expansion. By a straightforward calculation, it can be shown that

$$|b_{1,ij}| = \begin{cases} O(\varepsilon_{n,p}), & i=1 \text{ and } p-r+1 \le j \le p-r+2k-J, \\ & \text{or } j=1 \text{ and } p-r+1 \le i \le p-r+2k-J, \\ & \text{or } i=j=1, \\ O(\varepsilon_{n,p}^2), & p-r+1 \le i \le p-r+2k-J \text{ and} \\ & p-r+1 \le j \le p-r+2k-J, \\ 0, & \text{otherwise,} \end{cases}$$

and

$$0 \le b_{2,ij} = \begin{cases} O(J\varepsilon_{n,p}^2), & i=j=1, \\ \tau_{n,p}, & p-r+1 \le i \le p-r+k \text{ and} \\ & p-r+k-J \le j-1 \le p-r+2k-J, \\ 0, & \text{otherwise,} \end{cases}$$

where $\tau_{n,p} = O(\varepsilon_{n,p}^2)$, which implies

$$|a_{1,ij}| = \begin{cases} O(k\varepsilon_{n,p}^3), & i = 1 \text{ and } p - r + k - J \le j - 1 \le p - r + 2k - J, \\ & \text{or } i = j = 1, \\ O(J\varepsilon_{n,p}^3), & j = 1 \text{ and } p - r + 1 \le i \le p - r + 2k - J, \\ O(k\varepsilon_{n,p}^4), & p - r + 1 \le i \le p - r + 2k - J \text{ and} \\ & p - r + k - J \le j - 1 \le p - r + 2k - J, \\ 0, & \text{otherwise.} \end{cases}$$

Note that $\operatorname{rank}(A_1) \le 2$ due to the simple structure of $(\Omega_2 - \Omega_0)(\Omega_1 - \Omega_0)$. Let $A_2 = (a_{2,ij})$ with

$$|a_{2,ij}| = \begin{cases} O(k\varepsilon_{n,p}^3), & i = 1 \text{ and } j = 1, \\ O(k\varepsilon_{n,p}^4 + Jk\varepsilon_{n,p}^6), & p - r + 1 \le i \le p - r + 2k - J \text{ and} \\ & p - r + k - J \le j - 1 \le p - r + 2k - J, \\ 0, & \text{otherwise,} \end{cases}$$

and $\operatorname{rank}(A_2) \le 4$ by eliminating the nonzero off-diagonal elements of the first row and column of A_1, and

$$\exp\left[-\frac{n}{2}\log\det(I - A_1)\right] = \exp\left[-\frac{n}{2}\log\det(I - A_2)\right].$$

We can show that all eigenvalues of A_2 are $O(Jk^2\varepsilon_{n,p}^6 + k^2\varepsilon_{n,p}^4 + k\varepsilon_{n,p}^3)$. Since $k\varepsilon_{n,p} = o(1/\log p)$, then

$$nk\varepsilon_{n,p}^3 = \upsilon^3 \frac{k(\log p)^{1/2}}{\sqrt{n}} \log p = o(1)$$

which implies

$$n\left(Jk^2\varepsilon_{n,p}^6 + k^2\varepsilon_{n,p}^4 + k\varepsilon_{n,p}^3\right) = o(1).$$

Thus,

$$\exp\left[-\frac{n}{2}\log\det(I - A_1)\right] = 1 + o(1).$$

Now we establish equation (7.20), which, together with equation (7.19), yields equation (7.13) and thus Lemma 7.6 is established. Write

$$\begin{aligned} &(I - A_1)^{-1}(I - A) - I \\ &\quad = (I - A_1)^{-1}\left[(I - A) - (I - A_1)\right] \\ &\quad = (I - A_1)^{-1}(A_1 - A) \\ &\quad = (I - A_1)^{-1}\left[\begin{array}{c} (\Omega_1 - \Omega_0 + I)^{-1}(\Omega_2 - \Omega_0)(\Omega_2 - \Omega_0 + I)^{-1}(\Omega_1 - \Omega_0) \\ -\Omega_1^{-1}(\Omega_2 - \Omega_0)\Omega_2^{-1}(\Omega_1 - \Omega_0) \end{array}\right] \\ &\qquad \times \left[I - (\Omega_2 - \Omega_0)(\Omega_1 - \Omega_0)\right]^{-1}, \end{aligned}$$

where

$$
\begin{aligned}
&(\Omega_1-\Omega_0+I)^{-1}(\Omega_2-\Omega_0)(\Omega_2-\Omega_0+I)^{-1} \\
&\quad\times(\Omega_1-\Omega_0)-\Omega_1^{-1}(\Omega_2-\Omega_0)\Omega_2^{-1}(\Omega_1-\Omega_0) \\
&\quad=\bigl[(\Omega_1-\Omega_0+I)^{-1}-\Omega_1^{-1}\bigr](\Omega_2-\Omega_0)(\Omega_2-\Omega_0+I)^{-1}(\Omega_1-\Omega_0) \\
&\qquad+\Omega_1^{-1}(\Omega_2-\Omega_0)\bigl[(\Omega_2-\Omega_0+I)^{-1}-\Omega_2^{-1}\bigr](\Omega_1-\Omega_0) \\
&\quad=\Omega_1^{-1}(\Omega_0-I)(\Omega_1-\Omega_0+I)^{-1}(\Omega_2-\Omega_0)(\Omega_2-\Omega_0+I)^{-1}(\Omega_1-\Omega_0) \\
&\qquad+\Omega_1^{-1}(\Omega_2-\Omega_0)\Omega_2^{-1}(\Omega_0-I)(\Omega_2-\Omega_0+I)^{-1}(\Omega_1-\Omega_0) \\
&\quad=\Omega_1^{-1}(\Omega_0-I)(\Omega_1-\Omega_0+I)^{-1}(\Omega_2-\Omega_0)(\Omega_2-\Omega_0+I)^{-1}(\Omega_1-\Omega_0) \\
&\qquad+\Omega_1^{-1}(\Omega_2-\Omega_0)\Omega_2^{-1}(\Omega_0-I)\bigl[(\Omega_2-\Omega_0+I)^{-1}-I\bigr](\Omega_1-\Omega_0) \\
&\qquad+\Omega_1^{-1}(\Omega_2-\Omega_0)\Omega_2^{-1}(\Omega_0-I)(\Omega_1-\Omega_0).
\end{aligned}
$$

It is important to observe that $\operatorname{rank}((I-A_1)^{-1}(I-A)-I)\le 2$ again due to the simple structure of $(\Omega_2-\Omega_0)$ and $(\Omega_1-\Omega_0)$, then $-\log\det[(I-A_1)^{-1}(I-A)]$ is determined by at most two nonzero eigenvalues, which can be shown to be bounded by

$$
\begin{aligned}
&\bigl\|(I-A_1)^{-1}(I-A)-I\bigr\| \\
&\quad=\bigl(1+o(1)\bigr)\begin{pmatrix} o(1/n)+\bigl\|(I-\Omega_0)(\Omega_2-\Omega_0)(\Omega_1-\Omega_0)\bigr\| \\ +\bigl\|(\Omega_2-\Omega_0)\Omega_2^{-1}(\Omega_0-I)(\Omega_1-\Omega_0)\bigr\| \end{pmatrix}.
\end{aligned}
\tag{7.21}
$$

Note that $\|(I-A_1)^{-1}(I-A)-I\|=o(1)$, and

$$\bigl|\log(1-x)\bigr|\le 2|x| \qquad \text{for } |x|<1/3,$$

which implies

$$\bigl|-\log\det\bigl[(I-A_1)^{-1}(I-A)\bigr]\bigr|\le 2\bigl\|(I-A_1)^{-1}(I-A)-I\bigr\|,$$

that is,

$$\exp\biggl(\frac{n}{2}\cdot-\log\det\bigl[(I-A_1)^{-1}(I-A)\bigr]\biggr)\le\exp\bigl(n\bigl\|(I-A_1)^{-1}(I-A)-I\bigr\|\bigr).$$

Define

$$
\begin{aligned}
A_*&=(I-\Omega_0)(\Omega_2-\Omega_0)(\Omega_1-\Omega_0) \quad \text{and} \\
B_*&=(\Omega_2-\Omega_0)\Omega_2^{-1}(\Omega_0-I)(\Omega_1-\Omega_0),
\end{aligned}
$$

then

$$
\begin{aligned}
&\exp\biggl(\frac{n}{2}\cdot-\log\det\bigl[(I-A_1)^{-1}(I-A)\bigr]\biggr) \\
&\quad\le\bigl(1+o(1)\bigr)\exp\bigl[\bigl(1+o(1)\bigr)n\bigl(\|A_*\|+\|B_*\|\bigr)\bigr]
\end{aligned}
$$

from equations (7.21). By the Cauchy–Schwarz inequality, it is then sufficient to show

$$\operatorname*{Average}_{(\gamma_{-1},\lambda_{-1})\in\Theta_{-1}(\lambda_1,\lambda_1')} \exp(4n\|A_*\|) = 1+o(1) \quad \text{and}$$
$$\operatorname*{Average}_{(\gamma_{-1},\lambda_{-1})\in\Theta_{-1}(\lambda_1,\lambda_1')} \exp(4n\|B_*\|) = 1+o(1),$$

where $\|A_*\|$ and $\|B_*\|$ depends on the values of λ_1, λ_1' and $(\gamma_{-1},\lambda_{-1})$. We dropped the indices λ_1, λ_1' and $(\gamma_{-1},\lambda_{-1})$ from A_* and B_* to simplify the notation. Due space limit, we only show that

$$\operatorname*{Average}_{(\gamma_{-1},\lambda_{-1})\in\Theta_{-1}(\lambda_1,\lambda_1')} \exp(4n\|A_*\|) = 1+o(1),$$

while the bound for $\|B_*\|$ can be shown similarly by reducing it to study $\|(\Omega_2-\Omega_0)(I-\Omega_0)^l(\Omega_1-\Omega_0)\|$ for integers $l\ge 1$. Let $E_m=\{1,2,\dots,r\}/\{1,m\}$. Let $n_{\lambda_{E_m}}$ be the number of columns of λ_{E_m} with column sum at least $2k-2$ for which two rows cannot freely take value 0 or 1 in this column. Then we have $p_{\lambda_{E_m}}=\lceil p/2\rceil - n_{\lambda_{E_m}}$. Without loss of generality, we assume that $k\ge 3$. Since $n_{\lambda_{E_m}}\cdot(2k-2)\le\lceil p/2\rceil\cdot k$, the total number of 1's in the upper triangular matrix by the construction of the parameter set, we thus have $n_{\lambda_{E_m}}\le\lceil p/2\rceil\cdot\frac{3}{4}$, which immediately implies $p_{\lambda_{E_m}}=\lceil p/2\rceil - n_{\lambda_{E_m}}\ge\lceil p/2\rceil\frac{1}{4}\ge p/8$. Thus, we have

$$\begin{aligned}\mathbb{P}(\|A_*\|\ge 2t\cdot\varepsilon_{n,p}\cdot k\varepsilon_{n,p}^2) &\le \mathbb{P}(\|A_*\|_1\ge 2t\cdot\varepsilon_{n,p}\cdot k\varepsilon_{n,p}^2)\\ &\le \sum_m \operatorname*{Average}_{\lambda_{E_m}} \frac{\binom{k}{t}\binom{p_{\lambda_{E_m}}}{k-t}}{\binom{p_{\lambda_{E_m}}}{k}} \le p\left(\frac{k^2}{p/8-k}\right)^t\end{aligned}$$

from equation (7.14), which immediately implies

$$\begin{aligned}&\operatorname*{Average}_{(\gamma_{-1},\lambda_{-1})\in\Theta_{-1}(\lambda_1,\lambda_1')} \exp(4n\|A_*\|)\\ &\quad\le \exp\left(4n\cdot 2\cdot\frac{2(\beta-1)}{\beta}\cdot\varepsilon_{n,p}\cdot k\varepsilon_{n,p}^2\right)\\ &\qquad+\int_{(2(\beta-1))/\beta}^{\infty}\exp(4n\cdot 2t\cdot\varepsilon_{n,p}\cdot k\varepsilon_{n,p}^2)p\left(\frac{k^2}{p/8-k}\right)^t dt\\ &\quad=\exp\left(\frac{16(\beta-1)}{\beta}nk\varepsilon_{n,p}^3\right)\\ &\qquad+\int_{2(\beta-1)/\beta}^{\infty}\exp\left[\log p+t\left(8nk\varepsilon_{n,p}^3-\log\frac{k^2}{p/8-k}\right)\right]dt\\ &\quad=1+o(1),\end{aligned}$$

where the last step is an immediate consequence of the following two equations:

$$nk\varepsilon_{n,p}^3 = o(1)$$

and

$$(1+o(1))2\log p \le t \log \frac{p/8-1-k}{k^2} \qquad \text{for } t \ge \frac{2(\beta-1)}{\beta}$$

which follow from $k^2 = O(n) = O(p^{1/\beta})$ from equation (4.11) and the condition $p > c_1 n^\beta$ for some $\beta > 1$.

REFERENCES

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

CAI, T. T. and JIANG, T. (2011). Limiting laws of coherence of random matrices with applications to testing covariance structure and construction of compressed sensing matrices. *Ann. Statist.* **39** 1496–1525. MR2850210

CAI, T. and LIU, W. (2011). Adaptive thresholding for sparse covariance matrix estimation. *J. Amer. Statist. Assoc.* **106** 672–684. MR2847949

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. and YUAN, M. (2012). Adaptive covariance matrix estimation through block thresholding. *Ann. Statist.* **40** 2014–2042. MR3059075

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

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, T. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics* **9** 432–441.

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

LIU, H., LAFFERTY, J. and WASSERMAN, L. (2009). The nonparanormal: Semiparametric estimation of high dimensional undirected graphs. *J. Mach. Learn. Res.* **10** 2295–2328. MR2563983

LIU, W.-D., LIN, Z. and SHAO, Q.-M. (2008). The asymptotic distribution and Berry-Esseen bound of a new test for independence in high dimension with an application to stochastic optimization. *Ann. Appl. Probab.* **18** 2337–2366. MR2474539

LIU, H., HAN, F., YUAN, M., LAFFERTY, J. and WASSERMAN, L. (2012). High-dimensional semiparametric Gaussian copula graphical models. *Ann. Statist.* **40** 2293–2326. MR3059084

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

PETROV, V. V. (1975). *Sums of Independent Random Variables*. Springer, New York. Translated from the Russian by A. A. Brown, Ergebnisse der Mathematik und ihrer Grenzgebiete, Band 82. MR0388499

PETROV, V. V. (1995). *Limit Theorems of Probability Theory*: *Sequences of Independent Random Variables*. *Oxford Studies in Probability* **4**. Oxford Univ. Press, New York. MR1353441

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

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

SAON, G. and CHIEN, J. T. (2011). Bayesian sensing hidden Markov models for speech recognition. In *Acoustics*, *Speech and Signal Processing* (*ICASSP*), 2011 *IEEE International Conference* 5056–5059. Prague.

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

XUE, L. and ZOU, H. (2012). Regularized rank-based estimation of high-dimensional nonparanormal graphical models. *Ann*. *Statist*. **40** 2541–2571. MR3097612

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

T. T. CAI
DEPARTMENT OF STATISTICS
THE WHARTON SCHOOL
UNIVERSITY OF PENNSYLVANIA
PHILADELPHIA, PENNSYLVANIA 19104
USA
E-MAIL: tcai@wharton.upenn.edu
URL: http://www-stat.wharton.upenn.edu/~tcai

W. LIU
DEPARTMENT OF MATHEMATICS
AND INSTITUTE OF NATURAL SCIENCES
SHANGHAI JIAO TONG UNIVERSITY
SHANGHAI 200240
CHINA
E-MAIL: weidongl@sjtu.edu.cn
URL: http://www.math.sjtu.edu.cn/faculty/weidongl

H. H. ZHOU
DEPARTMENT OF STATISTICS
YALE UNIVERSITY
NEW HAVEN, CONNECTICUT 06511
USA
E-MAIL: huibin.zhou@yale.edu
URL: http://www.stat.yale.edu/~hz68