Minimax Estimation of Large Covariance Matrices under ℓ_1-Norm

T. Tony Cai[1] and Harrison H. Zhou[2]
University of Pennsylvania and Yale University

Abstract

Driven by a wide range of applications in high-dimensional data analysis, there has been significant recent interest in the estimation of large covariance matrices. In this paper, we consider optimal estimation of a covariance matrix as well as its inverse over several commonly used parameter spaces under the matrix ℓ_1 norm. Both minimax lower and upper bounds are derived.

The lower bounds are established by using hypothesis testing arguments, where at the core are a novel construction of collections of least favorable multivariate normal distributions and the bounding of the affinities between mixture distributions. The lower bound analysis also provides insight into where the difficulties of the covariance matrix estimation problem arise. A specific thresholding estimator and tapering estimator are constructed and shown to be minimax rate optimal. The optimal rates of convergence established in the paper can serve as a benchmark for the performance of covariance matrix estimation methods.

Keywords: Covariance matrix, ℓ_1 norm, minimax lower bound, operator norm, optimal rate of convergence, tapering, thresholding.
AMS 2000 Subject Classification: Primary 62H12; secondary 62F12, 62G09.

[1]Department of Statistics, The Wharton School, University of Pennsylvania, Philadelphia, PA 19104. The research of Tony Cai was supported in part by NSF Grant DMS-0604954 and NSF FRG Grant DMS-0854973.
[2]Department of Statistics, Yale University, New Haven, CT 06511. The research of Harrison Zhou was supported in part by NSF Career Award DMS-0645676 and NSF FRG Grant DMS-0854975.

1 Introduction

Estimating covariance matrices is essential for a wide range of statistical applications. With high-dimensional data becoming readily available, one is frequently faced with the problem of estimating large covariance matrices. It is now well understood that in such a setting the standard sample covariance matrix does not provide satisfactory performance and regularization is needed. Many regularization methods, including banding, tapering, thresholding and penalization, have been proposed. See, for example, Wu and Pourahmadi (2003), Zou, Hastie, and Tibshirani (2006), Bickel and Levina (2008a, b), El Karoui (2008), Lam and Fan (2009), Johnstone and Lu (2009), Cai, Zhang, and Zhou (2010), and Cai and Liu (2011). However, the fundamental properties of the covariance matrix estimation problems are still largely unknown.

The minimax risk, which quantifies the difficulty of an estimation problem, is one of the most commonly used benchmark. It is often used as the basis for the evaluation of performance of an estimation method. Cai, Zhang, and Zhou (2010) were the first to derive the minimax rates of convergence for estimating a class of large covariance matrices under the spectral norm and the Frobenius norm. Rate-sharp minimax lower bounds were derived and specific tapering estimators were constructed and shown to achieve the optimal rates of convergence. It was noted that the minimax behavior of the estimation problem critically depends on the norm under which the estimation error is measured.

It is of significant interest to understand how well covariance matrices can be estimated under different settings. Suppose we observe independent and identically distributed p-variate random variables $\mathbf{X}_1, \ldots, \mathbf{X}_n$ and wish to estimate their unknown covariance matrix $\Sigma_{p\times p}$ based on the sample $\{\mathbf{X}_l\}$. For a given collection $\mathcal{B}$ of distributions of $\mathbf{X}_1$ with a certain class of covariance matrices, the minimax risk of estimating Σ over $\mathcal{B}$ under a given matrix norm $\|\cdot\|$ is defined as

$$R(\mathcal{B}) = \inf_{\hat{\Sigma}} \sup_{\mathcal{B}} \mathbb{E}\|\hat{\Sigma} - \Sigma\|^2.$$

In the present paper, we establish the optimal rates of convergence for estimating the covariance matrix $\Sigma = (\sigma_{ij})_{1\le i,j\le p}$ as well as its inverse over several commonly used parameter spaces under the matrix ℓ_1 norm. For a matrix $A = (a_{ij})$, its ℓ_1 norm is the maximum absolute column sum, $\|A\|_1 = \max_j \sum_i |a_{i,j}|$.

In the high-dimensional setting, structural assumptions are needed in order to estimate the covariance matrix consistently. One widely used assumption is that the covariance matrix is sparse, i.e., most of the entries in each row/column are zero or negligible. An-

other common assumption used in the literature is that the variables exhibit a certain ordering structure, which is often the case for time series data. Under this assumption, the magnitude of the elements in the covariance matrix decays as they move away from the diagonal. We consider both cases in the present paper and study three different types of parameter spaces.

The first class of parameter spaces models sparse covariance matrices in which each column (or row) $(\sigma_{ij})_{1\le i\le p}$ is assumed to be in a sparse weak ℓ_q ball, as used in many applications including gene expression array analysis. More specifically, denote by $\left|\sigma_{[k]j}\right|$ the k-th largest element in magnitude of the jth column $(\sigma_{ij})_{1\le i\le p}$. For $0 \le q < 1$, define

$$\mathcal{G}_q(\rho, c_{n,p}) = \left\{\Sigma = (\sigma_{ij})_{1\le i,j\le p} : \max_{1\le j\le p} \left\{\left|\sigma_{[k]j}\right|^q\right\} \le c_{n,p}/k \ , \forall k, \text{ and } \max_i (\sigma_{ii}) \le \rho\right\}. \quad (1)$$

In the special case $q = 0$, a matrix in $\mathcal{G}_0(\rho, c_{n,p})$ has at most $c_{n,p}$ nonzero elements in each column. The weak ℓ_q ball has been used in Abramovich, Benjamini, Donoho, and Johnstone (2006) for the sparse normal means problem. The parameter space $\mathcal{G}_q$ contains the uniformity class of covariance matrices in Bickel and Levina (2008b, page 5) as a special case. The second class of parameter spaces under study is

$$\mathcal{F}_\alpha(\rho, M) = \left\{\Sigma : \max_j \sum_i |\sigma_{ij}| \{i : |i-j| > k\} \le Mk^{-\alpha}, \forall k, \text{ and } \max_i (\sigma_{ii}) \le \rho\right\} \quad (2)$$

where $\alpha > 0$, $M > 0$, and $\rho > 0$. The parameter α in (2), which essentially specifies the rate of decay for the covariances σ_{ij} as they move away from the diagonal, can be viewed as an analog of the smoothness parameter in nonparametric spectral density estimation. This class of covariance matrices is motivated by time series analysis for applications such as on-line modeling and forecasting. Note that the smallest eigenvalue of a covariance matrix in the parameter space $\mathcal{F}_\alpha$ is allowed to be 0, which is more general than the assumption at (5) of Bickel and Levina (2008a). The third parameter space is a subclass of $\mathcal{F}_\alpha$:

$$\mathcal{H}_\alpha(\rho, M) = \left\{\Sigma : |\sigma_{ij}| \le M|i-j|^{-(\alpha+1)} \text{ for } i \ne j \text{ and } \max_i (\sigma_{ii}) \le \rho\right\}. \quad (3)$$

This parameter space has been considered in Bickel and Levina (2008a) and Cai, Zhang, and Zhou (2010).

We assume that the distribution of the X_i's is subgaussian in the sense that, for all $t > 0$ and all $\mathbf{v} \in \mathbb{R}^p$ with $\|\mathbf{v}\|_2 = 1$,

$$\mathbb{P}\{|\mathbf{v}'(\mathbf{X}_1 - \mathbb{E}\mathbf{X}_1)| > t\} \le e^{-\frac{t^2}{2\rho}}. \quad (4)$$

Let $\mathcal{P}\left(\mathcal{G}_q(\rho, c_{n,p})\right)$ denote the set of distributions of $\mathbf{X}_1$ satisfying (1) and (4). The distribution classes $\mathcal{P}\left(\mathcal{F}_\alpha(\rho, M)\right)$ and $\mathcal{P}\left(\mathcal{H}_\alpha(\rho, M)\right)$ are defined similarly. Our analysis establishes the minimax rates of convergence for estimating the covariance matrices over the three distribution classes $\mathcal{P}\left(\mathcal{G}_q(\rho, c_{n,p})\right)$, $\mathcal{P}\left(\mathcal{F}_\alpha(\rho, M)\right)$, and $\mathcal{P}\left(\mathcal{H}_\alpha(\rho, M)\right)$. By combining the minimax lower and upper bounds developed in later sections, the main results on the optimal rates of convergence for estimating the covariance matrix under the ℓ_1 norm can be summarized as follows.

Theorem 1 *The minimax risk of estimating the covariance matrix Σ over the distribution class $\mathcal{P}\left(\mathcal{G}_q(\rho, c_{n,p})\right)$ satisfies*

$$\inf_{\hat{\Sigma}} \sup_{\mathcal{P}(\mathcal{G}_q(\rho, c_{n,p}))} \mathbb{E} \left\| \hat{\Sigma} - \Sigma \right\|_1^2 \asymp c_{n,p}^2 \left(\frac{\log p}{n} \right)^{1-q} \tag{5}$$

under assumptions (7) and (8), and the minimax risks of estimating the covariance matrix Σ over the distribution classes $\mathcal{P}\left(\mathcal{F}_\alpha(\rho, M)\right)$ and $\mathcal{P}\left(\mathcal{H}_\alpha(\rho, M)\right)$ satisfy

$$\inf_{\hat{\Sigma}} \sup_{\mathcal{A}} \mathbb{E} \left\| \hat{\Sigma} - \Sigma \right\|_1^2 \asymp \min \left\{ n^{-\frac{\alpha}{\alpha+1}} + \left(\frac{\log p}{n} \right)^{\frac{2\alpha}{2\alpha+1}}, \frac{p^2}{n} \right\}, \tag{6}$$

where $\mathcal{A} = \mathcal{P}\left(\mathcal{F}_\alpha(\rho, M)\right)$ or $\mathcal{P}\left(\mathcal{H}_\alpha(\rho, M)\right)$.

A key step in obtaining the optimal rates of convergence is the derivation of sharp minimax lower bounds. As noted in Cai, Zhang, and Zhou (2010), the lower bound analysis for covariance matrix estimation has quite distinct features from those used in the more conventional function/sequence estimation problems. We establish the lower bounds by using several different hypothesis testing arguments including Le Cam's method, Assouad's Lemma, and a version of Fano's Lemma, where at the core are a novel construction of collections of least favorable multivariate normal distributions and the bounding of the affinities between mixture distributions. An important technical step is to bound the affinity between pairs of probability measures in the collection; this is quite involved in matrix estimation problems. We shall see that, although the general principles remain the same, the specific technical analysis used to obtain the lower bounds under the ℓ_1 norm loss is rather different from those used in the cases of the spectral norm and Frobenius norm losses.

We then show that the minimax lower bounds are rate optimal by constructing explicit estimators that attain the same rates of convergence as those of the minimax lower bounds. In the sparse case, it is shown that a thresholding estimator attains the optimal rate of

convergence over the distribution class $\mathcal{P}\left(\mathcal{G}_q(\rho, c_{n,p})\right)$ under the ℓ_1 norm. The thresholding estimator was originally introduced in Bickel and Levina (2008b) for estimating sparse covariance matrices under the spectral norm; here we show that the estimator is rate-optimal over the distribution class $\mathcal{P}\left(\mathcal{G}_q(\rho, c_{n,p})\right)$ under the matrix ℓ_1 norm. For the other two distribution classes $\mathcal{P}\left(\mathcal{F}_\alpha(\rho, M)\right)$ and $\mathcal{P}\left(\mathcal{H}_\alpha(\rho, M)\right)$, we construct a tapering estimator that is closely related to the recent work in Cai, Zhang, and Zhou (2010), though the choice of the optimal tapering parameter is quite different. This phenomenon is important in practical tuning parameter selection. For covariance matrix estimation under the spectral norm, Bickel and Levina (2008a) suggested selecting the tuning parameter by cross-validation and minimizing ℓ_1 norm loss for convenience. However, even if the cross-validation method selects the ideal tuning parameter for optimal estimation under the ℓ_1 norm, the resulting banding estimator can be far from optimal for estimation under the spectral norm.

The rest of the paper is organized as follows. Section 2 focuses on minimax lower bounds for covariance matrix estimation under the ℓ_1 norm. We then establish the minimax rates of convergence by showing that the lower bounds are in fact rate sharp. This is accomplished in Section 3 by constructing thresholding and tapering estimators and proving that they attain the same convergence rates as those given in the lower bounds. Section 4 considers optimal estimation of the inverse covariance matrices under the ℓ_1 norm. Section 5 discusses connections and differences of the results with other related work. The proofs of the technical lemmas that are used to prove the main results are given in Section 6.

2 Minimax lower bounds under the ℓ_1 norm

A key step in establishing the optimal rate of convergence is the derivation of the minimax lower bounds. In this section, we consider the minimax lower bounds for the three distribution classes given earlier. The upper bounds derived in Section 3 show that these lower bounds are minimax rate optimal.

We work with various matrix operator norms. For $1 \le r \le \infty$, the matrix ℓ_r norm of a matrix A is defined as

$$\|A\|_r = \max_{x \neq 0} \frac{\|Ax\|_r}{\|x\|_r} = \max_{\|x\|_r = 1} \|Ax\|_r.$$

The spectral norm is the matrix ℓ_2 norm; the ℓ_1 norm is the "maximum absolute column sum", i.e., for a matrix $A = (a_{ij})$, $\|A\|_1 = \max_j \sum_i |a_{i,j}|$; the matrix ℓ_∞ norm is the

"maximum absolute row sum", $\|A\|_\infty = \max_i \sum_j |a_{i,j}|$. Note that for covariance matrices the ℓ_1 norm coincides with the ℓ_∞ norm and the spectral norm is the maximum eigenvalue.

Since every Gaussian random variable is subgaussian, it is sufficient to derive minimax lower bounds under the Gaussian assumption. In this section, we consider independent and identically distributed p-variate Gaussian random variables $\mathbf{X}_1, \ldots, \mathbf{X}_n$ and wish to estimate their unknown covariance matrix $\Sigma_{p\times p}$ under the ℓ_1 norm based on the sample $\{\mathbf{X}_l\}$.

Throughout the paper we denote by $C,\ c, C_1, c_1, C_2, c_2, \ldots$ etc. generic constants, not depending on n or p, which may vary from place to place.

2.1 Minimax lower bound over $\mathcal{P}\left(\mathcal{G}_q(\rho, c_{n,p})\right)$

We begin by considering the parameter space $\mathcal{G}_q = \mathcal{G}_q(\rho, c_{n,p})$ at (1). The goal is to derive a good lower bound for the minimax risk over $\mathcal{G}_q(\rho, c_{n,p})$. We focus on the high-dimensional case where

$$p \geq n^\nu \text{ with } \nu > 1 \tag{7}$$

and assume that

$$c_{n,p} \leq M \left(\frac{n}{\log p}\right)^{\frac{1}{2}-\frac{q}{2}} \tag{8}$$

for $0 \leq q < 1$ and some $M > 0$. Theorem 2 below implies that the assumption $c_{n,p}^2 \left(\frac{\log p}{n}\right)^{1-q} \to 0$ is necessary to obtain a consistent estimator. See Remark 1 for more details.

Our strategy for deriving the minimax lower bound is to carefully construct a finite collection of multivariate normal distributions and to calculate the total variation affinity between pairs of probability measures in the collection. The construction is as follows. Let $\lfloor x \rfloor$ denote the largest integer less than or equal x. Set $k = \left\lfloor c_{n,p}\left(n/\log p\right)^{q/2} \right\rfloor$. We construct matrices whose off-diagonal elements are equal to 0 except the first row/column. Denote by $\mathcal{H}$ the collection of all $p \times p$ symmetric matrices with exactly k off-diagonal elements equal to 1 on the first row and the rest all zeros. (The first column of a matrix in $\mathcal{H}$ is obtained by reflecting the first row.) Define

$$\mathcal{G}_0 = \left\{\Sigma : \Sigma = I_p \text{ or } \Sigma = I_p + aH, \text{ for some } H \in \mathcal{H}\right\}, \tag{9}$$

where $a = \sqrt{\frac{\tau_1 \log p}{n}}$ for some constant τ_1. Without loss of generality we assume that $\rho > 1$ in (1). It is easy to see that $\mathcal{G}_0 \subset \mathcal{G}_q(\rho, c_{n,p})$ when τ_1 is small. We pick the constant τ_1 such that $0 < \tau_1 < \min\left\{1, \frac{1}{4\nu}\left(\nu - 1\right), \frac{1}{2M^2}\right\}$. It is straightforward to check that with such a choice of τ_1, $\mathcal{G}_0 \subset \mathcal{G}_q(\rho, c_{n,p})$.

We use Le Cam's method to establish the lower bound by showing that there exists some constant $C_1 > 0$ such that for any estimator $\hat{\Sigma}$,

$$\sup_{\mathcal{G}_0} \mathbb{E} \left\| \hat{\Sigma} - \Sigma \right\|_1^2 \geq C_1 c_{n,p}^2 \left(\frac{\log p}{n} \right)^{1-q}, \tag{10}$$

which leads immediately to the following result.

Theorem 2 *Suppose we observe independent and identically distributed p-variate Gaussian random variables $\mathbf{X}_1, \ldots, \mathbf{X}_n$ with covariance matrix $\Sigma_{p\times p} \in \mathcal{G}_q(\rho, c_{n,p})$. Under assumptions (7) and (8), the minimax risk of estimating the covariance matrix $\Sigma_{p\times p}$ satisfies*

$$\inf_{\hat{\Sigma}} \sup_{\mathcal{G}_q(\rho, c_{n,p})} \mathbb{E} \left\| \hat{\Sigma} - \Sigma \right\|_1^2 \geq C_1 c_{n,p}^2 \left(\frac{\log p}{n} \right)^{1-q} \tag{11}$$

where C_1 is a positive constant.

Remark 1 In Theorem 2, $c_{n,p}$ is assumed to satisfy $c_{n,p} \leq M \left(\frac{n}{\log p} \right)^{\frac{1}{2}-\frac{q}{2}}$ for some constant $M > 0$. This assumption is necessary to obtain a consistent estimator. If $c_{n,p} > M \left(\frac{n}{\log p} \right)^{\frac{1}{2}-\frac{q}{2}}$, we have

$$\inf_{\hat{\Sigma}} \sup_{\mathcal{G}_q(\rho, c_{n,p})} \mathbb{E} \left\| \hat{\Sigma} - \Sigma \right\|_1^2 \geq \inf_{\hat{\Sigma}} \sup_{\mathcal{G}_q(\rho, M\left(\frac{n}{\log p} \right)^{\frac{1}{2}-\frac{q}{2}})} \mathbb{E} \left\| \hat{\Sigma} - \Sigma \right\|_1^2 \geq C_1 M^2$$

where the last inequality follows from (11). Furthermore by a similar argument as above, we need the condition $c_{n,p}^2 \left(\frac{\log p}{n} \right)^{1-q} \to 0$ to estimate Σ consistently under the ℓ_1 norm.

Results in Section 3 show that the lower bound given in (11) is minimax rate optimal. A threshold estimator is shown to attain the convergence rate given in (11).

Before we prove the theorem, we need to introduce some notation. Denote by m_* the number of non-identity covariance matrices in $\mathcal{G}_0$. Then $m_* = \text{Card}\,(\mathcal{G}_0) - 1 = \binom{p-1}{k}$. Let $\Sigma_m, 1 \leq m \leq m_*$, denote a non-identity covariance matrix in $\mathcal{G}_0$, and let Σ_0 be the identity matrix I_p. We denote the joint distribution and density of $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$ with $\mathbf{X}_l \sim N(0, \Sigma_m)$ by $\mathbb{P}_{\Sigma_m}$ and f_m, respectively, and take $\bar{\mathbb{P}} = \frac{1}{m_*} \sum_{m=1}^{m_*} \mathbb{P}_{\Sigma_m}$.

For two probability measures $\mathbb{P}$ and $\mathbb{Q}$ with density p and q with respect to any common dominating measure μ, write the total variation affinity $\|\mathbb{P} \wedge \mathbb{Q}\| = \int p \wedge q d\mu$. A major tool for the proof of Theorem 2 is the following lemma which is a direct consequence of Le Cam's lemma (cf. Le Cam (1973), Yu (1997)).

Lemma 1 *Let $\hat{\Sigma}$ be any estimator of Σ_m based on an observation from a distribution in the collection $\{\mathbb{P}_{\Sigma_m}, m = 0, 1, \ldots, m_*\}$, then*

$$\sup_{0\le m\le m_*} \mathbb{E}\left\|\hat{\Sigma} - \Sigma_m\right\|_1 \ge \frac{1}{2}\left\|\mathbb{P}_{\Sigma_0} \wedge \bar{\mathbb{P}}\right\| \cdot \inf_{1\le m\le m_*} \|\Sigma_m - \Sigma_0\|_1 .$$

Proof of Theorem 2: It is easy to see that

$$\inf_{1\le m\le m_*} \|\Sigma_m - \Sigma_0\|_1^2 = k^2a^2 \ge C_2 c_{n,p}^2 \left(\frac{\log p}{n}\right)^{1-q}$$

for some $C_2 > 0$. To prove the theorem, it thus suffices to show that there is a constant $C_3 > 0$ such that

$$\left\|\mathbb{P}_{\Sigma_0} \wedge \bar{\mathbb{P}}\right\| \ge C_3. \tag{12}$$

That immediately implies

$$\sup_{0\le m\le m_*} \mathbb{E}\left\|\hat{\Sigma} - \Sigma_m\right\|_1^2 \ge \sup_{0\le m\le m_*} \left(\mathbb{E}\left\|\hat{\Sigma} - \Sigma_m\right\|_1\right)^2 \ge \frac{1}{4}\cdot C_2 c_{n,p}^2 \left(\frac{\log p}{n}\right)^{1-q} \cdot C_3^2$$

which matches the lower bound in (11) up to a constant factor.

Now we establish the lower bound (12) for the total variation affinity. Since the affinity $\int q_0 \wedge q_1 d\mu = 1 - \frac{1}{2}\int |q_0 - q_1|\, d\mu$ for any two densities q_0 and q_1, Jensen's Inequality implies

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

Hence $\int q_0 \wedge q_1 d\mu \ge 1 - \frac{1}{2}\left[\int \left(\frac{q_1^2}{q_0} - 1\right) d\mu\right]^{1/2}$. To establish (12), it thus suffices to show that

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

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

Lemma 2 *Let g_s be the density function of $N(0, \Sigma_s)$, $s = 0, m, l$. Then*

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

Lemma 2 implies

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

Let $J(m, l)$ be the number of overlapping nonzero off-diagonal elements between Σ_m and Σ_l in the first row. Elementary calculations yield that $\|\Sigma_m - \Sigma_l\|_1 = 2(k - J)a$ and

$$\left[\det\left(I - (\Sigma_m - I_p)(\Sigma_l - I_p)\right)\right]^{1/2} = 1 - Ja^2,$$

which is 1 when $J = 0$. It is easy to see that the total number of pairs (Σ_m, Σ_l) such that $J(m,l) = j$ is $\binom{p-1}{k}\binom{k}{j}\binom{p-1-k}{k-j}$. Hence,

$$\begin{aligned}\Delta &= \frac{1}{m_*^2}\sum_{0\le j\le k}\sum_{J(m,l)=j}\int\left(\frac{f_m f_l}{f_0}-1\right) = \frac{1}{m_*^2}\sum_{0\le j\le k}\sum_{J(m,l)=j}\left[\left(1-ja^2\right)^{-n}-1\right] \\ &\le \frac{1}{m_*^2}\sum_{1\le j\le k}\binom{p-1}{k}\binom{k}{j}\binom{p-1-k}{k-j}(1-ja^2)^{-n}. \qquad (13)\end{aligned}$$

Note that

$$\left(1-ja^2\right)^{-n} \le \left(1+2ja^2\right)^n \le \exp\left(n2ja^2\right) = p^{2\tau_1 j}$$

where the first inequality follows from the fact that $ja^2 \le ka^2 \le \tau_1 M^2 < 1/2$. Hence,

$$\Delta \le \sum_{1\le j\le k}\frac{\binom{k}{j}\binom{p-1-k}{k-j}}{\binom{p-1}{k}}p^{2\tau_1 j} \le 2\sum_{1\le j\le k}\left(\frac{k^2 p^{2\tau_1}}{p-k}\right)^j.$$

Recall that $k = \left\lfloor c_{n,p}\left(\frac{n}{\log p}\right)^{q/2}\right\rfloor$ and $c_{n,p} \le M\left(n/\log p\right)^{\frac{1}{2}-\frac{q}{2}}$. So we have

$$\begin{aligned}k^2\frac{p^{2\tau_1}}{p-k} &\le c_{n,p}^2\left(\frac{n}{\log p}\right)^q\cdot\frac{p^{2\tau_1}}{p-k} \\ &\le M^2\left(\frac{n}{\log p}\right)^{1-q}\left(\frac{n}{\log p}\right)^q\cdot\frac{p^{2\tau_1}}{p-k} \\ &\le 2M^2\left(\frac{n}{\log p}\right)\cdot\frac{p^{2\tau_1}}{p} \le 2M^2 n^{(1-\nu)/2},\end{aligned}$$

where the last step follows from the fact that $\tau_1 \le (\nu-1)/(4\nu)$. Thus $\Delta \le Cn^{(1-\nu)/2} \to 0$, which immediately implies (12). ∎

2.2 Minimax lower bounds over $\mathcal{P}(\mathcal{F}_\alpha(\rho, M))$ and $\mathcal{P}(\mathcal{H}_\alpha(\rho, M))$

We now consider minimax lower bounds for the parameter spaces $\mathcal{F}_\alpha(\rho, M)$ and $\mathcal{H}_\alpha(\rho, M)$. We show that the minimax rates of convergence over these two parameter spaces are the same under the ℓ_1 norm. Since $\mathcal{H}_\alpha(\rho, M) \subset \mathcal{F}_\alpha(\rho, 2M/\alpha)$, it thus suffices to establish the minimax lower bound for $\mathcal{H}_\alpha(\rho, M)$.

As in Section 2.1, the basic strategy remains to carefully construct a finite collection of multivariate normal distributions such that the covariance matrices are "far apart" in ℓ_1 norm and yet it is still "sufficiently difficult" to test between them based on the observed sample. However, the specific construction and the technical tools used in the analysis are quite different from those in Section 2.1. Here we mainly rely on Assouad's Lemma and a version of Fano's Lemma given in Tsybakov (2009) to obtain the desired lower bound.

We define the parameter spaces that are appropriate for the minimax lower bound argument. In this section we assume $p \geq n^{\frac{1}{2\alpha+2}}$. The case $p < n^{\frac{1}{2\alpha+2}}$ is similar and slightly easier. Both lower bound and upper bound for this case will be discussed in Section 3.2.1.

We construct parameter spaces separately for the cases $p \leq \exp\left(n^{\frac{1}{2\alpha+2}}\right)$ and $p > \exp\left(n^{\frac{1}{2\alpha+2}}\right)$. For $p \leq \exp\left(n^{\frac{1}{2\alpha+2}}\right)$, set $k = \left\lfloor n^{\frac{1}{2\alpha+2}} \right\rfloor$. Without loss of generality let $\rho > 1$. Let τ_2 be a small constant to be specified later. Take the parameter space $\mathcal{F}_{11}$ of 2^{k-1} covariance matrices to consist of all $p \times p$ symmetric matrices with diagonal elements 1 and the first $(k-1)$ off-diagonal elements in the first row (and first column by symmetry) equal to either 0 or $\tau_2 n^{-1/2}$, with all other elements 0. Formally,

$$\mathcal{F}_{11} = \left\{ \Sigma(\theta) : \Sigma(\theta) = I_p + \tau_2 n^{-1/2} \sum_{s=2}^{k} \theta_s \begin{bmatrix} \left(I\{i=1, j=s\}\right)_{p\times p} \\ + \left(I\{i=s, j=1\}\right)_{p\times p} \end{bmatrix}, \theta = (\theta_s) \in \{0,1\}^{k-1} \right\}, \tag{14}$$

where I_p is the $p \times p$ identity matrix.

We pick τ_2 such that $0 < \tau_2 < \min\left\{M, M^2, 1/16\right\}$. It is then easy to see that for any $\Sigma = (\sigma_{i,j}) \in \mathcal{F}_{11}$,

$$|\sigma_{1,j}| \leq \tau_2 n^{-\frac{1}{2}} \leq \tau_2 k^{-(\alpha+1)} \leq M j^{-(\alpha+1)}$$

for all $2 \leq j \leq k$, and consequently $|\sigma_{i,j}| \leq M|i-j|^{-(\alpha+1)}$ for all $1 \leq i \neq j \leq p$. In addition, we have $\max_i(\sigma_{ii}) = 1 < \rho$. Hence, the collection $\mathcal{F}_{11} \subset \mathcal{H}_\alpha(\rho, M)$.

For $p \geq \exp\left(n^{\frac{1}{2\alpha+2}}\right)$, we set $k = \left\lfloor \left(\frac{n}{\log p}\right)^{\frac{1}{2\alpha+1}} \right\rfloor$. Define the $p \times p$ matrix $B_m = (b_{ij})_{p\times p}$ by

$$b_{ij} = I\{i = m \text{ and } m+1 \leq j \leq m+k-1, \text{ or } j = m \text{ and } m+1 \leq i \leq m+k-1\}.$$

In addition to $\mathcal{F}_{11}$ we take

$$\mathcal{F}_{12} = \left\{ \Sigma_m : \Sigma_m = I_p + b\sqrt{\tau_2 \log p} B_m,\ 1 \leq m \leq m_* \right\}, \tag{15}$$

where $b = (nk)^{-1/2}$ and $m_* = \lfloor p/k \rfloor - 1$. It is easy to see that

$$(bk)^2 \log p = \frac{k}{n} \log p \leq k^{-2\alpha}$$

which implies

$$b\sqrt{\tau_2 \log p} \leq M k^{-\alpha-1}$$

as long as $\tau_2 < M^2$, and $\sup_i(\sigma_{ii}) = 1 < \rho$. Then the collection $\mathcal{F}_{12} \subset \mathcal{H}_\alpha(\rho, M)$.

Let $\mathcal{F}_0 = \mathcal{F}_{11} \cup \mathcal{F}_{12}$. It is clear that $\mathcal{F}_0 \subset \mathcal{H}_\alpha(\rho, M)$. It will be shown below separately that for some constant $C_4 > 0$,

$$\inf_{\hat{\Sigma}} \sup_{\mathcal{F}_{11}} \mathbb{E} \left\| \hat{\Sigma} - \Sigma \right\|_1^2 \geq C_4 n^{-\frac{\alpha}{\alpha+1}}, \tag{16}$$

$$\inf_{\hat{\Sigma}} \sup_{\mathcal{F}_{12}} \mathbb{E} \left\| \hat{\Sigma} - \Sigma \right\|_1^2 \geq C_4 \left(\frac{\log p}{n} \right)^{\frac{2\alpha}{2\alpha+1}}. \tag{17}$$

Equations (16) and (17) together imply

$$\inf_{\hat{\Sigma}} \sup_{\mathcal{F}_0} \mathbb{E} \left\| \hat{\Sigma} - \Sigma \right\|_1^2 \geq \frac{C_4}{2} \left[n^{-\frac{\alpha}{\alpha+1}} + \left(\frac{\log p}{n} \right)^{\frac{2\alpha}{2\alpha+1}} \right], \tag{18}$$

which yields the follow result.

Theorem 3 *Suppose we observe independent and identically distributed p-variate Gaussian random variables $\mathbf{X}_1, \ldots, \mathbf{X}_n$ with covariance matrix $\Sigma_{p\times p} \in \mathcal{F}_\alpha(\rho, M)$ or $\mathcal{H}_\alpha(\rho, M)$. The minimax risks of estimating the covariance matrix Σ satisfy, for some $C > 0$,*

$$\inf_{\hat{\Sigma}} \sup_{\Sigma \in \mathcal{A}} \mathbb{E} \left\| \hat{\Sigma} - \Sigma \right\|_1^2 \geq C \left[n^{-\frac{\alpha}{\alpha+1}} + \left(\frac{\log p}{n} \right)^{\frac{2\alpha}{2\alpha+1}} \right] \tag{19}$$

where $\mathcal{A} = \mathcal{F}_\alpha(\rho, M)$ or $\mathcal{H}_\alpha(\rho, M)$.

It is shown in Section 3 that the rate of convergence given in the lower bound (19) is optimal. A specific tapering estimator is constructed and shown to attain the minimax rate of convergence $n^{-\frac{\alpha}{\alpha+1}} + \left(\frac{\log p}{n} \right)^{\frac{2\alpha}{2\alpha+1}}$.

We establish the lower bound (16) by using Assouad's Lemma and the lower bound (17) by using a version of Fano's Lemma given in Tsybakov (2009).

2.2.1 Proof of the lower bound (16)

The key technical tool to establish (16) is the lemma in Assouad (1983). It gives a lower bound for the maximum risk over the parameter set $\Theta = \{0,1\}^m$ for the problem of estimating an arbitrary quantity $\psi(\theta)$ belonging to a metric space space with metric d. Let $H(\theta, \theta') = \sum_{i=1}^m |\theta_i - \theta_i'|$ be the Hamming distance on $\{0,1\}^m$, which counts the number of positions at which θ and θ' differ. Assouad's Lemma provides a minimax lower bound.

Lemma 3 (Assouad) *Let $\Theta = \{0,1\}^m$ and let T be an estimator based on an observation from a distribution in the collection $\{P_\theta, \theta \in \Theta\}$. Then for all $s > 0$*

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

Assouad's Lemma is connected to multiple comparisons. In total there are m comparisons. The lower bound has three terms. The first term is basically the loss one would incur for each incorrect comparison, the last term is the lower bound for the total probability of type one and type two errors for each comparison, and $m/2$ is the expected number of mistakes one would make when $\mathbb{P}_\theta$ and $\mathbb{P}_{\theta'}$ are not distinguishable from each other when $H(\theta, \theta') = 1$.

We now prove (16). Let $\mathbf{X}_1, \ldots, \mathbf{X}_n \overset{iid}{\sim} N(0, \Sigma(\theta))$ with $\Sigma(\theta) \in \mathcal{F}_{11}$. Denote the joint distribution by P_θ. Applying Assouad's Lemma to the parameter space $\mathcal{F}_{11}$ with $m = k - 1$, we have

$$\inf_{\hat{\Sigma}} \max_{\theta \in \{0,1\}^k} 2^2 E_\theta \left\| \hat{\Sigma} - \Sigma(\theta) \right\|_1 \geq \min_{H(\theta,\theta') \geq 1} \frac{\|\Sigma(\theta) - \Sigma(\theta')\|_1}{H(\theta, \theta')} \frac{k-1}{2} \min_{H(\theta,\theta')=1} \|P_\theta \wedge P_{\theta'}\|. \tag{20}$$

We state the bounds for the two factors on the right hand of (20) in two lemmas.

Lemma 4 *Let $\Sigma(\theta)$ be defined as in (14). Then*

$$\min_{H(\theta,\theta') \geq 1} \frac{\|\Sigma(\theta) - \Sigma(\theta')\|_1}{H(\theta, \theta')} \geq cn^{-1/2} \tag{21}$$

for some $c > 0$.

The proof of Lemma 4 is straightforward and is thus omitted here.

Lemma 5 *Let $\mathbf{X}_1, \ldots, \mathbf{X}_n \overset{iid}{\sim} N(0, \Sigma(\theta))$ with $\Sigma(\theta) \in \mathcal{F}_{11}$. Denote the joint distribution by P_θ. Then for some constant $c_1 > 0$*

$$\min_{H(\theta,\theta')=1} \|P_\theta \wedge P_{\theta'}\| \geq c_1.$$

The proof of Lemma 5 is deferred to Section 6. It follows from Lemma 5, using $k = n^{\frac{1}{2\alpha+2}}$, that

$$\inf_{\hat{\Sigma}} \sup_{\Sigma(\theta) \in \mathcal{F}_{11}} 2^2 E_\theta \left\| \hat{\Sigma} - \Sigma(\theta) \right\|_1^2 \geq c_2 k^2 \left(n^{-1/2} \right)^2 = c_2 k^2 n^{-1} = c_2 n^{-\frac{\alpha}{\alpha+1}}. \quad \blacksquare$$

2.2.2 Proof of the lower bound (17)

Consider the parameter space $\mathcal{F}_{12}$ defined in (15). Denote by Σ_0 the $p \times p$ identity matrix. Let f_m, $1 \leq m \leq m_* = \lfloor p/k \rfloor - 1$, be the joint density of $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$ with $\mathbf{X}_l \sim N(0, \Sigma_m)$ where $\Sigma_m \in \mathcal{F}_{12}$. For two probability measures $\mathbb{P}$ and $\mathbb{Q}$ with density p and q with respect to a common dominating measure μ, write the Kullback-Leibler divergence as $K(\mathbb{P}, \mathbb{Q}) = \int p \log \frac{p}{q} d\mu$.

The following lemma, which can be viewed as a version of Fano's Lemma, gives a lower bound for the minimax risk over the parameter set $\Theta = \{\theta_0, \theta_1, \ldots, \theta_{m_*}\}$.

Lemma 6 *Let $\Theta = \{\theta_m : m = 0, ..., m_*\}$ be a parameter set satisfying $d(\theta_i, \theta_j) \geq 2s$ for all $0 \leq i \neq j \leq m_*$, where d is a distance over Θ. Let $\{\mathbb{P}_\theta : \theta \in \Theta\}$ be a collection of probability measures defined on a common probability space satisfying*

$$\frac{1}{m_*} \sum_{1 \leq m \leq m_*} K(\mathbb{P}_{\theta_m}, \mathbb{P}_{\theta_0}) \leq c \log m_*$$

with $0 < c < 1/8$. Let $\hat{\theta}$ be any estimator based on an observation from a distribution in the collection $\{\mathbb{P}_\theta, \theta \in \Theta\}$. Then

$$\sup_{\theta \in \Theta} \mathbb{E} d^2\left(\hat{\theta}, \theta\right) \geq s^2 \frac{\sqrt{m_*}}{1 + \sqrt{m_*}} \left(1 - 2c - \sqrt{\frac{2c}{\log m_*}}\right).$$

We refer to Tsybakov (2009, Section 2.6) for more detailed discussions. Now let $\Theta = \mathcal{F}_{12}$, $\theta_m = \Sigma_m$ for $0 \leq m \leq m_*$, and let the distance d be the ℓ_1 norm. It is easy to see that

$$d(\theta_i, \theta_j) = \|\Sigma_i - \Sigma_j\|_1 = b\sqrt{\tau_2 \log p}\,(k-1) \geq \sqrt{\frac{1}{2}\tau_2 \frac{k \log p}{n}} \text{ for all } 0 \leq i \neq j \leq m_*. \quad (22)$$

The next lemma, proved in Section 6, gives a bound for the Kullback-Leibler divergence.

Lemma 7 *For all $1 \leq m \leq m_*$, distributions in the collection $\{\mathbb{P}_\theta, \theta \in \Theta\}$ satisfy*

$$K(\mathbb{P}_{\theta_m}, \mathbb{P}_{\theta_0}) \leq 2\tau_2 \log p.$$

By taking the constant τ_2 sufficiently small, Lemma 7 yields that

$$\frac{1}{m_*} \sum_{1 \leq m \leq m_*} K(\mathbb{P}_{\theta_m}, \mathbb{P}_{\theta_0}) \leq c \log m_*$$

for some positive constant $0 < c < 1/8$. Then the lower bound (17) follows immediately from Lemma 6 and (22),

$$\inf_{\hat{\Sigma}} \sup_{\Sigma_m \in \mathcal{F}_{12}} \mathbb{E} \left\| \hat{\Sigma} - \Sigma_m \right\|_1^2 \geq C \left(\frac{\log p}{n}\right)^{\frac{2\alpha}{2\alpha+1}}$$

for some constant $C > 0$. ∎

3 Optimal estimation under the ℓ_1 norm

In this section we consider the upper bounds for the minimax risk and construct specific rate optimal estimators for estimation over the three distribution classes. These upper bounds show that the rates of convergence given in the lower bounds established in Section 2 are sharp. More specifically, we show that a thresholding estimator attains the optimal rate of convergence over the distribution class $\mathcal{P}\left(\mathcal{G}_q(\rho, c_{n,p})\right)$ and a tapering estimator is minimax rate optimal over the distribution classes $\mathcal{P}\left(\mathcal{F}_\alpha(\rho, M)\right)$ and $\mathcal{P}\left(\mathcal{H}_\alpha(\rho, M)\right)$. The two estimators are introduced and analyzed separately in Sections 3.1 and 3.2.

Given a random sample $\{\mathbf{X}_1, \ldots, \mathbf{X}_n\}$ from a population with covariance matrix $\Sigma = \Sigma_{p\times p}$, 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

$$\Sigma^* = (\sigma_{ij}^*)_{1\le i,j\le p} = \frac{1}{n}\sum_{l=1}^{n}\left(\mathbf{X}_l - \bar{\mathbf{X}}\right)\left(\mathbf{X}_l - \bar{\mathbf{X}}\right)^T \tag{23}$$

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

3.1 Optimal estimation over $\mathcal{P}\left(\mathcal{G}_q(\rho, c_{n,p})\right)$

Theorem 2 shows that the minimax risk of estimating the covariance matrix $\Sigma_{p\times p}$ over the distribution class $\mathcal{P}\left(\mathcal{G}_q(\rho, c_{n,p})\right)$ has a lower bound of order $c_{n,p}^2\left(\frac{\log p}{n}\right)^{1-q}$. We now prove that this rate is optimal by constructing a thresholding estimator and by showing that this estimator attains the rate given in the lower bound.

Under the subgaussian assumption (4), the sample covariance $\sigma_{i,j}^*$ is an average of n random variables with a finite exponential moment, so $\sigma_{i,j}^*$ satisfies the large deviation result that there exist constants $C_1 > 0$ and $\gamma > 0$ such that

$$\mathbb{P}\left(\left|\sigma_{ij}^* - \sigma_{ij}\right| > v\right) \le C_1 \exp\left(-\frac{8}{\gamma^2}nv^2\right) \tag{24}$$

for $|v| \le \delta$, where C_1, γ and δ are constants that depend only on ρ. See, for example, Saulis and Statulevičius (1991) and Bickel and Levina (2008a). The inequality (24) implies that σ_{ij}^* behaves like a subgaussian random variable. In particular for $v = \gamma\sqrt{\frac{\log p}{n}}$ we have

$$\mathbb{P}\left(\left|\sigma_{ij}^* - \sigma_{ij}\right| > v\right) \le C_1 p^{-8}. \tag{25}$$

We define a thresholding estimator as

$$\hat{\sigma}_{ij} = \sigma^*_{ij} \cdot I\left(|\sigma^*_{ij}| \geq \gamma\sqrt{\frac{\log p}{n}}\right) \tag{26}$$

and set $\hat{\Sigma} = (\hat{\sigma}_{ij})_{p\times p}$.

The following theorem shows that the thresholding estimator at (26) is rate optimal over the distribution class $\mathcal{P}\left(\mathcal{G}_q(\rho, c_{n,p})\right)$.

Theorem 4 *The thresholding estimator $\hat{\Sigma}$ satisfies*

$$\sup_{\mathcal{P}(\mathcal{G}_q(\rho,c_{n,p}))} \mathbb{E}\left\|\hat{\Sigma} - \Sigma\right\|_1^2 \leq Cc_{n,p}^2\left(\frac{\log p}{n}\right)^{1-q}, \tag{27}$$

for some constant $C > 0$. Consequently, the minimax risk of estimating the covariance matrix Σ the distribution classes $\mathcal{P}\left(\mathcal{G}_q(\rho, c_{n,p})\right)$ satisfies

$$\inf_{\hat{\Sigma}} \sup_{\mathcal{P}(\mathcal{G}_q(\rho,c_{n,p}))} \mathbb{E}\left\|\hat{\Sigma} - \Sigma\right\|_1^2 \asymp c_{n,p}^2\left(\frac{\log p}{n}\right)^{1-q}. \tag{28}$$

A main technical tool for the proof of Theorem 4 is the next lemma, which is proved in Section 6.

Lemma 8 *Define the event A_{ij} by $A_{ij} = \left\{|\hat{\sigma}_{ij} - \sigma_{ij}| \leq 4\min\left\{|\sigma_{ij}|, \gamma\sqrt{\frac{\log p}{n}}\right\}\right\}$. Then*

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

Lemma 8 will be applied to show that the thresholding estimator defined in (26) is rate optimal over the distribution class $\mathcal{P}\left(\mathcal{G}_q(\rho, c_{n,p})\right)$.

Proof of Theorem 4: Let $D = (d_{ij})_{1\leq i,j\leq p}$ with $d_{ij} = (\hat{\sigma}_{ij} - \sigma_{ij})\, I(A_{ij}^c)$. Then

$$\begin{aligned}
\mathbb{E}\left\|\hat{\Sigma} - \Sigma\right\|_1^2 &\leq 2\mathbb{E}\left\|\hat{\Sigma} - \Sigma - D\right\|_1^2 + 2\mathbb{E}\left\|D\right\|_1^2 \\
&\leq 2\mathbb{E}\left[\sup_j \sum_i |\hat{\sigma}_{ij} - \sigma_{ij}|\, I(A_{ij})\right]^2 + 2\mathbb{E}\left\|D\right\|_1^2 \\
&\leq 32\left[\sup_j \sum_i \min\left\{|\sigma_{ij}|, \gamma\sqrt{\frac{\log p}{n}}\right\}\right]^2 + 2\mathbb{E}\left\|D\right\|_1^2. \qquad (29)
\end{aligned}$$

We will see that the first term in (29) is dominating and bounded by $Cc_{n,p}^2\left(\frac{\log p}{n}\right)^{1-q}$, while the second term, $\mathbb{E}\left\|D\right\|_1^2$, is negligible.

Pick a k^* such that $(c_{n,p}/k^*)^{1/q} \geq \sqrt{\frac{\log p}{n}} \geq [c_{n,p}/(k^*+1)]^{1/q}$, which implies $k^* \left(\frac{\log p}{n}\right)^{q/2} = (1+o(1))\, c_{n,p}$. Then we have

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

which gives (27) if $\mathbb{E}\|D\|_1^2 = O\left(\frac{1}{n}\right)$; this can be shown as follows. Note that

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

where

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

since $\mathbb{P}\left(A_{ij}^c\right) \leq 2C_1 p^{-9/2}$ from Lemma 8, and

$$\begin{aligned}
R_2 &= p\sum_{ij} \mathbb{E}\sigma_{ij}^2 I\left(A_{ij}^c \cap \{\hat{\sigma}_{ij} = 0\}\right) = p\sum_{ij} \mathbb{E}\sigma_{ij}^2 I(|\sigma_{ij}| \geq 4\gamma\sqrt{\frac{\log p}{n}}) I(|\sigma_{ij}^*| \leq \gamma\sqrt{\frac{\log p}{n}}) \\
&\leq p\sum_{ij} \sigma_{ij}^2 \mathbb{E} I(|\sigma_{ij}| \geq 4\gamma\sqrt{\frac{\log p}{n}}) I(|\sigma_{ij}| - \left|\sigma_{ij}^* - \sigma_{ij}\right| \leq \gamma\sqrt{\frac{\log p}{n}}) \\
&\leq p\sum_{ij} \sigma_{ij}^2 \mathbb{E} I(|\sigma_{ij}| \geq 4\gamma\sqrt{\frac{\log p}{n}}) I(\left|\sigma_{ij}^* - \sigma_{ij}\right| > \frac{3}{4}|\sigma_{ij}|) \\
&\leq \frac{p}{n}\sum_{ij} n\sigma_{ij}^2 C_1 \exp\left(-\frac{9}{2\gamma^2} n\sigma_{ij}^2\right) I(|\sigma_{ij}| \geq 4\gamma\sqrt{\frac{\log p}{n}}) \\
&= \frac{p}{n}\sum_{ij} \left[n\sigma_{ij}^2 \cdot C_1 \exp\left(-\frac{1}{2\gamma^2} n\sigma_{ij}^2\right)\right] \cdot \exp\left(-\frac{4}{\gamma^2} n\sigma_{ij}^2\right) I(|\sigma_{ij}| \geq 4\gamma\sqrt{\frac{\log p}{n}}) \\
&\leq C_9 \frac{p}{n} \cdot p^2 \cdot p^{-64} \leq C_9/n. \quad \blacksquare
\end{aligned}$$

3.2 Optimal estimation over $\mathcal{P}\left(\mathcal{F}_\alpha(\rho, M)\right)$ and $\mathcal{P}\left(\mathcal{H}_\alpha(\rho, M)\right)$

We now turn to optimal estimation over the distribution classes $\mathcal{P}\left(\mathcal{F}_\alpha(\rho, M)\right)$ and $\mathcal{P}\left(\mathcal{H}_\alpha(\rho, M)\right)$. We construct estimators of the covariance matrix Σ by tapering the maximum likelihood estimator Σ^*. For a given even integer k with $1 \le k \le p$, we define a tapering estimator as

$$\hat{\Sigma} = \hat{\Sigma}_k = \left(w_{ij}\sigma^*_{ij}\right)_{p\times p} \tag{30}$$

where σ^*_{ij} are the entries in the maximum likelihood estimator Σ^* and the weights are

$$w_{ij} = k_h^{-1}\{(k - |i-j|)_+ - (k_h - |i-j|)_+\} \tag{31}$$

with $k_h = k/2$. Without loss of generality we assume that k is even. Note that the weights w_{ij} can be rewritten as

$$w_{ij} = \begin{cases} 1 & \text{when } |i-j| \le k_h \\ 2 - \frac{|i-j|}{k_h} & \text{when } k_h < |i-j| < k \\ 0 & \text{otherwise.} \end{cases} .$$

See Figure 1 for a plot of the weights w_{ij} as a function of $|i-j|$.

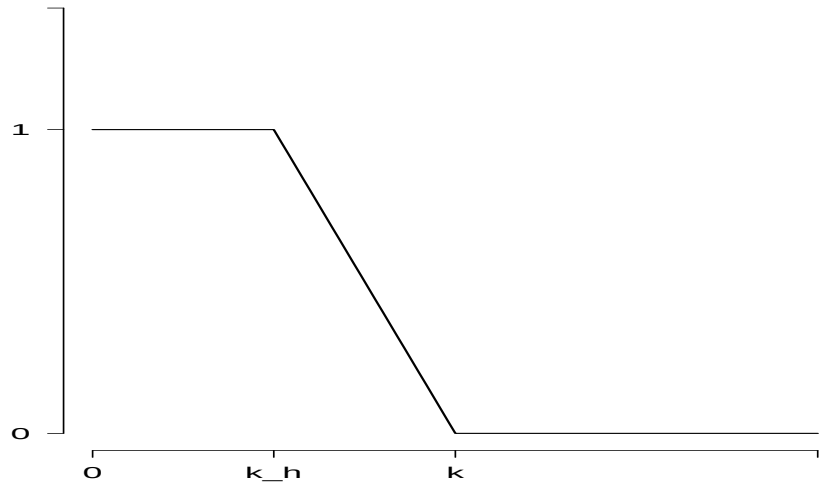

Figure 1: The weights as a function of $|i-j|$.

This class of tapering estimators was introduced in Cai, Zhang, and Zhou (2010) for covariance matrix estimation over the distribution class $\mathcal{P}\left(\mathcal{F}_\alpha(\rho, M)\right)$, and was shown to be minimax rate optimal under the spectral norm and Frobenius norm with appropriately

chosen tapering parameter k. The optimal choice of k critically depends on the norm under which the estimation error is measured. We shall see that the optimal choice of the tuning parameter under the ℓ_1 norm loss is different from that under either the spectral norm or the Frobenius norm. The tapering estimator defined in (30)has an important property: it can be rewritten as a sum of many small block matrices along the diagonal. This special property is useful for our technical arguments. Define the block matrices

$$U_l^{*(m)} = \left(\sigma_{ij}^* I\left\{l \le i < l+m, l \le j < l+m\right\}\right)_{p\times p}$$

and set

$$S^{*(m)} = \sum_{l=1-m}^{p} U_l^{*(m)}$$

for all integers $1-m \le l \le p$ and $m \ge 1$.

Lemma 9 *The tapering estimator $\hat{\Sigma}_k$ given in (30) can be written as*

$$\hat{\Sigma}_k = k_h^{-1}\left(S^{*(k)} - S^{*(k_h)}\right). \tag{32}$$

We now consider the performance of the tapering estimator under the ℓ_1 norm and establish the minimax upper bounds for the parameter spaces $\mathcal{P}\left(\mathcal{F}_\alpha(\rho, M)\right)$ and $\mathcal{P}\left(\mathcal{H}_\alpha(\rho, M)\right)$. We will show that the minimax rates of convergence over these two parameter spaces are the same under the ℓ_1 norm. Since $\mathcal{P}\left(\mathcal{H}_\alpha(\rho, M)\right) \subset \mathcal{P}\left(\mathcal{F}_\alpha(\rho, 2M/\alpha)\right)$, it thus suffices to establish the minimax upper bound for $\mathcal{P}\left(\mathcal{F}_\alpha(\rho, M)\right)$.

We focus on the case $p \ge n^{\frac{1}{2\alpha+2}}$. The case $p < n^{\frac{1}{2\alpha+2}}$, to be discussed in Section 3.2.1, is similar and slightly easier.

Theorem 5 *Suppose $p \ge n^{\frac{1}{2\alpha+2}}$. The tapering estimator $\hat{\Sigma}_k$ at (32) satisfies*

$$\sup_{\mathcal{A}} \mathbb{E}\left\|\hat{\Sigma}_k - \Sigma\right\|_1^2 \le C\frac{k^2 + k\log p}{n} + Ck^{-2\alpha} \tag{33}$$

for $k = o(n)$, $\log p = o(n)$, and some constant $C > 0$, where $\mathcal{A} = \mathcal{P}\left(\mathcal{F}_\alpha(\rho, M)\right)$ or $\mathcal{P}\left(\mathcal{H}_\alpha(\rho, M)\right)$. In particular, the estimator $\hat{\Sigma} = \hat{\Sigma}_k$ with

$$k = \min\left\{n^{\frac{1}{2\alpha+2}}, \left(\frac{n}{\log p}\right)^{\frac{1}{2\alpha+1}}\right\} \tag{34}$$

satisfies

$$\sup_{\mathcal{A}} \mathbb{E}\left\|\hat{\Sigma} - \Sigma\right\|_1^2 \le C\left[n^{-\frac{\alpha}{\alpha+1}} + \left(\frac{\log p}{n}\right)^{\frac{2\alpha}{2\alpha+1}}\right], \tag{35}$$

where $\mathcal{A} = \mathcal{P}\left(\mathcal{F}_\alpha(\rho, M)\right)$ and $\mathcal{P}\left(\mathcal{H}_\alpha(\rho, M)\right)$.

Together with Theorem 3, Theorem 5 shows that the tapering estimator with the optimal choice of the tapering parameter k given in (34) attains the optimal rate of convergence over both $\mathcal{P}\left(\mathcal{F}_\alpha(\rho, M)\right)$ and $\mathcal{P}\left(\mathcal{H}_\alpha(\rho, M)\right)$.

Proof of Theorem 5: It is easy to see that the minimum of $\frac{k^2+k\log p}{n} + k^{-2\alpha}$ is attained at $k \asymp n^{\frac{1}{2\alpha+2}}$ with the minimum value of order $n^{-\frac{\alpha}{\alpha+1}}$ when $p \le \exp\left(n^{\frac{1}{2\alpha+2}}\right)$. For $p \ge \exp\left(n^{\frac{1}{2\alpha+2}}\right)$, the minimum is attained at $k \asymp \left(\frac{n}{\log p}\right)^{\frac{1}{2\alpha+1}}$ and the minimum value is of order $\left(\frac{\log p}{n}\right)^{\frac{2\alpha}{2\alpha+1}}$.

Note that Σ^* is translation invariant and so is $\hat{\Sigma}$. We assume $\mathbb{E}\mathbf{X}_l = 0$ hereafter. Write

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

where $\bar{\mathbf{X}}\bar{\mathbf{X}}^T$ is a higher order term. Denote $\bar{\mathbf{X}}\bar{\mathbf{X}}^T$ by $G = (g_{ij})$. Since $\mathbb{E}g_{ij} \le C/n$, it is easy to see that

$$E\left\|\left(w_{ij}g_{ij}\right)_{p\times p}\right\|_1^2 \le C\frac{k^2\log p}{n^2} \le C\frac{k\log p}{n}, \quad \text{for } k \le n.$$

In what follows we ignore this negligible term and focus on the dominating term $\frac{1}{n}\sum_{l=1}^{n}\mathbf{X}_l\mathbf{X}_l^T$. Set $\tilde{\Sigma} = \frac{1}{n}\sum_{l=1}^{n}\mathbf{X}_l\mathbf{X}_l^T$ and write $\tilde{\Sigma} = (\tilde{\sigma}_{ij})_{1\le i,j\le p}$. Let

$$\breve{\Sigma} = (w_{ij}\tilde{\sigma}_{ij})_{1\le i,j\le p} \tag{36}$$

with w_{ij} given in (31). To prove Theorem 5, it suffices to show

$$\sup_{\mathcal{F}_\alpha(\rho,M)} \mathbb{E}\left\|\breve{\Sigma} - \Sigma\right\|_1^2 \le Cn^{-\frac{\alpha}{\alpha+1}} + C\left(\frac{\log p}{n}\right)^{\frac{2\alpha}{2\alpha+1}}. \tag{37}$$

Let $\mathbf{X}_l = \left(X_1^l, X_2^l, \ldots, X_p^l\right)^T$. We then write $\tilde{\sigma}_{ij} = \frac{1}{n}\sum_{l=1}^{n}X_i^lX_j^l$. It is easy to see

$$\mathbb{E}\tilde{\sigma}_{ij} = \sigma_{ij}, \tag{38}$$

$$\mathbb{V}ar(\tilde{\sigma}_{ij}) \le \frac{C_1}{n}, \tag{39}$$

for some $C_1 > 0$.

It is easy to bound the bias part,

$$\left\|\mathbb{E}\breve{\Sigma} - \Sigma\right\|_1^2 \le \left[\max_{i=1,\ldots,p}\sum_{j:|i-j|>k}|\sigma_{ij}|\right]^2 \le M^2k^{-2\alpha}. \tag{40}$$

We show that the variance

$$\mathbb{E}\left\|\breve{\Sigma}-\mathbb{E}\breve{\Sigma}\right\|_1^2 \leq C_2\frac{k^2+k\log p}{n}. \tag{41}$$

It then follows immediately that

$$\mathbb{E}\left\|\breve{\Sigma}-\Sigma\right\|_1^2 \leq 2\mathbb{E}\left\|\breve{\Sigma}-\mathbb{E}\breve{\Sigma}\right\|_1^2+2\left\|\mathbb{E}\breve{\Sigma}-\Sigma\right\|_1^2 \leq 2C_2\left(\frac{k^2+k\log p}{n}+k^{-2\alpha}\right).$$

This proves (37) and (33) then follows. Since $p \geq n^{\frac{1}{2\alpha+2}}$, we can set

$$k=\begin{cases}\left\lfloor n^{\frac{1}{2\alpha+2}}\right\rfloor, & \text{for } p \leq \exp\left(n^{\frac{1}{2\alpha+2}}\right)\\ \left\lfloor\left(\frac{n}{\log p}\right)^{\frac{1}{2\alpha+1}}\right\rfloor, & \text{otherwise}\end{cases} \tag{42}$$

and the estimator $\hat{\Sigma}$ with k given in (42) satisfies

$$\mathbb{E}\left\|\hat{\Sigma}-\Sigma\right\|_1^2 \leq 4C_2\left[n^{-\frac{\alpha}{\alpha+1}}+\left(\frac{\log p}{n}\right)^{\frac{2\alpha}{2\alpha+1}}\right].$$

Theorem 5 is then proved.

It remains to show (41). The key idea in the proof is to write the whole matrix as an average of a large number of small block matrices, and for each small block matrix the classical random matrix theory can be applied. The following lemma shows that the ℓ_1 norm of the random matrix $\breve{\Sigma}-\mathbb{E}\breve{\Sigma}$ is controlled by the maximum of p number of the ℓ_1 norms of $k\times k$ random matrices.

The next lemmas are proved in Section 6. Define

$$U_l^{(m)}=\left(\tilde{\sigma}_{ij}I\left\{l\leq i<l+m, l\leq j<l+m\right\}\right)_{p\times p} \tag{43}$$

for all integers $1-m\leq l\leq p$ and $m\geq 1$.

Lemma 10 *Let $\breve{\Sigma}$ be defined as in (32). Then*

$$\left\|\breve{\Sigma}-\mathbb{E}\breve{\Sigma}\right\|_1 \leq 3\max_{1\leq l\leq p-k+1}\left\|U_l^{(k)}-\mathbb{E}U_l^{(k)}\right\|_1.$$

Lemma 11 *There exists a constant $c_0>0$ such that*

$$\mathbb{P}\left\{\left\|U_l^{(m)}-\mathbb{E}U_l^{(m)}\right\|_1^2 > c_0\left(\frac{m^2}{n}+x^2\frac{m}{n}\right)\right\} \leq \exp\left(-2x^2\right) \tag{44}$$

for all $x>0$ and $1\leq l\leq p$.

It follows from Lemmas 10 and 11 that

$$\mathbb{E}\left\|\breve{\Sigma}-\mathbb{E}\breve{\Sigma}\right\|_1^2 \leq C_3\left(\frac{k^2+k\log p}{n}\right)+C_3k^{-2\alpha}$$

by plugging $x^2=C_4\max\{m,\log p\}$ into (44), for some $C_4>0$. ∎

The lower bound given in Theorem 3 and the upper bound given in Theorem 5 together show that the minimax risks over the distribution classes $\mathcal{P}(\mathcal{F}_\alpha(\rho,M))$ and $\mathcal{P}(\mathcal{H}_\alpha(\rho,M))$ when $p\geq n^{\frac{1}{2\alpha+2}}$, satisfy

$$\inf_{\hat{\Sigma}}\sup_{\mathcal{P}(\mathcal{F}_\alpha(\rho,M))}\mathbb{E}\left\|\hat{\Sigma}-\Sigma\right\|_1^2 \asymp \inf_{\hat{\Sigma}}\sup_{\mathcal{P}(\mathcal{H}_\alpha(\rho,M))}\mathbb{E}\left\|\hat{\Sigma}-\Sigma\right\|_1^2 \asymp n^{-\frac{\alpha}{\alpha+1}}+\left(\frac{\log p}{n}\right)^{\frac{2\alpha}{2\alpha+1}}. \quad (45)$$

3.2.1 Optimal estimation over $\mathcal{P}(\mathcal{F}_\alpha(\rho,M))$ and $\mathcal{P}(\mathcal{H}_\alpha(\rho,M))$: the case of $p<n^{\frac{1}{2\alpha+2}}$

For estimation over the distribution classes $\mathcal{P}(\mathcal{F}_\alpha(\rho,M))$ and $\mathcal{P}(\mathcal{H}_\alpha(\rho,M))$, for both the minimax lower and upper bounds, we have so far focused on the high dimensional case with $p\geq n^{\frac{1}{2\alpha+2}}$. In this section we consider the case $p<n^{\frac{1}{2\alpha+2}}$ and show that the minimax risk of estimating the covariance matrix Σ over the distribution classes $\mathcal{P}(\mathcal{F}_\alpha(\rho,M))$ and $\mathcal{P}(\mathcal{H}_\alpha(\rho,M))$ satisfies

$$\inf_{\hat{\Sigma}}\sup_{\mathcal{A}}\mathbb{E}\left\|\hat{\Sigma}-\Sigma\right\|_1^2 \asymp \frac{p^2}{n}$$

where $\mathcal{A}=\mathcal{P}(\mathcal{F}_\alpha(\rho,M))$ and $\mathcal{P}(\mathcal{H}_\alpha(\rho,M))$, when $p<n^{\frac{1}{2\alpha+2}}$.

This case is relatively easy. The upper bound can be attained by the sample covariance matrix $\hat{\Sigma}$. By (41) with $k=p$ we have,

$$\inf_{\hat{\Sigma}}\sup_{\mathcal{P}(\mathcal{F}_\alpha(\rho,M))}\mathbb{E}\left\|\hat{\Sigma}-\Sigma\right\|_1^2 \leq C\frac{p^2+p\log p}{n}\leq 2C\frac{p^2}{n}. \quad (46)$$

The lower bound can also be obtained by the application of Assouad's Lemma and by using the same parameter space $\mathcal{F}_{11}$ with $k=p$, i.e.,

$$\mathcal{F}_{11}=\left\{\Sigma(\theta):\Sigma(\theta)=I_p+\tau_2 n^{-1/2}\sum_{s=2}^{p}\theta_s\begin{bmatrix}(I\{i=1,j=s\})_{p\times p}\\+(I\{i=s,j=1\})_{p\times p}\end{bmatrix},\theta=(\theta_s)\in\{0,1\}^{p-1}\right\}$$

as in Section 2.2, where τ_2 satisfies $0<\tau_2<\min\{M,1/16\}$ such that the collection $\mathcal{F}_{11}\subset\mathcal{H}_\alpha(\rho,M)$. We obtain, as at (20) in Section 2.2.1,

$$\begin{aligned}\inf_{\hat{\Sigma}}\sup_{\mathcal{F}_{11}}\mathbb{E}\left\|\hat{\Sigma}-\Sigma\right\|_1^2 &\geq \min_{H(\theta,\theta')\geq 1}\frac{\|\Sigma(\theta)-\Sigma(\theta')\|_1}{H(\theta,\theta')}\frac{p-1}{2}\min_{H(\theta,\theta')=1}\|P_\theta\wedge P_{\theta'}\|\\ &\geq c\left(pn^{-1/2}\right)^2\geq c_1\frac{p^2}{n}. \quad (47)\end{aligned}$$

Inequalities (46) and (47) together yield the minimax rate of convergence for the case $p \leq n^{\frac{1}{2\alpha+2}}$,

$$\inf_{\hat{\Sigma}} \sup_{\mathcal{P}(\mathcal{F}_\alpha(\rho,M))} \mathbb{E}\left\|\hat{\Sigma}-\Sigma\right\|_1^2 \asymp \inf_{\hat{\Sigma}} \sup_{\mathcal{P}(\mathcal{H}_\alpha(\rho,M))} \mathbb{E}\left\|\hat{\Sigma}-\Sigma\right\|_1^2 \asymp \frac{p^2}{n}. \tag{48}$$

Combining (45) with (48), the optimal rate of convergence over two distribution classes $\mathcal{P}\left(\mathcal{F}_\alpha(\rho,M)\right)$ and $\mathcal{P}\left(\mathcal{H}_\alpha(\rho,M)\right)$ can be summarized as

$$\inf_{\hat{\Sigma}} \sup_{\mathcal{P}(\mathcal{F}_\alpha(\rho,M))} \mathbb{E}\left\|\hat{\Sigma}-\Sigma\right\|_1^2 \asymp \inf_{\hat{\Sigma}} \sup_{\mathcal{P}(\mathcal{H}_\alpha(\rho,M))} \mathbb{E}\left\|\hat{\Sigma}-\Sigma\right\|_1^2 \asymp \min\left\{n^{-\frac{\alpha}{\alpha+1}}+\left(\frac{\log p}{n}\right)^{\frac{2\alpha}{2\alpha+1}}, \frac{p^2}{n}\right\}.$$

4 Estimation of the inverse covariance matrix

In addition to the covariance matrix, the inverse covariance matrix Σ^{-1} is also of significant interest in many applications. The technical analysis given in the previous sections can be applied to obtain the minimax rate for estimating Σ^{-1} under the ℓ_1 norm.

For estimating the inverse covariance matrix Σ^{-1} it is necessary to require the ℓ_1 norm of Σ^{-1} to be bounded. For a positive constant $M_1 > 0$, set

$$\mathcal{G}_q(\rho, c_{n,p}, M_1) = \mathcal{G}_q(\rho, c_{n,p}) \cap \left\{\Sigma : \left\|\Sigma^{-1}\right\|_1 \leq M_1\right\}, \tag{49}$$

$$\mathcal{F}_\alpha(\rho, M, M_1) = \mathcal{F}_\alpha(\rho, M) \cap \left\{\Sigma : \left\|\Sigma^{-1}\right\|_1 \leq M_1\right\}, \tag{50}$$

$$\mathcal{H}_\alpha(\rho, M, M_1) = \mathcal{H}_\alpha(\rho, M) \cap \left\{\Sigma : \left\|\Sigma^{-1}\right\|_1 \leq M_1\right\}, \tag{51}$$

and define $\mathcal{P}\left(\mathcal{G}_q(\rho, c_{n,p}, M_1)\right)$ to be the set of distributions of $\mathbf{X}_1$ that satisfy both (4) and (49). The parameter spaces $\mathcal{P}\left(\mathcal{F}_\alpha(\rho, M, M_1)\right)$ and $\mathcal{P}\left(\mathcal{H}_\alpha(\rho, M, M_1)\right)$ are defined similarly.

Assume that

$$c_{n,p}^2\left(\frac{\log p}{n}\right)^{1-q} \to 0, \tag{52}$$

which is necessary to obtain a consistent estimator of Σ under ℓ_1 norm.

The following theorem gives the minimax rates of convergence for estimating Σ^{-1} over the three parameter spaces.

Theorem 6 *The minimax risk of estimating the inverse covariance matrix Σ^{-1} over the distribution class $\mathcal{P}\left(\mathcal{G}_q(\rho, c_{n,p}, M_1)\right)$ satisfies*

$$\inf_{\hat{\Omega}} \sup_{\mathcal{P}(\mathcal{G}_q(\rho,c_{n,p},M_1))} \mathbb{E}\left\|\hat{\Omega}-\Sigma^{-1}\right\|_1^2 \asymp c_{n,p}^2\left(\frac{\log p}{n}\right)^{1-q} \tag{53}$$

under assumptions (7) and (52), and the minimax risks of estimating the covariance matrix Σ *over the distribution classes* $\mathcal{P}\left(\mathcal{F}_\alpha(\rho, M, M_1)\right)$ *and* $\mathcal{P}\left(\mathcal{H}_\alpha(\rho, M, M_1)\right)$ *satisfy*

$$\inf_{\hat{\Omega}} \sup_{\mathcal{A}} \mathbb{E} \left\| \hat{\Omega} - \Sigma^{-1} \right\|_1^2 \asymp \min \left\{ n^{-\frac{\alpha}{\alpha+1}} + \left(\frac{\log p}{n} \right)^{\frac{2\alpha}{2\alpha+1}}, \frac{p^2}{n} \right\}, \tag{54}$$

where $\mathcal{A}$ *is* $\mathcal{P}\left(\mathcal{F}_\alpha(\rho, M, M_1)\right)$ *or* $\mathcal{P}\left(\mathcal{H}_\alpha(\rho, M, M_1)\right)$.

Remark 2 For estimating the inverse covariance matrix Σ^{-1}, we have assumed the ℓ_1 norm of Σ^{-1} to be uniformly bounded. This condition is satisfied if the variances σ_{ii} on the diagonal of Σ are bounded from below by some constant $c_0 > 0$ and the correlation matrix is diagonally dominant in the sense that

$$\max_{1 \le i \le p} \sum_{j, j \ne i} \frac{|\sigma_{ij}|}{\sqrt{\sigma_{ii}\sigma_{jj}}} \le 1 - \varepsilon \tag{55}$$

for some $\varepsilon > 0$. This can be seen as follows. Define $W_{p\times p} = \operatorname{diag}(\sigma_{11}, \ldots, \sigma_{pp})$, and write

$$\Sigma^{-1} = (W - (W - \Sigma))^{-1} = W^{-1/2} (I - V)^{-1} W^{-1/2},$$

where $V = W^{-1/2}(W - \Sigma) W^{-1/2}$. The assumption (55) implies that $\|V\|_1 \le 1 - \varepsilon$, so

$$(I - V)^{-1} = \sum_{i=0} V^i,$$

which implies

$$\left\|\Sigma^{-1}\right\|_1 \le \left\|W^{-1/2}\right\|_1^2 \left\|(I - V)^{-1}\right\|_1 \le c_0^{-1} \sum_{i=0} \|V\|_1^i \le (c_0 \varepsilon)^{-1}.$$

Proof of Theorem 6: The proof is similar to those for estimating the covariance matrix Σ. We only sketch the main steps below.

(I). Upper bounds. Let

$$\hat{\Omega} = \begin{cases} \hat{\Sigma}^{-1} & \text{if } \hat{\Sigma}^{-1} \text{ exists, and } \left\|\hat{\Sigma}^{-1}\right\|_1 \le n \\ I & \text{otherwise} \end{cases}.$$

Define the event $A_2 = \left\{ \hat{\Sigma}^{-1} \text{ exists, and } \left\|\hat{\Sigma}^{-1}\right\|_1 \le n \right\}$. On the event A_2 we write

$$\hat{\Sigma}^{-1} - \Sigma^{-1} = \hat{\Sigma}^{-1} \left(\Sigma - \hat{\Sigma} \right) \Sigma^{-1}$$

so that

$$\left\|\hat{\Sigma}^{-1} - \Sigma^{-1}\right\|_1 = \left\|\hat{\Sigma}^{-1} \left(\Sigma - \hat{\Sigma} \right) \Sigma^{-1}\right\|_1 \le \left\|\hat{\Sigma}^{-1}\right\|_1 \left\|\Sigma - \hat{\Sigma}\right\|_1 \left\|\Sigma^{-1}\right\|_1.$$

Note that

$$\left\|\hat{\Sigma}^{-1}\right\|_1 \leq \left\|\left(I+\left(\hat{\Sigma}-\Sigma\right)\Sigma^{-1}\right)^{-1}\right\|_1 \left\|\Sigma^{-1}\right\|_1 \leq \left\|\Sigma^{-1}\right\|_1 [1+\sum_{k=1}^{\infty}(\|H\|_1)^k], \tag{56}$$

where $H=\left(\hat{\Sigma}-\Sigma\right)\Sigma^{-1}$. Define

$$A_3=\left\{\left\|\hat{\Sigma}-\Sigma\right\|_1 \leq \varepsilon\right\}$$

for some $0<\varepsilon<\frac{1}{2M_1}$. It is easy to show that

$$\mathbb{P}\left(A_3^c\right) \leq C_D n^{-D} \tag{57}$$

for every $D>0$, using (24), (44), and (52). On A_3 we see that

$$\|H\|_1=\left\|\left(\hat{\Sigma}-\Sigma\right)\Sigma^{-1}\right\|_1 \leq \varepsilon\left\|\Sigma^{-1}\right\|_1<1/2.$$

Since $\left\|\Sigma^{-1}\right\|_1 \leq M_1$, which implies $\left\|\hat{\Sigma}^{-1}\right\|_1 \leq 2M_1$ on A_3 by (56), we have

$$\left\|\hat{\Omega}-\Sigma^{-1}\right\|_1^2=\left\|\hat{\Sigma}^{-1}-\Sigma^{-1}\right\|_1^2 \leq C\left\|\hat{\Sigma}-\Sigma\right\|_1^2$$

on $A_2\cap A_3$. It is actually easy to see $A_3\subset A_2$ and

$$\left\|\hat{\Omega}-\Sigma^{-1}\right\|_1^2 \leq Cn^2.$$

Let $\mathcal{B}$ be one of the three parameter spaces $\mathcal{P}\left(\mathcal{G}_q(\rho,c_{n,p},M_1)\right)$, $\mathcal{P}\left(\mathcal{F}_\alpha(\rho,M,M_1)\right)$, and $\mathcal{P}\left(\mathcal{H}_\alpha(\rho,M,M_1)\right)$. We have

$$\begin{aligned}\sup_{\mathcal{B}}\mathbb{E}\left\|\hat{\Omega}-\Sigma^{-1}\right\|_1^2 &= \sup_{\mathcal{B}}\mathbb{E}\left\{\left\|\hat{\Sigma}^{-1}-\Sigma^{-1}\right\|_1^2 I(A_3)\right\}+\sup_{\mathcal{B}}\mathbb{E}\left\{\left\|\hat{\Omega}-\Sigma^{-1}\right\|_1^2 I(A_3^c)\right\}\\ &\leq C\sup_{\mathcal{B}}\mathbb{E}\left\|\hat{\Sigma}-\Sigma\right\|_1^2+Cn^2\sup_{\mathcal{B}}\mathbb{P}\left(A_3^c\right) \leq C\sup_{\mathcal{B}}\mathbb{E}\left\|\hat{\Sigma}-\Sigma\right\|_1^2,\end{aligned}$$

where the last step follows from (57).

(II). Lower bounds. We use an elementary and unified argument to derive the lower bounds for estimating the inverse covariance matrices for all three parameter spaces. The basic strategy is to directly carry over the minimax lower bounds for estimating Σ to the ones for estimating Σ^{-1}. The following is a simple but very useful observation. Note that

$$\left\|\Sigma_1-\Sigma_2\right\|_1=\left\|\Sigma_1\left(\Sigma_1^{-1}-\Sigma_2^{-1}\right)\Sigma_2\right\|_1 \leq \left\|\Sigma_1\right\|_1\left\|\Sigma_1^{-1}-\Sigma_2^{-1}\right\|_1\left\|\Sigma_2\right\|_1,$$

which implies

$$\left\|\Sigma_1^{-1}-\Sigma_2^{-1}\right\|_1 \geq \left\|\Sigma_1\right\|_1^{-1}\left\|\Sigma_2\right\|_1^{-1}\left\|\Sigma_1-\Sigma_2\right\|_1.$$

If $\|\Sigma_1\|_1 \leq C$ and $\|\Sigma_2\|_1 \leq C$ for some $C > 0$, we have

$$\left\|\Sigma_1^{-1} - \Sigma_2^{-1}\right\|_1 \geq C^{-2} \|\Sigma_1 - \Sigma_2\|_1 . \tag{58}$$

Equation (58) shows that a lower bound for estimating Σ yields one for estimating Σ^{-1} over the same parameter space.

We first consider the lower bounds for $\mathcal{P}(\mathcal{H}_\alpha(\rho, M, M_1))$. Set $\mathcal{F}_0 = \mathcal{F}_{11} \cup \mathcal{F}_{12}$, where $\mathcal{F}_{11}$ and $\mathcal{F}_{12}$ are defined in (14) and (15), respectively. Over the parameter space $\mathcal{F}_{11}$ the proof is almost identical to the proof of the lower bound (16) in Section 2.2 except that here we need to show

$$\min_{H(\theta,\theta') \geq 1} \frac{\left\|\Sigma^{-1}(\theta) - \Sigma^{-1}(\theta')\right\|_1}{H(\theta, \theta')} \geq ca$$

instead of (21), for some $c > 0$. Actually the inequality follows from (21) together with (58), since $\|\Sigma(\theta)\|_1$ and $\|\Sigma(\theta')\|_1$ are bounded above by a finite constant. For $\mathcal{F}_{12}$ the lower bound argument is almost identical to the proof of the lower bound (17) by using a version of Fano's Lemma, except that we need

$$\|\Sigma_i^{-1} - \Sigma_j^{-1}\|_1 \geq \sqrt{c\frac{k \log p}{n}}$$

for some $c > 0$ and all $0 \leq i \neq j \leq m_*$ instead of (22). The inequality follows from (22) and (58).

The proof for the lower bound for the parameter space $\mathcal{P}(\mathcal{G}_q(\rho, c_{n,p}, M_1))$ is almost identical to that of Theorem 2. The only different argument in the proof is that

$$\inf_m \left\|\Sigma_m^{-1} - \Sigma_0^{-1}\right\|_1^2 \geq C c_{n,p}^2 \left(\frac{\log p}{n}\right)^{1-q}$$

for some $C > 0$; this is true since $\|\Sigma_m\|_1$ is uniformly bounded from above by a fixed constant. ∎

5 Discussions

In this paper we have established the optimal rates of convergence for estimating the covariance matrices over the three commonly used parameter spaces under the matrix ℓ_1 norm. Deriving the minimax lower bounds requires a careful construction of collections of least favorable multivariate normal distributions and the application of different lower bound techniques in various settings. The lower bound arguments also provide insight into where the difficulties of the covariance matrix estimation problem arise.

It is shown that the thresholding estimator originally introduced in Bickel and Levina (2008b) for estimating sparse covariance matrices under the spectral norm attains the optimal rate of convergence over the parameter space $\mathcal{P}\left(\mathcal{G}_q(\rho, c_{n,p})\right)$ under the matrix ℓ_1 norm. For minimax estimation over the other two parameter spaces $\mathcal{P}\left(\mathcal{F}_\alpha(\rho, M)\right)$ and $\mathcal{P}\left(\mathcal{H}_\alpha(\rho, M)\right)$, a tapering estimator is constructed and shown to be rate optimal. For estimation over these two parameter spaces, compared to the optimal tapering estimators under the spectral and Frobenius norms given in Cai, Zhang, and Zhou (2010), the best choice of the tapering parameter is different under the ℓ_1 norm. Consider the case $p \geq n$. The optimal choice of k under the ℓ_1 norm is

$$k_1 = \min\left\{ n^{\frac{1}{2\alpha+2}},\ \left(\frac{n}{\log p}\right)^{\frac{1}{2\alpha+1}} \right\}.$$

In contrast, the best choice of k under the spectral norm is $k_2 = n^{\frac{1}{2\alpha+1}}$, which is always larger than k_1. For estimation under the Frobenius norm, the optimal choice of k over $\mathcal{P}\left(\mathcal{H}_\alpha(\rho, M)\right)$ is $k_F = n^{\frac{1}{2\alpha+2}}$. This coincides with k_1 when $\log p \leq n^{\frac{1}{2\alpha+2}}$, and $k_F > k_1$ when $\log p \gg n^{\frac{1}{2\alpha+2}}$.

For estimation over the parameter spaces $\mathcal{P}\left(\mathcal{F}_\alpha(\rho, M)\right)$ and $\mathcal{P}\left(\mathcal{H}_\alpha(\rho, M)\right)$, it is also interesting to compare with the banding estimator introduced in Bickel and Levina (2008a). They considered the estimator

$$\hat{\Sigma}_B = \left(\sigma^*_{ij} I\left\{|i-j| \leq k\right\}\right)$$

and proposed the banding parameter

$$k = \left(\frac{n}{\log p}\right)^{\frac{1}{2\alpha+2}}.$$

Although this estimator was originally introduced for estimation under the spectral norm, it is still interesting to consider its performance under the matrix ℓ_1 norm. The estimator achieves the rate of convergence $\left(\frac{\log p}{n}\right)^{\frac{\alpha}{\alpha+1}}$ under the matrix ℓ_1 norm, which is inferior to the optimal rate $\min\left\{ n^{-\frac{\alpha}{\alpha+1}} + \left(\frac{\log p}{n}\right)^{\frac{2\alpha}{2\alpha+1}},\ \frac{p^2}{n} \right\}$ given at (6). Take for example $\alpha = 1/2$ and $p = e^{\sqrt{n}}$. In this case $\left(\frac{\log p}{n}\right)^{\frac{\alpha}{\alpha+1}} = n^{-\frac{1}{6}}$, while the optimal rate is $n^{-\frac{1}{4}}$. On the other hand, it can be shown by using (44) that the banding estimator with the same optimal k for the tapering estimator described at (34) of Section 3.2 is also rate optimal. In this sense there is no fundamental differences between the tapering and banding estimators for estimation over these two parameter spaces. We leave the detailed technical argument to the readers.

Our technical analysis also shows that covariance matrix estimation has quite different characteristics from those in the classical Gaussian sequence estimation problems. Johnstone (2011) gives a comprehensive treatment of minimax and adaptive estimation under the Gaussian sequence models. See also Abramovich, Benjamini, Donoho and Johnstone (2006) for Gaussian sequence estimation in the context of wavelet thresholding. In the matrix estimation problems, with the exception of the squared Frobenius norm loss, the loss functions are typically not separable as in the sequence estimation problems. For example, in this paper the loss function is not the usual squared vector ℓ_2 norm or vector ℓ_1 norm, which are sums of elementwise losses, but is the matrix ℓ_1 norm,

$$L\left(\hat{\Sigma}, \Sigma\right) = \max_i \sum_j \left|\hat{\sigma}_{ij} - \sigma_{ij}\right|.$$

This loss can be viewed as the maximum of p number of ℓ_1 losses for vectors and it cannot be decomposed as a sum of elementwise losses. Similarly the spectral norm loss is also not separable. This makes the theoretical analysis of the matrix estimation problems more involved. In addition, each element σ_{ij}^* of the sample covariance matrix is asymptotically normal with the mean σ_{ij} and the standard deviation of order $1/\sqrt{n}$, but the σ_{ij}^*'s are neither exactly normal nor homoskedastic as in the classical Gaussian sequence estimation problems. In addition, the σ_{ij}^*'s are dependent. These create additional technical complications and more care is thus needed.

In Cai and Zhou (2011) and Cai, Liu, and Zhou (2011), we considered the problems of optimal estimation of sparse covariance and sparse precision matrices under the spectral norm. The spectral norm is bounded from above by the matrix ℓ_1 norm, but is often much smaller than the matrix ℓ_1 norm. The lower bounds in this paper are not sufficient for optimal estimation in those settings. New and much more involved lower bounds arguments are developed in Cai and Zhou (2011) and Cai, Liu, and Zhou (2011) to overcome the technical difficulties there.

6 Proofs of technical lemmas

We prove the technical lemmas that are used in the proofs of the main results in the previous sections.

Proof of Lemma 5: When $H\left(\theta, \theta'\right) = 1$, Pinsker's Inequality (see, e.g., Csiszár (1967)) implies

$$\left\|\mathbb{P}_{\theta'} - \mathbb{P}_{\theta}\right\|_1^2 \leq 2K\left(\mathbb{P}_{\theta'} | \mathbb{P}_{\theta}\right) = n\left[tr\left(\Sigma\left(\theta'\right)\Sigma\left(\theta\right)^{-1}\right) - \log\det\left(\Sigma\left(\theta'\right)\Sigma\left(\theta\right)^{-1}\right) - p\right].$$

For a matrix $A = (a_{ij})$, let $\|A\|_F = \sqrt{\sum_{ij} a_{i,j}^2}$. It is easy to see that

$$tr\left(\Sigma\left(\theta'\right)\Sigma\left(\theta\right)^{-1}\right) - \log\det\left(\Sigma\left(\theta'\right)\Sigma\left(\theta\right)^{-1}\right) - p \leq \left\|\Sigma\left(\theta'\right) - \Sigma\left(\theta\right)\right\|_F^2 \tag{59}$$

when $\|\Sigma(\theta) - I\|_2 \leq 1/4$ and $\|\Sigma(\theta') - I\|_2 \leq 1/4$, and

$$\|\Sigma(\theta) - I\|_2 \leq \|\Sigma(\theta) - I\|_1 \leq \tau_2 k n^{-1/2} \leq \tau_2 < 1/4 \tag{60}$$

for $\tau_2 < 1/16$. Inequalities (59) and (60) imply

$$\|\mathbb{P}_{\theta'} - \mathbb{P}_\theta\|_1^2 \leq n\left\|\Sigma\left(\theta'\right) - \Sigma\left(\theta\right)\right\|_F^2 = n \cdot 2\tau_2^2\left(n^{-1/2}\right)^2 = 2\tau_2^2 < 1,$$

and the lemma follows immediately. ∎

Proof of Lemma 7: When $\tau_2 < 1/16$,

$$\|\Sigma(\theta_j) - I\|_2 \leq \|\Sigma(\theta_j) - I\|_1 \leq \sqrt{\tau_2 \log p}kb = \sqrt{\frac{\tau_2 k \log p}{n}} = \sqrt{\tau_2}\left(\frac{\log p}{n}\right)^{-\frac{\alpha}{2\alpha+1}} < 1/4.$$

Inequality (59) gives

$$K\left(\mathbb{P}_{\theta_j}, \mathbb{P}_{\theta_0}\right) \leq n\left\|\Sigma(\theta_j) - \Sigma(\theta_0)\right\|_F^2 \leq n \cdot 2\tau_2 k b^2 \log p \leq 2\tau_2 \log p. \quad ∎$$

Proof of Lemma 8: Let $A_1 = \left\{\left|\sigma_{ij}^*\right| \geq \gamma\sqrt{\frac{\log p}{n}}\right\}$. From the definition of $\hat{\sigma}_{ij}$ we have

$$|\hat{\sigma}_{ij} - \sigma_{ij}| = |\sigma_{ij}| \cdot I(A_1) + |\sigma_{ij}^* - \sigma_{ij}| \cdot I(A_1^c).$$

It is easy to see

$$\begin{aligned} A_1 &= \left\{\left|\sigma_{ij}^* - \sigma_{ij} + \sigma_{ij}\right| \geq \gamma\sqrt{\frac{\log p}{n}}\right\} \subset \left\{\left|\sigma_{ij}^* - \sigma_{ij}\right| \geq \gamma\sqrt{\frac{\log p}{n}} - |\sigma_{ij}|\right\}, \\ \text{and } A_1^c &= \left\{\left|\sigma_{ij}^* - \sigma_{ij} + \sigma_{ij}\right| < \gamma\sqrt{\frac{\log p}{n}}\right\} \subset \left\{\left|\sigma_{ij}^* - \sigma_{ij}\right| > |\sigma_{ij}| - \gamma\sqrt{\frac{\log p}{n}}\right\} \end{aligned}$$

by the triangle inequality. Note that (25) implies

$$\begin{aligned} \mathbb{P}(A_1) &\leq \mathbb{P}\left(\left|\sigma_{ij}^* - \sigma_{ij}\right| > \frac{3\gamma}{4}\sqrt{\frac{\log p}{n}}\right) \leq C_1 p^{-9/2}, \quad \text{when } |\sigma_{ij}| < \tfrac{\gamma}{4}\sqrt{\tfrac{\log p}{n}}, \\ \mathbb{P}(A_1^c) &\leq \mathbb{P}\left(\left|\sigma_{ij}^* - \sigma_{ij}\right| > \gamma\sqrt{\frac{\log p}{n}}\right) \leq C_1 p^{-8}, \quad \text{when } |\sigma_{ij}| > 2\gamma\sqrt{\tfrac{\log p}{n}}. \end{aligned}$$

Thus

$$|\hat{\sigma}_{ij} - \sigma_{ij}| = \begin{cases} |\sigma_{ij}| & |\sigma_{ij}| < \frac{\gamma}{4}\sqrt{\frac{\log p}{n}} \\ \left|\sigma_{ij}^* - \sigma_{ij}\right| \text{ or } |\sigma_{ij}| & \frac{\gamma}{4}\sqrt{\frac{\log p}{n}} \leq |\sigma_{ij}| \leq 2\gamma\sqrt{\frac{\log p}{n}} \\ \left|\sigma_{ij}^* - \sigma_{ij}\right| & |\sigma_{ij}| > 2\gamma\sqrt{\frac{\log p}{n}} \end{cases}$$

with a probability of at least $1 - C_1 p^{-9/2}$ for all settings. Since

$$\mathbb{P}\left(\left|\sigma_{ij}^* - \sigma_{ij}\right| \leq \gamma \sqrt{\frac{\log p}{n}} \right) \geq 1 - C_1 p^{-8},$$

it then is easy to see that for each of the three settings above we have

$$|\hat{\sigma}_{ij} - \sigma_{ij}| \leq 4 \min \left\{ |\sigma_{ij}|, \gamma \sqrt{\frac{\log p}{n}} \right\}$$

with a probability of at least $1 - 2C_1 p^{-9/2}$. ∎

Proof of Lemma 9: It is easy to see

$$\begin{aligned} kw_{ij} &= \#\{l : (i,j) \subset \{l, \ldots, l+2k-1\}\} - \#\{l : (i,j) \subset \{l, \ldots, l+k-1\}\} \\ &= (2k - |i-j|)_+ - (k - |i-j|)_+, \end{aligned}$$

which takes value in $[0, k]$. Clearly from the above, $kw_{ij} = k$ for $|i-j| \leq k$.∎

Proof of Lemma 10: Set $S^{(m)} = \sum_{l=1-m}^{p} U_l^{(m)}$. Without loss of generality we assume that p can be divided by m. Set $\delta_l^{(m)} = U_l^{(m)} - \mathbb{E}U_l^{(m)}$. By (32)

$$\left\| S^{(m)} - \mathbb{E}S^{(m)} \right\|_1 \leq \sum_{l=1}^{m} \left\| \sum_{-1 \leq j \, < \, p/m} \delta_{jm+l}^{(m)} \right\|_1 . \tag{61}$$

Since the $\delta_{jm+l}^{(m)}$ are diagonal blocks of their sum over $-1 \leq j < p/m$, we have

$$\left\| S^{(m)} - \mathbb{E}S^{(m)} \right\|_1 \leq m \max_{1 \leq l \leq m} \left\| \sum_{0 \leq j \, < \, p/m} \delta_{jm+l}^{(m)} \right\|_1 \leq m \max_{2-m \leq l \leq p} \left\| \delta_l^{(m)} \right\|_1 .$$

This and (32) imply the conclusion, since $\delta_l^{(k)}$ and $\delta_l^{(2k)}$ are all sub-blocks of a certain matrix $\delta_l^{(2k)}$ with $1 \leq l \leq p - 2k + 1$. ∎

Proof of Lemma 11: A key technical tool for the extension is the following lemma which was established in Section 7 of Cai, Zhang, and Zhou (2010).

Lemma 12 *There is a constant $\rho_1 > 0$ such that*

$$\mathbb{P}\left\{ \left\| U_l^{(m)} - \mathbb{E}U_l^{(m)} \right\| > x \right\} \leq 5^m \exp\left(-nx^2\rho_1\right)$$

for all $0 < x < \rho_1$ and $1 - m \leq l \leq p$.

Set $c_0 = 2/\rho_1$. From the fact $\|A_{m\times m}\|_1^2 \le m\,\|A_{m\times m}\|^2$ for any symmetric matrix $A_{m\times m}$ and Lemma 12, we have

$$\begin{aligned}\mathbb{P}\left\{\left\|U_l^{(m)} - \mathbb{E}U_l^{(m)}\right\|_1^2 > c_0\left(\frac{m^2}{n} + x^2\frac{m}{n}\right)\right\} &\le \mathbb{P}\left\{\left\|U_l^{(m)} - \mathbb{E}U_l^{(m)}\right\|^2 > c_0\left(\frac{m}{n} + \frac{x^2}{n}\right)\right\} \\ &\le 5^m \exp\left(-c_0\left(m + x^2\right)\rho_1\right) \\ &= \left(\frac{5}{e^2}\right)^m \exp\left(-2x^2\right) \le \exp\left(-2x^2\right). \quad \blacksquare\end{aligned}$$

Acknowledgment We thank the two referees for constructive comments that have helped to improve the presentation of the paper. We also thank Weidong Liu for helpful discussion which led to a concise proof of Lemma 12.

References

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.

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

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

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

Cai, T. T. and Liu, W. (2011). Adaptive thresholding for sparse covariance matrix estimation. *J. Amer. Statist. Assoc.*, to appear.

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

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

Cai, T. T. and Zhou, H. H. (2011). Optimal rates of convergence for sparse covariance matrix estimation. Technical report.

Csiszár, I. (1967). Information-type measures of difference of probability distributions and indirect observation. *Studia Scientiarum Mathematicarum Hungarica* **2**, 229–318.

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

Johnstone, I. M. (2011). *Gaussian Estimation: Sequence And Multiresolution Models.* Manuscript. Available at: `www-stat.stanford.edu/~imj/Book030811.pdf`.

Johnstone, I. M. and Lu, A. Y. (2009). On consistency and sparsity for principal components analysis in high dimensions. *J. Amer. Statist. Assoc.* **104**, 682-693.

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

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

Saulis, L. and Statulevičius, V. A. (1991). *Limit Theorems for Large Deviations.* Kluwer Academic Publishers, Dordrecht.

Tsybakov, A. B. (2009). *Introduction to Nonparametric Estimation.* Springer-Verlag.

Wu, W. B. and Pourahmadi, M. (2003). Nonparametric estimation of large covariance matrices of longitudinal data. *Biometrika* **90**, 831-844.

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

Zou, H., Hastie, T., and Tibshirani, R. (2006). Sparse principal components analysis. *Journal of Computational and Graphical Statistics* **15**, 265-286.