Submitted to the Annals of Statistics

OPTIMAL RATES OF CONVERGENCE FOR SPARSE COVARIANCE MATRIX ESTIMATION

By T. Tony Cai * and Harrison H. Zhou†

University of Pennsylvania and Yale University

This paper considers estimation of sparse covariance matrices and establishes the optimal rate of convergence under a range of matrix operator norm and Bregman divergence losses. A major focus is on the derivation of a rate sharp minimax lower bound. The problem exhibits new features that are significantly different from those that occur in the conventional nonparametric function estimation problems. Standard techniques fail to yield good results and new tools are thus needed.

We first develop a lower bound technique that is particularly well suited for treating "two-directional" problems such as estimating sparse covariance matrices. The result can be viewed as a generalization of Le Cam's method in one direction and Assouad's Lemma in another. This lower bound technique is of independent interest and can be used for other matrix estimation problems.

We then establish a rate sharp minimax lower bound for estimating sparse covariance matrices under the spectral norm by applying the general lower bound technique. A thresholding estimator is shown to attain the optimal rate of convergence under the spectral norm. The results are then extended to the general matrix ℓ_w operator norms for $1 \leqslant w \leqslant \infty$. In addition, we give a unified result on the minimax rate of convergence for sparse covariance matrix estimation under a class of Bregman divergence losses.

1. Introduction. Minimax risk is one of the most widely used benchmarks for optimality and substantial efforts have been made on developing minimax theories in the statistics literature. A key step in establishing a minimax theory is the derivation of minimax lower bounds and several effective lower bound arguments based on hypothesis testing have been introduced in the literature. Well known techniques include Le Cam's method, Assouad's Lemma and Fano's Lemma. See Le Cam (1986) and Tsybakov (2009) for

*The research of Tony Cai was supported in part by NSF FRG Grant DMS-0854973.

†The research of Harrison Zhou was supported in part by NSF Career Award DMS-0645676 and NSF FRG Grant DMS-0854975.

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

Keywords and phrases: Assouad's Lemma, Bregman divergence, covariance matrix estimation, Frobenius norm, Le Cam's method, minimax lower bound, spectral norm, optimal rate of convergence, thresholding.

more detailed discussions on minimax lower bound arguments.

Driven by a wide range of applications in high dimensional data analysis, estimation of large covariance matrices has drawn considerable recent attention. See, for example, Bickel and Levina (2008a, b), El Karoui (2008), Ravikumar, Wainwright, Raskutti and Yu (2008), Lam and Fan (2009), Cai and Zhou (2009), Cai, Zhang and Zhou (2010), and Cai and Liu (2011). Many theoretical results, including consistency and rates of convergence, have been obtained. However, the optimality question remains mostly open in the context of covariance matrix estimation under the spectral norm, mainly due to the technical difficulty in obtaining good minimax lower bounds.

In this paper we consider optimal estimation of sparse covariance matrices and establish the minimax rate of convergence under a range of matrix operator norm and Bregman divergence losses. A major focus is on the derivation of a rate sharp lower bound under the spectral norm loss. Conventional lower bound techniques such as the ones mentioned earlier are designed and well suited for problems with parameters that are scalar or vector-valued. They have achieved great successes in solving many nonparametric function estimation problems which can be treated exactly or approximately as estimation of a finite or infinite dimensional vector and can thus be viewed as "one-directional" in terms of the lower bound arguments. In contrast, the problem of estimating a sparse covariance matrix under the spectral norm can be regarded as a truly "two-directional" problem where one direction is along the rows and another along the columns. It cannot be essentially reduced to a problem of estimating a single or multiple vectors. As a consequence, standard lower bound techniques fail to yield good results for this matrix estimation problem. New and more general technical tools are thus needed.

In the present paper we first develop a minimax lower bound technique that is particularly well suited for treating "two-directional" problems such as estimating sparse covariance matrices. The result can be viewed as a simultaneous generalization of Le Cam's method in one direction and Assouad's Lemma in another. This general technical tool is of independent interest and is useful for solving other matrix estimation problems such as optimal estimation of sparse precision matrices.

We then consider specifically the problem of optimal estimation of sparse covariance matrices under the spectral norm. Let $\mathbf{X}_1, \ldots, \mathbf{X}_n$ be a random sample from a p-variate distribution with covariance matrix $\Sigma = (\sigma_{ij})_{1 \leqslant i,j \leqslant p}$. We wish to estimate the unknown matrix Σ based on the sample $\{\mathbf{X}_1, ..., \mathbf{X}_n\}$. In this paper we shall use the weak ℓ_q ball with $0 \leqslant q < 1$ to model the sparsity of the covariance matrix Σ. The weak ℓ_q ball was originally used in

Abramovich, Benjamini, Donoho and Johnstone (2006) for a sparse normal means problem. A weak ℓ_q ball of radius c in $\mathbb{R}^m$ contains elements with fast decaying ordered magnitudes of components:

$$B_q^m(c) = \left\{\xi \in \mathbb{R}^m : \ |\xi|_{(k)}^q \leqslant ck^{-1}, \quad \text{for all } k = 1, ..., m\right\}$$

where $|\xi|_{(k)}$ denotes the k-th largest element in magnitude of the vector ξ. For a covariance matrix $\Sigma = (\sigma_{ij})_{1\leqslant i,j\leqslant p}$, denote by $\sigma_{-j,j}$ the jth column of Σ with σ_{jj} removed. We shall assume that $\sigma_{-j,j}$ is in a weak ℓ_q ball for all $1 \leqslant j \leqslant p$. More specifically, for $0 \leqslant q < 1$, we define the parameter space $\mathcal{G}_q(c_{n,p})$ of covariance matrices by

$$\mathcal{G}_q(c_{n,p}) = \left\{\Sigma = (\sigma_{ij})_{1\leqslant i,j\leqslant p} : \ \sigma_{-j,j} \in B_q^{p-1}(c_{n,p}), \ 1 \leqslant j \leqslant p\right\}. \tag{1}$$

In the special case of $q = 0$, a matrix in $\mathcal{G}_0(c_{n,p})$ has at most $c_{n,p}$ nonzero off-diagonal elements on each column.

The problem of estimating sparse covariance matrices under the spectral norm has been considered, for example, in El Karoui (2008), Bickel and Levina (2008b), Rothman, Levina and Zhu (2009), and Cai and Liu (2011). Thresholding methods were introduced and rates of convergence in probability were obtained for the thresholding estimators. The parameter space $\mathcal{G}_q(c_{n,p})$ given in (1) also contains the uniformity class of covariance matrices considered in Bickel and Levina (2008b) as a special case. We assume that the distribution of the X_i's is subgaussian in the sense that there is $\tau > 0$ such that

$$\mathbb{P}\{|\mathbf{v}^T(\mathbf{X}_1 - \mathbb{E}\mathbf{X}_1)| > t\} \leqslant e^{-t^2/(2\tau)} \text{ for all } t > 0 \text{ and } \|\mathbf{v}\|_2 = 1. \tag{2}$$

Let $\mathcal{P}_q(\tau, c_{n,p})$ denote the set of distributions of $\mathbf{X}_1$ satisfying (2) and with covariance matrix $\Sigma \in \mathcal{G}_q(c_{n,p})$.

Our technical analysis used in establishing a rate-sharp minimax lower bound has three major steps. The first step is to reduce the original problem to a simpler estimation problem over a carefully chosen subset of the parameter space without essentially decreasing the level of difficulty. The second is to apply the general minimax lower bound technique to this simplified problem and the final key step is to bound the total variation affinities between pairs of mixture distributions with specially designed sparse covariance matrices. The technical analysis requires ideas that are quite different from those used in the typical function/sequence estimation problems.

The minimax upper bound is obtained by studying the risk properties of thresholding estimators. It will be shown that the optimal rate of convergence under mean squared spectral norm error is achieved by a thresholding

estimator introduced in Bickel and Levina (2008b). We write $a_n \asymp b_n$ if there are positive constants c and C independent of n such that $c \leqslant a_n/b_n \leqslant C$. For $1 \leqslant w \leqslant \infty$, 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 $|||\cdot|||_2$. (Throughout the paper, we shall write $|||\cdot|||$ without a subscript for the matrix spectral norm.) For a symmetric matrix A, it is known that the spectral norm $|||A|||$ is equal to the largest magnitude of the eigenvalues of A. Throughout the paper we shall assume that $1 < n^\beta \leqslant p$ for some constants $\beta > 1$. Combining the results given in Sections 3 and 4, we have the following optimal rate of convergence for estimating sparse covariance matrices under the spectral norm.

Theorem 1 *Assume that*

$$c_{n,p} \leqslant M n^{\frac{1-q}{2}} (\log p)^{-\frac{3-q}{2}} \tag{3}$$

for $0 \leqslant q < 1$. The minimax risk of estimating the covariance matrix Σ under the spectral norm over the class $\mathcal{P}_q(\tau, c_{n,p})$ satisfies

$$\inf_{\hat{\Sigma}} \sup_{\theta \in \mathcal{P}_q(\tau, c_{n,p})} \mathbb{E}_{\mathbf{X}|\theta} \left|\left|\left| \hat{\Sigma} - \Sigma \right|\right|\right|^2 \asymp c_{n,p}^2 \left(\frac{\log p}{n} \right)^{1-q} + \frac{\log p}{n} \tag{4}$$

where θ denotes a distribution in $\mathcal{P}_q(\tau, c_{n,p})$ with the covariance matrix Σ. Furthermore, (4) holds under the squared ℓ_w operator norm loss for all $1 \leqslant w \leqslant \infty$.

We shall focus the discussions on the spectral norm loss. The extension to the general matrix ℓ_w operator norm is given in Section 6. In addition, we also consider optimal estimation under a class of Bregman matrix divergences which include the Stein's loss, squared Frobenius norm, and von Neumann entropy as special cases. Bregman matrix divergences provide a flexible class of dissimilarity measures between symmetric matrices and have been used for covariance and precision matrix estimation as well as matrix approximation problems. See, for example, Dhillon and Tropp (2007), Ravikumar, et al. (2008), and Kulis, Sustik and Dhillon (2009). We give a unified result on the minimax rate of convergence in Section 5.

Besides the sparsity assumption considered in this paper, another commonly used structural assumption in the literature is that the covariance matrix is "bandable" where the entries decay as they move away from the diagonal. This is particularly suitable in the setting where the variables exhibit a certain ordering structure, which is often the case for time series data. Various regularization methods have been proposed and studied under this

assumption. Bickel and Levina (2008a) proposed a banding estimator and obtained rate of convergence for the estimator. Cai, Zhang and Zhou (2010) established the minimax rates of convergence and introduced a rate-optimal tapering estimator. In particular, Cai, Zhang and Zhou (2010) derived rate sharp minimax lower bounds for estimating bandable matrices. It should be noted that the lower bound techniques used there do not lead to a good result for estimating sparse covariance matrices under the spectral norm.

The rest of the paper is organized as follows. Section 2, introduces a general technical tool for deriving minimax lower bounds on the minimax risk. Section 3 establishes the minimax lower bound for estimating sparse covariance matrices under the spectral norm. The upper bound is obtained in Section 4 by studying the risk properties of thresholding estimators. Section 5 considers optimal estimation under the Bregman divergences. A uniform optimal rate of convergence is given for a class of Bregman divergence losses. Section 6 discusses extensions to estimation under the general ℓ_w norm for $1 \leqslant w \leqslant \infty$ and connections to other related problems including optimal estimation of sparse precision matrices. The proofs are given in Section 7.

2. General Lower Bound for Minimax Risk. In this section we develop a new general minimax lower bound technique that is particularly well suited for treating "two-directional" problems such as estimating sparse covariance matrices. The new method can be viewed as a generalization of both Le Cam's method and Assouad's Lemma. To help motivate and understand the new lower bound argument, it is useful to briefly review the Le Cam's method and Assouad's Lemma.

The Le Cam's method is based on a two-point testing argument and is particularly well used in estimating linear functionals. See Le Cam (1973) and Donoho and Liu (1991). Let X be an observation from a distribution $\mathbb{P}_\theta$ where θ belongs to a parameter set Θ. 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 $\|\mathbb{P} \wedge \mathbb{Q}\| = \int p \wedge q d\mu$. Le Cam's method works with a finite parameter set $\Theta = \{\theta_0, \theta_1, \ldots, \theta_D\}$. Let L be a loss function. Define $l_{\min} = \min_{1 \leqslant i \leqslant D} \inf_t \left[L\left(t, \theta_0\right) + L\left(t, \theta_i\right)\right]$ and denote $\bar{\mathbb{P}} = \frac{1}{D}\sum_{i=1}^{D} \mathbb{P}_{\theta_i}$. Le Cam's method gives a lower bound for the maximum estimation risk over the parameter set Θ.

Lemma 1 (Le Cam) *Let T be any estimator of θ based on an observation X from a distribution $\mathbb{P}_\theta$ with $\theta \in \Theta = \{\theta_0, \theta_1, \ldots, \theta_D\}$, then*

$$\sup_{\theta \in \Theta} \mathbb{E}_{\mathbf{X}|\theta} L\left(T, \theta\right) \geqslant \frac{1}{2} l_{\min} \left\|\mathbb{P}_{\theta_0} \wedge \bar{\mathbb{P}}\right\|. \tag{5}$$

Write $\Theta_1 = \{\theta_1, \ldots, \theta_D\}$. One can view the lower bound in (5) as obtained from testing the simple hypothesis $H_0 : \theta = \theta_0$ against the composite alternative $H_1 : \theta \in \Theta_1$.

Assouad's Lemma works with a hypercube $\Theta = \{0,1\}^r$. It is based on testing a number of pairs of simple hypotheses and is connected to multiple comparisons. For a parameter $\theta = (\theta_1, ..., \theta_r)$ where $\theta_i \in \{0,1\}$, one tests whether $\theta_i = 0$ or 1 for each $1 \leqslant i \leqslant r$ based on the observation X. For each pair of simple hypotheses, there is a certain loss for making an error in the comparison. The lower bound given by Assouad's Lemma is a combination of losses from testing all pairs of simple hypotheses. Let

$$H\left(\theta, \theta'\right) = \sum_{i=1}^{r} \left|\theta_i - \theta_i'\right| \tag{6}$$

be the Hamming distance on Θ. Assouad's Lemma gives a lower bound for the maximum risk over the hypercube Θ of estimating an arbitrary quantity $\psi(\theta)$ belonging to a metric space with metric d.

Lemma 2 (Assouad) *Let $X \sim \mathbb{P}_\theta$ with $\theta \in \Theta = \{0,1\}^r$ and let $T = T(X)$ be an estimator of $\psi(\theta)$ based on X. Then for all $s > 0$*

(7)

$$\max_{\theta \in \Theta} 2^s \mathbb{E}_{\mathbf{X}|\theta} d^s\left(T, \psi(\theta)\right) \geqslant \min_{H(\theta,\theta') \geqslant 1} \frac{d^s\left(\psi(\theta), \psi(\theta')\right)}{H\left(\theta, \theta'\right)} \cdot \frac{r}{2} \cdot \min_{H(\theta,\theta')=1} \left\|\mathbb{P}_\theta \wedge \mathbb{P}_{\theta'}\right\|.$$

We now introduce our new lower bound technique. Again, let $X \sim \mathbb{P}_\theta$ where $\theta \in \Theta$. The parameter space Θ of interest has a special structure which can be viewed as the Cartesian product of two components Γ and Λ. For a given positive integer r and a finite set $B \subset \mathbb{R}^p \backslash \{\mathbf{0}_{1\times p}\}$, let $\Gamma = \{0,1\}^r$ and $\Lambda \subseteq B^r$. Define

$$\Theta = \Gamma \otimes \Lambda = \{\theta = (\gamma, \lambda) : \gamma \in \Gamma \text{ and } \lambda \in \Lambda\}. \tag{8}$$

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 the Le Cam's method typically applies to a parameter set like Λ with $r = 1$. For $\theta = (\gamma, \lambda) \in \Theta$, denote the projection of θ to Γ by $\gamma(\theta) = \gamma$ and to Λ by $\lambda(\theta) = \lambda$.

It is important to understand the structure of the parameter space Θ. One can view an element $\lambda \in \Lambda$ as an $r \times p$ matrix with each row coming from the set B and view Γ as a set of parameters along the rows indicating whether a given row of λ is present or not. Let D_Λ =Card(Λ). For a given $a \in \{0,1\}$ and $1 \leqslant i \leqslant r$, denote $\Theta_{i,a} = \{\theta \in \Theta : \gamma_i(\theta) = a\}$ where $\theta = (\gamma, \lambda)$

and $\gamma_i(\theta)$ is the $i-$th coordinate of of the first component of θ. It is easy to see that $\mathrm{Card}(\Theta_{i,a}) = 2^{r-1}D_\Lambda$. Define the mixture distribution $\bar{\mathbb{P}}_{i,a}$ by

$$\bar{\mathbb{P}}_{i,a} = \frac{1}{2^{r-1}D_\Lambda} \sum_{\theta \in \Theta_{i,a}} \mathbb{P}_\theta \tag{9}$$

So $\bar{\mathbb{P}}_{i,a}$ is the mixture distribution over all $\mathbb{P}_\theta$ with $\gamma_i(\theta)$ fixed to be a while all other components of θ vary over all possible values in Θ.

The following lemma gives a lower bound for the maximum risk over the parameter set Θ of estimating a functional $\psi(\theta)$ belonging to a metric space with metric d.

Lemma 3 *For any $s > 0$ and any estimator T of $\psi(\theta)$ based on an observation from the experiment $\{\mathbb{P}_\theta, \theta \in \Theta\}$ where Θ is given in (8),*

$$\max_{\Theta} 2^s \mathbb{E}_{\mathbf{X}|\theta} d^s\left(T, \psi\left(\theta\right)\right) \geqslant \alpha \frac{r}{2} \min_{1 \leqslant i \leqslant r} \left\| \bar{\mathbb{P}}_{i,0} \wedge \bar{\mathbb{P}}_{i,1} \right\| \tag{10}$$

where $\bar{\mathbb{P}}_{i,a}$ is defined in Equation (9) and α is given by

$$\alpha = \min_{\{(\theta,\theta'): H(\gamma(\theta),\gamma(\theta')) \geqslant 1\}} \frac{d^s(\psi(\theta), \psi(\theta'))}{H(\gamma(\theta), \gamma(\theta'))}. \tag{11}$$

The idea behind this new lower bound argument is similar to the one for Assouad's Lemma, but in a more complicated setting. Based on an observation $X \sim \mathbb{P}_\theta$ where $\theta = (\gamma, \lambda) \in \Theta = \Gamma \otimes \Lambda$, we wish to test whether $\gamma_i = 0$ or 1 for each $1 \leqslant i \leqslant r$. The first factor α in the lower bound (10) is the minimum cost of making an error per comparison. The second factor $r/2$ is the expected number of errors one makes to estimate γ when $\mathbb{P}_\theta$ and $\mathbb{P}_{\theta'}$ are indistinguishable from each other in the case $H\left(\gamma(\theta), \gamma(\theta')\right) = r$, and the last factor is the lower bound for the total probability of making type I and type II errors for each comparison. A major difference is that in this third factor the distributions $\bar{\mathbb{P}}_{i,0}$ and $\bar{\mathbb{P}}_{i,1}$ are both complicated mixture distributions instead of the typically simple ones in Assouad's Lemma. This makes the lower bound argument more generally applicable, while the calculation of the affinity becomes much more difficult.

In applications of Lemma 3, for a $\gamma = (\gamma_1, ..., \gamma_r) \in \Gamma$ where γ_i takes value 0 or 1, and a $\lambda = (\lambda_1, ..., \lambda_r) \in \Lambda$ where each $\lambda_i \in B$ is a p-dimensional nonzero row vector, the element $\theta = (\gamma, \lambda) \in \Theta$ can be equivalently viewed as an $r \times p$ matrix

$$\begin{pmatrix} \gamma_1 \cdot \lambda_1 \\ \gamma_2 \cdot \lambda_2 \\ \vdots \\ \gamma_r \cdot \lambda_r \end{pmatrix} \tag{12}$$

where the product $\gamma_i \cdot \lambda_i$ is taken elementwise: $\gamma_i \cdot \lambda_i = \lambda_i$ if $\gamma_i = 1$ and the ith row of θ is the zero vector if $\gamma_i = 0$. The term $\|\bar{\mathbb{P}}_{i,0} \wedge \bar{\mathbb{P}}_{i,1}\|$ of Equation (10) is then the lower bound for the total probability of making type I and type II errors for testing whether or not the ith row of θ is zero.

Note that the lower bound (10) reduces to the classical Assouad's Lemma when Λ contains only one matrix for which every row is nonzero, and becomes a two-point argument of Le Cam with one point against a mixture when $r = 1$. The proof of this lemma is given in Section 7. The technical argument is an extension of that of Assouad's Lemma. See Assouad (1983), Yu (1997) and van der Vaart (1998).

The advantage of this method is to break down the lower bound calculations for the whole matrix estimation problem into calculations for individual rows so that the overall analysis is simplified and more tractable. Although the tool is introduced here for the purpose of estimating a sparse covariance matrix, it is of independent interest and is expected to be useful for solving other matrix estimation problems as well.

Bounding the total variation affinity between two mixture distributions in (10) is quite challenging in general. The following well known result on the affinity is helpful in some applications. It provides lower bounds for the affinity between two mixture distributions in terms of the affinities between simpler distributions in the mixtures.

Lemma 4 *Let* $\overline{\mathbb{P}}_m = \sum_{i=1}^m w_i \mathbb{P}_i$ *and* $\overline{\mathbb{Q}}_m = \sum_{i=1}^m w_i \mathbb{Q}_i$ *where* $w_i \geqslant 0$ *and* $\sum_{i=1}^m w_i = 1$. *Then*

$$\left\|\overline{\mathbb{P}}_m \wedge \overline{\mathbb{Q}}_m\right\| \geqslant \sum_{i=1}^m w_i \left\|\mathbb{P}_i \wedge \mathbb{Q}_i\right\| \geqslant \min_{1 \leqslant i \leqslant m} \left\|\mathbb{P}_i \wedge \mathbb{Q}_i\right\|.$$

More specifically, in our construction of the parameter set for establishing the minimax lower bound, r is the number of possibly non-zero 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. Recall that the projection of $\theta \in \Theta$ to Γ is $\gamma(\theta) = \gamma = (\gamma_i(\theta))_{1 \leqslant i \leqslant r}$ and the projection of θ to Λ is $\lambda(\theta) = \lambda = (\lambda_i(\theta))_{1 \leqslant i \leqslant r}$. More generally, for a subset $A \subseteq \{1, 2, \ldots, r\}$, we define a projection of θ to a subset of Γ by $\gamma_A(\theta) = (\gamma_i(\theta))_{i \in A}$. A particularly useful example of set A is

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

for which $\gamma_{\{-i\}}(\theta) = (\gamma_1(\theta), \ldots, \gamma_{i-1}(\theta), \gamma_{i+1}(\theta), \gamma_r(\theta))$ and in this case for convenience we set $\gamma_{-i} = \gamma_{\{-i\}}$. $\lambda_A(\theta)$ and $\lambda_{-i}(\theta)$ are defined similarly. We also define the set $\Lambda_A = \{\lambda_A(\theta) : \theta \in \Theta\}$. A special case is $A = \{-i\}$.

Now we define a subset of Θ to reduce the problem of estimating Θ to a problem of estimating $\lambda_i(\theta)$. For $a \in \{0,1\}$, $b \in \{0,1\}^{r-1}$, and $c \in \Lambda_{-i} \subseteq B^{r-1}$, let

$$\Theta_{(i,a,b,c)} = \{\theta \in \Theta : \gamma_i(\theta) = a, \gamma_{-i}(\theta) = b \text{ and } \lambda_{-i}(\theta) = c\}$$

and $D_{(i,b,c)} = \text{Card}(\Theta_{(i,a,b,c)})$. Note that the cardinality of $\Theta_{(i,a,b,c)}$ on the right hand side does not depend on the value of a due to the Cartesian product structure of $\Theta = \Gamma \otimes \Lambda$. Define the mixture distribution

$$\bar{\mathbb{P}}_{(i,a,b,c)} = \frac{1}{D_{(i,b,c)}} \sum_{\theta \in \Theta_{(i,a,b,c)}} \mathbb{P}_\theta. \tag{13}$$

In other words, $\bar{\mathbb{P}}_{(i,a,b,c)}$ 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. It is helpful to observe that when $a = 0$ we have $\gamma_i(\theta) \cdot \lambda_i(\theta) = 0$ for which $\bar{\mathbb{P}}_{(i,a,b,c)}$ is degenerate in the sense that it is an average of identical distributions.

Lemmas 3 and 4 together immediately imply the following result which is based on the total variation affinities between slightly less complicated mixture distributions. We need to introduce a new notation $\tilde{\mathbb{E}}_\theta$ to denote the average of a function g over Θ, i.e.,

$$\tilde{\mathbb{E}}_\theta g(\theta) = \sum_{\theta \in \Theta} \frac{1}{2^{r-1} D_\Lambda} g(\theta).$$

The parameter θ is seen uniformly distributed over Θ. Let

$$\begin{aligned} \Theta_{-i} &= \{0,1\}^{r-1} \otimes \Lambda_{-i} \\ &= \{(b,c) : \exists\, \theta \in \Theta \text{ such that } \gamma_{-i}(\theta) = b \text{ and } \lambda_{-i}(\theta) = c\}, \end{aligned}$$

and an average of $h(\gamma_{-i}, \lambda_{-i})$ over the set Θ_{-i} is defined as follows

$$\tilde{\mathbb{E}}_{(\gamma_{-i},\lambda_{-i})} h(\gamma_{-i}, \lambda_{-i}) = \sum_{(b,c) \in \Theta_{-i}} \frac{D_{i,b,c}}{2^{r-1} D_\Lambda} h(b,c)$$

where the distribution of $(\gamma_{-i}, \lambda_{-i})$ is induced by the uniform distribution over Θ.

Corollary 1 *For any $s > 0$ and any estimator T of $\psi(\theta)$ based on an observation from the experiment $\{\mathbb{P}_\theta, \theta \in \Theta\}$ where the parameter space Θ is*

given in (8),

$$\begin{aligned} & \max_{\Theta} 2^s \mathbb{E}_{\mathbf{X}|\theta} d^s\left(T, \psi\left(\theta\right)\right) \\ (14) \qquad \geqslant\ & \alpha \frac{r}{2} \min_i \tilde{\mathbb{E}}_{(\gamma_{-i},\lambda_{-i})} \left\| \bar{\mathbb{P}}_{(i,0,\gamma_{-i},\lambda_{-i})} \wedge \bar{\mathbb{P}}_{(i,1,\gamma_{-i},\lambda_{-i})} \right\| \\ (15) \qquad \geqslant\ & \alpha \frac{r}{2} \min_i \min_{\gamma_{-i},\lambda_{-i}} \left\| \bar{\mathbb{P}}_{(i,0,\gamma_{-i},\lambda_{-i})} \wedge \bar{\mathbb{P}}_{(i,1,\gamma_{-i},\lambda_{-i})} \right\|, \end{aligned}$$

where α and $\bar{\mathbb{P}}_{i,a,b,c}$ are defined in Equations (11) and (13) respectively.

Remark 1 A key technical step in applying Lemma 3 in a typical application is to show that the affinity $\left\| \bar{\mathbb{P}}_{i,0} \wedge \bar{\mathbb{P}}_{i,1} \right\|$ is uniformly bounded away from 0 by a constant for all i. Then the term αr on the right hand side of Equation (10) in Lemma 3 gives the lower bound for the minimax rate of convergence. As mentioned earlier, the affinity calculations for two mixture distributions can be very much involved. Corollary 1 gives two lower bounds in terms of the affinities. As noted earlier, $\bar{\mathbb{P}}_{(i,0,\gamma_{-i},\lambda_{-i})}$ in the affinity in Equations (14) and (15) is in fact a single normal distribution, not a mixture. Thus the lower bounds given in Equations (14) and (15) require simpler, although still involved, calculations. In this paper we will apply Equation (14), which has an average of affinities on the right hand side.

3. Lower Bound for Estimating Sparse Covariance Matrix under the Spectral Norm. We now turn to the minimax lower bound for estimating sparse covariance matrices under the spectral norm. We shall apply the lower bound technique developed in the previous section to establish rate sharp results. The same lower bound also holds under the general ℓ_w norm for $1 \leqslant w \leqslant \infty$. Upper bounds are discussed in Section 4 and optimal estimation under Bregman divergence losses is considered in Section 5.

In this section we shall focus on the Gaussian case and wish to estimate the covariance matrix $\Sigma_{p\times p}$ under the spectral norm based on the sample $\mathbf{X}_1, \ldots, \mathbf{X}_n \overset{iid}{\sim} N(\mu, \Sigma_{p\times p})$. The parameter space $\mathcal{G}_q(c_{n,p})$ for sparse covariance matrices is defined as in (1). In the special case of $q = 0$, $\mathcal{G}_0(c_{n,p})$ contains matrices with at most $c_{n,p} + 1$ nonzero elements on each row/column. The parameter space $\mathcal{G}_q(c_{n,p})$ also contains the uniformity class $\mathcal{G}_q^*(c_{n,p})$ considered in Bickel and Levina (2008b) as a special case, where $\mathcal{G}_q^*(c_{n,p})$ is defined as, for $0 \leqslant q < 1$,

$$(16) \qquad \mathcal{G}_q^*(c_{n,p}) = \left\{ \Sigma = (\sigma_{ij})_{1\leqslant i,j\leqslant p} : \max_{j\leqslant p, j\neq i} \sum_{i\neq j} |\sigma_{ij}|^q \leqslant c_{n,p} \right\}.$$

The columns of $\Sigma \in \mathcal{G}_q^*(c_{n,p})$ are assumed to belong to a strong ℓ_q ball.

We now state and prove the minimax lower bound for estimating a sparse covariance matrix over the parameter space $\mathcal{G}_q(c_{n,p})$ under the spectral norm. The derivation of the lower bounds relies heavily on the general lower bound technique developed in the previous section. It also requires a careful construction of a finite subset of the parameter space and detailed calculations of an effective lower bound for the total variation affinities between mixtures of multivariate Gaussian distributions.

Theorem 2 *Let $\mathbf{X}_1, \ldots, \mathbf{X}_n \overset{iid}{\sim} N(\mu, \Sigma_{p\times p})$. The minimax risk for estimating the covariance matrix Σ over the parameter space $\mathcal{G}_q(c_{n,p})$ with $c_{n,p} \leqslant Mn^{\frac{1-q}{2}} (\log p)^{-\frac{3-q}{2}}$ satisfies*

$$\inf_{\hat{\Sigma}} \sup_{\Sigma\in\mathcal{G}_q(c_{n,p})} \mathbb{E}_{\mathbf{X}|\Sigma} \left\|\hat{\Sigma} - \Sigma\right\|^2 \geqslant c \left(c_{n,p}^2 \left(\frac{\log p}{n}\right)^{1-q} + \frac{\log p}{n} \right). \tag{17}$$

for some constant $c > 0$, where $\|\cdot\|$ denotes the matrix spectral norm.

Theorem 2 yields immediately a minimax lower bound for the more general subgaussian case under the assumption (2),

$$\inf_{\hat{\Sigma}} \sup_{\theta\in\mathcal{P}_q(\tau,c_{n,p})} \mathbb{E}_{\mathbf{X}|\theta} \left\|\hat{\Sigma} - \Sigma\right\|^2 \geqslant c \left(c_{n,p}^2 \left(\frac{\log p}{n}\right)^{1-q} + \frac{\log p}{n} \right).$$

It has been shown in Cai, Zhang and Zhou (2010) that

$$\inf_{\hat{\Sigma}} \sup_{\theta\in\mathcal{P}_q(\tau,c_{n,p})} \mathbb{E}_{\mathbf{X}|\theta} \left\|\hat{\Sigma} - \Sigma\right\|^2 \geqslant c\frac{\log p}{n}$$

by constructing a parameter space with only diagonal matrices. It then suffices to show that

$$\inf_{\hat{\Sigma}} \sup_{\theta\in\mathcal{P}_q(\tau,c_{n,p})} \mathbb{E}_{\mathbf{X}|\theta} \left\|\hat{\Sigma} - \Sigma\right\|^2 \geqslant c \cdot c_{n,p}^2 \left(\frac{\log p}{n}\right)^{1-q}$$

to establish Theorem 2.

The proof of Theorem 2 contains three major steps. In the first step we construct in detail a finite subset $\mathcal{F}_*$ of the parameter space $\mathcal{G}_q(c_{n,p})$ such that the difficulty of estimation over $\mathcal{F}_*$ is essentially the same as that of estimation over $\mathcal{G}_q(c_{n,p})$. The second step is the application of Lemma 3 to the carefully constructed parameter set $\mathcal{F}_*$. Finally in the third step we calculate the factor α defined in (11) and the total variation affinity between two multivariate normal mixtures. Bounding the affinity is technically involved. The main ideas of the proof are outlined here and detailed proofs of some technical lemmas used here are deferred to Section 7.

Proof of Theorem 2: The proof is divided into three main steps.

Step 1: Constructing the parameter set. Let $r = \lfloor p/2 \rfloor$, where $\lfloor x \rfloor$ denotes the largest integer less than or equal to x, and let B be the collection of all row vectors $b = (v_j)_{1 \leqslant j \leqslant p}$ such that $v_j = 0$ for $1 \leqslant j \leqslant p-r$ and $v_j = 0$ or 1 for $p-r+1 \leqslant j \leqslant p$ under the constraint the total number of 1's is $\|b\|_0 = k$, where the value of k will be specified later. We shall treat each $(b_1, ..., b_r) \in B^r$ as an $r \times p$ matrix with the ith row equal to b_i.

Set $\Gamma = \{0,1\}^r$. Define $\Lambda \subset B^r$ to be the set of all elements in B^r such that each column sum is less than or equal to $2k$. For each component λ_m, $1 \leqslant m \leqslant r$, of $\lambda = (\lambda_1, ..., \lambda_r) \in \Lambda$, define a $p \times p$ symmetric matrix $A_m(\lambda_m)$ by making the m-th row of $A_m(\lambda_m)$ equal to λ_m, the m-th column equal to λ_m^T, and the rest of the entries 0. Note that for each $\lambda = (\lambda_1, ..., \lambda_r) \in \Lambda$, each column/row sum of the matrix $\sum_{m=1}^r A_m(\lambda_m)$ is less than or equal to $2k$.

Define

$$\Theta = \Gamma \otimes \Lambda \tag{18}$$

and let $\epsilon_{n,p} \in \mathbb{R}$ be fixed. (The exact value of $\epsilon_{n,p}$ will be chosen later.) For each $\theta = (\gamma, \lambda) \in \Theta$ with $\gamma = (\gamma_1, ..., \gamma_r) \in \Gamma$ and $\lambda = (\lambda_1, ..., \lambda_r) \in \Lambda$, we associate θ with a covariance matrix $\Sigma(\theta)$ by

$$\Sigma(\theta) = I_p + \epsilon_{n,p} \sum_{m=1}^{r} \gamma_m A_m(\lambda_m). \tag{19}$$

It is easy to see that in the Gaussian case $\|\!|\Sigma_{p\times p}|\!\| \leqslant \tau$ is a sufficient condition for (2). Without loss of generality we assume that $\tau > 1$ in the subgaussianity assumption (2), otherwise we replace I_p in (19) by cI_p with a small constant $c > 0$. Finally we define a collection $\mathcal{F}_*$ of covariance matrices as

$$\mathcal{F}_* = \left\{ \Sigma(\theta) : \Sigma(\theta) = I_p + \epsilon_{n,p} \sum_{m=1}^{r} \gamma_m A_m(\lambda_m),\ \theta = (\gamma, \lambda) \in \Theta \right\}. \tag{20}$$

Note that each $\Sigma \in \mathcal{F}_*$ has value 1 along the main diagonal, and contains an $r \times r$ submatrix, say, A, at the upper right corner, A^T at the lower left corner, and 0 elsewhere. Each row of A is either identically 0 (if the corresponding γ value is 0) or has exactly k nonzero elements with value $\epsilon_{n,p}$.

We now specify the values of $\epsilon_{n,p}$ and k to ensure $\mathcal{F}_* \subset \mathcal{G}_q(c_{n,p})$. Set $\epsilon_{n,p} = \upsilon\sqrt{\frac{\log p}{n}}$ for a fixed small constant υ, and let $k = \max\left(\left\lceil \frac{1}{2} c_{n,p} \epsilon_{n,p}^{-q} \right\rceil - 1,\ 0\right)$ which implies

$$\max_{1 \leqslant j \leqslant p} \sum_{i \neq j} |\sigma_{ij}|^q \leqslant 2k\epsilon_{n,p}^q \leqslant c_{n,p}.$$

We require

$$0 < \upsilon < \left[\min\left\{\frac{1}{3}, \tau - 1\right\}\frac{1}{M}\right]^{\frac{1}{1-q}} \quad \text{and} \quad \upsilon^2 < \frac{\beta - 1}{54\beta}. \tag{21}$$

Note that $\epsilon_{n,p}$ and k satisfy

$$2k\epsilon_{n,p} \leqslant c_{n,p}\epsilon_{n,p}^{1-q} \leqslant M\upsilon^{1-q} < \min\left\{\frac{1}{3}, \tau - 1\right\} \tag{22}$$

and consequently every $\Sigma(\theta)$ is diagonally dominant and positive definite, and $\|\Sigma(\theta)\| \leqslant \|\Sigma(\theta)\|_1 \leqslant 2k\epsilon_{n,p} + 1 < \tau$. Thus we have $\mathcal{F}_* \subset \mathcal{G}_q(c_{n,p})$, and the subgaussianity assumption (2) is satisfied.

Step 2: Applying the general lower bound argument. Let $\mathbf{X}_1, \ldots, \mathbf{X}_n \overset{iid}{\sim} N(0, \Sigma(\theta))$ with $\theta \in \Theta$ and denote the joint distribution by $\mathbb{P}_\theta$. Applying Lemma 3 to the parameter space Θ with $s = 2$, we have

$$\inf_{\hat{\Sigma}} \max_{\theta\in\Theta} 2^2 \mathbb{E}_{\mathbf{X}|\theta} \left\|\hat{\Sigma} - \Sigma(\theta)\right\|^2 \geqslant \alpha \cdot \frac{r}{2} \cdot \min_{1\leqslant i\leqslant r} \left\|\bar{\mathbb{P}}_{i,0} \wedge \bar{\mathbb{P}}_{i,1}\right\| \tag{23}$$

where

$$\alpha \equiv \min_{\{(\theta,\theta'):H(\gamma(\theta),\gamma(\theta'))\geqslant 1\}} \frac{\|\Sigma(\theta) - \Sigma(\theta')\|^2}{H(\gamma(\theta), \gamma(\theta'))} \tag{24}$$

and $\bar{\mathbb{P}}_{i,0}$ and $\bar{\mathbb{P}}_{i,1}$ are defined as in (9).

Step 3: Bounding the affinity and per comparison loss. We shall now bound the two factors α and $\min_i \left\|\bar{\mathbb{P}}_{i,0} \wedge \bar{\mathbb{P}}_{i,1}\right\|$ in (23). This is done separately in the next two lemmas which are proved in detailed in Section 7. Lemma 5 gives a lower bound to the per comparison loss and it is easy to prove.

Lemma 5 *For α defined in Equation (24) we have*

$$\alpha \geqslant \frac{(k\epsilon_{n,p})^2}{p}.$$

The key technical difficulty is in bounding the affinity between the Gaussian mixtures. The proof is quite involved.

Lemma 6 *Let $\mathbf{X}_1, \ldots, \mathbf{X}_n \overset{iid}{\sim} N(0, \Sigma(\theta))$ with $\theta \in \Theta$ defined in Equation (18) and denote the joint distribution by $\mathbb{P}_\theta$. For $a \in \{0, 1\}$ and $1 \leqslant i \leqslant r$, define $\bar{\mathbb{P}}_{i,a}$ as in (9). Then there exists a constant $c_1 > 0$ such that*

$$\min_{1\leqslant i\leqslant r} \left\|\bar{\mathbb{P}}_{i,0} \wedge \bar{\mathbb{P}}_{i,1}\right\| \geqslant c_1.$$

Finally, the minimax lower bound for estimation over $\mathcal{G}_q(c_{n,p})$ is obtained by putting together the bounds given in Lemmas 5 and 6,

$$\begin{aligned}\inf_{\hat{\Sigma}} \sup_{\Sigma\in\mathcal{G}_q(c_{n,p})} \mathbb{E}_{\mathbf{X}|\Sigma} \left\|\hat{\Sigma}-\Sigma\right\|^2 &\geqslant \inf_{\hat{\Sigma}} \max_{\Sigma(\theta)\in\mathcal{F}_*} \mathbb{E}_{\mathbf{X}|\theta} \left\|\hat{\Sigma}-\Sigma\left(\theta\right)\right\|^2 \geqslant \frac{\left(k\epsilon_{n,p}\right)^2}{p}\cdot\frac{r}{8}\cdot c_1 \\ &\geqslant c_2 c_{n,p}^2\left(\frac{\log p}{n}\right)^{1-q},\end{aligned}$$

for some constant $c_2 > 0$. ∎

Remark 2 It is easy to check that the proof of Theorem 2 also yields a lower bound for estimation under the general matrix ℓ_w operator norm for any $1 \leqslant w \leqslant \infty$,

$$\inf_{\hat{\Sigma}} \sup_{\theta\in\mathcal{P}_q(\tau,c_{n,p})} \mathbb{E}_{\mathbf{X}|\theta} \left\|\hat{\Sigma}-\Sigma\right\|_w^2 \geqslant c\left(c_{n,p}^2\left(\frac{\log p}{n}\right)^{1-q}+\frac{\log p}{n}\right)$$

by applying Lemma 3 with $s = 1$.

4. Minimax Upper Bound under the Spectral Norm. Section 3 developed a minimax lower bound for estimating a sparse covariance matrix under the spectral norm over $\mathcal{G}_q(c_{n,p})$. In this section we shall show that the lower bound is rate-sharp and therefore establish the optimal rate of convergence. To derive a minimax upper bound, we shall consider the properties of a thresholding estimator introduced in Bickel and Levina (2008b). Given a random sample $\{\mathbf{X}_1,\ldots,\mathbf{X}_n\}$ of p -variate observations drawn from a distribution in $\mathcal{P}_q(\tau, c_{n,p})$, the sample covariance matrix is

$$\frac{1}{n-1}\sum_{l=1}^{n}\left(\mathbf{X}_l-\bar{\mathbf{X}}\right)\left(\mathbf{X}_l-\bar{\mathbf{X}}\right)^T,$$

which is an unbiased estimate of Σ, and the maximum likelihood estimator of Σ is

$$\Sigma^* = (\sigma_{ij}^*)_{1\leqslant i,j\leqslant p} = \frac{1}{n}\sum_{l=1}^{n}\left(\mathbf{X}_l-\bar{\mathbf{X}}\right)\left(\mathbf{X}_l-\bar{\mathbf{X}}\right)^T \tag{25}$$

when $\mathbf{X}_l$'s are normally distributed. These two estimators are close to each other for large n. We shall construct estimators of the covariance matrix Σ by thresholding the maximum likelihood estimator Σ^*.

Note that the subgaussianity condition (2) implies

$$\|\Sigma\| = \sup_{\mathbf{v}:\|\mathbf{v}\|=1} \mathrm{Var}\left[\mathbf{v}^T(\mathbf{X}_1-\mathbb{E}\mathbf{X}_1)\right] \leqslant \int_0^\infty e^{-x/(2\tau)}dx = 2\tau.$$

Then the empirical covariance $\sigma^*_{i,j}$ satisfies the following large deviation result that there exist constants $C_1 > 0$ and $\gamma > 0$ such that

$$\mathbb{P}\left(\left|\sigma^*_{ij} - \sigma_{ij}\right| > t\right) \leqslant C_1 \exp\left(-\frac{8}{\gamma^2} n t^2\right) \tag{26}$$

for $|t| \leqslant \delta$, where C_1, γ and δ are constants and depend only on τ. See Saulis and Statulevičius (1991) and Bickel and Levina (2008a). The inequality (26) implies σ^*_{ij} behaves like a subgaussian random variable. In particular for $t = \gamma\sqrt{\frac{\log p}{n}}$ we have

$$\mathbb{P}\left(\left|\sigma^*_{ij} - \sigma_{ij}\right| > t\right) \leqslant C_1 p^{-8}. \tag{27}$$

Define the thresholding estimator $\hat{\Sigma} = (\hat{\sigma}_{ij})_{p\times p}$ by

$$\hat{\sigma}_{ij} = \sigma^*_{ij} \cdot I\left(|\sigma^*_{ij}| \geqslant \gamma\sqrt{\frac{\log p}{n}}\right). \tag{28}$$

This thresholding estimator was first proposed in Bickel and Levina (2008b) in which a rate of convergence of the loss function in probability was given over the uniformity class $\mathcal{G}^*_q(c_{n,p})$. Here we provide an upper bound for mean squared spectral norm error over the parameter space $\mathcal{G}_q(c_{n,p})$.

Throughout the rest of the paper we denote by C a generic positive constant which may vary from place to place. The following theorem shows that the thresholding estimator defined in (28) is rate optimal over the parameter space $\mathcal{G}_q(c_{n,p})$.

Theorem 3 *The thresholding estimator $\hat{\Sigma}$ given in (28) satisfies, for some constant $C > 0$,*

$$\sup_{\theta\in\mathcal{P}_q(\tau,c_{n,p})} \mathbb{E}_{\mathbf{X}|\theta} \left\|\hat{\Sigma} - \Sigma\right\|^2 \leqslant C\left[c^2_{n,p}\left(\frac{\log p}{n}\right)^{1-q} + \frac{\log p}{n}\right]. \tag{29}$$

Consequently, the minimax risk of estimating the sparse covariance matrix Σ over $\mathcal{G}_q(c_{n,p})$ satisfies

$$\inf_{\hat{\Sigma}} \sup_{\theta\in\mathcal{P}_q(\tau,c_{n,p})} \mathbb{E}_{\mathbf{X}|\theta} \left\|\hat{\Sigma} - \Sigma\right\|^2 \asymp c^2_{n,p}\left(\frac{\log p}{n}\right)^{1-q} + \frac{\log p}{n}. \tag{30}$$

Remark 3 A similar argument to the proof of Equation (29) in Section 7.4 yields the following upper bound for estimation under the matrix ℓ_1 norm,

$$\sup_{\theta\in\mathcal{P}_q(\tau,c_{n,p})} \mathbb{E}_{\mathbf{X}|\theta} \left\|\hat{\Sigma} - \Sigma\right\|^2_1 \leqslant C\left[c^2_{n,p}\left(\frac{\log p}{n}\right)^{1-q} + \frac{\log p}{n}\right].$$

Theorem 3 shows that the optimal rate of convergence for estimating a sparse covariance matrix over $\mathcal{G}_q(c_{n,p})$ under the squared spectral norm is $c_{n,p}^2\left(\frac{\log p}{n}\right)^{1-q}$. In Bickel and Levina (2008b) the uniformity class $\mathcal{G}_q^*(c_{n,p})$ defined in (16) was considered. We shall now show that the same minimax rate of convergence holds for estimation over $\mathcal{G}_q^*(c_{n,p})$. It is easy to check in the proof of the lower bound that for every $\Sigma \in \mathcal{F}_*$ defined in (20) we have

$$\max_{1\leqslant j\leqslant p}\sum_{i\neq j}|\sigma_{ij}|^q \leqslant 2k\epsilon_{n,p}^q \leqslant c_{n,p}$$

and consequently $\mathcal{F}_* \subset \mathcal{G}_q^*(c_{n,p})$. Thus the lower bound established for $\mathcal{F}_*$ automatically yields a lower bound for $\mathcal{G}_q^*(c_{n,p})$. On the other hand, since a strong ℓ_q ball is always contained in a weak ℓ_q ball by the Markov inequality, the upper bound in Equation (29) for the parameter space $\mathcal{G}_q$ also holds for $\mathcal{G}_q^*(c_{n,p})$. Let $\mathcal{P}_q^*(\tau, c_{n,p})$ denote the set of distributions of $\mathbf{X}_1$ satisfying (2) and with covariance matrix $\Sigma \in \mathcal{G}_q^*(c_{n,p})$. Then we have the following result.

Proposition 1 *The minimax risk for estimating the covariance matrix under the spectral norm over the uniformity class $\mathcal{G}_q^*(c_{n,p})$ satisfies*

$$\inf_{\hat{\Sigma}}\sup_{\theta\in\mathcal{P}_q^*(\tau,c_{n,p})}\mathbb{E}_{\mathbf{X}|\theta}\left\|\hat{\Sigma}-\Sigma\right\|^2 \asymp c_{n,p}^2\left(\frac{\log p}{n}\right)^{1-q}+\frac{\log p}{n}.$$

The thresholding estimator $\hat{\Sigma}$ defined by (28) is positive definite with high probability, but it is not guaranteed to be positive definite. A simple additional step can make the final estimator positive semi-definite and achieve the optimal rate of convergence. Write the eigen-decomposition of $\hat{\Sigma}$ as

$$\hat{\Sigma} = \sum_{i=1}^p \hat{\lambda}_i v_i v_i^T$$

where $\hat{\lambda}_i$'s and v_i's are the eigenvalues and eigenvectors of $\hat{\Sigma}$, respectively. Let $\hat{\lambda}_i^+ = \max(\hat{\lambda}_i, 0)$ be the positive part of $\hat{\lambda}_i$ and define

$$\hat{\Sigma}^+ = \sum_{i=1}^p \hat{\lambda}_i^+ v_i v_i^T.$$

Then

$$\begin{aligned}\left\|\hat{\Sigma}^+-\Sigma\right\| &\leqslant \left\|\hat{\Sigma}^+-\hat{\Sigma}\right\|+\left\|\hat{\Sigma}-\Sigma\right\| \leqslant \max_{i:\hat{\lambda}_i\leqslant 0}\left|\hat{\lambda}_i\right|+\left\|\hat{\Sigma}-\Sigma\right\| \\ &\leqslant \max_{i:\hat{\lambda}_i\leqslant 0}\left|\hat{\lambda}_i-\lambda_i\right|+\left\|\hat{\Sigma}-\Sigma\right\| \leqslant 2\left\|\hat{\Sigma}-\Sigma\right\|.\end{aligned}$$

The resulting estimator $\hat{\Sigma}^+$ is positive semi-definite and attains the same rate as the original thresholding estimator $\hat{\Sigma}$. This method can be applied to the tapering estimator in Cai, Zhang and Zhou (2010) as well to make the estimator positive semi-definite, while still achieving the optimal rate.

5. Optimal Estimation under Bregman Divergences. We have so far focused on the optimal rate of convergence under the spectral norm. In this section we turn to minimax estimation of sparse covariance matrices under a class of Bregman divergence losses which include the Stein's loss, Frobenius norm, and von Neumann entropy as special cases. Bregman matrix divergences have been used for matrix estimation and matrix approximation problems, see, e.g., Dhillon and Tropp (2007), Ravikumar, et al. (2008), and Kulis, Sustik and Dhillon (2009). In this section we establish the optimal rate of convergence uniformly for a class of Bregman divergence losses.

Bregman (1967) introduced the Bregman divergence as a dissimilarity measure between vectors,

$$D_\phi(\mathbf{x},\mathbf{y}) = \phi(\mathbf{x}) - \phi(\mathbf{y}) - (\nabla\phi(\mathbf{y}))^T(\mathbf{x}-\mathbf{y})$$

where ϕ is a differentiable, real-valued, and strictly convex function defined over a convex set in a Euclidean space $\mathbb{R}^m$ and $\nabla\phi$ is the gradient of ϕ. The well known Mahalanobis distance is a Bregman divergence. This concept can be naturally extended to the space of real and symmetric matrices as

$$D_\phi(X,Y) = \phi(X) - \phi(Y) - tr\left[(\nabla\phi(Y))^T(X-Y)\right],$$

where X and Y are real symmetric matrices and ϕ is a differentiable strictly convex function over the space. See Censor and Zenios (1997) and Kulis, Sustik and Dhillon (2009). A particularly interesting class of ϕ is

$$\phi(X) = \sum_{i=1}^{p} \varphi(\lambda_i) \tag{31}$$

where λ_i's are the eigenvalues of X and φ is a differentiable, real-valued, and strictly convex function over a convex set in $\mathbb{R}$. See Dhillon and Tropp (2007) and Kulis, Sustik and Dhillon (2009). Examples of this class of Bregman divergences include:

- $\varphi(\lambda) = -\log\lambda$, or equivalently $\phi(X) = -\log\det(X)$. The corresponding Bregman divergence can be written as

  $$D_\phi(X,Y) = tr\left(XY^{-1}\right) - \log\det\left(XY^{-1}\right) - p$$

  which is often called the Stein's loss in the statistical literature.

- $\varphi(\lambda) = \lambda\log\lambda - \lambda$, or equivalently $\phi(X) = tr(X\log X - X)$, where X is positive definite such that $\log X$ is well defined. The corresponding Bregman divergence is the von Neumann divergence

$$D_\phi(X,Y) = tr(X\log X - X\log Y - X + Y)$$

- $\varphi(\lambda) = \lambda^2$, or equivalently $\phi(X) = tr\left(X^2\right)$. The resulting Bregman divergence is the squared Frobenius norm

$$D_\phi(X,Y) = tr\left[(X-Y)^2\right] = \|\|X-Y\|\|_F^2 = \sum_{i,j}(x_{ij}-y_{ij})^2$$

for $X = (x_{ij})_{1\leqslant i,j\leqslant p}$ and $Y = (y_{ij})_{1\leqslant i,j\leqslant p}$.

Define a class Ψ of functions φ satisfying the following conditions:

1. φ is twice differentiable, real-valued, and strictly convex over $\lambda \in (0,\infty)$;
2. $|\varphi(\lambda)| \leqslant C\lambda^r$ for some $C > 0$ and some real number r uniformly over $\lambda \in (0,\infty)$;
3. For every positive constants ϵ_2 and M_2 there are some positive constants c_L and c_u depending on ϵ_2 and M_2 such that $c_L \leqslant \varphi''(\lambda) \leqslant c_u$ for all $\lambda \in [\epsilon_2, M_2]$.

In this paper, we shall consider the following class of Bregman divergences:

$$\Phi = \left\{\phi(\Sigma) = \sum_{i=1}^{p}\varphi(\lambda_i) : \varphi \in \Psi\right\}. \tag{32}$$

It is easy to see that the Stein's loss, von Neumann divergence and squared Frobenius norm are in this class.

Let $\epsilon_1 > 0$ be a positive constant. Let $\mathcal{P}_q^B(\tau, c_{n,p})$ denote the set of distributions of $\mathbf{X}_1$ satisfying (2) and with covariance matrix

$$\Sigma \in \mathcal{G}_q^B(c_{n,p}) = \mathcal{G}_q(c_{n,p}) \cap \{\Sigma : \lambda_{\min} \geqslant \epsilon_1\}.$$

Here $\lambda_{\min}$ denotes the minimum eigenvalue of Σ. The assumption that all eigenvalues are bounded away from 0 is necessary when $\varphi(\lambda)$ is not well defined at 0. An example is the Stein's loss where $\varphi(\lambda) = -\log\lambda$. Under this assumption all losses D_ϕ are equivalent to the squared Frobenious norm.

The following theorem gives a unified result on the minimax rate of convergence for estimating the covariance matrix over the parameter space $\mathcal{P}_q^B(\tau, c_{n,p})$ for all Bregman divergences $\phi \in \Phi$ defined in (32).

Theorem 4 *Assume that* $c_{n,p} \leqslant Mn^{\frac{1-q}{2}}(\log p)^{-\frac{3-q}{2}}$ *for some* $M > 0$ *and* $0 \leqslant q < 1$. *The minimax risk over* $\mathcal{P}_q^B(\tau, c_{n,p})$ *under the loss function*

$$\mathrm{L}_\phi\left(\hat{\Sigma}, \Sigma\right) = \frac{1}{p} D_\phi\left(\hat{\Sigma}, \Sigma\right)$$

for all Bregman divergences $\phi \in \Phi$ *defined in (32) satisfies*

$$\inf_{\hat{\Sigma}} \sup_{\phi\in\Phi} \sup_{\theta\in\mathcal{P}_q^B(\tau,c_{n,p})} \mathbb{E}_{\mathbf{X}|\theta}\mathrm{L}_\phi\left(\hat{\Sigma}, \Sigma\right) \asymp c_{n,p}\left(\frac{\log p}{n}\right)^{1-\frac{q}{2}} + \frac{1}{n}. \tag{33}$$

Note that Theorem 4 gives the minimax rate of convergence uniformly under all Bregman divergences defined in (32). For an individual Bregman divergence loss, the condition that all eigenvalues are bounded away from 0 is not needed if the function φ is well behaved at 0. For example, such is the case for the Frobenius norm.

The optimal rate of convergence is attained by a modified thresholding estimator. Let $\hat{\Sigma} = (\hat{\sigma}_{ij})_{1\leqslant i,j\leqslant p}$ be the thresholding estimator given in (28). Define the final estimator of Σ by

$$\hat{\Sigma}_B = \begin{cases} \hat{\Sigma}, & \text{if } \frac{1}{\max\{\log n, \log p\}} \leqslant \lambda_{\min}(\hat{\Sigma}) \leqslant \max\{\log n, \log p\} \\ I, & \text{otherwise.} \end{cases} \tag{34}$$

It will be proved in Section 7.5 that the estimator $\hat{\Sigma}_B$ given in (34) is rate optimal uniformly under all Bregman divergences satisfying (32). Note that the modification of $\hat{\Sigma}$ given in (34) is needed. Without it, the loss $\mathrm{L}_\phi(\hat{\Sigma}, \Sigma)$ may not be well behaved under some Bregman divergences such as the Stein's loss and von Neumann divergence.

Remark 4 Let $\mathcal{P}_q^{*B}(\tau, c_{n,p})$ denote the set of distributions of $\mathbf{X}_1$ satisfying (2) and with covariance matrix $\Sigma \in \mathcal{G}_q^{*B}(c_{n,p}) = \mathcal{G}_q^*(c_{n,p}) \cap \{\Sigma : \lambda_{min} \geqslant \epsilon_1\}$. Then under the same conditions as in Theorem 4,

$$\inf_{\hat{\Sigma}} \sup_{\phi\in\Phi} \sup_{\theta\in\mathcal{P}_q^{*B}(\tau,c_{n,p})} \mathbb{E}_{\mathbf{X}|\theta}\mathrm{L}_\phi\left(\hat{\Sigma}, \Sigma\right) \asymp c_{n,p}\left(\frac{\log p}{n}\right)^{1-\frac{q}{2}} + \frac{1}{n}.$$

6. Discussions. The focus of this paper is mainly on the optimal estimation under the spectral norm. However, both the lower and upper bounds can be easily extended to the general matrix ℓ_w norm for $1 \leqslant w \leqslant \infty$ by using similar arguments given in Sections 3 and 4.

Theorem 5 *Under the assumptions in Theorem 1, the minimax risk of estimating the covariance matrix Σ under the matrix ℓ_w-norm for $1 \leqslant w \leqslant \infty$ over the class $\mathcal{P}_q(\tau, c_{n,p})$ satisfies*

$$\inf_{\hat{\Sigma}} \sup_{\theta \in \mathcal{P}_q(\tau, c_{n,p})} \mathbb{E}_{\mathbf{X}|\theta} \left\| \hat{\Sigma} - \Sigma \right\|_w^2 \asymp c_{n,p}^2 \left(\frac{\log p}{n} \right)^{1-q} + \frac{\log p}{n}. \tag{35}$$

Moreover, the thresholding estimator $\hat{\Sigma}$ defined in (28) is rate-optimal.

As noted in Section 3, a rate-sharp lower bound for the minimax risk under the ℓ_w norm can be obtained by using essentially the same argument with the same parameter space $\mathcal{F}_*$ and a slightly modified version of Lemma 5. The upper bound can be proved by applying the Riesz-Thorin Interpolation Theorem, which yields $\|A\|_w \leqslant \max\{\|A\|_1, \|A\|_2, \|A\|_\infty\}$ for all $w \in [1, \infty)$, and by using the facts $\|A\|_1 = \|A\|_\infty$ and $\|A\|_2 \leqslant \|A\|_1$, when A is symmetric. In Section 4 we have in fact established the same risk bound for both the spectral norm and matrix ℓ_1-norm.

The spectral norm of a matrix depends on the entries in a subtle way and the "interactions" among different rows/columns must be taken into account. The lower bound argument developed in this paper is aimed at treating "two-directional" problems by mixing over both rows and columns. It can be viewed as a simultaneous application of Le Cam's method in one direction and Assouad's Lemma in another. In contrast, for sequence estimation problems we typically need one or the other, but not both at the same time. The lower bound techniques developed in this paper can be used to solve other matrix estimation problems. For example, Cai, Liu and Zhou (2011) applied the general lower bound argument to the problem of estimating sparse precision matrices under the spectral norm and established the optimal rate of convergence. This problem is closely connected to graphical model selection. The derivations of both the lower and upper bounds are involved. For reasons of space, we shall report the results elsewhere.

In this paper we also developed a unified result on the minimax rate of convergence for estimating sparse covariance matrices under a class of Bregman divergence losses which include the commonly used Frobenius norm as a special case. The optimal rate of convergence given in Theorem 4 is identical to the minimax rate for estimating a row/column as a vector with the weak ℓ_q ball constraint under the squared error loss. Our result shows that this class of Bregman divergence losses are essentially the same and thus can be studied simultaneously in terms of the minimax rate of convergence.

Estimating a sparse covariance matrix is intrinsically a heteroscedastic problem in the sense that the variances of the entries of the sample covariance

matrix are not equal and can vary over a wide range. A natural approach is to adaptively threshold the entries according to their individual variabilities. Cai and Liu (2011) considered such an adaptive approach for estimation over the weighted ℓ_q balls which contains the strong ℓ_q balls as subsets. The lower bound given in Proposition 1 in the present paper immediately yields a lower bound for estimation over the weighted ℓ_q balls. A data-driven thresholding procedure was introduced and shown to adaptively achieve the optimal rate of convergence over a large collection of the weighted ℓ_q balls under the spectral norm. In contrast, universal thresholding estimators are sub-optimal over the same parameter spaces.

In addition to the hard thresholding estimator used in Bickel and Levina (2008b), Rothman, Levina and Zhu (2009) considered a class of thresholding rules with more general thresholding functions including soft thresholding and adaptive Lasso. It is straightforward to show that these thresholding estimators with the same choice of threshold level used in (28) also attains the optimal rate of convergence over the parameter space $\mathcal{G}_q(c_{n,p})$ under mean squared spectral norm error as well as under the class of Bregman divergence losses considered in Section 5 with the same modification as in (34). Therefore, the choice of the thresholding function is not important as far as the rate optimality is concerned.

7. Proofs. In this section we prove the general lower bound result given in Lemma 3, Theorems 3 and 4 as well as some of the important technical lemmas used in the proof of Theorem 2 given in Section 3. The proofs of a few technical results used in this section are deferred to the Appendix. Throughout this section, we denote by C a generic constant that may vary from place to place.

7.1. *Proof of Lemma 3.* We first bound the maximum risk by the average over the whole parameter set,

(36)
$$\max_{\Theta} 2^s \mathbb{E}_{\mathbf{X}|\theta} d^s\left(T, \psi\left(\theta\right)\right) \geqslant \frac{1}{2^r D_\Lambda} \sum_{\theta} 2^s \mathbb{E}_{\mathbf{X}|\theta} d^s\left(T, \psi\left(\theta\right)\right) = \frac{1}{2^r D_\Lambda} \sum_{\theta} \mathbb{E}_{\mathbf{X}|\theta}\left[2d\left(T, \psi\left(\theta\right)\right)\right]^s.$$

Set $\hat{\theta} = \arg\min_{\theta\in\Theta} d^s\left(T, \psi\left(\theta\right)\right)$. Note that the minimum is not necessarily unique. When it is not unique, pick $\hat{\theta}$ to be any point in the minimum set. Then the triangle inequality for the metric d gives

(37)
$$\mathbb{E}_{\mathbf{X}|\theta} d^s\left(\psi\left(\hat{\theta}\right), \psi\left(\theta\right)\right) \leqslant \mathbb{E}_{\mathbf{X}|\theta}\left[d\left(\psi\left(\hat{\theta}\right), T\right) + d\left(T, \psi\left(\theta\right)\right)\right]^s \leqslant \mathbb{E}_{\mathbf{X}|\theta}\left[2d\left(T, \psi\left(\theta\right)\right)\right]^s$$

where the last inequality is due to the fact $d\left(\psi\left(\hat{\theta}\right), T\right) = d\left(T, \psi\left(\hat{\theta}\right)\right) \leqslant d\left(T, \psi\left(\theta\right)\right)$ from the definition of $\hat{\theta}$. Equations (36) and (37) together yield

$$\begin{aligned}
\max_{\Theta} 2^s \mathbb{E}_{\mathbf{X}|\theta} d^s\left(T, \psi\left(\theta\right)\right) &\geqslant \frac{1}{2^r D_\Lambda} \sum_{\theta} \mathbb{E}_{\mathbf{X}|\theta} d^s\left(\psi\left(\hat{\theta}\right), \psi\left(\theta\right)\right) \\
&\geqslant \frac{1}{2^r D_\Lambda} \sum_{\theta} \mathbb{E}_{\mathbf{X}|\theta} \frac{d^s\left(\psi\left(\hat{\theta}\right), \psi\left(\theta\right)\right)}{H\left(\gamma\left(\hat{\theta}\right), \gamma\left(\theta\right)\right) \vee 1} \cdot H\left(\gamma\left(\hat{\theta}\right), \gamma\left(\theta\right)\right) \\
&\geqslant \alpha \cdot \frac{1}{2^r D_\Lambda} \sum_{\theta} \mathbb{E}_{\mathbf{X}|\theta} H\left(\gamma\left(\hat{\theta}\right), \gamma\left(\theta\right)\right), \qquad (38)
\end{aligned}$$

where the last step follows from the definition of α in Equation (11).

We now show

$$\frac{1}{2^r D_\Lambda} \sum_{\theta} \mathbb{E}_{\mathbf{X}|\theta} H\left(\gamma\left(\hat{\theta}\right), \gamma\left(\theta\right)\right) \geqslant \frac{r}{2} \min_i \left\| \bar{\mathbb{P}}_{i,0} \wedge \bar{\mathbb{P}}_{i,1} \right\| \qquad (39)$$

which immediately implies $\max_\Theta 2^s \mathbb{E}_{\mathbf{X}|\theta} d^s\left(T, \psi\left(\theta\right)\right) \geqslant \alpha \frac{r}{2} \min_i \left\| \bar{\mathbb{P}}_{i,0} \wedge \bar{\mathbb{P}}_{i,1} \right\|$, and Lemma 3 follows. From the definition of H in Equation (6) we write

$$\frac{1}{2^r D_\Lambda} \sum_{\theta} \mathbb{E}_{\mathbf{X}|\theta} H\left(\gamma\left(\hat{\theta}\right), \gamma\left(\theta\right)\right) = \frac{1}{2^r D_\Lambda} \sum_{\theta} \sum_{i=1}^{r} \mathbb{E}_{\mathbf{X}|\theta} \left| \gamma_i\left(\hat{\theta}\right) - \gamma_i\left(\theta\right) \right|.$$

The right hand side can be further written as

$$\begin{aligned}
&\sum_{i=1}^{r} \frac{1}{2^r D_\Lambda} \sum_{\rho \in \Gamma} \left[\sum_{\{\theta : \gamma(\theta) = \rho\}} \mathbb{E}_{\mathbf{X}|\theta} \left| \gamma_i(\hat{\theta}) - \gamma_i(\theta) \right| \right] \\
= \; & \frac{1}{2} \sum_{i=1}^{r} \left[\frac{1}{2^{r-1} D_\Lambda} \sum_{\{\rho : \rho_i = 0\}} \sum_{\{\theta : \gamma(\theta) = \rho\}} \int \gamma_i(\hat{\theta}) d\mathbb{P}_\theta + \frac{1}{2^{r-1} D_\Lambda} \sum_{\{\rho : \rho_i = 1\}} \sum_{\{\theta : \gamma(\theta) = \rho\}} \int (1 - \gamma_i(\hat{\theta})) d\mathbb{P}_{\theta'} \right] \\
= \; & \frac{1}{2} \sum_{i=1}^{r} \left[\int \gamma_i(\hat{\theta}) \left(\frac{1}{2^{r-1} D_\Lambda} \sum_{\{\rho : \rho_i = 0\}} \sum_{\{\theta : \gamma(\theta) = \rho\}} d\mathbb{P}_\theta \right) + \int (1 - \gamma_i(\hat{\theta})) \left(\frac{1}{2^{r-1} D_\Lambda} \sum_{\{\rho : \rho_i = 1\}} \sum_{\{\theta : \gamma(\theta) = \rho\}} d\mathbb{P}_\theta \right) \right] \\
= \; & \frac{1}{2} \sum_{i=1}^{r} \left[\int \gamma_i(\hat{\theta}) d\bar{\mathbb{P}}_{i,0} + \int (1 - \gamma_i(\hat{\theta})) d\bar{\mathbb{P}}_{i,1} \right].
\end{aligned}$$

The following elementary result is useful to establish the lower bound for the minimax risk. See, for example, page 40 of Le Cam (1973).

Lemma 7 *The total variation affinity satisfies*

$$\|\mathbb{P} \wedge \mathbb{Q}\| = \inf_{0 \leqslant f \leqslant 1} \left\{ \int f d\mathbb{P} + \int (1-f) d\mathbb{Q} \right\}.$$

It follows immediately from Lemma 7 that

$$\frac{1}{2} \sum_{i=1}^{r} \left[\int \gamma_i(\hat{\theta}) d\bar{\mathbb{P}}_{i,0} + \int (1-\gamma_i(\hat{\theta})) d\bar{\mathbb{P}}_{i,1} \right] \geqslant \frac{1}{2} \sum_{i=1}^{r} \left\| \bar{\mathbb{P}}_{i,0} \wedge \bar{\mathbb{P}}_{i,1} \right\| \geqslant \frac{r}{2} \min_i \left\| \bar{\mathbb{P}}_{i,0} \wedge \bar{\mathbb{P}}_{i,1} \right\|$$

and so Equation (39) is established. ∎

7.2. *Proof of Lemma 5.* Let $v = (v_i)$ be a column p-vector with $v_i = 0$ for $1 \leqslant i \leqslant p-r$ and $v_i = 1$ for $p-r+1 \leqslant i \leqslant p$, i.e., $v = (1\{p - r + 1 \leqslant i \leqslant p\})_{p \times 1}$. Set $w = (w_i) = [\Sigma(\theta) - \Sigma(\theta')] v$. Note that for each i, if $|\gamma_i(\theta) - \gamma_i(\theta')| = 1$, we have $|w_i| = k\epsilon_{n,p}$. Then there are at least $H(\gamma(\theta), \gamma(\theta'))$ number of elements w_i with $|w_i| = k\epsilon_{n,p}$, which implies

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

Since $\|v\|^2 = r \leqslant p$, the equation above yields

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

i.e.,

$$\frac{\left\| \Sigma(\theta) - \Sigma(\theta') \right\|^2}{H(\gamma(\theta), \gamma(\theta'))} \geqslant \frac{(k\epsilon_{n,p})^2}{p}$$

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

7.3. *Proof of Lemma 6.* The proof of the bound for the affinity given in Lemma 6 is involved. We break the proof into a few major technical lemmas which are proved in Section 7.3 and the Appendix. Without loss of generality we consider only the case $i = 1$ and prove that there exists a constant $c_1 > 0$ such that $\left\| \bar{\mathbb{P}}_{1,0} \wedge \bar{\mathbb{P}}_{1,1} \right\| \geqslant c_1$. The following lemma is the key step which turns the problem of bounding the total variation affinity into a chi-squared distance calculation on Gaussian mixtures.

Lemma 8 *(i). There exists a constant $c_2 < 1$ such that*

$$\tilde{\mathbb{E}}_{(\gamma_{-1}, \lambda_{-1})} \left\{ \int \left(\frac{d\bar{\mathbb{P}}_{(1,1,\gamma_{-1},\lambda_{-1})}}{d\bar{\mathbb{P}}_{(1,0,\gamma_{-1},\lambda_{-1})}} \right)^2 d\bar{\mathbb{P}}_{(1,0,\gamma_{-1},\lambda_{-1})} - 1 \right\} \leqslant c_2^2. \tag{40}$$

(ii). Moreover, Equation (40) *implies that* $\left\| \bar{\mathbb{P}}_{1,0} \wedge \bar{\mathbb{P}}_{1,1} \right\| \geqslant 1 - c_2 > 0$.

The proof of Lemma 8 (ii) is relatively easy and is given in the Appendix. Our goal in the remainder of this proof is to establish (40), which requires detailed understanding of $\bar{\mathbb{P}}_{(1,0,\gamma_{-1},\lambda_{-1})}$ and the mixture distribution $\bar{\mathbb{P}}_{(1,1,\gamma_{-1},\lambda_{-1})}$ as well as a careful analysis of the cross-product terms in the chi-squared distances on the left hand of (40).

From the definition of θ in Equation (12) and $\bar{\mathbb{P}}_{(1,0,\gamma_{-1},\lambda_{-1})}$ in Equation (13), $\gamma_1 = 0$ implies $\bar{\mathbb{P}}_{(1,0,\gamma_{-1},\lambda_{-1})}$ is a single multivariate normal distribution with a covariance matrix,

$$\Sigma_0 = \begin{pmatrix} 1 & \mathbf{0}_{1\times(p-1)} \\ \mathbf{0}_{(p-1)\times 1} & \mathbf{S}_{(p-1)\times(p-1)} \end{pmatrix}. \tag{41}$$

Here $\mathbf{S}_{(p-1)\times(p-1)} = (s_{ij})_{2\leqslant i,j\leqslant p}$ is a symmetric matrix uniquely determined by $(\gamma_{-1}, \lambda_{-1}) = ((\gamma_2, ..., \gamma_r), (\lambda_2, ..., \lambda_r))$ where for $i \leqslant j$,

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

Let

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

which gives the set of all possible values of the first row with the rest of the rows fixed, i.e., $\lambda_{-1}(\theta) = c$. Let $n_{\lambda_{-1}}$ be the number of columns of λ_{-1} with the column sum equal to $2k$ for which the first row has no choice but to take value 0 in this column. Set $p_{\lambda_{-1}} = r - n_{\lambda_{-1}}$. It is helpful to observe that $p_{\lambda_{-1}} \geqslant p/4 - 1$. Since $n_{\lambda_{-1}} \cdot 2k \leqslant r \cdot k$, the total number of 1's in the upper triangular matrix by the construction of the parameter set, we thus have $n_{\lambda_{-1}} \leqslant r/2$, which immediately implies $p_{\lambda_{-1}} = r - n_{\lambda_{-1}} \geqslant r/2 \geqslant p/4 - 1$. It follows $\text{Card}(\Lambda_1(\lambda_{-1})) = \binom{p_{\lambda_{-1}}}{k}$. Then, from the definitions in Equations (12) and (13), $\bar{\mathbb{P}}_{(1,1,\gamma_{-1},\lambda_{-1})}$ is an average of $\binom{p_{\lambda_{-1}}}{k}$ multivariate normal distributions with covariance matrices of the following form

$$\begin{pmatrix} 1 & \mathbf{r}_{1\times(p-1)} \\ \left(\mathbf{r}_{1\times(p-1)}\right)^T & \mathbf{S}_{(p-1)\times(p-1)} \end{pmatrix} \tag{42}$$

where $\|\mathbf{r}\|_0 = k$ with nonzero elements of r equal $\epsilon_{n,p}$ and the submatrix $\mathbf{S}_{(p-1)\times(p-1)}$ is the same as the one for Σ_0 given in (41).

Recall that for each $\theta \in \Theta$, $\mathbb{P}_\theta$ is the joint distribution of the n i.i.d. multivariate normal variables $\mathbf{X}_1, \ldots, \mathbf{X}_n$. So each term in the chi-squared distance on the left hand side of (40) is of the form $\left(\int \frac{g_1 g_2}{g_0}\right)^n$ where g_i are

the density function of $N(0, \Sigma_i)$ for $i = 0, 1$ and 2, with Σ_0 defined in (41) and Σ_1 and Σ_2 of the form (42).

The following lemma is useful for calculating the cross product terms in the chi-squared distance between Gaussian mixtures. The proof of the lemma is straightforward and is thus omitted.

Lemma 9 *Let g_i be the density function of $N(0, \Sigma_i)$ for $i = 0, 1$ and 2, respectively. Then*

$$\int \frac{g_1 g_2}{g_0} = \left[\det\left(I - \Sigma_0^{-2}(\Sigma_1 - \Sigma_0)(\Sigma_2 - \Sigma_0)\right)\right]^{-\frac{1}{2}}.$$

Let Σ_0 be defined in (41) and determined by $(\gamma_{-1}, \lambda_{-1})$. Let Σ_1 and Σ_2 be of the form (42) with the first row λ_1 and λ_1' respectively. Set

$$R^{\gamma_{-1},\lambda_{-1}}_{\lambda_1,\lambda_1'} = -\log\det\left(I - \Sigma_0^{-2}(\Sigma_0 - \Sigma_1)(\Sigma_0 - \Sigma_2)\right). \tag{43}$$

We sometimes drop the indices (λ_1, λ_1') and $(\gamma_{-1}, \lambda_{-1})$ from Σ_i to simplify the notations whenever there is no ambiguity. Then each term in the chi-squared distance on the left hand side of (40) can be expressed as in the form of

$$\exp\left(\frac{n}{2} \cdot R^{\gamma_{-1},\lambda_{-1}}_{\lambda_1,\lambda_1'}\right) - 1.$$

Define

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

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 9 the average of the chi-squared distance on the left hand side of Equation (40) can now be written as

$$\begin{aligned}& \tilde{\mathbb{E}}_{(\gamma_{-1},\lambda_{-1})}\left\{\tilde{\mathbb{E}}_{(\lambda_1,\lambda_1')|\lambda_{-1}}\left[\exp(\frac{n}{2} \cdot R^{\gamma_{-1},\lambda_{-1}}_{\lambda_1,\lambda_1'}) - 1\right]\right\} \\ = \quad & \tilde{\mathbb{E}}_{(\lambda_1,\lambda_1')}\left\{\tilde{\mathbb{E}}_{(\gamma_{-1},\lambda_{-1})|(\lambda_1,\lambda_1')}\left[\exp(\frac{n}{2} \cdot R^{\gamma_{-1},\lambda_{-1}}_{\lambda_1,\lambda_1'}) - 1\right]\right\}\end{aligned} \tag{44}$$

where λ_1 and λ_1' are independent and uniformly distributed over $\Lambda_1(\lambda_{-1})$ (not over B) for given λ_{-1}, and the distribution of $(\gamma_{-1}, \lambda_{-1})$ given (λ_1, λ_1') is uniform over $\Theta_{-1}(\lambda_1, \lambda_1')$, but the marginal distribution of λ_1 and λ_1' are not independent and uniformly distributed over B.

Let Σ_1 and Σ_2 be two covariance matrices of the form (42). Note that Σ_1 and Σ_2 differ from each other only in the first row/column. Then $\Sigma_i - \Sigma_0$, $i = 1$ or 2, has a very simple structure. The nonzero elements only appear

in the first row/column, and in total there are at most $2k$ nonzero elements. This property immediately implies the following lemma which makes the problem of studying the determinant in Lemma 9 relatively easy. The proof of Lemma 10 below is given in the Appendix.

Lemma 10 *Let Σ_0 be defined in (41) and let Σ_1 and Σ_2 be two covariance matrices of the form (42). Define J to be the number of overlapping $\epsilon_{n,p}$'s between Σ_1 and Σ_2 on the first row, and*

$$Q \stackrel{\triangle}{=} (q_{ij})_{1\leqslant i,j\leqslant p} = (\Sigma_1 - \Sigma_0)(\Sigma_2 - \Sigma_0).$$

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

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

and the matrix $(\Sigma_0 - \Sigma_1)(\Sigma_0 - \Sigma_2)$ has rank 2 with two identical nonzero eigenvalues $J\epsilon_{n,p}^2$ when $J > 0$.

The matrix Q is determined by two interesting parts, the first element $q_{11} = J\epsilon_{n,p}^2$ and a very special $k \times k$ square matrix $(q_{ij} : i \in I_r \text{ and } j \in I_c)$ with all elements equal to $\epsilon_{n,p}^2$. The following result, which is proved in the Appendix, shows that $R_{\lambda_1,\lambda_1'}^{\gamma_{-1},\lambda_{-1}}$ is approximately equal to

$$-\log\det(I - (\Sigma_0 - \Sigma_1)(\Sigma_0 - \Sigma_2)) = -2\log\left(1 - J\epsilon_{n,p}^2\right),$$

where J is defined in Lemma 10. Define

$$\Lambda_{1,J} = \left\{(\lambda_1, \lambda_1') \in B \otimes B : \text{the number of overlapping } \epsilon_{n,p}\text{'s between } \lambda_1 \text{and } \lambda_1' \text{is } J\right\}.$$

Lemma 11 *Let $R_{\lambda_1,\lambda_1'}^{\gamma_{-1},\lambda_{-1}}$ be defined in Equation (43) . Then*

$$R_{\lambda_1,\lambda_1'}^{\gamma_{-1},\lambda_{-1}} = -2\log\left(1 - J\epsilon_{n,p}^2\right) + R_{1,\lambda_1,\lambda_1'}^{\gamma_{-1},\lambda_{-1}}, \tag{45}$$

where $R_{1,\lambda_1,\lambda_1'}^{\gamma_{-1},\lambda_{-1}}$ satisfies, uniformly over all J,

$$\tilde{\mathbb{E}}_{(\lambda_1,\lambda_1')|J}\left[\tilde{\mathbb{E}}_{(\gamma_{-1},\lambda_{-1})|(\lambda_1,\lambda_1')}\exp\left(\frac{n}{2}R_{1,\lambda_1,\lambda_1'}^{\gamma_{-1},\lambda_{-1}}\right)\right] \leqslant \frac{3}{2}. \tag{46}$$

With the preparations given above, we are now ready to establish Equation (40) and thus complete the proof of Lemma 6.

Proof of Equation (40). Equation (45) in Lemma 11 yields that

$$\begin{aligned}
&\tilde{\mathbb{E}}_{(\lambda_1,\lambda_1')}\left\{\tilde{\mathbb{E}}_{(\gamma_{-1},\lambda_{-1})|(\lambda_1,\lambda_1')}\left[\exp(\frac{n}{2}R^{\gamma_{-1},\lambda_{-1}}_{\lambda_1,\lambda_1'})-1\right]\right\}\\
= \;& \tilde{\mathbb{E}}_J\left\{\exp\left[-n\log\left(1-J\epsilon_{n,p}^2\right)\right]\cdot\tilde{\mathbb{E}}_{(\lambda_1,\lambda_1')|J}\left[\tilde{\mathbb{E}}_{(\gamma_{-1},\lambda_{-1})|(\lambda_1,\lambda_1')}\exp\left(\frac{n}{2}R^{\gamma_{-1},\lambda_{-1}}_{1,\lambda_1,\lambda_1'}\right)\right]-1\right\}.
\end{aligned}$$

Recall that J is the number of overlapping $\epsilon_{n,p}$'s between Σ_1 and Σ_2 on the first row. It is easy to see that J has the hypergeometric distribution as λ_1 and λ_1' vary in B for each given λ_{-1}. For $0\leqslant j\leqslant k$,

$$\begin{aligned}
(47)\quad \tilde{\mathbb{E}}_J(\mathbf{1}\{J=j\}|\lambda_{-1}) &= \binom{k}{j}\binom{p_{\lambda_{-1}}-k}{k-j}\Big/\binom{p_{\lambda_{-1}}}{k}\\
&= \frac{\left(\frac{k!}{(k-j)!}\right)^2}{\frac{p_{\lambda_{-1}}!(p_{\lambda_{-1}}-2k+j)!}{\left[(p_{\lambda_{-1}}-k)!\right]^2}}\cdot\frac{1}{j!}\leqslant\left(\frac{k^2}{p_{\lambda_{-1}}-k}\right)^j,
\end{aligned}$$

where $\frac{k!}{(k-j)!}$ is a product of j term with each term $\leqslant k$ and for $\frac{p_{\lambda_{-1}}!(p_{\lambda_{-1}}-2k+j)!}{\left[(p_{\lambda_{-1}}-k)!\right]^2}$ it is bounded below by a product of j term with each term $\geqslant p_{\lambda_{-1}}-j$. Since $p_{\lambda_{-1}}\geqslant p/4-1$ for all λ_{-1}, we have

$$\tilde{\mathbb{E}}(\mathbf{1}\{J=j\})=\tilde{\mathbb{E}}_{\lambda_{-1}}\left[\tilde{\mathbb{E}}_J\left(\mathbf{1}\{J=j\}|\lambda_{-1}\right)\right]\leqslant\left(\frac{k^2}{p/4-1-k}\right)^j.$$

Thus

$$\begin{aligned}
&\tilde{\mathbb{E}}_{(\gamma_{-1},\lambda_{-1})}\left\{\int\left(\frac{d\bar{\mathbb{P}}_{(1,1,\gamma_{-1},\lambda_{-1})}}{d\bar{\mathbb{P}}_{(1,0,\gamma_{-1},\lambda_{-1})}}\right)^2 d\bar{\mathbb{P}}_{(1,0,\gamma_{-1},\lambda_{-1})}-1\right\}\\
\leqslant\;& \sum_{j\geqslant 0}\left(\frac{k^2}{p/4-1-k}\right)^j\left\{\exp\left[-n\log\left(1-j\epsilon_{n,p}^2\right)\right]\cdot\frac{3}{2}-1\right\}\\
(48)\qquad =\;& \frac{3}{2}\sum_{j\geqslant 1}\left(\frac{k^2}{p/4-1-k}\right)^j\exp\left[2j\left(\upsilon^2\log p\right)\right]+\\
&+\left(\frac{k^2}{p/4-1-k}\right)^0\left\{\exp\left[-n\log\left(1-0\cdot\epsilon_{n,p}^2\right)\right]\cdot\frac{3}{2}-1\right\}\\
\leqslant\;& C\sum_{j\geqslant 1}\left(p^{\frac{\beta-1}{\beta}}\cdot p^{-2\upsilon^2}\right)^{-j}+\frac{1}{2}<C\sum_{j\geqslant 1}\left(p^{\frac{\beta-1}{2\beta}}\right)^{-j}+\frac{1}{2}<c_2^2
\end{aligned}$$

by setting $c_2^2=3/4$, where the last step follows from $\upsilon^2<\frac{\beta-1}{54\beta}$ and $k^2=O\left(\frac{n}{\log p}\right)=O\left(\frac{p^{1/\beta}}{\log p}\right)$ as defined in Section 3. ∎

Remark 5 The condition $p \geqslant n^{\beta}$ for some $\beta > 1$ is assumed so that

$$\frac{k^2}{p_{\lambda_{-1}} - k} \leqslant \frac{k^2}{p/4 - k} = \frac{O\left(n/\log p\right)}{p/4 - k} = o\left(p^{-\varepsilon}\right)$$

for some $\varepsilon > 0$ to make the term (48) to be $o\left(1\right)$.

7.4. *Proof of Theorem 3.* The following lemma, which is proved in Cai and Zhou (2009), is now useful to prove Theorem 3.

Lemma 12 *Define the event A_{ij} by*

$$A_{ij} = \left\{ \left|\hat{\sigma}_{ij} - \sigma_{ij}\right| \leqslant 4 \min\left\{ \left|\sigma_{ij}\right|, \gamma\sqrt{\frac{\log p}{n}} \right\} \right\}. \tag{49}$$

Then $\mathbb{P}\left(A_{ij}\right) \geqslant 1 - 2C_1 p^{-9/2}$.

Let $D = \left(d_{ij}\right)_{1 \leqslant i,j \leqslant p}$ with $d_{ij} = \left(\hat{\sigma}_{ij} - \sigma_{ij}\right) I(A_{ij}^c)$. Then

$$\begin{aligned}
\mathbb{E}_{\mathbf{X}|\theta} \left\| \hat{\Sigma} - \Sigma \right\|^2 &\leqslant 2\mathbb{E}_{\mathbf{X}|\theta} \left[\sup_j \sum_{i \neq j} \left|\hat{\sigma}_{ij} - \sigma_{ij}\right| I(A_{ij}) \right]^2 + 2\mathbb{E}_{\mathbf{X}|\theta} \left\| D \right\|_1^2 + C\frac{\log p}{n} \\
&\leqslant 32 \left[\sup_j \sum_{i \neq j} \min\left\{ \left|\sigma_{ij}\right|, \gamma\sqrt{\frac{\log p}{n}} \right\} \right]^2 + 2\mathbb{E}_{\mathbf{X}|\theta} \left\| D \right\|_1^2 + C\frac{\log p}{n}. \qquad (50)
\end{aligned}$$

We will see that the first term in Equation (50) is dominating and is bounded by $Cc_{n,p}^2 \left(\frac{\log p}{n}\right)^{1-q}$, while the second term $\mathbb{E}_{\mathbf{X}|\theta} \left\| D \right\|_1^2$ is negligible.

Set $k^* = \left\lfloor c_{n,p} \left(\frac{n}{\log p}\right)^{q/2} \right\rfloor$. Then we have

$$\begin{aligned}
\sum_{i \neq j} \min\left\{ \left|\sigma_{ij}\right|, \gamma\sqrt{\frac{\log p}{n}} \right\} &\leqslant \gamma \left(\sum_{i \leqslant k^*} + \sum_{i > k^*} \right) \min\left\{ \left|\sigma_{[i]j}\right|, \sqrt{\frac{\log p}{n}} \right\} \\
&\leqslant C_5 k^* \sqrt{\frac{\log p}{n}} + C_5 \sum_{i > k^*} \left(\frac{c_{n,p}}{i}\right)^{\frac{1}{q}} \\
&\leqslant C_6 \left[k^* \sqrt{\frac{\log p}{n}} + c_{n,p}^{\frac{1}{q}} \cdot \left(k^*\right)^{1 - \frac{1}{q}} \right] \leqslant C_7 c_{n,p} \left(\frac{\log p}{n}\right)^{\frac{1-q}{2}} \qquad (51)
\end{aligned}$$

which immediately implies Equation (29) if $\mathbb{E}_{\mathbf{X}|\theta} \|\!| D |\!\|_1^2 = O\left(\frac{1}{n}\right)$. We shall now show that $\mathbb{E}_{\mathbf{X}|\theta} \|\!| D |\!\|_1^2 = O\left(\frac{1}{n}\right)$. Note that

$$\begin{aligned}
\mathbb{E}_{\mathbf{X}|\theta} \|\!| D |\!\|_1^2 &\leqslant p \sum_{ij} \mathbb{E}_{\mathbf{X}|\theta} d_{ij}^2 = p \sum_{ij} \mathbb{E}_{\mathbf{X}|\theta} \left\{ \left[d_{ij}^2 I(A_{ij}^c \cap \left\{ \hat{\sigma}_{ij} = \sigma_{ij}^* \right\}) + d_{ij}^2 I(A_{ij}^c \cap \{\hat{\sigma}_{ij} = 0\} \right] \right\} \\
&= p \sum_{ij} \mathbb{E}_{\mathbf{X}|\theta} \left\{ \left(\sigma_{ij}^* - \sigma_{ij} \right)^2 I(A_{ij}^c) \right\} + p \sum_{ij} \mathbb{E}_{\mathbf{X}|\theta} \sigma_{ij}^2 I(A_{ij}^c \cap \{\hat{\sigma}_{ij} = 0\} \\
&\equiv R_1 + R_2.
\end{aligned}$$

Lemma 12 yields that $\mathbb{P}\left(A_{ij}^c\right) \leqslant 2C_1 p^{-9/2}$, and the Whittle's inequality implies $\sigma_{ij}^* - \sigma_{ij}$ has all finite moments (cf. Whittle (1960)) under the subgaussianity condition (2). Hence

$$\begin{aligned}
R_1 &= p \sum_{ij} \mathbb{E}_{\mathbf{X}|\theta} \left\{ \left(\sigma_{ij}^* - \sigma_{ij} \right)^2 I(A_{ij}^c) \right\} \leqslant p \sum_{ij} \left[\mathbb{E}_{\mathbf{X}|\theta} \left(\sigma_{ij}^* - \sigma_{ij} \right)^6 \right]^{1/3} \mathbb{P}^{2/3} \left(A_{ij}^c \right) \\
&\leqslant C_8 p \cdot p^2 \cdot \frac{1}{n} \cdot p^{-3} = C_8 / n.
\end{aligned}$$

On the other hand,

$$\begin{aligned}
R_2 &= p \sum_{ij} \mathbb{E}_{\mathbf{X}|\theta} \sigma_{ij}^2 I(A_{ij}^c \cap \{\hat{\sigma}_{ij} = 0\} = p \sum_{ij} \mathbb{E}_{\mathbf{X}|\theta} \sigma_{ij}^2 I(|\sigma_{ij}| \geqslant 4\gamma \sqrt{\frac{\log p}{n}}) I(|\sigma_{ij}^*| \leqslant \gamma \sqrt{\frac{\log p}{n}}) \\
&\leqslant p \sum_{ij} \sigma_{ij}^2 \mathbb{E}_{\mathbf{X}|\theta} I(|\sigma_{ij}| \geqslant 4\gamma \sqrt{\frac{\log p}{n}}) I(|\sigma_{ij}| - \left|\sigma_{ij}^* - \sigma_{ij}\right| \leqslant \gamma \sqrt{\frac{\log p}{n}}) \\
&\leqslant p \sum_{ij} \sigma_{ij}^2 \mathbb{E}_{\mathbf{X}|\theta} I(\left|\sigma_{ij}^* - \sigma_{ij}\right| > \frac{3}{4} |\sigma_{ij}|) I(|\sigma_{ij}| \geqslant 4\gamma \sqrt{\frac{\log p}{n}}) \\
&\leqslant \frac{p}{n} \sum_{ij} n\sigma_{ij}^2 C_1 \exp\left(-\frac{9}{2\gamma^2} n\sigma_{ij}^2 \right) I(|\sigma_{ij}| \geqslant 4\gamma \sqrt{\frac{\log p}{n}}) \\
&= \frac{p}{n} \sum_{ij} \left[n\sigma_{ij}^2 \cdot C_1 \exp\left(-\frac{1}{2\gamma^2} n\sigma_{ij}^2 \right) \right] \cdot \exp\left(-\frac{4}{\gamma^2} n\sigma_{ij}^2 \right) I(|\sigma_{ij}| \geqslant 4\gamma \sqrt{\frac{\log p}{n}}) \\
&\leqslant C_9 \frac{p}{n} \cdot p^2 \cdot p^{-16} \leqslant C_9 / n.
\end{aligned}$$

Putting R_1 and R_2 together yields that for some constant $C > 0$

$$\mathbb{E}_{\mathbf{X}|\theta} \|\!| D |\!\|_1^2 \leqslant \frac{C}{n}. \tag{52}$$

Theorem 3 is proved by combining equations (50), (51) and (52). ∎

7.5. *Proof of Theorem 4.* We establish separately the lower and upper bounds under the Bregman divergence losses. The following lemma relates a general Bregman divergence to the squared Frobenius norm.

Lemma 13 *Assume that all eigenvalues of two symmetric matrices X and Y belong to $[\epsilon_2, M_2]$. Then there exist constants $c_2 > c_1 > 0$ depending only on ϵ_2 and M_2 such that for all $\phi \in \Phi$ defined in (32),*

$$c_1 \left\| X - Y \right\|_F^2 \leqslant D_\phi (X, Y) \leqslant c_2 \left\| X - Y \right\|_F^2 .$$

Proof of Lemma 13: Let the eigen decompositions of X and Y be

$$X = \sum_{i=1}^{p} \lambda_i v_i^T v_i \text{ and } Y = \sum_{i=1}^{p} \gamma_i u_i^T u_i.$$

For every $\phi (X) = \sum_{i=1}^{p} \varphi (\lambda_i)$ it is easy to see that

$$D_\phi (X, Y) = \sum_{i,j} \left(v_i^T u_i \right)^2 \left[\varphi (\lambda_i) - \varphi (\gamma_j) - \varphi'(\gamma_j) \cdot (\lambda_i - \gamma_j) \right]. \tag{53}$$

See Kulis, Sustik and Dhillon (2009, Lemma 1). The Taylor expansion gives

$$D_\phi (X, Y) = \sum_{i,j} \left(v_i^T u_i \right)^2 \frac{1}{2} \varphi'' (\xi_{ij}) (\lambda_i - \gamma_j)^2$$

where ξ_{ij} is in between λ_i and γ_j and then contained in $[\epsilon_2, M_2]$. From the assumption in (32), there are constants c_L and c_u such that $c_L \leqslant \varphi'' (\lambda) \leqslant c_u$ for all λ in $[\epsilon_2, M_2]$, which immediately implies

$$\frac{1}{2} c_L \sum_{i,j} \left(v_i^T u_i \right)^2 (\lambda_i - \gamma_j)^2 \leqslant D_\phi (X, Y) \leqslant \frac{1}{2} c_u \sum_{i,j} \left(v_i^T u_i \right)^2 (\lambda_i - \gamma_j)^2 = \left\| X - Y \right\|_F^2$$

or equivalently

$$\frac{1}{2} c_L \left\| X - Y \right\|_F^2 \leqslant D_\phi (X, Y) \leqslant \frac{1}{2} c_u \left\| X - Y \right\|_F^2 . \quad \blacksquare$$

Lower Bound under Bregman Matrix Divergences. It is trivial to see that

$$\inf_{\hat{\Sigma}} \sup_{\theta \in \mathcal{P}_q(\tau, c_{n,p})} \mathbb{E}_{\mathbf{X}|\theta} \mathrm{L}_\phi \left(\hat{\Sigma}, \Sigma (\theta) \right) \geqslant c \frac{1}{n}$$

by constructing a parameter space with only diagonal matrices. It is then enough to show that there exists some constant $c > 0$ such that

$$\inf_{\hat{\Sigma}} \max_{\mathcal{F}^*} \mathbb{E}_{\mathbf{X}|\theta} \mathrm{L}_\phi \left(\hat{\Sigma}, \Sigma (\theta) \right) \geqslant c c_{n,p} \left(\frac{\log p}{n} \right)^{1 - q/2}$$

for all $\phi \in \Phi$ defined in (32). Equation (53) implies

$$\inf_{\hat{\Sigma}} \max_{\mathcal{F}^*} \mathbb{E}_{\mathbf{X}|\theta} \mathrm{L}_\phi \left(\hat{\Sigma}, \Sigma(\theta) \right) = \inf_{\hat{\Sigma}: \epsilon_1 I < \hat{\Sigma} < 2\tau I} \max_{\mathcal{F}^*} \mathbb{E}_{\mathbf{X}|\theta} \mathrm{L}_\phi \left(\hat{\Sigma}, \Sigma(\theta) \right) \tag{54}$$

Convexity of φ implies $\varphi(\lambda_i) - \varphi(\gamma_j) - \varphi'(\gamma_j) \cdot (\lambda_i - \gamma_j)$ is nonnegative and increasing when λ_i moves away from the range $[\epsilon_1, 2\tau]$ of those eigenvalues γ_j's of $\Sigma(\theta)$. From Lemma 13 there is a universal constant c_L such that

$$\begin{aligned} \inf_{\hat{\Sigma}: \epsilon_1 I < \hat{\Sigma} < 2\tau I} \max_{\mathcal{F}^*} \mathbb{E}_{\mathbf{X}|\theta} \mathrm{L}_\phi \left(\hat{\Sigma}, \Sigma(\theta) \right) &\geqslant c_L \inf_{\hat{\Sigma}: \epsilon_1 I < \hat{\Sigma} < 2\tau I} \max_{\mathcal{F}^*} \mathbb{E}_{\mathbf{X}|\theta} \frac{1}{p} \left\| \hat{\Sigma} - \Sigma(\theta) \right\|_F^2 \\ &= \frac{c_L}{p} \inf_{\hat{\Sigma}} \max_{\mathcal{F}^*} \mathbb{E}_{\mathbf{X}|\theta} \left\| \hat{\Sigma} - \Sigma(\theta) \right\|_F^2 \end{aligned}$$

where the last equality is from the same argument for Equation (54).

It then suffices to study the lower bound under the Frobenius norm. Similar to the lower bound under the spectral norm one has

$$\inf_{\hat{\Sigma}} \max_{\theta \in \mathcal{F}^*} 2_\theta^2 \mathbb{E}_{\mathbf{X}|\theta} \left\| \hat{\Sigma} - \Sigma(\theta) \right\|_F^2 \geqslant \min_{\{(\theta, \theta'): H(\gamma(\theta), \gamma(\theta')) \geqslant 1\}} \frac{\left\| \Sigma(\theta) - \Sigma(\theta') \right\|_F^2}{H(\gamma(\theta), \gamma(\theta'))} \frac{p}{2} \min_i \left\| \bar{\mathbb{P}}_{i,0} \wedge \bar{\mathbb{P}}_{i,1} \right\|.$$

It is easy to see

$$\min_{\{(\theta, \theta'): H(\gamma(\theta), \gamma(\theta')) \geqslant 1\}} \frac{\left\| \Sigma(\theta) - \Sigma(\theta') \right\|_F^2}{H(\gamma(\theta), \gamma(\theta'))} \asymp c_{n,p} \left(\frac{\log p}{n} \right)^{1-q/2}$$

and it follows from Lemma 6 that there is a constant $c > 0$ such that

$$\min_i \left\| \bar{\mathbb{P}}_{i,0} \wedge \bar{\mathbb{P}}_{i,1} \right\| \geqslant c. \quad \blacksquare$$

Upper Bound under Bregman Matrix Divergences. We now show that there exists an estimator $\hat{\Sigma}$ such that

$$\mathbb{E}_{\mathbf{X}|\theta} \mathrm{L}_\phi \left(\hat{\Sigma}, \Sigma \right) \leqslant c \left[c_{n,p} \left(\frac{\log p}{n} \right)^{1-q/2} + \frac{1}{n} \right] \tag{55}$$

some constant $c > 0$, uniformly over all $\phi \in \Phi$ and $\Sigma \in \mathcal{P}_q^B(\tau, c_{n,p})$. Let $A_0 = \cap_{i,j} A_{ij}$, where A_{ij} is defined in (49). Lemma 12 yields that

$$\mathbb{P}(A_0) \geqslant 1 - 2C_1 p^{-5/2}. \tag{56}$$

Lemma 14 *Let $\hat{\Sigma}_B$ be defined in Equation (34). Then for all $\Sigma \in \mathcal{G}_q^B(\rho, c_{n,p})$*

$$\mathbb{P}\left(\frac{\epsilon_1}{2} I \prec \hat{\Sigma}_B \prec 3\tau I \right) \geqslant 1 - C_1 p^{-5/2}.$$

Proof of Lemma 14: Write $\hat{\Sigma}_B = \Sigma + \left(\hat{\Sigma}_B - \Sigma\right)$. Since $\left\|\hat{\Sigma}_B - \Sigma\right\| \leqslant \left\|\hat{\Sigma}_B - \Sigma\right\|_1$, the lemma is then a direct consequence of Lemma 12 and Equation (51) which implies $\left\|\hat{\Sigma}_B - \Sigma\right\|_1 \leqslant Cc_{n,p}\left(\frac{\log p}{n}\right)^{(1-q)/2} \to 0$ over A_0. ∎

Lemma 13 implies

$$\begin{aligned}
\mathbb{E}_{\mathbf{X}|\theta}\mathrm{L}_\phi\left(\hat{\Sigma}_B, \Sigma\right) &= \mathbb{E}_{\mathbf{X}|\theta}\left\{\mathrm{L}_\phi\left(\hat{\Sigma}_B, \Sigma\right) I(A_0)\right\} + \mathbb{E}_{\mathbf{X}|\theta}\left\{\mathrm{L}_\phi\left(\hat{\Sigma}_B, \Sigma\right) I(A_0^c)\right\} \\
&\leqslant C\mathbb{E}_{\mathbf{X}|\theta}\left\{\frac{1}{p}\left\|\hat{\Sigma} - \Sigma\right\|_F^2 I(A_0)\right\} + \mathbb{E}_{\mathbf{X}|\theta}\left\{\mathrm{L}_\phi\left(\hat{\Sigma}_B, \Sigma\right) I(A_0^c)\right\} \\
(57) \leqslant\ & 16C \sup_j \sum_{i\neq j} \min\left\{|\sigma_{ij}|^2, \gamma\frac{\log p}{n}\right\} + \mathbb{E}_{\mathbf{X}|\theta}\left\{\mathrm{L}_\phi\left(\hat{\Sigma}_B, \Sigma\right) I(A_0^c)\right\} + C\frac{1}{n}.
\end{aligned}$$

The second term in (57) is negligible since

$$\begin{aligned}
\mathbb{E}_{\mathbf{X}|\theta}\mathrm{L}_\phi\left(\hat{\Sigma}_B, \Sigma\right)\{A_0^c\} &\leqslant C \cdot \left[\max\left\{\log n, \log p\right\}\right]^{|r|} \cdot \mathbb{P}\left(A_0^c\right) \\
&\leqslant C \cdot \left[\max\left\{\log n, \log p\right\}\right]^{|r|} C_1 p^{-5/4} = o\left(c_{n,p}\left(\frac{\log p}{n}\right)^{1-q/2}\right)
\end{aligned}$$

by applying the Cauchy–Schwarz inequality twice. We now consider the first term in Equation (57). Set $k^* = \lfloor c_{n,p}\left(\frac{n}{\log p}\right)^{q/2} \rfloor$. Then we have

$$\begin{aligned}
\sum_{i\neq j} \min\left\{|\sigma_{ij}|^2, \gamma\frac{\log p}{n}\right\} &\leqslant \gamma^2\left(\sum_{i\leqslant k^*} + \sum_{i>k^*}\right)\min\left\{\left|\sigma_{[i]j}\right|^2, \frac{\log p}{n}\right\} \\
&\leqslant C_3 k^* \frac{\log p}{n} + C_3 \sum_{i>k^*}\left(\frac{c_{n,p}}{i}\right)^{2/q} \\
&\leqslant C_4\left[k^*\frac{\log p}{n} + c_{n,p}^{2/q} k^* \cdot \left(k^*\right)^{-2/q}\right] \leqslant C_5 c_{n,p}\left(\frac{\log p}{n}\right)^{1-q/2}
\end{aligned}$$

which immediately yields Equation (55). ∎

REFERENCES

[1] Abramovich, F., Benjamini, Y., Donoho, D. L. and Johnstone, I. M. (2006). Adapting to unknown sparsity by controlling the false discovery rate. *Ann. Statist.* **34**, 584-653.

[2] Assouad, P. (1983). Deux remarques sur l'estimation. *C. R. Acad. Sci. Paris* **296**, 1021-1024.

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

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

[5] Bregman, L. M. (1967). The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming. *USSR Comput. Math. Math. Phys.* **7**, 200-217.

[6] Cai, T. T. and Liu, W. (2011). Adaptive thresholding for sparse covariance matrix estimation. *J. Amer. Statist. Assoc.* **106**, 672-684.

[7] Cai, T. T., Liu, W. and Zhou, H. H. (2011). Optimal estimation of large sparse precision matrices. Manuscript.

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

[9] Cai, T. T. and Zhou, H. H. (2009). Covariance matrix estimation under the ℓ_1 norm (with discussion). *Statist. Sinica*, to appear.

[10] Censor, Y. and Zenios, S. A. (1997). *Parallel Optimization: Theory, Algorithms, and Applications.* Oxford University Press.

[11] Dhillon, I. S. and Tropp, J. A. (2007). Matrix nearness problems with Bregman divergences. *SIAM Journal on Matrix Analysis and Applications* **29**, 1120-1146.

[12] Donoho, D. L. and Liu, R. C. (1991). Geometrizing Rates of Convergence, II. *Ann. Statist.* **19**, 633-667.

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

[14] Kulis, B., Sustik, M. A. and Dhillon I. S.(2009). Low-rank kernel learning with Bregman matrix divergences. *Journal of Machine Learning Research* **10**, 341–376.

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

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

[17] Le Cam, L. (1986). *Asymptotic Methods in Statistical Decision Theory.* Springer-Verlag, New York.

[18] Ravikumar, P., Wainwright, M., Raskutti, G. and Yu, B. (2008). High-dimensional covariance estimation by minimizing l_1-penalized log-determinant divergence. Technical Report 797, UC Berkeley, Statistics Department, Nov. 2008.

[19] Rothman, A.J., Levina, E., and Zhu, J. (2009). Generalized thresholding of large covariance matrices. *J. Amer. Statist. Assoc.* **104**, 177-186.

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

[21] Saulis, L and Statulevicius, V.A. (1991). *Limit Theorems For Large Deviations.* Kluwer Academic Publishers.

[22] Tsybakov, A. B. (2009). *Introduction to Nonparametric Estimation*. Springer-Verlag, New York.

[23] van der Vaart, A.W. (1998). *Asymptotic Statistics*. Cambridge University Press, Cambridge.

[24] Whittle, P. (1960). Bounds for the moments of linear and quadratic forms in independent variables. *Theory Probab. Appl.* **5**, 302-305.

[25] Yu, B. (1997). Assouad, Fano, and Le Cam. *Festschrift for Lucien Le Cam*. D. Pollard, E. Torgersen, and G. Yang (eds), 423-435, Springer-Verlag.

Department of Statistics
The Wharton School
University of Pennsylvania
Philadelphia, PA 19104.
E-mail: tcai@wharton.upenn.edu
URL: http://www-stat.wharton.upenn.edu/ tcai

Department of Statistics
Yale University
New Haven, CT 06511.
E-mail: huibin.zhou@yale.edu
URL: http://www.stat.yale.edu/ hz68