Statistica Sinica **22** (2012), 1319-1378
doi:http://dx.doi.org/10.5705/ss.2010.253

MINIMAX ESTIMATION OF LARGE COVARIANCE MATRICES UNDER ℓ_1-NORM

T. Tony Cai and Harrison H. Zhou

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.

Key words and phrases: Covariance matrix, ℓ_1 norm, minimax lower bound, operator norm, optimal rate of convergence, tapering, thresholding.

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 $\Sigma = (\sigma_{ij})_{1\leq i,j\leq p}$ 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. Another 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) $(\sigma_{ij})_{1\leq i\leq p}$ 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 $(\sigma_{ij})_{1\leq i\leq p}$. For $0 \leq q < 1$, define

$$\mathcal{G}_q(\rho, c_{n,p}) = \left\{ \Sigma = (\sigma_{ij})_{1\leq i,j\leq p} : \max_{1\leq j\leq p} \left\{ \left|\sigma_{[k]j}\right|^q \right\} \leq \frac{c_{n,p}}{k},\ \forall k, \text{ and } \max_i (\sigma_{ii}) \leq \rho \right\}. \tag{1.1}$$

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 without loss a of generality, we shall assume $c_{n,p} \geq 1$. The weak ℓ_q ball has been used in Abramovich et al. (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, p.5) as a special case. The second class of parameter spaces under study is

$$\mathcal{F}_\alpha(\rho, M) = \left\{ \Sigma : \max_j \sum_i |\sigma_{ij}| \{i : |i-j| > k\} \le M k^{-\alpha}, \forall k, \text{ and } \max_i (\sigma_{ii}) \le \rho \right\}, \tag{1.2}$$

where $\alpha > 0$, $M > 0$, and $\rho > 0$. The parameter α in (1.2), 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 $\mathcal{F}_\alpha$:

$$\mathcal{H}_\alpha(\rho, M) = \left\{ \Sigma : |\sigma_{ij}| \le M |i-j|^{-(\alpha+1)} \text{ for } i \ne j \text{ and } \max_i (\sigma_{ii}) \le \rho \right\}. \tag{1.3}$$

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 $\|\mathbf{v}\|_2 = 1$,

$$\mathbb{P}\{|\mathbf{v}'(\mathbf{X}_1 - \mathbb{E}\mathbf{X}_1)| > t\} \le e^{-t^2/(2\rho)}. \tag{1.4}$$

Let $\mathcal{P}(\mathcal{G}_q(\rho, c_{n,p}))$ denote the set of distributions of $\mathbf{X}_1$ satisfying (1.1) and (1.4). The distribution classes $\mathcal{P}(\mathcal{F}_\alpha(\rho, M))$ and $\mathcal{P}(\mathcal{H}_\alpha(\rho, M))$ are defined similarly. Our analysis establishes the minimax rates of convergence for estimating the covariance matrices over the three distribution classes $\mathcal{P}(\mathcal{G}_q(\rho, c_{n,p}))$, $\mathcal{P}(\mathcal{F}_\alpha(\rho, M))$, and $\mathcal{P}(\mathcal{H}_\alpha(\rho, M))$. 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 $\mathcal{P}(\mathcal{G}_q(\rho, c_{n,p}))$ satisfies*

$$\inf_{\hat{\Sigma}} \sup_{\mathcal{P}(\mathcal{G}_q(\rho, c_{n,p}))} \mathbb{E} \left\| \hat{\Sigma} - \Sigma \right\|_1^2 \asymp c_{n,p}^2 \left(\frac{\log p}{n} \right)^{1-q} \tag{1.5}$$

under assumptions (2.1) *and* (2.2), *and the minimax risks of estimating the covariance matrix Σ over the distribution classes $\mathcal{P}(\mathcal{F}_\alpha(\rho, M))$ and $\mathcal{P}(\mathcal{H}_\alpha(\rho, M))$*

satisfy

$$\inf_{\hat{\Sigma}} \sup_{\mathcal{A}} \mathbb{E} \left\| \hat{\Sigma} - \Sigma \right\|_1^2 \asymp \min \left\{ n^{-\alpha/(\alpha+1)} + \left(\frac{\log p}{n} \right)^{2\alpha/(2\alpha+1)}, \frac{p^2}{n} \right\}, \tag{1.6}$$

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

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 under operator norm losses 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 $\mathcal{P}(\mathcal{G}_q(\rho, c_{n,p}))$ 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 $\mathcal{P}(\mathcal{G}_q(\rho, c_{n,p}))$ under the matrix ℓ_1 norm. For the other two distribution classes $\mathcal{P}(\mathcal{F}_\alpha(\rho, M))$ and $\mathcal{P}(\mathcal{H}_\alpha(\rho, M))$, 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 (1.1). 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

$$p \ge n^\nu \text{ with } \nu > 1, \tag{2.1}$$

$\log p \le n$ and assume that

$$c_{n,p} \le M \left(\frac{n}{\log p}\right)^{1/2 - q/2} \tag{2.2}$$

for $0 \leq q < 1$ and some $M > 0$. Theorem 2 below implies that the assumption $c_{n,p}^2 (\log p/n)^{1-q} \to 0$ 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 $k = \left\lfloor c_{n,p} (n/\log p)^{q/2} \right\rfloor$. 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

$$\mathcal{G}_0 = \left\{\Sigma : \Sigma = I_p \text{ or } \Sigma = I_p + aH, \text{ for some } H \in \mathcal{H}\right\}, \tag{2.3}$$

where $a = \sqrt{\tau_1 \log p/n}$ for some constant τ_1. Without loss of generality we assume that $\rho > 1$ in (1.1). 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 $0 < \tau_1 < \min\left\{1, (\nu - 1)/4\nu, 1/(2M^2)\right\}$. 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 $\hat{\Sigma}$,

$$\sup_{\mathcal{G}_0} \mathbb{E} \left\| \hat{\Sigma} - \Sigma \right\|_1^2 \geq C_1 c_{n,p}^2 \left(\frac{\log p}{n}\right)^{1-q}, \tag{2.4}$$

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* (2.1) *and* (2.2)*, the minimax risk of estimating the covariance matrix $\Sigma_{p\times p}$ satisfies*

$$\inf_{\hat{\Sigma}} \sup_{\mathcal{G}_q(\rho, c_{n,p})} \mathbb{E} \left\| \hat{\Sigma} - \Sigma \right\|_1^2 \geq C_1 c_{n,p}^2 \left(\frac{\log p}{n}\right)^{1-q}, \tag{2.5}$$

where C_1 is a positive constant.

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

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

where the last inequality follows from (2.5). Furthermore by a similar argument as above, we need the condition $c_{n,p}^2 (\log p/n)^{1-q} \to 0$ to estimate Σ consistently under the ℓ_1 norm.

Results in Section 3 show that the lower bound given in (2.5) is minimax rate optimal. A threshold estimator is shown to attain the convergence rate given in (2.5).

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 $\Sigma_m, 1 \le m \le m_*$, 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 $\bar{\mathbb{P}} = (1/m_*) \sum_{m=1}^{m_*} \mathbb{P}_{\Sigma_m}$.

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*

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

Proof of Theorem 2. It is easy to see that

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

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

$$\left\| \mathbb{P}_{\Sigma_0} \wedge \bar{\mathbb{P}} \right\| \ge C_3. \tag{2.6}$$

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 (2.5) up to a constant factor.

Now we establish the lower bound (2.6) for the total variation affinity. Since the affinity $\int q_0 \wedge q_1 d\mu = 1 - (1/2) \int |q_0 - q_1| \, d\mu$ 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 $\int q_0 \wedge q_1 d\mu \geq 1 - (1/2) \left[\int \left(q_1^2/q_0 - 1\right) d\mu\right]^{1/2}$. To establish (2.6), 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*

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

Lemma 2 implies

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

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

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

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,

$$\Delta = \frac{1}{m_*^2}\sum_{0\leq j\leq k}\sum_{J(m,l)=j}\int\left(\frac{f_m f_l}{f_0} - 1\right) = \frac{1}{m_*^2}\sum_{0\leq j\leq k}\sum_{J(m,l)=j}\left[\left(1 - ja^2\right)^{-n} - 1\right]$$
$$\leq \frac{1}{m_*^2}\sum_{1\leq j\leq k}\binom{p-1}{k}\binom{k}{j}\binom{p-1-k}{k-j}(1 - ja^2)^{-n}. \tag{2.7}$$

Note that

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

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

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

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

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

where the last step follows from the fact that $\tau_1 \leq (\nu - 1)/(4\nu)$. Thus $\Delta \leq Cn^{(1-\nu)/2} \to 0$, which immediately implies (2.6).

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 $p \geq n^{1/(2\alpha+2)}$. The case $p < n^{1/(2\alpha+2)}$ 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 $p \leq \exp\left(n^{1/(2\alpha+2)}\right)$ and $p > \exp\left(n^{1/(2\alpha+2)}\right)$. For $p \leq \exp\left(n^{1/(2\alpha+2)}\right)$, set $k = \left\lfloor n^{1/(2\alpha+2)} \right\rfloor$. 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,

$$\mathcal{F}_{11} = \left\{ \Sigma(\theta) : \Sigma(\theta) = I_p + \tau_2 n^{-1/2} \sum_{s=2}^{k} \theta_s \begin{bmatrix} (I\{i = 1, j = s\})_{p\times p} \\ + (I\{i = s, j = 1\})_{p\times p} \end{bmatrix}, \right.$$
$$\left. \theta = (\theta_s) \in \{0,1\}^{k-1} \right\}, \quad (2.8)$$

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 $\Sigma = (\sigma_{i,j}) \in \mathcal{F}_{11}$,

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

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

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

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

$$\mathcal{F}_{12} = \left\{\Sigma_m : \Sigma_m = I_p + b\sqrt{\tau_2 \log p} B_m,\ 1 \leq m \leq m_*\right\}, \tag{2.9}$$

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 $C_4 > 0$,

$$\inf_{\hat{\Sigma}} \sup_{\mathcal{F}_{11}} \mathbb{E}\left\|\hat{\Sigma} - \Sigma\right\|_1^2 \geq C_4 n^{-\alpha/(\alpha+1)}, \tag{2.10}$$

$$\inf_{\hat{\Sigma}} \sup_{\mathcal{F}_{12}} \mathbb{E}\left\|\hat{\Sigma} - \Sigma\right\|_1^2 \geq C_4 \left(\frac{\log p}{n}\right)^{2\alpha/(2\alpha+1)}. \tag{2.11}$$

Equations (2.10) and (2.11) together imply

$$\inf_{\hat{\Sigma}} \sup_{\mathcal{F}_0} \mathbb{E}\left\|\hat{\Sigma} - \Sigma\right\|_1^2 \geq \frac{C_4}{2} \left[n^{-\alpha/(\alpha+1)} + \left(\frac{\log p}{n}\right)^{2\alpha/(2\alpha+1)}\right], \tag{2.12}$$

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 $C > 0$,*

$$\inf_{\hat{\Sigma}} \sup_{\Sigma \in \mathcal{A}} \mathbb{E}\left\|\hat{\Sigma} - \Sigma\right\|_1^2 \geq C \left[n^{-\alpha/(\alpha+1)} + \left(\frac{\log p}{n}\right)^{2\alpha/(2\alpha+1)}\right], \tag{2.13}$$

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 (2.13) is optimal. A specific tapering estimator is constructed and shown to attain the minimax rate of convergence $n^{-\alpha/(\alpha+1)} + (\log p/n)^{2\alpha/(2\alpha+1)}$.

We establish the lower bound (2.10) by using Assouad's Lemma and the lower bound (2.11) by using a version of Fano's Lemma given in Tsybakov (2009).

2.2.1. Proof of the lower bound (2.10)

The key technical tool to establish (2.10) 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 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 (2.10). Let $\mathbf{X}_1, \ldots, \mathbf{X}_n \overset{i.i.d.}{\sim} N(0, \Sigma(\theta))$ 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

$$\inf_{\hat{\Sigma}} \max_{\theta\in\{0,1\}^{k-1}} 2^2 E_\theta \left\| \hat{\Sigma} - \Sigma(\theta) \right\|_1 \geq \min_{H(\theta,\theta')\geq 1} \frac{\left\|\Sigma(\theta) - \Sigma(\theta')\right\|_1}{H(\theta,\theta')} \frac{k-1}{2} \min_{H(\theta,\theta')=1} \left\| P_\theta \wedge P_{\theta'} \right\|. \tag{2.14}$$

We state the bounds for the two factors on the right hand of (2.14) in two lemmas.

Lemma 4. *Let $\Sigma(\theta)$ be defined as in* (2.8)*. Then*

$$\min_{H(\theta,\theta')\geq 1} \frac{\left\|\Sigma(\theta) - \Sigma(\theta')\right\|_1}{H(\theta,\theta')} \geq cn^{-1/2} \tag{2.15}$$

for some $c > 0$.

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

Lemma 5. *Let $\mathbf{X}_1, \ldots, \mathbf{X}_n \overset{i.i.d.}{\sim} N(0, \Sigma(\theta))$ 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 $k = n^{1/(2\alpha+2)}$, that

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

2.2.2. Proof of the lower bound (2.11)

Consider the parameter space $\mathcal{F}_{12}$ defined in (2.9). 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 $\Theta = \{\theta_0, \ldots, \theta_{m_*}\}$.

Lemma 6. *Let $\Theta = \{\theta_m : m = 0, \ldots, m_*\}$ 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*

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

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, Sec. 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

$$d(\theta_i, \theta_j) = \|\Sigma_i - \Sigma_j\|_1 = b\sqrt{\tau_2 \log p}(k-1) \geq \sqrt{\frac{1}{2}\tau_2 \frac{k \log p}{n}} \quad \text{for all } 0 \leq i \neq j \leq m_*. \tag{2.16}$$

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*

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

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

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

for some positive constant $0 < c < 1/8$. Then the lower bound (2.11) follows immediately from Lemma 6 and (2.16),

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

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

$$\Sigma^* = (\sigma^*_{ij})_{1 \leq i,j \leq p} = \frac{1}{n} \sum_{l=1}^{n} \left(\mathbf{X}_l - \bar{\mathbf{X}}\right) \left(\mathbf{X}_l - \bar{\mathbf{X}}\right)^T \tag{3.1}$$

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 $c_{n,p}^2 \left(\log p/n\right)^{1-q}$. 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 (1.4), 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

$$\mathbb{P}\left(\left|\sigma_{ij}^* - \sigma_{ij}\right| > v\right) \le C_1 \exp\left(-\frac{8}{\gamma^2} n v^2\right) \tag{3.2}$$

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 (3.2) implies that σ_{ij}^* behaves like a subgaussian random variable. In particular for $v = \gamma\sqrt{\log p/n}$ we have

$$\mathbb{P}\left(\left|\sigma_{ij}^* - \sigma_{ij}\right| > v\right) \le C_1 p^{-8}. \tag{3.3}$$

We define a thresholding estimator as

$$\hat{\sigma}_{ij} = \sigma_{ij}^* \cdot I\left(|\sigma_{ij}^*| \ge \gamma\sqrt{\frac{\log p}{n}}\right) \tag{3.4}$$

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

The following theorem shows that the thresholding estimator at (3.4) 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*

$$\sup_{\mathcal{P}(\mathcal{G}_q(\rho, c_{n,p}))} \mathbb{E}\left\|\hat{\Sigma} - \Sigma\right\|_1^2 \le C c_{n,p}^2 \left(\frac{\log p}{n}\right)^{1-q}, \tag{3.5}$$

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

$$\inf_{\hat{\Sigma}} \sup_{\mathcal{P}(\mathcal{G}_q(\rho, c_{n,p}))} \mathbb{E}\left\|\hat{\Sigma} - \Sigma\right\|_1^2 \asymp c_{n,p}^2 \left(\frac{\log p}{n}\right)^{1-q}. \tag{3.6}$$

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* $A_{ij} = \{|\hat{\sigma}_{ij} - \sigma_{ij}| \leq 4\min\{|\sigma_{ij}|, \gamma\sqrt{\log p/n}\}\}$. *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 (3.4) 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

$$\begin{aligned}\mathbb{E}\left\|\hat{\Sigma} - \Sigma\right\|_1^2 &\leq 2\mathbb{E}\left\|\hat{\Sigma} - \Sigma - D\right\|_1^2 + 2\mathbb{E}\left\|D\right\|_1^2 \\ &\leq 2\mathbb{E}\left[\sup_j \sum_i |\hat{\sigma}_{ij} - \sigma_{ij}|\, I(A_{ij})\right]^2 + 2\mathbb{E}\left\|D\right\|_1^2 \\ &\leq 32\left[\sup_j \sum_i \min\left\{|\sigma_{ij}|, \gamma\sqrt{\frac{\log p}{n}}\right\}\right]^2 + 2\mathbb{E}\left\|D\right\|_1^2. \qquad (3.7)\end{aligned}$$

We will see that the first term in (3.7) is dominating and bounded by $Cc_{n,p}^2 \left(\log p/n\right)^{1-q}$, while the second term, $\mathbb{E}\left\|D\right\|_1^2$, is negligible.

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

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

which gives (3.5) if $\mathbb{E}\left\|D\right\|_1^2 = O\left(1/n\right)$; this can be shown as follows. Note that

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

where

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

since $\mathbb{P}\left(A_{ij}^c\right)\le 2C_1p^{-9/2}$ from Lemma 8, and

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

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

$$\hat\Sigma=\hat\Sigma_k=\left(w_{ij}\sigma_{ij}^*\right)_{p\times p}, \tag{3.8}$$

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

$$w_{ij}=k_h^{-1}\{(k-|i-j|)_+-(k_h-|i-j|)_+\} \tag{3.9}$$

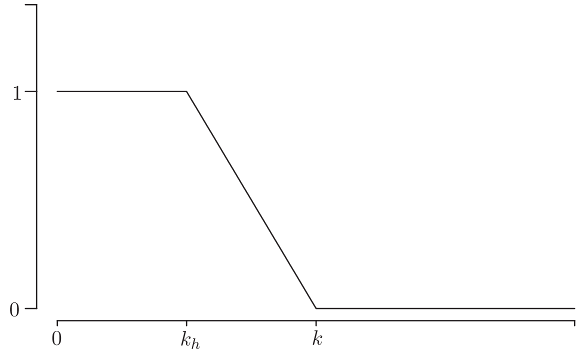

Figure 1. The weights as a function of $|i-j|$.

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

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

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

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 (3.8)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

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

and set

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

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

Lemma 9. *The tapering estimator* $\hat{\Sigma}_k$ *given in* (3.8) *can be written as*

$$\hat{\Sigma}_k = k_h^{-1}\left(S^{*(k)} - S^{*(k_h)}\right). \tag{3.10}$$

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

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

Theorem 5. *Suppose $p \geq n^{1/(2\alpha+2)}$. The tapering estimator $\hat{\Sigma}_k$ at* (3.10) *satisfies*

$$\sup_{\mathcal{A}} \mathbb{E} \left\| \hat{\Sigma}_k - \Sigma \right\|_1^2 \leq C \frac{k^2 + k \log p}{n} + C k^{-2\alpha} \tag{3.11}$$

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

$$k = \min \left\{ n^{1/(2\alpha+2)}, \left(\frac{n}{\log p} \right)^{1/(2\alpha+1)} \right\} \tag{3.12}$$

satisfies

$$\sup_{\mathcal{A}} \mathbb{E} \left\| \hat{\Sigma} - \Sigma \right\|_1^2 \leq C \left[n^{-\alpha/(\alpha+1)} + \left(\frac{\log p}{n} \right)^{2\alpha/(2\alpha+1)} \right], \tag{3.13}$$

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

Together with Theorem 3, Theorem 5 shows that the tapering estimator with the optimal choice of the tapering parameter k given in (3.12) attains the optimal rate of convergence over both $\mathcal{P}(\mathcal{F}_\alpha(\rho, M))$ and $\mathcal{P}(\mathcal{H}_\alpha(\rho, M))$.

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

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 $\mathbb{E} g_{ij} \leq C/n$, it is easy to see that

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

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

$$\breve{\Sigma}=(w_{ij}\tilde{\sigma}_{ij})_{1\le i,j\le p} \tag{3.14}$$

with w_{ij} given in (3.9). To prove Theorem 5, it suffices to show

$$\sup_{\mathcal{F}_\alpha(\rho,M)}\mathbb{E}\left\|\breve{\Sigma}-\Sigma\right\|_1^2\le Cn^{-\alpha/(\alpha+1)}+C\left(\frac{\log p}{n}\right)^{2\alpha/(2\alpha+1)}. \tag{3.15}$$

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

$$\mathbb{E}\tilde{\sigma}_{ij}=\sigma_{ij}, \tag{3.16}$$

$$\mathbb{V}ar(\tilde{\sigma}_{ij})\le\frac{C_1}{n}, \tag{3.17}$$

for some $C_1>0$.

It is easy to bound the bias part,

$$\left\|\mathbb{E}\breve{\Sigma}-\Sigma\right\|_1^2\le\left[\max_{i=1,\ldots,p}\sum_{j:|i-j|>k}|\sigma_{ij}|\right]^2\le M^2k^{-2\alpha}. \tag{3.18}$$

We show that the variance

$$\mathbb{E}\left\|\breve{\Sigma}-\mathbb{E}\breve{\Sigma}\right\|_1^2\le C_2\frac{k^2+k\log p}{n}. \tag{3.19}$$

It then follows immediately that

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

This proves (3.15) and (3.11) then follows. Since $p\ge n^{1/(2\alpha+2)}$, we can set

$$k=\begin{cases}\left\lfloor n^{1/(2\alpha+2)}\right\rfloor, & \text{for } p\le\exp\left(n^{1/(2\alpha+2)}\right)\\ \left\lfloor\left(\frac{n}{\log p}\right)^{1/(2\alpha+1)}\right\rfloor, & \text{otherwise}\end{cases} \tag{3.20}$$

and the estimator $\hat{\Sigma}$ with k given in (3.20) satisfies

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

Theorem 5 is then proved.

It remains to show (3.19). 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

$$U_l^{(m)} = \left(\tilde{\sigma}_{ij} I\left\{l \le i < l+m, l \le j < l+m\right\}\right)_{p\times p} \tag{3.21}$$

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

Lemma 10. *Let* $\breve{\Sigma}$ *be defined as in* (3.10)*. Then*

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

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

$$\mathbb{P}\left\{\left\|U_l^{(m)} - \mathbb{E}U_l^{(m)}\right\|_1^2 > c_0\left(\frac{m^2}{n} + x^2\frac{m}{n}\right)\right\} \le \exp\left(-2x^2\right) \tag{3.22}$$

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

It follows from Lemmas 10 and 11 that

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

by plugging $x^2 = C_4 \max\{m, \log p\}$ into (3.22), 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 $p \ge n^{1/(2\alpha+2)}$, satisfy

$$\begin{aligned}\inf_{\hat{\Sigma}} \sup_{\mathcal{P}(\mathcal{F}_\alpha(\rho,M))} \mathbb{E}\left\|\hat{\Sigma} - \Sigma\right\|_1^2 &\asymp \inf_{\hat{\Sigma}} \sup_{\mathcal{P}(\mathcal{H}_\alpha(\rho,M))} \mathbb{E}\left\|\hat{\Sigma} - \Sigma\right\|_1^2 \\ &\asymp n^{-\alpha/(\alpha+1)} + \left(\frac{\log p}{n}\right)^{2\alpha/(2\alpha+1)} .\end{aligned} \tag{3.23}$$

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

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 $p \ge n^{1/(2\alpha+2)}$. In this section we consider the case

$p < n^{1/(2\alpha+2)}$ 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 $p < n^{1/(2\alpha+2)}$.

This case is relatively easy. The upper bound can be attained by the sample covariance matrix $\hat{\Sigma}$. By (3.19) with $k = p$ we have,

$$\inf_{\hat{\Sigma}} \sup_{\mathcal{P}(\mathcal{F}_\alpha(\rho,M))} \mathbb{E} \left\| \hat{\Sigma} - \Sigma \right\|_1^2 \leq C \frac{p^2 + p \log p}{n} \leq 2C \frac{p^2}{n}. \tag{3.24}$$

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.,

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

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 (2.14) in Section 2.2.1,

$$\begin{aligned} \inf_{\hat{\Sigma}} \sup_{\mathcal{F}_{11}} \mathbb{E} \left\| \hat{\Sigma} - \Sigma \right\|_1^2 &\geq \min_{H(\theta,\theta') \geq 1} \frac{\|\Sigma(\theta) - \Sigma(\theta')\|_1}{H(\theta, \theta')} \frac{p-1}{2} \min_{H(\theta,\theta')=1} \|P_\theta \wedge P_{\theta'}\| \\ &\geq c \left(p n^{-1/2} \right)^2 \geq c_1 \frac{p^2}{n}. \end{aligned} \tag{3.25}$$

Inequalities (3.24) and (3.25) together yield the minimax rate of convergence for the case $p \leq n^{1/(2\alpha+2)}$,

$$\inf_{\hat{\Sigma}} \sup_{\mathcal{P}(\mathcal{F}_\alpha(\rho,M))} \mathbb{E} \left\| \hat{\Sigma} - \Sigma \right\|_1^2 \asymp \inf_{\hat{\Sigma}} \sup_{\mathcal{P}(\mathcal{H}_\alpha(\rho,M))} \mathbb{E} \left\| \hat{\Sigma} - \Sigma \right\|_1^2 \asymp \frac{p^2}{n}. \tag{3.26}$$

Combining (3.23) with (3.26), the optimal rate of convergence over two distribution classes $\mathcal{P}(\mathcal{F}_\alpha(\rho, M))$ and $\mathcal{P}(\mathcal{H}_\alpha(\rho, M))$ can be summarized as

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

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

$$\mathcal{G}_q(\rho, c_{n,p}, M_1) = \mathcal{G}_q(\rho, c_{n,p}) \cap \left\{\Sigma : \left\|\Sigma^{-1}\right\|_1 \leq M_1\right\}, \tag{4.1}$$

$$\mathcal{F}_\alpha(\rho, M, M_1) = \mathcal{F}_\alpha(\rho, M) \cap \left\{\Sigma : \left\|\Sigma^{-1}\right\|_1 \leq M_1\right\}, \tag{4.2}$$

$$\mathcal{H}_\alpha(\rho, M, M_1) = \mathcal{H}_\alpha(\rho, M) \cap \left\{\Sigma : \left\|\Sigma^{-1}\right\|_1 \leq M_1\right\}, \tag{4.3}$$

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 (1.4) and (4.1). 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

$$c_{n,p}^2 \left(\frac{\log p}{n}\right)^{1-q} \to 0, \tag{4.4}$$

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*

$$\inf_{\hat{\Omega}} \sup_{\mathcal{P}(\mathcal{G}_q(\rho, c_{n,p}, M_1))} \mathbb{E}\left\|\hat{\Omega} - \Sigma^{-1}\right\|_1^2 \asymp c_{n,p}^2 \left(\frac{\log p}{n}\right)^{1-q} \tag{4.5}$$

under assumptions (2.1) *and* (4.4), *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*

$$\inf_{\hat{\Omega}} \sup_{\mathcal{A}} \mathbb{E}\left\|\hat{\Omega} - \Sigma^{-1}\right\|_1^2 \asymp \min\left\{n^{-\alpha/(\alpha+1)} + \left(\frac{\log p}{n}\right)^{2\alpha/(2\alpha+1)}, \frac{p^2}{n}\right\}, \tag{4.6}$$

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

$$\max_{1 \leq i \leq p} \sum_{j, j \neq i} \frac{|\sigma_{ij}|}{\sqrt{\sigma_{ii}\sigma_{jj}}} \leq 1 - \varepsilon \tag{4.7}$$

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 (4.7) 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

$$\left\|\hat{\Sigma}^{-1}\right\|_1 \le \left\|\left(I + \left(\hat{\Sigma} - \Sigma\right)\Sigma^{-1}\right)^{-1}\right\|_1 \left\|\Sigma^{-1}\right\|_1 \le \left\|\Sigma^{-1}\right\|_1 [1 + \sum_{k=1}^{\infty} (\|H\|_1)^k], \quad (4.8)$$

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

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

for some $0 < \varepsilon < 1/(2M_1)$. It is easy to show that

$$\mathbb{P}\left(A_3^c\right) \le C_D n^{-D} \quad (4.9)$$

for every $D > 0$, using (3.2), (3.22), and (4.4). On A_3 we see that

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

Since $\left\|\Sigma^{-1}\right\|_1 \le M_1$, which implies $\left\|\hat{\Sigma}^{-1}\right\|_1 \le 2M_1$ on A_3 by (4.8), we have

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

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

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

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

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

where the last step follows from (4.9).

(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

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

which implies

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

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

$$\left\|\Sigma_1^{-1} - \Sigma_2^{-1}\right\|_1 \ge C^{-2} \|\Sigma_1 - \Sigma_2\|_1. \tag{4.10}$$

Equation (4.10) 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 $\mathcal{P}\left(\mathcal{H}_\alpha(\rho, M, M_1)\right)$. Set $\mathcal{F}_0 = \mathcal{F}_{11} \cup \mathcal{F}_{12}$, where $\mathcal{F}_{11}$ and $\mathcal{F}_{12}$ are defined in (2.8) and (2.9), respectively. Over the parameter space $\mathcal{F}_{11}$ the proof is almost identical to the proof of the lower bound (2.10) 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 (2.15), for some $c > 0$. Actually the inequality follows from (2.15) together with (4.10), 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 (2.11) 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 (2.16). The inequality follows from (2.16) and (4.10).

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

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

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 $\mathcal{P}(\mathcal{G}_q(\rho, c_{n,p}))$ under the matrix ℓ_1 norm. For minimax estimation over the other two parameter spaces $\mathcal{P}(\mathcal{F}_\alpha(\rho, M))$ and $\mathcal{P}(\mathcal{H}_\alpha(\rho, M))$, 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

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

In contrast, the best choice of k under the spectral norm is $k_2 = n^{1/(2\alpha+1)}$, 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 $k_F = n^{1/(2\alpha+2)}$. This coincides with k_1 when $\log p \leq n^{1/(2\alpha+2)}$, and $k_F > k_1$ when $\log p \gg n^{1/(2\alpha+2)}$.

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

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

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 $(\log p/n)^{\alpha/(\alpha+1)}$ under the matrix ℓ_1 norm, which is inferior to the optimal rate $\min\{n^{-\alpha/(\alpha+1)} + (\log p/n)^{2\alpha/(2\alpha+1)}, p^2/n\}$ given at (1.6). Take for example $\alpha = 1/2$ and $p = e^{\sqrt{n}}$. In this case $(\log p/n)^{\alpha/(\alpha+1)} = n^{-1/6}$, while the optimal rate is $n^{-1/4}$. On the other hand, it can be shown by using (3.22) that the banding estimator with the same optimal k for the tapering estimator described at (3.12) 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 Cai, Liu and Zhou (2011) 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 $H(\theta, \theta') = 1$, Pinsker's Inequality (see, e.g., Csiszár (1967)) implies

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

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

$$tr\left(\Sigma\left(\theta'\right)\Sigma\left(\theta\right)^{-1}\right) - \log\det\left(\Sigma\left(\theta'\right)\Sigma\left(\theta\right)^{-1}\right) - p \le \left\|\Sigma\left(\theta'\right) - \Sigma\left(\theta\right)\right\|_F^2 \quad (4.11)$$

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

$$\left\|\Sigma\left(\theta\right) - I\right\|_2 \le \left\|\Sigma\left(\theta\right) - I\right\|_1 \le \tau_2 k n^{-1/2} \le \tau_2 < \frac{1}{4} \quad (4.12)$$

for $\tau_2 < 1/16$. Inequalities (4.11) and (4.12) imply

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

and the lemma follows immediately.

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

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

Inequality (4.11) gives

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

Proof of Lemma 8. Let $A_1 = \left\{\left|\sigma_{ij}^*\right| \ge \gamma\sqrt{\frac{\log p}{n}}\right\}$. 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

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

by the triangle inequality. Note that (3.3) implies

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

Thus

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

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

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

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 (3.10)

$$\left\| S^{(m)} - \mathbb{E} S^{(m)} \right\|_1 \leq \sum_{l=1}^{m} \left\| \sum_{-1 \leq j \ < \ p/m} \delta_{jm+l}^{(m)} \right\|_1 . \tag{4.13}$$

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 (3.10) 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*

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

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

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

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

Acknowledgement

We thank the two referees for constructive comments that have helped to improve the presentation of the paper. We thank Sasha Tsybakov for pointing out some typos. We also thank Weidong Liu for helpful discussion which led to a concise proof of Lemma 12. The research of Tony Cai was supported in part by NSF Grant DMS-0604954 and 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.

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.* **106**, 672-684

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.* (Edited by 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. *J. Comput. Graph. Statist.* **15**, 265-286.

Department of Statistics, The Wharton School, University of Pennsylvania, Philadelphia, PA 19104, USA.

E-mail: tcai@wharton.upenn.edu

Department of Statistics, Yale University, New Haven, CT 06511, USA.

E-mail: huibin.zhou@yale.edu

(Received November 2010; accepted July 2011)

COMMENT

Lingzhou Xue and Hui Zou

University of Minnesota

We would like to first congratulate Professors Cai and Zhou for their path-breaking contributions to high-dimensional covariance matrix estimation. Their work greatly deepens our understandings about the nature of large covariance matrix estimation. The technical ideas developed in their work are very useful for studying many high-dimensional learning problems.

1. Geometric Decay Spaces

Throughout this discussion we assume $\log(p) \ll n < p$ and the loss function is the matrix ℓ_1 norm. In their paper, Cai and Zhou (2012) have shown that thresholding is minimax optimal for estimating Σ over the weak ℓ_q ball

$$\mathcal{G}_q(\rho, c_{n,p}) = \{\Sigma : \max_{1\le j\le p} |\sigma_{[k]j}|^q \le c_{n,p}k^{-1}, \ \forall\ k, \text{ and } \max_i \sigma_{ii} \le \rho, \ 0 \le q < 1\},$$

and that tapering/banding is minimax optimal for estimating Σ over

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

Beyond the polynomial decay space, it is natural to consider covariance matrices with a geometric decay rate. We introduce the parameter spaces

$$\mathcal{A}_\eta(\rho, M) = \{\Sigma : |\sigma_{ij}| \le M\eta^{|i-j|} \text{ for } i \ne j \text{ and } \max_i \sigma_{ii} \le \rho\},$$

$$\mathcal{B}_\eta(\rho, M) = \{\Sigma : \max_{1\le j\le p} |\sigma_{[k]j}| \le M\eta^k, \forall k \text{ and } \max_i \sigma_{ii} \le \rho\},$$

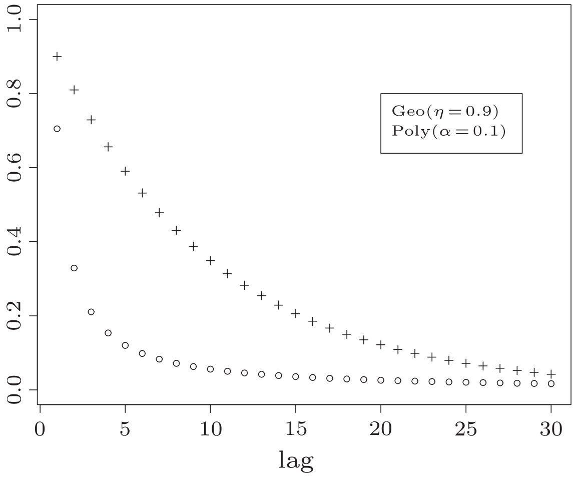

Figure A.1. A slow geometric decay curve versus a slow polynomial decay curve.

where $0 < \eta < 1$. The popular autoregressive matrices belong to geometric decay spaces such as $\mathcal{A}_\eta(\rho, M)$ and $\mathcal{B}_\eta(\rho, M)$. Figure A.1 compares a slow geometric decay curve (with $\eta = 0.9$) and a slow polynomial decay curve (with $\alpha = 0.1$). It is interesting to see that the geometric decay curve is well above the polynomial curve. The reason is that in the polynomial decay case the constant M should be less than 0.705 in order to keep the covariance matrix positive definite. This example suggests that the geometric decay space deserves some special consideration. An important technical contribution in Cai and Zhou (2012) is their carefully designed least favorable distributions for establishing the minimax lower bounds. We follow their idea and give minimax rates under the ℓ_1 norm for geometric decay spaces.

Theorem 1. *Thresholding attains the minimax risk of estimating Σ under the matrix ℓ_1-norm over $\mathcal{B}_\eta(\rho, M)$, with minimax rate*

$$\inf_{\hat{\Sigma}} \sup_{\mathcal{B}_\eta(\rho,M)} \mathrm{E}\left\|\hat{\Sigma} - \Sigma\right\|_1^2 \asymp \frac{\log p}{n} \cdot \log^2\left(\frac{n}{\log p}\right). \tag{A.1.1}$$

Tapering and banding both attain the minimax risk of estimating Σ under the ℓ_1-norm over $\mathcal{A}_\eta(\rho, M)$, with minimax rate

$$\inf_{\hat{\Sigma}} \sup_{\mathcal{A}_\eta(\rho,M)} \mathrm{E}\left\|\hat{\Sigma} - \Sigma\right\|_1^2 \asymp \frac{\log p}{n} \cdot \log\left(\frac{n}{\log p}\right). \tag{A.1.2}$$

Theorem 1 also indicates that thresholding is nearly minimax optimal for estimating Σ over $\mathcal{A}_\eta(\rho, M)$. To see that, for $\Sigma \in \mathcal{A}_\eta(\rho, M)$ we have

$$|\sigma_{[k]j}| = \min_{1 \le i \le k} |\sigma_{[i]j}| \le \min_{1 \le i \le k} M\eta^{|[i]-j|} \le M\eta^{k/2} = M\sqrt{\eta}^k,$$

which immediately implies that $\mathcal{A}_\eta(\rho, M) \subset \mathcal{B}_{\sqrt{\eta}}(\rho, M)$. By Theorem 1 we know that the thresholding estimator achieves a rate of convergence of $(\log p/n) \cdot \log^2(n/\log p)$ over $\mathcal{A}_\eta(\rho, M)$. This rate of convergence differs from the exact minimax lower bound by the factor $\log(n/\log p)$.

The above nearly minimax result is not true in the polynomial decay spaces. Note that $\mathcal{H}_\alpha(\rho, M) \subset \mathcal{G}_{1/(\alpha+1)}(\rho, 2M^{1/(\alpha+1)})$. If we apply the thresholding estimator to estimate Σ over $\mathcal{H}_\alpha(\rho, M)$, the rate of convergence is $(\log p/n)^{\alpha/(1+\alpha)}$. Comparing it to the minimax rate over $\mathcal{H}_\alpha(\rho, M)$ given in Theorem 1 of Cai and Zhou (2012), we see that tapering/banding is fundamentally better than thresholding for estimating bandable matrices over a polynomial decay space.

2. Double Thresholding?

Thresholding estimator is permutation-invariant, whereas banding/tapering estimator requires a natural ordering among variables. It is of interest to combine the strengths of banding and thresholding. This motivates us to consider the double thresholding estimator $\widehat{\Sigma}_{double} = (\hat{\sigma}_{ij}^{double})_{p\times p}$ by performing the entry-wise double thresholding rule

$$\hat{\sigma}_{ij}^{double} = \tilde{\sigma}_{ij}^{block} \cdot I(|\tilde{\sigma}_{ij}^{block}| \geq \lambda_2), \tag{A.2.1}$$

where

$$\tilde{\sigma}_{ij}^{block} = \hat{\sigma}_{ij} \cdot I\big(\max_{(s,t):s-t=i-j} |\hat{\sigma}_{st}| \geq \lambda_1\big).$$

We conducted a small simulation study to compare five regularized covariance matrix estimators (banding, tapering, simple thresholding, block thresholding, and double thresholding). In the simulation study, we considered four covariance models.

- Model 1: $\sigma_{ij} = (1 - |i-j|/\gamma)_+$ for $\gamma = 0.05p$.
- Model 2: $\sigma_{ij} = s_{ij} \cdot (s_{ii}s_{jj})^{-1/2}$, where $S = (I_{p\times p} + U)^T(I_{p\times p} + U) = (s_{ij})_{p\times p}$ with U being a sparse matrix with exactly κ nonzero entries equal to $+1$ or -1 with equal probability for $\kappa = p$.
- Model 3: $\sigma_{ij} = I_{\{i=j\}} + a_{ij}(1+\epsilon)^{-1/2} \cdot I_{\{i\neq j\}}$, where a_{ij} is equal to 0 or $0.6 \cdot |i-j|^{-1.3}$ with equal probability, and ϵ is chosen to be the absolute value of the minimal eigenvalue of $(I_{\{i=j\}} + a_{ij} \cdot I_{\{i\neq j\}})_{p\times p}$ plus 0.01.
- Model 4: $\sigma_{ij} = I_{\{i=j\}} + (1+\epsilon)^{-1/2} \cdot (b_{ij} \cdot I_{\{0<|i-j|\geq 0.5p\}} + c_{ij} \cdot I_{\{|i-j|>0.5p\}})$, where b_{ij} or c_{ij} equals 0 with probability 0.7 and equals $0.7^{|i-j|}$ or $0.7^{|i-j-0.5p|}$ with probability 0.3, and ϵ is chosen to be the absolute value of the minimal eigenvalue of $(I_{\{i=j\}} + b_{ij} \cdot I_{\{0<|i-j|\geq 0.5p\}} + c_{ij} \cdot I_{\{|i-j|>0.5p\}})_{p\times p}$ plus 0.01.

Table A.1. Comparison of banding, tapering, simple thresholding, block thresholding and double thresholding estimators. The standard errors are also shown in the bracket.

	Model 1	Model 2	Model 3	Model 4
Banding	5.55	4.60	1.98	2.15
	(0.08)	(0.00)	(0.01)	(0.01)
Tapering	5.66	4.60	1.98	2.19
	(0.10)	(0.00)	(0.01)	(0.01)
Simple Thresholding	10.61	3.38	2.19	2.40
	(0.19)	(0.03)	(0.02)	(0.02)
Block Thresholding	5.66	4.60	1.91	1.97
	(0.08)	(0.00)	(0.01)	(0.01)
Double Thresholding	5.68	3.38	1.87	1.82
	(0.08)	(0.03)	(0.01)	(0.02)

For each model we generated a training data set with $n = 100$ and $p = 500$ to construct the five estimators, and we also generated an independent validations set of size 100 to tune each estimator. The procedure was repeated 100 times. The estimation accuracy was measured by the matrix ℓ_1 norm averaged over 100 replications. The simulation results are summarized in Table A.1. Model 1 is designed for banding/tapering, and simple thresholding fails miserably there, while blockwise thresholding works as well as tapering. Model 2 is designed for thresholding, and it has a total of 1546 nonzero off-diagonal entries. Banding and tapering fail, as does blockwise thresholding. Model 3 and Model 4 are more interesting examples, because neither banding/tapering nor simple thresholding can give the best estimation. Blockwise thresholding does better than banding/tapering and simple thresholding. However, the best results are given by double thresholding: it significantly outperforms the other four estimators. This simulation study suggests that the double thresholding estimator deserves a more thorough theoretical investigation.

Appendix

For the sake of completeness we give the proof of Theorem 1.

Proof of Theorem 1. We first establish the upper bounds. Recall the thresholding estimator as defined in Cai and Zhou (2012), $\hat{\sigma}_{ij} = \sigma^*_{ij} \cdot I_{\{|\sigma^*_{ij}| \geq \gamma\sqrt{\log p/n}\}}$, where γ is chosen such that $\Pr(|\sigma^*_{ij} - \sigma_{ij}| > \gamma\sqrt{\log p/n}) \leq C_1 p^{-8}$. There exists some integer k^* such that $M\eta^{k^*} > \gamma\sqrt{\log p/n} \geq M\eta^{k^*+1}$. Then we have

$$\sum_i \min\left\{|\sigma_{ij}|, \gamma\sqrt{\frac{\log p}{n}}\right\} \leq k^* \cdot \gamma\sqrt{\frac{\log p}{n}} + M\sum_{i>k^*} \eta^i \leq C\sqrt{\frac{\log p}{n}} \log(\frac{n}{\log p}).$$

Applying (3.6) and $E\|D\|_1^2 = O(1/n)$ as in Cai and Zhou (2012) yields

$$\sup_{\mathcal{B}_\eta(\rho,M)} E\|\hat{\Sigma}-\Sigma\|_1^2 \le C\left[\frac{\log p}{n}\log^2(\frac{n}{\log p})+\frac{1}{n}\right] \le C\frac{\log p}{n}\cdot\log^2(\frac{n}{\log p}).$$

For the tapering estimator, we can use the steps for proving Theorem 5 in Cai and Zhou (2012) to obtain

$$\sup_{\mathcal{A}_\eta(\rho,M)} \mathrm{E}\left\|\hat{\Sigma}_k-\Sigma\right\|_1^2 \le C\frac{k^2+k\log p}{n}+C\frac{\eta^k}{(1-\eta)^2}.$$

Therefore the tapering estimator with $k=\log(n/\log p)/\log(1/\eta)$ can have the rate of convergence

$$\sup_{\mathcal{A}_\eta(\rho,M)} \mathrm{E}\left\|\hat{\Sigma}_k-\Sigma\right\|_1^2 \le C\frac{\log p}{n}\cdot\log(\frac{n}{\log p}).$$

We now prove the lower bounds. Let $\mathcal{H}=\{H_{1,k},H_{2,k},\cdots,H_{m_*,k}\}$ be the collection of symmetric matrices with exactly k elements equal to 1 in the first row/column and the rest zeros. To show (A.1.1) we consider

$$\mathcal{B}_0=\{\Sigma_0=\rho\cdot I_p \text{ and } \Sigma_m=\rho\cdot I_p+a\cdot H_{m,k}: 1\le m\le m_*\},$$

where $k=\lfloor\log(n/\log p)/2\log(1/\eta)\rfloor$ and $a=\sqrt{\tau\log p/n}$ for some small constant τ. Since $a\le\eta^k$ still holds, $\mathcal{B}_0$ is a subclass of $\mathcal{B}_\eta(\rho,M)$. Note that $\mathcal{B}_0$ is similar to the space defined in (2.2) in Cai and Zhou (2012) but with a differen k value. Then by Le Cam's lemma and arguments in Section 2.1 of Cai and Zhou (2012), we can show that

$$\sup_{0\le m\le m_*} E\|\hat{\Sigma}-\Sigma_m\|_1^2 \ge \frac{1}{2}\|\mathcal{P}_0\wedge\bar{\mathcal{P}}\|\cdot\inf_{1\le m\le m_*}\|\Sigma_m-\Sigma_0\|_1^2 \ge c\frac{\log p}{n}\cdot\log^2(\frac{n}{\log p}).$$

Thus the lower bound in (A.1.1) is proved.

To show (A.1.2) we consider

$$\mathcal{A}_0=\left\{\Sigma_m=\rho\cdot I_p+a\cdot B_{m,k}: 0\le m\le m_*=\left\lfloor\frac{p}{k}\right\rfloor-1\right\},$$

where $k=\log(n/\log p)/2\log(1/\eta)$, $a=\sqrt{\log p/16nk}$, and $B_{m,k}=(b_{ij})_{1\le i,j\le p}$ with

$$b_{ij}=I_{\{i=m \text{ and } k+1\le j\le m+k-1\}}+I_{\{j=m \text{ and } k+1\le i\le m+k-1\}}.$$

Since $a^2\le\log p/n=\eta^{2k}$, $a\le\eta^k$ obviously holds. Then it is easy to show that $\mathcal{A}_0$ is a subclass of $\mathcal{A}_\eta(\rho,M)$. Note that $\mathcal{A}_0$ is similar to the space defined in (2.8) in Cai and Zhou (2012), but with a differen k value. Then by Fano's lemma and the arguments in Section 2.2, we can have

$$\inf_{\hat{\Sigma}}\sup_{\mathcal{A}_0(k,a)} \mathrm{E}\left\|\hat{\Sigma}-\Sigma\right\|_1^2 \ge c\cdot\frac{\log p}{n}\cdot\log(\frac{n}{\log p}).$$

Thus the lower bound in (A.1.2) is proved.

References

Cai, T. and Zhou, H. (2012). Minimax estimation of large covariance matrices under ℓ_1-norm. *Statist. Sinica* **22**, 1319-1378.

School of Statistics, University of Minnesota, Minneapolis, MN 55455, USA.
E-mail: lzxue@stat.umn.edu

School of Statistics, University of Minnesota, Minneapolis, MN 55455, USA.
E-mail: zouxx019@umn.edu

(Received March 2012; accepted March 2012)

COMMENT

Tingni Sun and Cun-Hui Zhang

Rutgers University

The estimation of covariance matrices and their inverses is a problem of great practical value and theoretical interest. We congratulate the authors for making an important contribution to it by finding the rate of minimax risk with the ℓ_1 operator norm as the loss function.

A natural question arising from this interesting paper is the minimax rate when the loss is the ℓ_w operator norm $\|M\|_w = \max_{\|u\|_w=1} \|Mu\|_w$. For $\mathcal{G}_q(\rho, c_{n,p})$, the minimax rate has already been established in Cai and Zhou (2012). For $\mathcal{A} = \mathcal{F}_\alpha(\rho, M)$ or $\mathcal{A} = \mathcal{H}_\alpha(\rho, M)$, the upper bound for $w \in [1, 2]$,

$$\inf_{\widehat{\Sigma}} \sup_{\mathcal{A}} \mathrm{E}\big\|\widehat{\Sigma} - \Sigma\big\|_w^2 \lesssim \min\Big\{ n^{-2\alpha/(2\alpha+2/w)} + \Big(\frac{\log p}{n}\Big)^{2\alpha/(2\alpha+2/w-1)}, \frac{p^{2/w}}{n} \Big\}, \quad \text{(B.1)}$$

follows from the $\|\cdot\|_2$ bound on the variance term of the tapering estimator and the $\|\cdot\|_1$ bound on the bias term, since $\|M\|_w \le \min\big(\|M\|_1, k^{1/w-1/2}\|M\|_2\big)$ for symmetric $M \in \mathbb{R}^{k\times k}$. Since Σ is symmetric, (B.1) is also valid with w replaced by $w/(w-1) \in [2, \infty]$. For $w = 1$, this gives the minimax rate of the authors. However, it is unclear if the lower bound argument works for $w \in (1, 2)$.

Recent advances in high-dimensional data have been focused on the estimation of high-dimensional objects. However, the estimation of low-dimensional functionals of high-dimensional objects is also of interest. A rate minimax estimator of a high-dimensional parameter does not automatically yield rate minimax estimates of its low-dimensional functionals. For example, instead of the entire covariance matrix or its inverse, one might be more interested in the relationship between individual pairs of variables. In what follows, we consider efficient estimation of partial correlation with high-dimensional Gaussian data.

The partial correlation is of primary interest in Gaussian-Markov graphical models. Let $X = (X_1, \dots, X_p)^\top$ be a $N(0, \Sigma)$ random vector. The partial correlation between X_j and X_k, say r_{jk}, is their conditional correlation given all other variables. It can be also written as the error correlation in the linear regression of $X_{\{j,k\}}$ against $X_{\{j,k\}^c}$. In general, for any proper subset $A \subset \{1, \dots, p\}$, a multivariate linear regression model can be written as

$$X_A = X_{A^c}\beta_{A^c,A} + \varepsilon_A. \tag{B.2}$$

For $A = \{j, k\}$, the partial correlation r_{jk} is the correlation between the two entries of ε_A. Throughout the sequel, we consider sets A of bounded size.

It is well known that the conditional distribution of X_A given X_{A^c} is

$$X_A | X_{A^c} \sim N(X_{A^c}\Sigma^{-1}_{A^c,A^c}\Sigma_{A^c,A}, \Sigma_A - \Sigma_{A,A^c}\Sigma^{-1}_{A^c,A^c}\Sigma_{A^c,A}).$$

Thus, the coefficient matrix in (B.2) is $\beta_{A^c,A} = \Sigma^{-1}_{A^c,A^c}\Sigma_{A^c,A}$ and the residual ε_A follows the multivariate normal distribution $N(0, \Sigma_A - \Sigma_{A,A^c}\Sigma^{-1}_{A^c,A^c}\Sigma_{A^c,A})$. Let $\Theta = \Sigma^{-1}$ be the precision matrix. It follows easily from the block inversion formula that the covariance matrix for the residual ε_A is $\Theta_A^{-1} = \Sigma_A - \Sigma_{A,A^c}\Sigma^{-1}_{A^c,A^c}\Sigma_{A^c,A}$. Thus,

$$r_{jk} = -\frac{\Theta_{jk}}{(\Theta_{jj}\Theta_{kk})^{1/2}}. \tag{B.3}$$

We consider the slightly more general problem of estimating a smooth function of Θ_A^{-1}, say $\tau = \tau(\Theta_A^{-1})$.

Suppose we have a data matrix $\boldsymbol{X} \in \mathrm{R}^{n\times p}$ with iid rows from $N(0, \Sigma)$. An oracle expert observing both $\boldsymbol{X}$ and $\boldsymbol{\varepsilon}_A = \boldsymbol{X}_A - \boldsymbol{X}_{A^c}\beta_{A^c,A}$ can estimate τ by the oracle MLE

$$\tau^* = \tau\Big(\frac{\boldsymbol{\varepsilon}_A^\top \boldsymbol{\varepsilon}_A}{n}\Big) \tag{B.4}$$

due to the sufficiency of $\boldsymbol{\varepsilon}_A$ for Θ_A. Our idea is to find an estimator close to the oracle τ^*. For $|A| = 1$, this was done in Sun and Zhang (2011), where the scaled Lasso is used to jointly estimate the coefficient vector and the noise level in univariate linear regression. This noise level estimator was proven to be within $o(n^{-1/2})$ of the oracle τ^* under certain “large-p-smaller-n” settings. We extend their results to $|A| > 1$ as follows.

Let $\boldsymbol{X}_j \in \mathbb{R}^n$ be the j-th column of $\boldsymbol{X}$. For each $j \in A$, we apply the scaled Lasso to the univariate linear regression of $\boldsymbol{X}_j$ against $\boldsymbol{X}_{A^c}$ as follows:

$$\{\widehat{\beta}_{A^c,j},\ \widehat{\sigma}_j\} = \underset{b_{A^c},\sigma}{\arg\min}\Big\{\frac{\|\boldsymbol{X}_j - \boldsymbol{X}_{A^c}b_{A^c}\|^2}{2n\sigma} + \frac{\sigma}{2} + \lambda\sum_{k\in A^c}\|\boldsymbol{X}_k\|_2\frac{|b_k|}{\sqrt{n}}\Big\}. \tag{B.5}$$

Let $\widehat{\beta}_{A^c,j}, j \in A$, be the columns of $\widehat{\beta}_{A^c,A}$ and $\boldsymbol{z}_A = \boldsymbol{X}_A - \boldsymbol{X}_{A^c}\widehat{\beta}_{A^c,A}$. Define

$$\widehat{\tau} = \tau\Big(\frac{\boldsymbol{z}_A^\top \boldsymbol{z}_A}{n}\Big) \tag{B.6}$$

as the scaled Lasso estimator of $\tau(\Theta_A^{-1})$. The following theorem gives an error bound for the estimator $\widehat{\tau}$ in (B.6) by comparing it with the oracle MLE.

Theorem 1. *Suppose $\tau : \mathbb{R}^{A\times A} \to \mathbb{R}$ is a unit Lipschitz function in a neighborhood $\{M : \|M - \Theta_A^{-1}\|_2 \le \eta_0\}$. Let $\widehat{\tau}$ be given by* (B.6) *with $\lambda = \{3(\log p)/n\}^{1/2}$ in* (B.5). *Let $s_A = \max_{j\in A}\sum_{k\in A^c}\min(1, |\Theta_{jk}|/\lambda)$. Suppose that for a fixed M_0, $\|\Theta\|_2 + \|\Sigma\|_2 \le M_0$. Then there exist constants $a_0 > 0$ and $C_0 < \infty$, both depending on $\{\eta_0, M_0\}$ only, such that for $s_A \le a_0 n/\log p$,*

$$\mathbb{P}\Big\{\big|\widehat{\tau} - \tau^*\big| > \frac{C_0 s_A(\log p)}{n}\Big\} \le p^{-1/3},$$

where τ^* is the oracle MLE (B.4). *In particular, if $s_A(\log p)/\sqrt{n} = o(1)$, then*

$$\sqrt{nF_\tau}(\widehat{\tau} - \tau) \xrightarrow{D} N(0,1),$$

where F_τ is the minimum Fisher information for the estimation of τ.

The proof of Theorem 1 and additional related results will be presented in a forthcoming paper. Since the oracle MLE τ^* in (B.4) is based on an $|A|$-dimensional regular multivariate normal model, $\varepsilon_A \sim N(0, \Theta_A^{-1})$, and $|A|$ is bounded, τ^* is efficient. This gives the efficiency of $\widehat{\tau}$.

Now consider the estimation of the partial correlation (B.3). With $A = \{j,k\}$ in (B.5), the oracle and scaled Lasso estimators are

$$r_{jk}^* = \frac{\boldsymbol{\varepsilon}_j^\top \boldsymbol{\varepsilon}_k}{(\|\boldsymbol{\varepsilon}_j\|_2\|\boldsymbol{\varepsilon}_k\|_2)}, \quad \widehat{r}_{jk} = \frac{\boldsymbol{z}_j^\top \boldsymbol{z}_k}{(\|\boldsymbol{z}_j\|_2\|\boldsymbol{z}_k\|_2)}. \tag{B.7}$$

Corollary 1. *Let r_{jk}^* and $\widehat{r}_{jk}$ be given by* (B.7). *Suppose the conditions of Theorem 1 hold with $A = \{j,k\}$ and $s = s_A$. Then, $\widehat{r}_{jk} - r_{jk}^* = O_P(s(\log p)/n)$. Consequently, if $s(\log p)/\sqrt{n} \to 0$, then*

$$\frac{\sqrt{n}(\widehat{r}_{jk} - r_{jk})}{(1 - \widehat{r}_{jk}^2)} \xrightarrow{D} N(0,1).$$

Table B.1. Mean and standard error of the scaled Lasso estimator for the partial correlation and the ratio of the simulated and theoretical MSEs, κ =MSE$/\{(1-r_{jk}^2)^2/n\}$.

Example 1: five-diagonal precision matrix						
	$r_{12}=-0.6$		$r_{13}=-0.1$		$r_{14}=0$	
p	Mean±SE($\widehat{r}$)	κ	Mean±SE($\widehat{r}$)	κ	Mean±SE($\widehat{r}$)	κ
200	-0.626 ± 0.055	0.894	-0.042 ± 0.083	1.037	-0.010 ± 0.097	0.937
1000	-0.643 ± 0.056	1.214	-0.043 ± 0.088	1.104	-0.007 ± 0.089	0.797

Example 2: exponential decay precision matrix						
	$r_{12}=-0.6$		$r_{13}=-0.36$		$r_{14}=-0.216$	
p	Mean±SE($\widehat{r}$)	κ	Mean±SE($\widehat{r}$)	κ	Mean±SE($\widehat{r}$)	κ
200	-0.551 ± 0.064	1.602	-0.236 ± 0.079	2.846	-0.042 ± 0.100	4.412
1000	-0.539 ± 0.079	2.425	-0.224 ± 0.089	3.475	-0.029 ± 0.101	4.962

Since r_{jk}^* is the (oracle) MLE of the correlation based on iid bivariate normal observations, $\sqrt{n}(r_{jk}^*-r_{jk})$ converges to $N(0,(1-r_{jk}^2)^2)$ in distribution. Thus, Corollary 1 directly follows from Theorem 1.

A major difference between our theory and existing work based on variable selection is that Θ is allowed to have many elements of small and moderate magnitude in Theorem 1 and Corollary 1. This is similar to Zhang and Zhang (2011) where statistical inference of regression coefficients is considered.

We present some simulation results to demonstrate the performance of the scaled Lasso for partial correlation. Two examples are considered. The first example is a five-diagonal precision matrix with $\Theta_{jj}=1$, $\Theta_{j-1,j}=\Theta_{j,j-1}=0.6$, and $\Theta_{j-2,j}=\Theta_{j,j-2}=0.1$. In the second example, we set $\Theta_{jk}=0.6^{|j-k|}$ (no entry of the precision matrix is exactly zero). The partial correlations are computed by $r_{jk}=-\Theta_{jk}/(\Theta_{jj}\Theta_{kk})^{1/2}$. We generated a random sample of size $n=100$ from $N(0,\Sigma)$ with $\Sigma=\Theta^{-1}$. The scaled Lasso estimator was computed with $\lambda=\{(\log p)/n\}^{1/2}$. For each example, we took $p=200$ and $p=1,000$.

Table B.1 shows the scaled Lasso estimates for r_{12}, r_{13}, and r_{14} based on 100 replications. In Example 1, $\widehat{r}_{jk}$ is quite accurate, as the condition of small $s_A(\log p)/\sqrt{n}$ holds well with values 0.8, 1.3, and 1.5 for the estimation of r_{12}, r_{13}, and r_{14} when $p=200$, and with values 1.0, 1.6, and 1.9 when $p=1,000$. In Example 2, the scaled Lasso deteriorates as the condition $s_A(\log p)/\sqrt{n}$ starts to fail, with values 2.3, 2.8, and 3.4 for the estimation of r_{12}, r_{13}, and r_{14} when $p=200$, and with values 2.8, 3.5, and 4.2 when $p=1,000$.

Acknowledgement

This research is partially supported by the NSF Grants DMS 0906420, DMS-11-06753, DMS 1209014 and NSA Grant H98230-11-1-0205.

References

Cai, T. and Zhou, H. (2012). Minimax estimation of large covariance matrices under ℓ_1-norm. *Statist. Sinica* **22**, 1319-1378.

Sun, T. and Zhang, C.-H. (2011). Scaled sparse linear regression. arXiv:1104.4595v1.

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

Department of Statistics and Biostatistics, Hill Center, Busch Campus, Rutgers University, Piscataway, New Jersey 08854, U.S.A.

E-mail: tingni@stat.rutgers.edu

Department of Statistics and Biostatistics, Hill Center, Busch Campus, Rutgers University, Piscataway, New Jersey 08854, U.S.A.

E-mail: czhang@stat.rutgers.edu

(Received April 2012; accepted April 2012)

COMMENT

Philippe Rigollet and Alexandre B. Tsybakov

Princeton University and CREST-ENSAE

1. Introduction

Estimation of covariance matrices in various norms is an issue that finds applications in a wide range of statistical problems and especially in principal component analysis. It is well known that, without further assumptions, the empirical covariance matrix Σ^* is the best possible estimator in many ways, and in particular in a minimax sense. However, it is also well known that Σ^* is not an accurate estimator when the dimension p of the observations is high. The minimax analysis carried out by Tony Cai and Harry Zhou (Cai and Zhou (2012) in what follows) guarantees that for several classes of matrices with reasonable structure (sparse or banded matrices), the fully data-driven thresholding estimator achieves the best possible rates when p is much larger than the sample size n. This is done, in particular, by proving minimax lower bounds that ensure that no estimator can perform better than the hard thresholding estimator, uniformly over the sparsity classes $\mathcal{G}_q$ for each $0 \leq q < 1$. This result has a flavor of universality in the sense that one and the same estimator is minimax optimal for several classes of matrices.

Our comments focus on the sparsity classes of matrices.

(a) *Optimal rates.* Optimal rates are obtained in Cai and Zhou (2012) under the assumption that the dimension is very high: $p \geq n^{\nu}$, $\nu > 1$. Thus, the case of dimensions smaller than n, or even $p \approx n$, is excluded. This seems to be due to the technique employed to prove the lower bound (Theorem 2 in Cai and Zhou (2012)). Indeed, by a different technique, we show that the lower bound holds without this assumption, cf. Theorem 1 below. Furthermore, in general, our lower rate $\psi^{(1)}$ is different from that obtained in Cai and Zhou (2012) and has ingredients similar to the optimal rate for the Gaussian sequence model. We conjecture that it is optimal for all admissible configurations of n, p, and sparsity parameters.

(b) *Frobenius norm and global sparsity.* We argue that the Frobenius norm is naturally adapted to the structure of the problem, at least for Gaussian observations, and we derive optimal rates under the Frobenius risk and global sparsity assumption.

(c) *Approximate sparsity.* Again under the Frobenius risk, one can obtain not only the minimax results but also oracle inequalities. We demonstrate it for the soft-thresholding estimator. This allows us to deal with a more general setup where the covariance matrix is not necessarily sparse but can be well approximated by a sparse matrix.

Below we denote by $\|A\|$ the Frobenius norm of a matrix A:

$$\|A\|^2 = \operatorname{tr}(AA^{\top}) = \sum_{i,j} a_{ij}^2 \,,$$

where $\operatorname{tr}(B)$ stands for the trace of square matrix B. Moreover, for $q > 0$, we denote by $|v|_q$ the ℓ_q-norm of a vector v and by $|A|_q$ the ℓ_q norm of the off-diagonal entries of A. We set $|A|_0 = \sum_{i \neq j} I(a_{ij} \neq 0)$ (the number of non-zero off-diagonal entries of A). The operator $\ell_q \to \ell_q$ norm of A is denoted by $\|A\|_q$.

2. Frobenius Norm and Sparsity

The cone of positive semi-definite (PSD) matrices can be equipped with a variety of norms, even more so than a vector space. Cai and Zhou (2012) choose the $\|\cdot\|_1$ norm and consider classes of matrices that are essentially adapted to this metric. For example, the class $\mathcal{G}_q$ defined in (1) controls the largest ℓ_q norm of the columns of the covariance matrix Σ with $0 \leq q < 1$ while the $\|\cdot\|_1$ norm measures the largest ℓ_1 norm of the columns of $\hat{\Sigma} - \Sigma$. Theorem 1 below indicates that for $q = 1$ consistent estimators do not exist.

One may wonder whether faster rates can be obtained if, for example, Σ has one row/column with large ℓ_q norm and all other rows/columns have small ℓ_q norm. It is quite clear that the $\|\cdot\|_1$ norm fails to capture such a behavior and we need to resort to other norms. As we see below, this is achievable when the Frobenius norm is used.

The Frobenius norm is a rather weak norm on the PSD cone. Indeed, it is very much a vector norm unlike the $\|\cdot\|_1$ norm used by Cai and Zhou (2012) or the spectral norm, that are both operator norms. However, the choice of a norm is rather subjective but some general guidelines exist in a given statistical setup. It can be motivated by the idea of minimizing the Kullback-Leibler divergence between the true distribution and its estimator (see, e.g., Rigollet (2012)). This principle naturally gives rise to the use of the Frobenius norm in Gaussian covariance matrix estimation, as indicated by the following lemma.

Lemma 1. *Let I_p be the $p \times p$ identity matrix and Δ be a symmetric $p \times p$ matrix such that $I_p + \Delta$ is PSD. Denote by P_Σ the distribution of $\mathcal{N}_p(0, \Sigma)$ (a zero-mean normal random variable in $\mathbb{R}^p$ with covariance matrix $\Sigma > 0$). Then, for any $0 < \varepsilon < 1$, the Kullback-Leibler divergence between $P_{I_p+\varepsilon\Delta}$ and P_{I_p} satisfies*

$$\mathsf{KL}(P_{I_p+\varepsilon\Delta}, P_{I_p}) \leq \frac{g(-\varepsilon)}{2}\|\Delta\|^2 ,$$

where

$$g(\varepsilon) = \frac{\varepsilon - \log(1+\varepsilon)}{\varepsilon^2} .$$

Moreover if $\|\Delta\|_2 \leq 1$, we have

$$\mathsf{KL}(P_{I_p+\varepsilon\Delta}, P_{I_p}) \geq \frac{(1-\log 2)\varepsilon^2}{2}\|\Delta\|^2 . \tag{C.2.1}$$

Proof. Take $\Sigma = I_p + \varepsilon\Delta$ and observe that

$$\mathsf{KL}(P_\Sigma, P_{I_p}) = \mathrm{E}\log\left(\frac{\mathrm{d}P_\Sigma}{\mathrm{d}P_{I_p}}(X)\right) = \frac{1}{2}\mathrm{E}\log\left(\frac{1}{\det(\Sigma)}\right) + \frac{1}{2}\mathrm{E}[X^\top X - X^\top\Sigma^{-1}X] ,$$

where $X \sim \mathcal{N}_p(0, \Sigma)$. Let $\lambda_1, \ldots, \lambda_p$ denote the eigenvalues of Δ and recall that $\det(\Sigma) = \prod_j(1+\varepsilon\lambda_j)$. Moreover,

$$\mathrm{E}[X^\top X - X^\top\Sigma^{-1}X] = \mathrm{tr}(\mathrm{E}[XX^\top] - \Sigma^{-1}\mathrm{E}[XX^\top]) = \mathrm{tr}(\Sigma - I_p) = \sum_j \varepsilon\lambda_j .$$

Therefore,

$$\mathsf{KL}(P_\Sigma, P_{I_p}) = \frac{1}{2}\sum_{j=1}^p[\varepsilon\lambda_j - \log(1+\varepsilon\lambda_j)] \leq \frac{1}{2}\sum_{j=1}^p g(\varepsilon\lambda_j)\lambda_j^2 .$$

Note now that since $I_p + \Delta$ is PSD, then $\lambda_j \geq -1$ for all $j = 1, \ldots, p$. Therefore, since g is monotone decreasing on $(-1, \infty)$, it yields $g(\varepsilon\lambda_j) \leq g(-\varepsilon)$. The second statement of the lemma follows by observing that if $\|\Delta\|_2 \leq 1$, then $\varepsilon\lambda_j \leq \varepsilon \leq 1$ for all $j = 1, \ldots, p$.

3. Minimax Lower Bounds over Classes of Sparse Matrices

We denote by σ_{ij} the elements of Σ and by $\sigma_{(j)}$ the jth column of Σ with its jth component replaced by 0. For any $q > 0, R > 0$, we define the following classes of matrices:

$$\mathcal{G}_q^{(0)}(R) = \left\{ \Sigma \in \mathcal{C}_{>0} \,:\, |\Sigma|_q^q \le R\,,\ \sigma_{ii} = 1, \forall i \right\},$$
$$\mathcal{G}_q^{(1)}(R) = \left\{ \Sigma \in \mathcal{C}_{>0} \,:\, \max_{1\le j\le p} |\sigma_{(j)}|_q^q \le R\,,\ \sigma_{ii} = 1, \forall i \right\},$$

where $\mathcal{C}_{>0}$ is the set of all positive definite symmetric $p \times p$ matrices. For $q = 0$, we define the classes $\mathcal{G}_0^{(0)}(R)$ and $\mathcal{G}_0^{(1)}(R)$ analogously, with the respective constraints $|\Sigma|_0 \le R$ and $\max_{1\le j\le p} |\sigma_{(j)}|_0 \le R$. Here R is an integer for the class $\mathcal{G}_0^{(1)}(R)$, and an even integer for $\mathcal{G}_0^{(0)}(R)$ in view of the symmetry. We assume that $R = 2k \le p(p-1)$ for $\mathcal{G}_0^{(0)}(R)$ and $R = k \le p-1$ for $\mathcal{G}_0^{(1)}(R)$, where k is an integer. Set

$$\psi^{(0)} = R^{1/2}\left(\frac{1}{n}\log\left(1+\frac{c_0 p^2}{R n^{q/2}}\right)\right)^{1/2-q/4}, \psi^{(1)} = R\left(\frac{1}{n}\log\left(1+\frac{c_0 p}{R n^{q/2}}\right)\right)^{(1-q)/2},$$

for some positive constant c_0 that does not depend on the parameters p, n, R. The following minimax lower bounds hold.

Theorem 1. *Fix $R > 0$, $0 \le q \le 2$, $C_0 > 0$, and integers $n \ge 1$, $p \ge 2$. Consider the conditions*

$$R\left(\frac{\log p}{n}\right)^{1-q/2} \le C_0, \quad R\left(\frac{\log p}{n}\right)^{(1-q)/2} \le C_0, \quad R^{-1}\left(\frac{\log p}{n}\right)^{q/2} \le C_0. \tag{C.3.1}$$

Let $X_1, \dots, X_n$ be i.i.d. $\mathcal{N}_p(0,\Sigma)$ random vectors, and let $w : [0,\infty) \to [0,\infty)$ be a monotone non-decreasing function such that $w(0) = 0$ and $w \not\equiv 0$. Then there exist constants $c_0 > 0, c_1 > 0, c > 0$, depending only on C_0 such that, under the first and third conditions in (C.3.1),

$$\inf_{\hat\Sigma} \sup_{\Sigma \in \mathcal{G}_q^{(0)}(R)} E_\Sigma w\Big(\frac{\|\hat\Sigma - \Sigma\|}{c_1 \psi^{(0)}}\Big) \ge c, \tag{C.3.2}$$

and under the second and third conditions in (C.3.1),

$$\inf_{\hat\Sigma} \sup_{\Sigma \in \mathcal{G}_q^{(1)}(R)} E_\Sigma w\Big(\frac{\|\hat\Sigma - \Sigma\|_1}{c_1 \psi^{(1)}}\Big) \ge c, \quad \forall\ 0 \le q \le 1, \tag{C.3.3}$$

where E_Σ denotes the expectation with respect to the joint distribution of $X_1, \dots,$ X_n and the infimum is taken over all estimators based on $X_1, \dots, X_n$.

Proof. We first prove (C.3.2) with $q = 0$ and $R = 2k$. Assume first that $k \leq p^2/16$. We use Theorem 2.7 in Tsybakov (2009). It is enough to check that there exists a finite subset $\mathcal{N}$ of $\mathcal{G}_0^{(0)}(2k)$ such that, for some constant $C > 0$ and some $\psi \geq C\psi^{(0)}$, we have

(i) $\|\Sigma - \Sigma'\| \geq \psi\,,\ \forall\ \Sigma \neq \Sigma' \in \mathcal{N} \cup \{I_p\}$,
(ii) $n\,\mathsf{KL}(P_\Sigma, P_{I_p}) \leq 2^{-4}\log(\operatorname{card}\mathcal{N})\,,\ \forall\ \Sigma \in \mathcal{N}\,.$

We show that these conditions hold for

$$\psi = \left(\frac{k}{n}\log\left(1 + \frac{ep(p-1)}{2k}\right)\right)^{1/2}\ .$$

Let $\mathcal{B}$ be the family of all $p \times p$ symmetric binary matrices, banded such that for all $B \in \mathcal{B}$, $b_{ij} = 0$ if $|i - j| > \sqrt{k}$, with 0 on the diagonal and exactly k nonzero over-diagonal entries equal to 1. Let M be the number of elements in the over-diagonal band where the entry 1 can only appear. For $k \leq p^2/4$ we have $M \geq p\sqrt{k} - k \geq p\sqrt{k}/2$. Therefore for, $k \leq p\sqrt{k}/4$, Lemma A.3 in Rigollet and Tsybakov (2011) implies that there exists a subset $\mathcal{B}_0$ of $\mathcal{B}$ such that for any $B, B' \in \mathcal{B}_0, B \neq B'$, we have $\|B - B'\|^2 \geq (k+1)/4$, and

$$\log(\operatorname{card}\mathcal{B}_0) \geq C_1 k \log\left(1 + \frac{ep}{4\sqrt{k}}\right) \tag{C.3.4}$$

for some absolute constant $C_1 > 0$. Consider the family of matrices $\mathcal{N} = \{\Sigma = I_p + \frac{a}{2}B :\ B \in \mathcal{B}_0\}$ where

$$a = a_0\left(\frac{1}{n}\log\left(1 + \frac{ep}{4\sqrt{k}}\right)\right)^{1/2}$$

for some $a_0 > 0$. All matrices in $\mathcal{N}$ have at most $2\sqrt{k}$ nonzero elements equal to a in each row. Therefore, the first inequality in (C.3.1) guarantees that for a_0 small enough, matrices $I_p + aB$ with $B \in \mathcal{B}_0$ and, a fortiori, $\Sigma \in \mathcal{N}$ are diagonally dominant and hence PSD. Thus, $\mathcal{N} \subset \mathcal{G}_0^{(0)}(2k)$ for sufficiently small $a_0 > 0$. Also, for any $\Sigma, \Sigma' \in \mathcal{N}$, $\Sigma \neq \Sigma'$, we have

$$\|\Sigma - \Sigma'\|^2 \geq C_2 a^2 k$$

for some absolute constant $C_2 > 0$. It is easy to see that this inequality also holds with a different C_2 if Σ or Σ' is equal to I_p. The above display implies (i). To check (ii), observe first that since $I_p + aB$ is PSD, we can apply Lemma 1 with $\Delta = aB$, $\varepsilon = 1/2$, to get

$$n\mathsf{KL}(P_\Sigma, P_{I_p}) \leq \frac{na^2 g(-1/2)}{2}\|B\|^2 \leq a^2 kn\,, \quad \forall\ \Sigma \in \mathcal{N}\,.$$

To prove (ii), it suffices to take $a_0^2 < 2^{-4}C_1$, and to use (C.3.4). This proves (C.3.2) with $q = 0$ under the assumption $k \leq p^2/16$. The case $q = 0$, $k > p^2/16$ corresponds to a rate $\psi^{(0)}$ of order $\sqrt{p/n}$ and is easily treated via the Varshamov-Gilbert argument (we omit the details).

Next, observe that (C.3.2), for $0 < q \leq 2$, follows from the case $q = 0$. Indeed, let k be the maximal integer such that $2ka^q \leq R$ (we assume a_0 small enough to have $k \geq 1$, cf. the third inequality in (C.3.1)). Hence, $|\Sigma|_q^q = 2ka^q \leq R$ for any $\Sigma \in \mathcal{N}$. Also, $a\sqrt{k} \leq R^{1/2}a^{1-q/2}/\sqrt{2}$ and thus the first inequality in (C.3.1) ensures the positive definiteness of all $\Sigma \in \mathcal{N}$ for small a_0. For this choice of k, we have $k + 1 > Ra^{-q}/2$ and $k \leq C_3Rn^{q/2}$ with some constant $C_3 > 0$. It can be easily shown that (i) holds with

$$\psi^2 \geq CR\left(\frac{1}{n}\log\left(1 + \frac{ep^2}{Ra^{-q}}\right)\right)^{1-q/2} \geq CR\left(\frac{1}{n}\log\left(1 + \frac{c_0p^2}{Rn^{q/2}}\right)\right)^{1-q/2}.$$

The proof of (C.3.3) is quite analogous, with the only difference that $\mathcal{B}$ is now defined as the set of all symmetric binary matrices with exactly k off-diagonal entries equal to 1 in the first row and in the first column, and all other entries 0. Then, for $k \leq (p-1)/2$, Lemma A.3 in Rigollet and Tsybakov (2011) implies that there exists a subset $\mathcal{B}_1$ of $\mathcal{B}$ such that for any two distinct $B, B' \in \mathcal{B}_1$, we have $|b_{(1)} - b'_{(1)}|_1 \geq (k+1)/4$ (consequently, $\|B - B'\|_1 \geq (k+1)/4$) and

$$\log(\operatorname{card}\mathcal{B}_1) \geq C_1k\log\left(1 + \frac{e(p-1)}{k}\right). \tag{C.3.5}$$

Here, $b_{(1)}, b'_{(1)}$ are the first columns of B, B' with their first components replaced by 0. Thus, for any two distinct matrices Σ and Σ' belonging to the family $\mathcal{N}' = \{\Sigma = I_p + \frac{a}{2}B : B \in \mathcal{B}_1\}$, we have $\|\Sigma - \Sigma'\|_1^2 \geq C_4a^2k^2$ for some constant $C_4 > 0$. Here, $\mathcal{N}' \subset \mathcal{G}_0^{(1)}(k)$ thanks to the second inequality in (C.3.1). Also, by Lemma 1, $\mathsf{KL}(P_\Sigma, P_{I_p}) \leq a^2k$ for all $\Sigma \in \mathcal{N}'$. These remarks and (C.3.5) imply the suitably modified (i) and (ii) for the choice

$$a = a_0\left(\frac{1}{n}\log\left(1 + \frac{e(p-1)}{k}\right)\right)^{1/2}$$

with a_0 small enough. The rest of the proof follows the same lines as the proof of (C.3.2).

The lower bound (C.3.3) and Theorem 4 in Cai and Zhou (2012) imply that the rate $R((\log p)/n)^{(1-q)/2}$ is optimal on the class $\mathcal{G}_q^{(1)}(R)$ under the $\|\cdot\|_1$ norm if $Rn^{q/2} \leq p^\alpha$ with some $\alpha < 1$. In particular, for $q = 0$ this optimality holds under the quite natural condition $k = O(p^\alpha)$, and no lower bound on p

in terms of n is required. Clearly, this is also true when we drop the condition $\Sigma > 0$ in the definition of $\mathcal{G}_q^{(1)}(R)$ and consider a weak ℓ_q constraint as in Cai and Zhou (2012).

Note that the rate $\psi^{(0)}$ is very similar to the optimal rate in the Gaussian sequence model, cf. Section 11.5 in Johnstone (2011). This is due to the similarity between the vector ℓ_2 norm and the Frobenius norm. The rate $\psi^{(1)}$ is different but nevertheless has analogous ingredients. Observe also that, in contrast to the remark after Theorem 1 in Cai and Zhou (2012), we prove the Frobenius and the $\|\cdot\|_1$-norm lower bounds (C.3.2) and (C.3.3) by exactly the same technique. The key point is the use of the "k-selection lemma" (Lemma A.3 in Rigollet and Tsybakov (2011)). The lower bound (C.3.3) improves upon Theorem 2 in Cai and Zhou (2012) in two aspects. First, it does not need the assumption $p \geq n^{\nu}$, $\nu > 1$, and provides insight on the presumed optimal rate for any configuration of n, p, R. Second, it is established for general loss functions w, in particular for the "in probability" loss that we consider below. The technique used in Theorem 2 of Cai and Zhou (2012) is not adapted for this purpose as it applies to special losses derived from $w(t) = t$.

4. Approximate Sparsity and Optimal Rates

Along with the hard thresholding estimator considered by Cai and Zhou (2012), one can use the soft thresholding estimator $\tilde{\Sigma}$ defined as the matrix with off-diagonal elements

$$\tilde{\sigma}_{ij} = \operatorname{sign}(\sigma^*_{ij})(|\sigma^*_{ij}| - \tau)_+ \,,$$

where σ^*_{ij} are the elements of the sample covariance matrix Σ^*, $\tau > 0$ is a threshold, and $(\cdot)_+$ denotes the positive part. The diagonal elements of $\tilde{\Sigma}$ are all set to 1 since we consider the classes $\mathcal{G}_q^{(j)}(R)$, $j = 0, 1$. Then $\tilde{\Sigma} = I_p + \tilde{\Sigma}_{\text{off}}$ where $\tilde{\Sigma}_{\text{off}}$ admits the representation (the minimum is taken over all $p \times p$ matrices S with zero diagonal):

$$\tilde{\Sigma}_{\text{off}} = \operatorname*{arg\,min}_{S:\, \operatorname{diag}(S)=0} \left\{ |S - \Sigma^*|_2^2 + 2\tau |S|_1 \right\} .$$

Take the threshold

$$\tau = A\gamma \sqrt{\frac{\log p}{n}} \,, \tag{C.4.1}$$

where $A > 1$ and γ is the constant in the inequality (3.2) in Cai and Zhou (2012).

Theorem 2. *Let $X_1, \ldots, X_n$ be i.i.d. random vectors in $\mathbb{R}^p$ with covariance matrix Σ such that* (3.2) *in* Cai and Zhou (2012) *holds. Assume that p, n, and A are such that $\tau \leq \delta$, where δ is the constant introduced after* (3.2) *in* Cai and Zhou

(2012). *Then there exists* $C_* > 0$ *such that, with probability at least* $1-C_*p^{2-2A^2}$,

$$\left\|\tilde{\Sigma} - \Sigma\right\|^2 \leq \min_S \left\{ \|S - \Sigma\|^2 + \left(\frac{1+\sqrt{2}}{2}\right)^2 A^2\gamma^2 \frac{|S|_0 \log p}{n} \right\}, \tag{C.4.2}$$

where $\min_S$ *denotes the minimum over all* $p \times p$ *matrices.*

Proof. Write $\sigma^*_{ij} = \sigma_{ij} + \xi_{ij}$ where the $\xi_{ij} = \sigma^*_{ij} - \sigma_{ij}$ are zero-mean random variables, $i \neq j$. Thus, considering σ^*_{ij} as observations, we have a sequence model in dimension $p(p-1)$. It is easy to see that it is a special case of the trace regression model studied in Koltchinskii, Lounici, and Tsybakov (2011) where A_0 is a diagonal matrix with the $p(p-1)$ off-diagonal entries of Σ on the diagonal. In the notation of Koltchinskii, Lounici, and Tsybakov (2011), the corresponding matrices X_i are diagonalizations of canonical basis vectors, the norm $\|\cdot\|_{L_2(\Pi)}$ coincides with the norm $|\cdot|_2$, and rank(B) is equal to the number of non-zero entries of diagonal matrix B. Thus, Assumption 1 in Koltchinskii, Lounici, and Tsybakov (2011) is satisfied with $\mu = 1$, and we can apply Theorem 1 in Koltchinskii, Lounici, and Tsybakov (2011). It yields a deterministic statement:

$$\left|\tilde{\Sigma} - \Sigma\right|_2^2 \leq \min_S \left\{ |S - \Sigma|_2^2 + \left(\frac{1+\sqrt{2}}{2}\right)^2 \tau^2 |S|_0 \right\}$$

provided $\tau > 2\max_{i\neq j}|\sigma^*_{ij} - \sigma_{ij}|$. From (3.2) in Cai and Zhou (2012) and a union bound, we obtain that, for τ defined in (C.4.1), this inequality holds with probability greater than $1 - C_*p^{2-2A^2}$.

Corollary 2. *Under the assumptions of Theorem* 2, *for any* $0 < q < 2$, *there exist constants* $C', C_* > 0$ *such that with probability at least* $1 - C_*p^{2-2A^2}$,

$$\left\|\tilde{\Sigma} - \Sigma\right\|^2 \leq \min_S \left\{ 2\|S - \Sigma\|^2 + C'|S|_q^q \left(\frac{\log p}{n}\right)^{1-q/2} \right\}. \tag{C.4.3}$$

Proof. Let $|s_{[l]}|$, $l = 1, \ldots, p(p-1)$, denote the absolute values of the off-diagonal elements of S ordered in a decreasing order. Note that for any $p \times p$ matrix S and any $0 < q < 2$ we have $|s_{[l]}|^q \leq |S|_q^q/l$. Fix an integer $k \leq p(p-1)$. Taking $s'_{ij} = s_{ij}$ if $|s_{ij}| \geq |s_{[k]}|$ and $s'_{ij} = 0$ otherwise, we get that for any S there exists a $p \times p$ matrix S' with $|S'|_0 = k$ such that

$$|S - S'|_2^2 = \sum_{l>k} s_{[l]}^2 \leq |S|_q^2 \sum_{l>k} l^{-2/q} \leq \frac{|S|_q^2 k^{1-2/q}}{2/q - 1}.$$

Together with Theorem 2, this implies that for any integer $k \le p(p-1)$ we have

$$\big|\tilde{\Sigma} - \Sigma\big|_2^2 \le \min_S \left\{ 2|S - \Sigma|_2^2 + \frac{|S|_q^2 k^{1-2/q}}{2/q - 1} + \left(\frac{1+\sqrt{2}}{2}\right)^2 A^2 \gamma^2 \frac{k \log p}{n} \right\} .$$

Optimizing the right hand side over k completes the proof.

Note that the oracle inequalities (C.4.2) and (C.4.3) are satisfied for any covariance matrix Σ, not necessarily for sparse Σ. They quantify a trade-off between the approximation and sparisty terms. Their right-hand sides are small if Σ is well approximated by a matrix S with a small number of entries or with small ℓ_q norm of the off-diagonal elements. If the matrix Σ is sparse, $\Sigma \in \mathcal{G}_q^{(0)}(R)$, the oracle inequalities (C.4.2) and (C.4.3) imply that

$$\sup_{\Sigma \in \mathcal{G}_q^{(0)}(R)} P_\Sigma \left(\|\tilde{\Sigma} - \Sigma\| > C'' R^{1/2} \left(\frac{\log p}{n} \right)^{1/2 - q/4} \right) \le C_* p^{2-2A^2}$$

for some constant $C'' > 0$. This also holds when we drop the condition $\Sigma > 0$ in the definition of $\mathcal{G}_q^{(0)}(R)$. Combining this with Theorem 1, we find that the rate $R^{1/2}\left((\log p)/n\right)^{1/2-q/4}$ is optimal on the class $\mathcal{G}_q^{(0)}(R)$ under the Frobenius norm if $Rn^{q/2} \le p^{2\alpha}$ with some $\alpha < 1$. In particular, for $q = 0$ this optimality holds under the condition $k \le p^{2\alpha}$ with some $\alpha < 1$.

Acknowledgement

Philippe Rigollet is supported in part by NSF grants DMS-0906424 and CAREER-DMS-1053987. Alexandre B. Tsybakov is supported in part by ANR grant "Parcimonie".

References

Cai, T. and Zhou, H. (2012). Minimax estimation of large covariance matrices under ℓ_1-norm. *Statist. Sinica* **22**, 1319-1378.

Johnstone, I. M. (2011). *Gaussian Estimation: Sequence and Wavelet Models.* Unpublished Manuscript. `http://www-stat.stanford.edu/~imj/`.

Koltchinskii, V., Lounici, K. and Tsybakov, A. B. (2011). Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion. *Ann. Statist.* **39**, 2302–2329.

Rigollet, P. (2012). Kullback-Leibler aggregation and misspecified generalized linear models. *Ann. Statist.* **40**, 639-665.

Rigollet, P. and Tsybakov, A. B. (2011). Exponential screening and optimal rates of sparse estimation. *Ann. Statist.* **39**, 731-771.

Tsybakov, A. B. (2009). *Introduction to Nonparametric Estimation.* Springer, New York.

Department of Operations Research and Financial Engineering, Princeton University, Princeton, NJ 08544, USA.

E-mail: rigollet@princeton.edu

Laboratoire de Statistique, CREST-ENSAE, 3, av. Pierre Larousse, F-92240 Malakoff Cedex, France.

E-mail: Alexandre.Tsybakov@ensae.fr

(Received April 2012; accepted April 2012)

COMMENT

Peter J. Bickel[1], Elizaveta Levina[2], Adam J. Rothman[3] and Ji Zhu[2]

[1] *University of California, Berkeley,* [2] *University of Michigan*
and [3] *University of Minnesota*

The authors offer insightful results on minimax rates for large covariance matrix estimation under the matrix ℓ_1-norm that add to the previously known results on the matrix ℓ_2-norm. Incidentally, we expect that some version of the results on the ℓ_1 and ℓ_2 norms in this context can also be developed for the Wiener norm (see Bickel and Lindner (2011) for more details), defined by $\|\Sigma\|_W = \max_k \sum\{|\sigma_{ij}| : |i-j| = k\}$, particularly in the time series domain for which it was introduced by Wiener.

Minimax risk is often used as a benchmark for the evaluation of an estimation method, and having optimal tuning parameter rates is helpful for understanding the behavior of various methods. However, there is also the issue of selecting the tuning parameter in practice, mentioned in the paper as well, which cannot be done using the theoretical bounds of this kind and requires cross-validation. Since this paper studies the convergence in the matrix ℓ_1-norm, and most of the previous literature focuses on convergence in the matrix ℓ_2-norm, we decided to investigate the effect of using various norms for tuning parameter selection via cross-validation, focusing on the thresholding estimator and the parameter space $\mathcal{P}(\mathcal{G}_q(\rho, c_{n,p}))$. Our expectation was that the empirical risk calculated via a particular norm would be minimized by the tuning parameter selected by cross-validation using the same norm, but this turned out not to be the case.

Specifically, we evaluated the performance of the random splitting method for tuning parameter selection described in Bickel and Levina (2008a,b). The n observations are randomly partitioned M times into a validation set of size $n_{\text{va}} = n/\log n$ and a training set of size $n_{\text{tr}} = n - n_{\text{va}}$. Define the ℓ_1-norm empirical risk $\hat{R}_1$, ℓ_2-norm empirical risk $\hat{R}_2$, and Frobenius norm empirical risk $\hat{R}_F$ as follows:

$$\hat{R}_1(\lambda) = \frac{1}{M}\sum_{m=1}^{M} \|\hat{\Sigma}_\lambda^{(\mathrm{tr,m})} - \hat{\Sigma}^{(\mathrm{va,m})}\|_1,$$

$$\hat{R}_2(\lambda) = \frac{1}{M}\sum_{m=1}^{M} \|\hat{\Sigma}_\lambda^{(\mathrm{tr,m})} - \hat{\Sigma}^{(\mathrm{va,m})}\|_2,$$

$$\hat{R}_F(\lambda) = \frac{1}{M}\sum_{m=1}^{M} \|\hat{\Sigma}_\lambda^{(\mathrm{tr,m})} - \hat{\Sigma}^{(\mathrm{va,m})}\|_F^2,$$

where $\hat{\Sigma}_\lambda^{(\mathrm{tr,m})}$ is the sample covariance computed from the training set of the m-th split and thresholded at λ, and $\hat{\Sigma}^{(\mathrm{va,m})}$ is the sample covariance computed from the validation set of the m-th split.

We generated an i.i.d. sample of size n from $N_p(0, \Sigma)$, where Σ has entries $\sigma_{ij} = 0.4 \cdot I(|i-j| = 1) + I(i = j)$. Then we selected the tuning parameters $\hat{\lambda}_1$, $\hat{\lambda}_2$, and $\hat{\lambda}_F$ by minimizing the empirical risks $\hat{R}_1(\lambda)$, $\hat{R}_2(\lambda)$, $\hat{R}_F(\lambda)$, respectively. For each norm, we also computed the "oracle" tuning parameter $\hat{\lambda}_0 = \arg\min_\lambda \|\hat{\Sigma}_\lambda - \Sigma\|$. The performance of each of the tuning parameters was evaluated using the squared L_1 risk, the squared L_2 risk and the squared Frobenius risk, defined respectively as

$$\widehat{\mathrm{E}}\|\hat{\Sigma}_{\hat{\lambda}} - \Sigma\|_1^2, \quad \widehat{\mathrm{E}}\|\hat{\Sigma}_{\hat{\lambda}} - \Sigma\|_2^2, \quad \text{and} \quad \widehat{\mathrm{E}}\|\hat{\Sigma}_{\hat{\lambda}} - \Sigma\|_F^2\, p^{-1},$$

where $\widehat{\mathrm{E}}$ is the average over simulation replications.

We considered two scenarios, $n < p$ and $n > p$. In the $n < p$ scenario, we set $n = p/2$, where p=30, 50, 100, 200 and 500. We used M=10 random splits to estimate the empirical risk and a 200 point resolution for λ. We performed 500 independent replications for $p \leq 50$ and 100 independent replications for $p \geq 100$. In the $n > p$ scenario, everything was the same, except for n=60, 100, 200, 500, 1,000 and $p = n/4$.

In Figures D.1 (for $n < p$) and D.2 (for $n > p$) we plot the estimated empirical risks. Each plot corresponds to one evaluation criterion, and the curves on each plot correspond to different methods of selecting the tuning parameter. Surprisingly, the Frobenius norm tuning is always the closest to the oracle, regardless of the evaluation criterion. This is quite counter-intuitive as one would expect, and as was also argued in the paper, that for different evaluation criteria the optimal threshold should be different. Interestingly, however, the Frobenius norm cross-validation tuning is the only one that was analyzed theoretically, in Bickel and Levina (2008b). We may be observing a finite sample phenomenon, but it would be interesting to connect this practical observation to the authors' results on optimal thresholds.

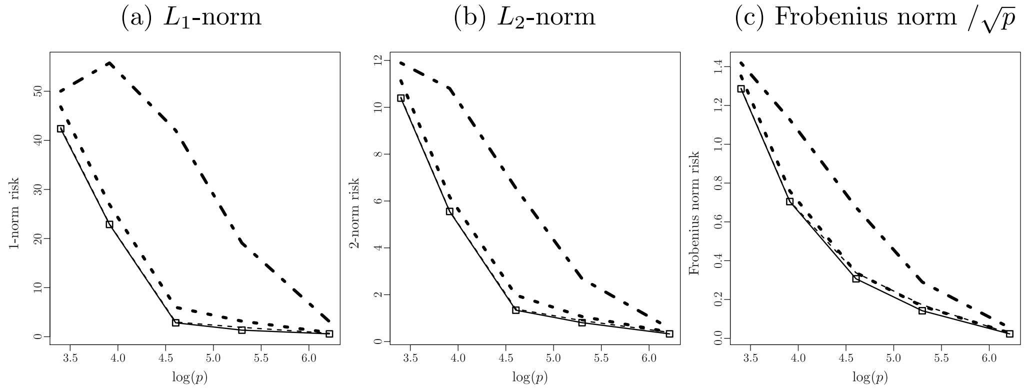
(a) L_1-norm (b) L_2-norm (c) Frobenius norm $/\sqrt{p}$

Figure D.1. The $n < p$ scenario. Simulated risk for hard thresholding of the sample covariance matrix with the threshold parameter $\hat{\lambda}_0$ (solid), $\hat{\lambda}_1$ (dots), $\hat{\lambda}_2$ (dash-dot), and $\hat{\lambda}_F$ (dashes).

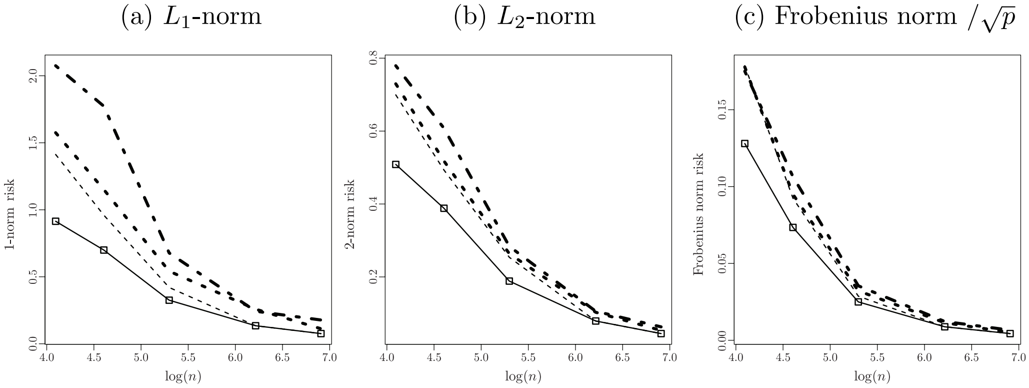
(a) L_1-norm (b) L_2-norm (c) Frobenius norm $/\sqrt{p}$

Figure D.2. The $n > p$ scenario. Simulated risk for hard thresholding of the sample covariance matrix with the threshold parameter $\hat{\lambda}_0$ (solid), $\hat{\lambda}_1$ (dots), $\hat{\lambda}_2$ (dash-dot), and $\hat{\lambda}_F$ (dashes).

References

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.

Bickel, P. J. and Lindner, M. (2011). Approximating the inverse of banded matrices by banded matrices with applications to probability and statistics. *Theory Probab. Appl.* **56**, 1-20.

Department of Statistics, University of California, Berkeley, CA 94710-3860, U.S.A.
E-mail: bickel@stat.berkeley.edu

Department of Statistics, University of Michigan, 459 West Hall, Ann Arbor, MI 48109-1107, U.S.A.
E-mail: elevina@umich.edu

School of Statistics, University of Minnesota, 313 Ford Hall, 224 Church Street, SE Minneapolis, MN 55455, U.S.A.
E-mail: arothman@umn.edu

Department of Statistics, University of Michigan, 459 West Hall, Ann Arbor, MI 48109-1107, U.S.A.
E-mail: jizhu@umich.edu

(Received April 2012; accepted April 2012)

COMMENT

Wei Biao Wu

University of Chicago

I congratulate Professor Cai and Professor Zhou for their timely and important contribution of sharp minimax convergence rates for estimating large covariance matrices. The argument for proving the lower bound is quite sophisticated and is of independent interest. As a useful property, for a class of sparse covariance matrices (cf $\mathcal{G}_q(\rho, c)$ in their (1.1)), the well-known thresholded covariance matrix estimate of Bickel and Levina (2008b) can achieve the minimax rate, while for a class of covariance matrices with weakly correlations (cf $\mathcal{F}_\alpha(\rho, M)$ in (1.2) and $\mathcal{H}_\alpha(\rho, M)$ in (1.3)), a tapered estimate can also have the minimax rate. The paper provides, in the minimax sense, a rigorous justification of the use of the thresholded and the tapered covariance matrix estimates.

My primary concern is the time series application of the large-p-small-n results from the multivariate setting of independent and identically distributed p-variate random vectors. In many time series applications, one has only one realization, $n = 1$. This covariance matrix estimation problem has been discussed by Wu and Pourahmadi (2009), McMurry and Politis (2010), Bickel and Gel (2011), and Xiao and Wu (2012). With $n = 1$, structural assumptions such as stationarity are needed so that the covariance matrix is estimable. Here we propose a possible link between these two settings via block sampling (Politis, Romano, and Wolf (1999)). With observations $X_1, \ldots, X_p$ from a stationary process $(X_i)_{i\in\mathbb{Z}}$, we can consider the $l = \lfloor p/b \rfloor$ blocks $\mathbf{X}_1 = (X_1, \ldots, X_b)'$, $\mathbf{X}_2 = (X_{b+1}, \ldots, X_{2b})'$, …, $\mathbf{X}_l = (X_{(l-1)b+1}, \ldots, X_{lb})'$, with b the block size. Consider the estimation of Σ_b, the $b \times b$ covariance matrix of $\mathbf{X}_1$. Assuming weak dependence, one would expect that results similar to (1.5) in their paper can hold.

As an alternative way to deal with stationary processes, one can use the sample auto-covariance function $\hat{\gamma}_k = p^{-1}\sum_{i=1+k}^{p}(X_i - \bar{X})(X_{i-k} - \bar{X})$, $0 \le k \le p-1$, where $\bar{X} = p^{-1}\sum_{i=1}^{p} X_i$. Using the operator or spectral norm, Xiao and Wu (2012) obtained a sharp convergence rate for banded and tapered sample covariance matrix estimators. That result parallels the optimal minimax rate derived in Cai, Zhang, and Zhou (2010), though the settings are different.

If one indeed has n i.i.d. realizations of $(X_i)_{i=1}^{p}$, and $\hat{\gamma}_m^{(l)}$ is the lag-m sample auto-covariance from the lth realization $(X_{li})_{i=1}^{p}$, $1 \le l \le n$, then it is expected that the rates (1.5) and (1.6) can be substantially improved if one uses the averaged sample auto-covariances

$$\bar{\gamma}_m = n^{-1}\sum_{l=1}^{n} \hat{\gamma}_m^{(l)}.$$

Specifically, assume that $E(X_i) = 0$ and that $(X_i)_{i\in\mathbb{Z}}$ is short-memory in the sense that its 4th order functional dependence measures $\delta_4(i)$ (Wu (2005)) are summable. For $\hat{\gamma}_m^{(l)} = p^{-1}\sum_{i=1+m}^{p} X_{li}X_{l,i-m}$, following Lemma 1 in Wu and Pourahmadi (2009), we have

$$E[(\hat{\gamma}_m^{(l)} - E\hat{\gamma}_m^{(l)})^2] \le 4p^{-1}\kappa^2, \text{ where } \kappa = [E(X_1^4)]^{1/4}\sum_{i=0}^{\infty}\delta_4(i). \tag{E.1}$$

Since $\gamma_m^{(l)}$, $l = 1, \ldots, n$, are i.i.d., (E.1) implies

$$E[(\bar{\gamma}_m - E\bar{\gamma}_m)^2] \le \frac{4\kappa^2}{np}. \tag{E.2}$$

Consider the tapered covariance matrix estimate $\hat{\Sigma} = (w_{ij}\bar{\gamma}_{i-j})_{1\le i,j\le p}$, where $w_{ij} = w_{i-j}$ are weights which can be chosen as are those in (3.9). Then

$$\begin{aligned}\|\hat{\Sigma} - E\hat{\Sigma}\|_1 &= \max_{i\le p}\sum_{j=1}^{p}|w_{i-j}\bar{\gamma}_{i-j} - \gamma_{i-j}| \\ &\le 2\sum_{j=0}^{p-1}|w_j\bar{\gamma}_j - w_jE\bar{\gamma}_j| + 2\sum_{j=0}^{p-1}|w_jE\bar{\gamma}_j - \gamma_j|.\end{aligned}$$

If one uses the weights in (3.9), by (1.2), the stochastic part here satisfies

$$E\Big[\sum_{j=0}^{p-1}|w_j\bar{\gamma}_j - w_jE\bar{\gamma}_j|\Big]^2 = O(\frac{k^2}{(np)}),$$

where we assume that $k := \lfloor (np)^{1/(2+2\alpha)} \rfloor = o(p)$, allowing the high-dimensional setting with $n = o(p)$. Under the classes $\mathcal{F}_\alpha(\rho, M)$ in (1.2), or $\mathcal{H}_\alpha(\rho, M)$ in (1.3) in Cai and Zhou's paper, the bias

$$\begin{aligned}\sum_{j=0}^{p-1} |w_j E\bar{\gamma}_j - \gamma_j| &= \sum_{j=0}^{k} |E\bar{\gamma}_j - \gamma_j| + \sum_{j=1+k}^{p-1} |w_j E\bar{\gamma}_j - \gamma_j| \\ &= \sum_{j=0}^{k} O(\frac{j|\gamma_j|}{p}) + O(k^{-\alpha}) = \sum_{j=1}^{k} O(\frac{j^{-\alpha}}{p}) + O(k^{-\alpha}).\end{aligned}$$

Hence $E(\|\hat{\Sigma} - E\hat{\Sigma}\|_1^2) = O((np)^{-\alpha/(1+\alpha)})$ if $\alpha \neq 1$ and $E(\|\hat{\Sigma} - E\hat{\Sigma}\|_1^2) = O((np)^{-1/2} + p^{-2}\log^2 p)$ if $\alpha = 1$. It is not clear whether the above bound is optimal. A minimax theory is needed and would be useful.

My other concern is the authors' assumption that X_i is sub-gaussian in the sense of (1.4). How do the minimax rates (1.5) and (1.6) change if one assumes that X_i has only exponential or polynomial decaying tails? For the latter case, Bickel and Levina (2008a) obtained convergence rates for banded covariance matrix estimates. In the setting of estimating auto-covariance matrices of stationary processes, Xiao and Wu (2012) showed that optimal convergence rates can be reached under the milder polynomial moment conditions.

References

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.

Bickel, P. J. and Gel, Y. R. (2011). Banded regularization of autocovariance matrices in application to parameter estimation and forecasting of time series. *J. Roy. Statist. Soc. Ser. B* **73**, 711-728.

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

McMurry, T. L. and Politis, D. N. (2010). Banded and tapered estimates for autocovariance matrices and the linear process bootstrap. *J. Time Ser. Anal.* **31**, 471-482.

Politis, D. N., Romano, J. P. and Wolf, M. (1999). *Subsampling.* Springer, New York.

Wu, W. B. (2005). Nonlinear system theory: another look at dependence. *Proc. Natl. Acad. Sci. USA* **102**, 14150-14154.

Wu, W. B. and Pourahmadi, M. (2009). Banding sample covariance matrices of stationary processes. *Statist. Sinica* **19**, 1755-1768.

Xiao, H. and Wu, W. B. (2012). Covariance matrix estimation for stationary time series. *Ann. Statist.* **40**, 466-493

Department of Statistics, The University of Chicago, Chicago, IL 60637, USA.

E-mail: wbwu@galton.uchicago.edu

(Received April 2012; accepted April 2012)

COMMENT

Ming Yuan

Georgia Institute of Technology

Professors Cai and Zhou are to be congratulated for making yet another important contribution to the development of theory and methodology for high-dimensional covariance matrix estimation. In this article, hereafter referred to as CZ, they considered large covariance matrix estimation under the matrix ℓ_1 loss for both sparse and bandable covariance matrices. As is common in the current literature, the results from CZ are derived under the subgaussian assumption as characterized by their (1.4). Thus far, it remains unknown how essential this assumption is. To partially address this intriguing question, I shall illustrate through a simple example that subgaussianity may not play a fundamental role in determining the difficulty of estimating a large covariance matrix.

Consider here the problem of estimating a large scale matrix for elliptically contoured distributions, a more general problem than estimating the covariance matrix for multiavariate normal distributions. Let $X \in \mathbb{R}^p$ have an elliptically contoured distribution in that there exist parameters $\mu \in \mathbb{R}^p$ and $\Sigma \in \mathbb{R}^{p\times p}$ such that

$$X =_d \mu + rAU,$$

where $r \geq 0$ is a random variable, U is uniformly distributed over the unit sphere in $\mathbb{R}^n$ and is independent of r, and $A \in \mathbb{R}^{p\times p}$ is a constant matrix such that $AA^{\mathcal{F}T} = \Sigma$. In particular when r has a density, the density of X is

$$f(\mathbf{x}) = |\Sigma|^{-1/2} g((\mathbf{x}-\mu)^{\mathcal{F}T}\Sigma^{-1}(\mathbf{x}-\mu)), \qquad \mathbf{x} \in \mathbb{R}^p,$$

where g is the so-called kernel function uniquely determined by the distribution of r. Notable examples of elliptically contoured distribution are the multivariate normal, t, and the stable distributions. Note that many elliptically contoured distributions are not subgaussian and some do not even have finite second moments. For brevity, we assume that $\mu = 0$ and that Σ is a correlation-like matrix with ones on its diagonal. Our goal is to estimate Σ given a sample $X_1, \ldots, X_n$ consisting of independent copies of X. To fix ideas, wel focus on estimating sparse matrices. Write

$$\tilde{\mathcal{G}}_q(\rho, c_{n,p}) = \{\Sigma \in \mathcal{G}_q(\rho, c_{n,p}) : \Sigma_{ii} = 1 \quad \forall i\}.$$

Denote by $\mathcal{E}(\tilde{\mathcal{G}}_q(\rho, c_{n,p}))$ the collection of centered elliptically contoured distributions with $\Sigma \in \tilde{\mathcal{G}}_q(\rho, c_{n,p})$. By the argument of CZ and Cai and Zhou (2011),

$$\inf_{\hat{\Sigma}} \sup_{\mathcal{L}(X)\in\mathcal{E}(\tilde{\mathcal{G}}_q(\rho,c_{n,p})} \|\hat{\Sigma}-\Sigma\|^2 \gtrsim c_{n,p}^2\left(\frac{\log p}{n}\right)^{1-q}, \tag{F.1}$$

where $\|\cdot\|$ is the matrix ℓ_α norm with any $\alpha \geq 1$. The question of interest here is whether or not this lower bound remains tight despite the lack of subgaussianity for many distributions from $\mathcal{E}(\tilde{\mathcal{G}}_q(\rho, c_{n,p})$. Interestingly, the answer is affirmative.

To this end, we need to construct a rate optimal estimator. We appeal to a useful property of elliptically contoured distributions. Let $Y = (Y_1, Y_2)^{\mathcal{F}T}$ follow an elliptically contoured distribution with

$$\Sigma = \begin{pmatrix} 1 \; \sigma \\ \sigma \; 1 \end{pmatrix}.$$

Let $\tau = \mathbb{P}\{(Y_1 - Y_1^*)(Y_2 - Y_2^*) > 0\} - \mathbb{P}\{(Y_1 - Y_1^*)(Y_2 - Y_2^*) < 0\}$ be the population version of Kendall's τ statistic, where $Y^* = (Y_1^*, Y_2^*)^{\mathcal{F}T}$ is an independent copy of Y. Then (see, e.g., Fang, Fang, and Kotz (2002))

$$\tau = \frac{2}{\pi}\arcsin(\sigma).$$

Using this fact, we can estimate Σ in three steps.

(1) Estimate $\tau(X_i, X_j)$ by the sample Kendall's τ, denoted by $\hat{\tau}_{ij}$.

(2) Estimate Σ_{ij} by

$$\tilde{\Sigma}_{ij} = \sin\left(\frac{\pi}{2}\hat{\tau}_{ij}\right), \qquad \forall i \neq j.$$

(3) Let $\tilde{\Sigma}_{ii} = 1$ and apply thresholding to $(\tilde{\Sigma}_{ij})$:

$$\hat{\Sigma}_{ij} = \tilde{\Sigma}_{ij} I\left(\left|\tilde{\Sigma}_{ij}\right| \geq c\sqrt{\frac{\log p}{n}}\right)$$

for some numerical constant $c > 0$.

We argue that the resulting estimate $\hat{\Sigma}$ is indeed rate optimal. A careful examination of the proof of CZ reveals that it suffices to establish bounds for $|\tilde{\Sigma}_{ij} - \Sigma_{ij}|$ similar to their (3.2). This, as shown in Liu et al. (2012), can be achieved using Hoeffding's inequality for U-statistics. More specifically, we have

$$\mathbb{P}(|\tilde{\Sigma}_{ij} - \Sigma_{ij}| \geq t) \leq \exp\left(-\frac{nt^2}{2\pi^2}\right).$$

Using this in place of (3.2) of CZ, it can then be shown that

$$\sup_{\mathcal{L}(X)\in\mathcal{E}(\tilde{\mathcal{G}}_q(\rho,c_{n,p})} \|\hat{\Sigma}-\Sigma\|^2 \leq \sup_{\mathcal{L}(X)\in\mathcal{E}(\tilde{\mathcal{G}}_q(\rho,c_{n,p})} \|\hat{\Sigma}-\Sigma\|_1^2 \lesssim c_{n,p}^2\left(\frac{\log p}{n}\right)^{1-q}. \tag{F.2}$$

Combining (F.1) and (F.2), we can conclude that

$$\inf_{\hat{\Sigma}} \sup_{\mathcal{L}(X)\in\mathcal{E}(\tilde{\mathcal{G}}_q(\rho,c_{n,p})} \|\hat{\Sigma}-\Sigma\|^2 \asymp c_{n,p}^2 \left(\frac{\log p}{n}\right)^{1-q}.$$

In this particular exercise, the subgaussian assumption is irrelevant. Of course, it is also a very specific example. The exact role of subgaussianity in high-dimensional covariance matrix estimation remains to be seen.

Acknowledgement

The research of Ming Yuan was supported in part by NSF Career Award DMS-0846234.

References

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

Fang, H., Fang, K. and Kotz, S. (2002). The meta-elliptical distributions with fixed marginals, *J. Multivariate Anal.* **82**, 1-16.

Liu, H., Han, F., Yuan, M., Lafferty, J. and Wasserman, L. (2012). High dimensional semiparametric Gaussian copula graphical models. Technical report.

H. Milton Stewart School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, GA 30332, USA.

E-mail: myuan@isye.gatech.edu

(Received May 2012; accepted May 2012)

REJOINDER

T. Tony Cai and Harrison H. Zhou

University of Pennsylvania and Yale University

We deeply appreciate the many thoughtful and constructive remarks and suggestions made by the discussants of this paper. The discussants raise a number of specific points including selection of the tuning parameters (Bickel, Levina, Rothman and Zhu), estimation under different norms (Bickel, Levina, Rothman and Zhu, Sun and Zhang, and Rigollet and Tsybakov), covariance matrix estimation for time series (Wu), subgaussian condition (Wu, and Yuan), and estimation of covariance matrices with geometrically decaying entries (Xue and Zou). The discussion suggests that much work is still needed to gain deeper understanding of various aspects of the covariance matrix estimation problems in the high dimensional setting.

Professors Bickel, Levina, Rothman and Zhu raise the important issue of tuning parameter selection. This issue arises in many statistical problems and and is particularly relevant for estimating bandable covariance matrices where the cutoff for the optimal tapering/banding estimators under the Frobenius norm, matrix ℓ_1 norm, and spectral norm are quite different. It is of significant interest to see if cross-validation or other methods can lead to a theoretically justified optimal choice of the tuning parameters under the ℓ_1 norm or spectral norm losses. An example is given in their discussion which shows that for estimating a sparse covariance matrix using element-wise thresholding, the optimal choices of the threshold for estimation under the Frobenius norm, matrix ℓ_1 norm, and spectral norm, are not significantly different from each other. This phenomenon is in fact to be expected from the theory for estimating sparse covariance matrices. It has been shown in Cai and Zhou (2012) and the present paper that a single entrywise thresholding estimator is rate-optimal for estimating sparse covariance matrices in $\mathcal{G}_q(\rho, c_{n,p})$ under a wide range of losses, including the matrix ℓ_w norm for all $1 \leq w \leq \infty$, which contain the ℓ_1 norm and the spectral norm as special cases, and a class of the Bregman divergence losses, which include as special cases the squared Frobenius norm, Stein's loss, and von Neumann divergence. In other words, for estimating a sparse covariance matrix, a single thresholding estimator can achieve the optimal rates convergence under a wide range of losses. The optimal choice of the threshold is however not specified. In a recent paper, Cai and Liu (2011) introduced a fully data driven adaptive thresholding estimator, without the need of choosing a tuning parameter, that achieves the optimal rate of convergence over a larger class of sparse covariance matrices.

Professor Wu raises an interesting question on estimating covariance matrices for stationary time series. This is indeed an important problem. The setting and the techniques used in the analysis are quite different from the problem considered in the present paper. See, for example, Cai, Ren, and Zhou (2012), where optimal estimation of Toeplitz covariance matrices is considered. In the case of observing n i.i.d. realizations of a stationary time series, Professor Wu suggests to construct an estimator by tapering the average of the n auto-covariance matrices. An upper bound of order $(np)^{-\alpha/(1+\alpha)}$ (for $\alpha \neq 1$) is obtained under the squared matrix ℓ_1 norm. This bound can indeed be shown to be rate optimal. A matching lower bound can be obtained by applying Assouad's Lemma to a suitably constructed subset of the parameter space. We shall report the technical details elsewhere.

Estimation under a variety of other norms including the Frobenius norm, matrix ℓ_w operator norm, and Wiener norm, is raised by several discussants. This is an interesting question. Among these norms, estimation under the squared Frobenius norm is perhaps technically the easiest in typical settings as the problem is very similar to the usual Gaussian sequence estimation problems. This is true both for the construction of the optimal procedures and for the lower bound techniques. What sets matrix estimation apart from the usual vector estimation is the problem of estimating under the matrix operator norm losses, especially under the spectral norm loss which is highly non-additive in terms of the entrywise errors. For estimation under the matrix ℓ_w operator norm for $1 \le w \le \infty$, Cai and Zhou (2012) considered this class of losses for estimating sparse covariance matrices and established the minimax rate of convergence for all $1 \le w \le \infty$.

The difficult and intriguing case is that of estimating bandable covariance matrices under the matrix ℓ_w norm for $1 \le w \le \infty$. The minimax rate is still unknown in this case, except for $w = 1, 2$ and ∞. The major technical difficulties appear to be in the derivation of a rate-sharp minimax lower bound. An upper bound can be easily obtained by applying the Riesz-Thorin Interpolation Theorem to the variance part together with the known results for $w = 1, 2$ and ∞. Whether this upper bound is optimal remains an interesting open problem. We believe that the upper bound is rate optimal under the ℓ_w operator norm loss for $1 \le w \le \infty$, but are so far not able to establish the matching lower bound for a general value of w.

In many statistical applications, the object of direct interest is often a low-dimensional functional of the covariance matrix instead of the whole matrix itself. As in many nonparametric function estimation problems, it is true that a rate-optimal estimate of a large covariance matrix does not automatically yield rate-optimal estimates of its low-dimensional functionals. Sun and Zhang consider estimation of a regular Lipschitz functional τ which includes the partial correlation as an example. In particular, asymptotic normality of the scale LASSO estimator is established. This is an interesting result. The optimal rate of convergence for estimating these regular functionals is the usual parametric rate $\sqrt{n}$. It is of significant interest to consider estimation of other functionals which has a nonparametric minimax rate.

The subgaussianity condition used in the paper is for technical convenience. Professor Yuan considers estimation of a covariance matrix for elliptically contoured distributions and shows that the same results hold for this more general class of distributions. Professor Wu asks if the same results continue to hold under weaker conditions on the tails of the distributions. The same lower bounds clearly hold for the larger class of distributions. The question is whether the upper bounds continue to hold. In many cases this is indeed true. It is interesting to characterize the precise conditions under which the same upper bounds remain valid. When the upper bound fails, it is important to establish the new minimax rate of convergence for the class of distributions with heavy tails.

In addition to the three parameter spaces considered in the present paper, other classes of matrices can be studied as well. Xue and Zou obtain the optimal rate of convergence for a class of covariance matrices with geometrically decaying entries. In this case the minimax rate is within a log factor of the parametric rate as the rate of decay of the entries is much faster than the polynomial rate of decay considered in our paper. As observed by Xue and Zou the "effective" model size for the geometric decay case is of order $\log n$ while for the polynomial rate of decay the "effective" model size is a power of the sample size n. The double thresholding procedure is intrinsically connected to a different class of matrices. It is not clear to us for what class of matrices the double thresholding estimator can be justified theoretically better than both simple entrywise thresholding and tapering/banding estimators.

When the covariance matrix is in the sparse class $\mathcal{G}_0(\rho, k)$, we are pleased to see that Professors Rigollet and Tsybakov extend our results by relaxing our assumption $p > n^v$ to $p > k^v$ for some $v > 1$. An alternative way to relax the assumption is to use the new lower bound argument developed in Cai and Zhou (2012, Lemma 3). One can use a subset of the parameter space defined by Equation (2.3) in Section 2.1 by dividing the first row of H into k blocks with the size of order p/k, then an application of Corollary 3 of Cai and Zhou (2012) leads to the desired lower bound. The proof is not very much involved due to the simplicity of the construction.

Finally, we would like to thank the discussants again for a constructive and engaging discussion of a number of important issues on covariance matrix estimation. We are grateful for the opportunity to have learned so much from these distinguished colleagues.

References

Cai, T. T. and Liu, W. (2011). Adaptive thresholding for sparse covariance matrix estimation. *J. Amer. Statist. Assoc.* **494**, 672-684.

Cai, T. T., Ren, Z. and Zhou, H. (2012). Optimal rates of convergence for estimating Toeplitz covariance matrices. *Probab. Theory Related Fields*, to appear.

Cai, T. T. and Zhou, H. (2012). Optimal rates of convergence for sparse covariance matrix estimation. *Ann. Statist.*, to appear.

Department of Statistics, The Wharton School, University of Pennsylvania, Philadelphia, PA 19104.

E-mail: tcai@wharton.upenn.edu

Department of Statistics, Yale University, New Haven, CT 06511.

E-mail: huibin.zhou@yale.edu

(Received May 2012; accepted May 2012)