Optimal Rates of Convergence for Covariance Matrix Estimation

T. Tony Cai[1], Cun-Hui Zhang[2] and Harrison H. Zhou[3]
University of Pennsylvania, Rutgers University and Yale University

Abstract

Covariance matrix plays a central role in multivariate statistical analysis. Existing estimation methods and theoretical analyses employ the strategy of reducing the matrix estimation problem to that of estimating vectors, with the aim of estimating each row/column optimally. In this paper we establish the optimal rates of convergence for estimating the covariance matrix under both the operator norm and Frobenius norm. It is shown that optimal procedures under the two norms are different and consequently matrix estimation under the operator norm is fundamentally different from vector estimation. The minimax upper bound is obtained by constructing a special class of tapering estimators and by studying their risk properties. A key step in obtaining the optimal rate of convergence is the derivation of the minimax lower bound. The technical analysis requires new ideas that are quite different from those used in the more conventional function/sequence estimation problems.

Keywords: Covariance matrix, Frobenius norm, minimax lower bound, operator norm, optimal rate of convergence, tapering.

AMS 2000 Subject Classification: Primary 62H12; secondary 62F12, 62G09.

[1]The research of Tony Cai was supported in part by NSF Grant DMS-0604954.

[2]The research of Cun-hui Zhang was supported in part by NSF Grants DMS 0504387 and DMS 0604571 and NSA Grant MDS 904-02-1-0063

[3]The research of Harrison Zhou was supported in part by NSF Career Award DMS-0645676.

1 Introduction

Suppose we observe independent and identically distributed p-variate random variables $\mathbf{X}_1, \ldots, \mathbf{X}_n$ with covariance matrix $\Sigma_{p\times p}$. The goal is to estimate the unknown matrix $\Sigma_{p\times p}$ based on the sample $\{\mathbf{X}_i\}$. This covariance matrix estimation problem is of fundamental importance in multivariate analysis. A wide range of statistical methodologies, including clustering analysis, principal component analysis, linear and quadratic discriminant analysis, regression analysis, require the estimation of the covariance matrices. The standard and most natural estimator is the sample covariance matrix. With dramatic advances in technology, large high-dimensional data are now routinely collected in scientific investigations. Examples include climate studies, gene expression arrays, functional magnetic resonance imaging, risk management and portfolio allocation and web search problems. In such settings, the sample covariance matrix often performs poorly. See, for example, Muirhead (1987), Johnstone (2001), Bickel and Levina (2008a, b) and Fan, Fan and Lv (2007).

Regularization methods, originally developed in nonparametric function estimation, have recently been applied to estimate large covariance matrices. These include "tapering" or "banding" method in Wu and Pourahmadi (2003), Furrer and Bengtsson (2007) and Bickel and Levina (2008a), thresholding method in Bickel and Levina (2008b) and El Karoui (2008), penalized estimation in Meinshausen and Bühlmann(2006), Huang, Liu, Pourahmadi and Liu (2006), Lam and Fan (2007) and Rothman, Bickel, Levina and Zhu (2008), regularizing principal components in Johnstone and Lu (2004) and Zou, Hastie, and Tibshirani (2006). Asymptotic properties and convergence results have been given in several papers. In particular, Bickel and Levina (2008a, 2008b), El Karoui (2008) and Lam and Fan (2007) showed consistency of their estimators in operator norm and even obtained explicit rates of convergence. However, it is not clear whether any of these rates of convergence are optimal.

Despite recent progress on covariance matrix estimation there has been remarkably little fundamental theoretical study on optimal estimation. Existing estimation methods and theoretical analyses essentially employ the strategy of reducing the matrix estimation problem to that of estimating vectors, with the aim of optimally estimating each row/column of the covariance matrix Σ separately. We will show that optimal estimation of the rows/columns does not in general lead to optimal estimation of the matrix under the commonly used operator norm.

In this paper, we establish the optimal rate of convergence for estimating the covariance matrix as well as its inverse over a wide range of classes of covariance matrices. Both the operator norm and Frobenius norm are considered. It is shown that optimal procedures for these two norms are different. In particular, our result implies that the banding estimator given in Bickel and Levina (2008a) is sub-optimal under the operator norm and the performance can be significantly improved.

We begin by considering optimal estimation of the covariance matrix Σ over a class of matrices that has been considered in Bickel and Levina (2008a). Both minimax lower and upper bounds are derived. We write $a_n \asymp b_n$ if there are positive constants c and C independent of n such that $c \le a_n/b_n \le C$. For a matrix A its operator norm is defined as $\|A\| = \sup_{\|x\|_2=1} \|Ax\|_2$. We assume that $p \le \exp(\gamma n)$ for some constant $\gamma > 0$. Combining the results given in Section 3, we have the following optimal rate of convergence for estimating the covariance matrix under the operator norm.

Theorem 1 *The minimax risk of estimating the covariance matrix Σ over the class $\mathcal{F}_\alpha$ given in (3) satisfies*

$$\inf_{\hat{\Sigma}} \sup_{\mathcal{F}_\alpha} \mathbb{E} \left\| \hat{\Sigma} - \Sigma \right\|^2 \asymp \min \left\{ n^{-\frac{2\alpha}{2\alpha+1}} + \frac{\log p}{n}, \; \frac{p}{n} \right\}. \tag{1}$$

The minimax upper bound is obtained by constructing a class of tapering estimators and by studying their risk properties. It is shown that the estimator with the optimal choice of the tapering parameter attains the optimal rate of convergence. In comparison to existing methods in the literature, the proposed procedure does not attempt to estimate each row/column optimally as a vector. In fact, our procedure does not optimally trade bias and variance for each row/column. As a vector estimator, it has larger variance than squared bias for each row/column. In other words, it is undersmoothed as a vector.

A key step in obtaining the optimal rate of convergence is the derivation of the minimax lower bound. The lower bound is established by using a testing argument, where at the core is a novel construction of a collection of least favorable multivariate normal distributions and the application of Assouad's Lemma and Le Cam's method. The technical analysis requires ideas that are quite different from those used in the more conventional function/sequence estimation problems.

The paper is organized as follows. In Section 2, after basic notations and definitions are introduced, we propose a tapering procedure for the covariance matrix estimation. Section 3 derives the optimal rate of convergence for estimation under the operator norm. The upper bound is obtained by studying the properties of the tapering estimators and the minimax lower bound is obtained by a testing argument. Section 4 considers optimal estimation under the Frobenius norm. The problem of estimating the inverse of a covariance matrix is treated in Section 5. The technical proofs of auxiliary lemmas are given in Section 6.

2 Methodology

In this section we will introduce a tapering procedure for estimating the covariance matrix $\Sigma_{p\times p}$ based on a random sample of p-variate observations $\mathbf{X}_1, \ldots, \mathbf{X}_n$. The properties of the tapering estimators under the operator norm and Frobenius norm will be studied and used to establish the minimax upper bounds in Sections 3 and 4.

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

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

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

$$\Sigma^* = (\sigma^*_{ij})_{1\le i,j\le p} = \frac{1}{n}\sum_{l=1}^{n}\left(\mathbf{X}_l - \bar{\mathbf{X}}\right)\left(\mathbf{X}_l - \bar{\mathbf{X}}\right)^T \tag{2}$$

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

Following Bickel and Levina (2008a) we consider estimating the covariance matrix $\Sigma_{p\times p} = (\sigma_{ij})_{1\le i,j\le p}$ over the following parameter space

$$\mathcal{F}_\alpha = \mathcal{F}_\alpha(\varepsilon, M) = \left\{ \begin{array}{c} \Sigma : \max_j \sum_i \{|\sigma_{ij}| : |i-j| > k\} \le M k^{-\alpha} \text{ for all } k \\ \text{and } 0 < \varepsilon \le \lambda_{\min}(\Sigma) \le \lambda_{\max}(\Sigma) \le 1/\varepsilon \end{array} \right\} \tag{3}$$

where $\lambda_{\max}(\Sigma)$ and $\lambda_{\min}(\Sigma)$ are respectively the maximum and minimum eigenvalues of the matrix Σ, and $\alpha > 0$, $M > 0$ and $0 < \varepsilon < 1$. The parameter α in (3), 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 function estimation problems. The optimal rate of convergence for estimating Σ over the parameter space $\mathcal{F}_\alpha(\varepsilon, M)$ critically depends on the value of α. Our estimators of the covariance matrix Σ are constructed by tapering the maximum likelihood estimator (2) as follows.

Estimation Procedure For a given integer k with $1 \leq k \leq p$, we define a tapering estimator as

$$\hat{\Sigma} = \hat{\Sigma}_k = \left(w_{ij}\sigma_{ij}^*\right)_{p\times p} \tag{4}$$

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

$$w_{ij} = k^{-1}\{(2k - |i-j|)_+ - (k - |i-j|)_+\}. \tag{5}$$

Note that the weights w_{ij} can be written as

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

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

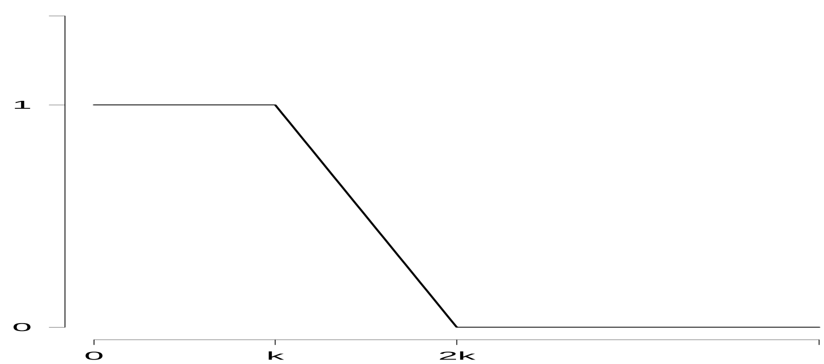

Figure 1: The weights as a function of $|i-j|$.

The tapering estimators are different from the banding estimators used in Bickel and Levina (2008a). It is important to note that the tapering estimator given in (4) can be rewritten as a sum of many small block matrices along the diagonal. This simple but important observation is very useful for our technical arguments. Define the block matrices

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

and set

$$S^{*(m)} = \sum_{l=2-m}^{p} M_l^{*(m)}$$

for all integers $2-m \leq l \leq p$ and $m \geq 1$.

Lemma 1 *The taping estimator $\hat{\Sigma}_k$ given in (4) can be written as*

$$\hat{\Sigma}_k = k^{-1}\left(S^{*(2k)} - S^{*(k)}\right). \tag{6}$$

It is clear that the performance of the estimator $\hat{\Sigma}_k$ depends on the choice of the tapering parameter k. The optimal choice of k critically depends on the norm under which the estimation error is measured. We will study in the next two sections the rate of convergence of the tapering estimator under both the operator norm and Frobenius norm. Together with the minimax lower bounds derived in Sections 3 and 4, the results show that a tapering estimator with the optimal choice of k attains the optimal rate of convergence under these two norms.

3 Rate Optimality under the Operator Norm

In this section we will establish the optimal rate of convergence under the operator norm. We shall focus on the Gaussian case where $\mathbf{X}_1, \ldots, \mathbf{X}_n \overset{iid}{\sim} N(\mu, \Sigma_{p\times p})$. For $1 \le q \le \infty$, the matrix ℓ_q-norm of a matrix A is defined by $\|A\|_q = \max_{\|x\|_q=1} \|Ax\|_q$. The commonly used operator norm $\|\cdot\|$ coincides with the matrix ℓ_2-norm $\|\cdot\|_2$. For a symmetric matrix A, it is known that the operator norm $\|A\|$ is equal to the magnitude of the largest eigenvalue of A. Hence it is also called the spectral norm. We will establish Theorem 1 by deriving a minimax upper bound using the tapering estimator and a matching minimax lower bound by a careful construction of a collection of multivariate normal distributions and the application of Assouad's Lemma and Le Cam's method. We shall focus on the case $p \ge 2n^{\frac{1}{2\alpha+1}}$ in Sections 3.1 and 3.2. The case of $p < 2n^{\frac{1}{2\alpha+1}}$, which will be discussed in Section 3.3, is similar and slightly easier.

3.1 Minimax Upper Bound under the Operator Norm

We derive in this section the risk upper bound for the tapering estimators defined in (6) under the operator norm. Throughout the paper we denote by C a generic positive constant which may vary from place to place but always depends only on indices α, ε and M of the matrix family.

Theorem 2 *The tapering estimator $\hat{\Sigma}_k$, defined in (6), of the covariance matrix $\Sigma_{p\times p}$ with $p \ge 2n^{\frac{1}{2\alpha+1}}$ satisfies*

$$\sup_{\mathcal{F}_\alpha} \mathbb{E}\left\|\hat{\Sigma}_k - \Sigma\right\|^2 \le C\frac{k+\log p}{n} + Ck^{-2\alpha} \tag{7}$$

for some constant $C > 0$. In particular, the estimator $\hat{\Sigma} = \hat{\Sigma}_k$ with $k = n^{\frac{1}{2\alpha+1}}$ satisfies

$$\sup_{\mathcal{F}_\alpha} \mathbb{E} \left\| \hat{\Sigma} - \Sigma \right\|^2 \le C n^{-\frac{2\alpha}{2\alpha+1}} + \frac{\log p}{n}. \tag{8}$$

From (7) it is clear that the optimal choice of k is of order $n^{\frac{1}{2\alpha+1}}$. The upper bound given in (8) is thus rate optimal among the class of the tapering estimators defined in (6). The minimax lower bound derived in Section 3.2 shows that the rate of convergence given in (8) is in fact optimal among all estimators.

Proof of Theorem 2: Note that Σ^* is translation invariant and so is $\hat{\Sigma}$. We shall thus assume $\mu = 0$ for the rest of the paper. 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, since the distribution of $\bar{\mathbf{X}} \bar{\mathbf{X}}^T$ is identical to $\frac{1}{n} \mathbf{X}_1 \mathbf{X}_1^T$. In what follows we shall ignore this negligible term and focus on the dominating term $\frac{1}{n} \sum_{l=1}^{n} \mathbf{X}_l \mathbf{X}_l^T$.

Set $\tilde{\Sigma} = \frac{1}{n} \sum_{l=1}^{n} \mathbf{X}_l \mathbf{X}_l^T$ and write $\tilde{\Sigma} = (\tilde{\sigma}_{ij})_{1 \le i,j \le p}$. Let

$$\breve{\Sigma} = (\breve{\sigma}_{ij})_{1 \le i,j \le p} = (w_{ij} \tilde{\sigma}_{ij})_{1 \le i,j \le p} \tag{9}$$

with w_{ij} given in (5). To prove Theorem 2, it suffices to show

$$\sup_{\mathcal{F}} \mathbb{E} \left\| \breve{\Sigma} - \Sigma \right\|^2 \le C n^{-\frac{2\alpha}{2\alpha+1}} + C \frac{\log p}{n}. \tag{10}$$

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

$$\mathbb{E} \tilde{\sigma}_{ij} = \sigma_{ij} \tag{11}$$

$$\mathbb{V}ar(\tilde{\sigma}_{ij}) = \frac{1}{n} \left(\sigma_{ii} \sigma_{jj} + \sigma_{ij}^2 \right) \tag{12}$$

i.e., $\tilde{\sigma}_{ij}$ is an unbiased estimator of σ_{ij} with a variance $\left(\sigma_{ii} \sigma_{jj} + \sigma_{ij}^2 \right) / n$.

We will first show that the variance part,

$$\mathbb{E} \left\| \breve{\Sigma} - \mathbb{E} \breve{\Sigma} \right\|^2 \le C \frac{k + \log p}{n} \tag{13}$$

and then the bias part,

$$\left\| \mathbb{E} \breve{\Sigma} - \Sigma \right\|^2 \le C k^{-2\alpha}. \tag{14}$$

It then follows immediately that

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

This proves equation (10) and equation (7) then follows. Since $p \geq 2n^{\frac{1}{2\alpha+1}}$, we may choose

$$k = n^{\frac{1}{2\alpha+1}} \tag{15}$$

and the estimator $\hat{\Sigma}$ with k given in (15) satisfies

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

Theorem 2 is then proved.

We first prove the risk upper bound (14) for the bias part. It is well known that the operator norm of a symmetric matrix $A = (a_{ij})_{p\times p}$ is bounded by its ℓ_1 norm, i.e.,

$$\|A\| \leq \|A\|_1 = \max_{i=1,\dots,p} \sum_{j=1}^{p} |a_{ij}|.$$

This fact can be argued easily as follows. Let λ be an eigenvalue of A, and $v = (v_i)_{1\leq i\leq p}$ be a corresponding eigenvector, i.e., $Av = \lambda v$. Let $|v_i| = \|v\|_\infty$. Then $\lambda = \sum_{j=1}^{p} a_{ij}\frac{v_j}{v_i}$ and hence $|\lambda| \leq \sum_{j=1}^{p} |a_{ij}| \left|\frac{v_j}{v_i}\right| \leq \sum_{j=1}^{p} |a_{ij}| \leq \max_{i=1,\dots,p} \sum_{j=1}^{p} |a_{ij}|$. We bound the operator norm of the bias part $\mathbb{E}\hat{\Sigma} - \Sigma$ by its l_1 norm. Since $\mathbb{E}\tilde{\sigma}_{ij} = \sigma_{ij}$, we have

$$\mathbb{E}\breve{\Sigma} - \Sigma = \left((w_{ij} - 1)\,\sigma_{ij}\right)_{p\times p}$$

where $w_{ij} \in [0,1]$ and is exactly 1 when $|i-j| \leq k$, then

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

Now we establish (13) which is relatively complicated. The key idea in the proof is to write the whole matrix as an average of matrices which are sum of a large number of small disjoint block matrices, and for each small block matrix the classical random matrix theory can be applied. The following lemma shows that the operator norm of the random matrix $\breve{\Sigma} - \mathbb{E}\breve{\Sigma}$ is controlled by the maximum of operator norms of p number of $2k \times 2k$ random matrices. Let $M_l^{(m)} = \left(\breve{\sigma}_{ij} I\left\{l \leq i < l+m, l \leq j < l+m\right\}\right)_{p\times p}$.

Lemma 2 *Let $\breve{\Sigma}$ be defined as in (6). Then*

$$\left\|\breve{\Sigma} - \mathbb{E}\breve{\Sigma}\right\| \leq 3 \max_{1\leq l\leq p} \left\|M_l^{(2k)} - \mathbb{E}M_l^{(2k)}\right\|.$$

For each small $m \times m$ random matrix with $m = 2k$, we control its operator norm by using classical results on random matrices. See, for example, Theorem II.13 of Davidson and Szarek (2001).

Lemma 3 *Let $Z \sim N(0, 1/n)$. If $m \le n$, then*

$$\mathbb{P}\left\{\left\|M_l^{(m)} - \mathbb{E}M_l^{(m)}\right\| / \left\|\mathbb{E}M_l^{(m)}\right\| > x\right\} \le \mathbb{P}\left\{(1 + \sqrt{m/n} + |Z|)^2 - 1 > x\right\} \qquad (16)$$

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

The proof of Lemmas 2 and 3 is given in Section 6. Lemma 3 implies that for each l there is a normal random variable $Z_l \sim N(0, 1/n)$ such that

$$\left\|M_l^{(m)} - \mathbb{E}M_l^{(m)}\right\| / \left\|\mathbb{E}M_l^{(m)}\right\| \le (1 + \sqrt{m/n} + |Z_l|)^2 - 1$$

by using quantile coupling.

With Lemmas 2 and 3 we are now ready to show the variance bound (13). The assumption $0 < \varepsilon \le \lambda_{\min}(\Sigma) \le \lambda_{\max}(\Sigma) \le 1/\varepsilon$ implies that $\left\|\mathbb{E}M_l^{(m)}\right\| \le 1/\varepsilon$. By Lemma 2 and the coupling result we have

$$\begin{aligned}
\mathbb{E}\left\|\breve{\Sigma} - \mathbb{E}\breve{\Sigma}\right\|^2 &\le 9\mathbb{E}\max_{1\le l\le p}\left\|M_l^{(2k)} - \mathbb{E}M_l^{(2k)}\right\|^2 \\
&\le 9C\mathbb{E}\max_{1\le l\le p}\left\{(1 + \sqrt{2k/n} + |Z_l|)^2 - 1\right\}^2 \\
&\le 9C\mathbb{E}\max_{1\le l\le p}\left\{2\sqrt{2k/n} + 2|Z_l| + 2|Z_l|\sqrt{2k/n} + 2k/n + Z_l^2\right\}^2 \\
&\le C_1\mathbb{E}\max_{1\le l\le p}\left\{\sqrt{k/n} + |Z_l|\right\}^2 \le 2C_1\mathbb{E}\max_{1\le l\le p}\left\{k/n + |Z_l|^2\right\}.
\end{aligned}$$

Equation (13) now follows from the well known fact that $\mathbb{E}\max_{1\le l\le p}|Z_l|^2 \le \frac{2\log p}{n}$ for p sufficiently large. ∎

3.2 Lower Bound under the Operator Norm

Theorem 2 in Section 3.1 shows that the optimal tapering estimator attains the rate of convergence $n^{-\frac{2\alpha}{2\alpha+1}} + \frac{\log p}{n}$. In this section we shall show that this rate of convergence is indeed optimal among all estimators by showing that the upper bound in equation (8) can not be improved. More specifically we shall show that the following minimax lower bound holds.

Theorem 3 *The minimax risk satisfies*

$$\inf_{\hat{\Sigma}}\sup_{\mathcal{F}_\alpha}\mathbb{E}\left\|\hat{\Sigma} - \Sigma\right\|^2 \ge cn^{-\frac{2\alpha}{2\alpha+1}} + c\frac{\log p}{n}.$$

The basic strategy underlying the proof of Theorem 3 is to carefully construct a finite collection of multivariate normal distributions and calculate the total variation affinity between pairs of probability measures in the collection.

We shall now define a parameter space that is appropriate for the minimax lower bound argument. For given positive integers k and m with $2k \le p$ and $1 \le m \le k$, define the $p \times p$ matrix $B(m,k) = (b_{ij})_{p\times p}$ with

$$b_{ij} = I\left\{i = m \text{ and } m+1 \le j \le 2k, \text{ or } j = m \text{ and } m+1 \le i \le 2k\right\}.$$

Set $k = n^{\frac{1}{2\alpha+1}}$ and $a = k^{-(\alpha+1)}$. We then define the collection of 2^k covariance matrices as

$$\mathcal{F}_{11} = \left\{\Sigma\left(\theta\right) : \Sigma\left(\theta\right) = I_p + \tau a \sum_{m=1}^{k} \theta_m B(m,k), \quad \theta = (\theta_m) \in \{0,1\}^k\right\} \tag{17}$$

where I_p is the $p \times p$ identity matrix and $0 < \tau < \min\left\{2^{-\alpha-1}M, \left(1-\varepsilon\right)/2\right\}$. It is easy to check that $\mathcal{F}_{11} \subset \mathcal{F}_\alpha(\varepsilon, M)$. In addition to $\mathcal{F}_{11}$ we also define a collection of diagonal matrices

$$\mathcal{F}_{12} = \left\{\Sigma_m : \Sigma_m = I_p + \left(\sqrt{\frac{\tau}{n}\log p_1} I\left\{i = j = m\right\}\right)_{p\times p}, 0 \le m \le p_1\right\}$$

where $p_1 = \min\left\{p, e^n\right\}$ and $0 < \tau < \min\left\{\left(1/\varepsilon - 1\right)^2, 1\right\}$. Let $\mathcal{F}_1 = \mathcal{F}_{11} \cup \mathcal{F}_{12}$. It is clear that $\mathcal{F}_1 \subset \mathcal{F}_\alpha(\varepsilon, M)$ for .

We shall show below separately that for some constant $c > 0$

$$\inf_{\hat{\Sigma}} \sup_{\mathcal{F}_{11}} \mathbb{E}\left\|\hat{\Sigma} - \Sigma\right\|^2 \ge c n^{-\frac{2\alpha}{2\alpha+1}} \tag{18}$$

and

$$\inf_{\hat{\Sigma}} \sup_{\mathcal{F}_{12}} \mathbb{E}\left\|\hat{\Sigma} - \Sigma\right\|^2 \ge c\frac{\log p}{n}. \tag{19}$$

Equations (18) and (19) together imply

$$\inf_{\hat{\Sigma}} \sup_{\mathcal{F}_{1}} \mathbb{E}\left\|\hat{\Sigma} - \Sigma\right\|^2 \ge \frac{c}{2}\left(n^{-\frac{2\alpha}{2\alpha+1}} + \frac{\log p}{n}\right) \tag{20}$$

We shall establish the lower bound (18) by using Assouad's Lemma and the lower bound (19) by using Le Cam's method and a two-point argument.

3.2.1 A Lower Bound by Assouad's Lemma

The key technical tool to establish equation (18) is the Assouad's lemma in Assouad (1983). It gives a lower bound for the maximum risk over the parameter set $\Theta = \{0,1\}^k$

to the problem of estimating an arbitrary quantity $\psi(\theta)$, belonging to a metric space with metric d. Let $H(\theta,\theta') = \sum_{i=1}^{k} |\theta_i - \theta_i'|$ be the Hamming distance on $\{0,1\}^k$, which counts the number of positions at which θ and θ' differ. For two probability measures P and Q with density p and q with respect to any common dominating measure μ, write the total variation affinity $\|P \wedge Q\| = \int p \wedge q d\mu$. Assouad's Lemma provides a minimax lower bound.

Lemma 4 (Assouad) *Let $\Theta = \{0,1\}^k$ and let T be an estimator based on an observation from a distribution in the collection $\{P_\theta, \theta \in \Theta\}$. Then for all $s > 0$*

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

The Assouad's lemma is connected to multiple comparisons. In total there are k comparisons. The lower bound has three factors. The first factor is basically the minimum cost of making a mistake per comparison, and the last factor is the lower bound for the total probability of making type I and type II errors for each comparison, and $k/2$ is the expected number of mistakes one makes when $\mathbb{P}_\theta$ and $\mathbb{P}_{\theta'}$ are not distinguishable from each other when $H(\theta,\theta') = 1$.

We now prove the lower bound (18). Let $\mathbf{X}_1, \ldots, \mathbf{X}_n \overset{iid}{\sim} N(0, \Sigma(\theta))$ with $\Sigma(\theta) \in \mathcal{F}_{11}$. Denote the joint distribution by P_θ. Applying Assouad's Lemma to the parameter space $\mathcal{F}_{11}$, we have

$$\inf_{\hat{\Sigma}} \max_{\theta\in\{0,1\}^k} 2^2 E_\theta \left\|\hat{\Sigma} - \Sigma(\theta)\right\|^2 \geq \min_{H(\theta,\theta')\geq 1} \frac{\|\Sigma(\theta) - \Sigma(\theta')\|^2}{H(\theta,\theta')} \frac{k}{2} \min_{H(\theta,\theta')=1} \|P_\theta \wedge P_{\theta'}\|. \quad (21)$$

We shall state the bounds for the the first and third factors on the right hand of (21) in two lemmas. The proof of these lemmas is given in Section 6.

Lemma 5 *Let $\Sigma(\theta)$ be defined as in (17). Then for some constant $c > 0$*

$$\min_{H(\theta,\theta')\geq 1} \frac{\|\Sigma(\theta) - \Sigma(\theta')\|^2}{H(\theta,\theta')} \geq cka^2.$$

Lemma 6 *Let $\mathbf{X}_1, \ldots, \mathbf{X}_n \overset{iid}{\sim} N(0, \Sigma(\theta))$ with $\Sigma(\theta) \in \mathcal{F}_{11}$. Denote the joint distribution by P_θ. Then for some constant $c > 0$*

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

It then follows from Lemmas 5 and 6 together with the fact $k = n^{\frac{1}{2\alpha+1}}$

$$\max_{\Sigma(\theta)\in\mathcal{F}_{11}} 2^2 E_\theta \left\|\hat{\Sigma} - \Sigma(\theta)\right\|^2 \geq \frac{c^2}{2} k^2 a^2 \geq c_1 n^{-\frac{2\alpha}{2\alpha+1}}$$

for some $c_1 > 0$. ∎

3.2.2 A Lower Bound Using Le Cam's Method

Now we prove the lower bound (19). For $1 \le m \le p_1$, let Σ_m be a diagonal covariance matrix with $\sigma_{mm} = 1 + \sqrt{\tau \frac{\log p_1}{n}}$, $\sigma_{ii} = 1$ for $i \neq m$, and let Σ_0 be the identity matrix. Let f_m, $1 \le m \le p_1$, be the joint density of $\mathbf{X}_1, \ldots, \mathbf{X}_n$ with $\mathbf{X}_i \sim N(0, \Sigma_m)$.

We will apply Le Cam's method to derive a lower bound for minimax risk. Let X be an observation from a distribution in the collection $\{P_\theta, \theta \in \Theta\}$ where $\Theta = \{\theta_0, \theta_1, \ldots, \theta_{p_1}\}$. Le Cam's method, which is based on a two-point testing argument, gives a lower bound for the maximum estimation risk over the parameter set Θ. More specifically, let L be the loss function. Define $r(\theta_0, \theta_m) = \inf_t [L(t, \theta_0) + L(t, \theta_m)]$ and $r_{\min} = \inf_{1 \le m \le p_1} r(\theta_0, \theta_m)$, and denote $\bar{\mathbb{P}} = \frac{1}{p_1} \sum_{m=1}^{p_1} \mathbb{P}_{\theta_m}$.

Lemma 7 *Let T be an estimator of θ based on an observation from a distribution in the collection $\{P_\theta, \theta \in \Theta = \{\theta_0, \theta_1, \ldots, \theta_{p_1}\}\}$, then*

$$\sup_\theta \mathbb{E} L(T, \theta) \ge \frac{1}{2} r_{\min} \left\| \mathbb{P}_{\theta_0} \wedge \bar{\mathbb{P}} \right\|$$

We refer to Yu (1997) for more detailed discussions on Le Cam's method. Let $\theta_m = \Sigma_m$ for $0 \le m \le p_1$ and the loss function L be the squared operator norm. It is easy to see $r(\theta_0, \theta_m) = \frac{1}{2} \tau \frac{\log p_1}{n}$ for all $1 \le m \le p_1$. Then the lower bound (19) follows immediately if there is a constant $c > 0$ such that

$$\left\| \mathbb{P}_{\theta_0} \wedge \bar{\mathbb{P}} \right\| \ge c. \tag{22}$$

Since $\int q_0 \wedge q_1 d\mu = 1 - \frac{1}{2} \int |q_0 - q_1| \, d\mu$ for any two densities q_0 and q_1, and the Jensen's inequality implies

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

Hence $\int q_0 \wedge q_1 d\mu \ge 1 - \frac{1}{2} \left[\int \left(\frac{q_0^2}{q_1} - 1 \right) d\mu \right]^{1/2}$. To establish equation (22), it thus suffices to show that $\int \left(\frac{1}{p_1} \sum_{m=1}^{p_1} f_m \right)^2 / f_0 - 1 \to 0$, i.e.,

$$\frac{1}{p_1^2} \sum_{m=1}^{p_1} \left(\frac{f_m^2}{f_0} - 1 \right) + \frac{1}{p_1^2} \sum_{m \neq j} \left(\frac{f_m f_j}{f_0} - 1 \right) \to 0. \tag{23}$$

We now calculate $\int \frac{f_m f_j}{f_0}$. For $m \neq j$ it is easy to see

$$\int \frac{f_m f_j}{f_0} - 1 = 0.$$

When $m = j$, we have

$$\int \frac{f_m^2}{f_0} = \frac{\left(\sqrt{2\pi\sigma_{mm}}\right)^{-2n}}{\left(\sqrt{2\pi}\right)^{-n}} \left[\int \exp\left(x^2\left(-\frac{1}{\sigma_{mm}}+\frac{1}{2}\right)\right)\right]^n$$
$$= \left[1-(1-\sigma_{mm})^2\right]^{-n/2} = \left(1-\tau\frac{\log p_1}{n}\right)^{-n/2}.$$

Thus

$$\int\left(\frac{1}{p_1}\sum_{m=1}^{p_1} f_m\right)^2 / f_0 - 1 = \frac{1}{p_1^2}\sum_{m=1}^{p_1}\left(\int \frac{f_m^2}{f_0}-1\right) \le \frac{1}{p_1}\left(1-\tau\frac{\log p_1}{n}\right)^{-n/2} - \frac{1}{p_1} \to 0$$

for $0 < \tau < 1$, which immediately implies equation (19).

Remark 1 *In Bickel and Levina (2008a) it was assumed that* $\frac{\log p}{n} \to 0$*. Here we show that this assumption is necessary to estimate the covariance matrix consistently under the operator norm.*

3.3 Discussion

Theorems 2 and 3 together shows that the minimax risk for estimating the covariance matrices over the parameter space $\mathcal{F}_\alpha$ satisfies, for $p \ge 2n^{\frac{1}{2\alpha+1}}$,

$$\inf_{\hat\Sigma}\sup_{\mathcal{F}} \mathbb{E}\left\|\hat\Sigma - \Sigma\right\|^2 \asymp n^{-\frac{2\alpha}{2\alpha+1}} + \frac{\log p}{n}. \tag{24}$$

The results also show that the tapering estimator $\hat\Sigma_k$ with tapering parameter $k = n^{\frac{1}{2\alpha+1}}$ attains the optimal rate of convergence $n^{-\frac{2\alpha}{2\alpha+1}} + \frac{\log p}{n}$.

A few interesting points can be made on the optimal rate of convergence $n^{-\frac{2\alpha}{2\alpha+1}} + \frac{\log p}{n}$. When the dimension p is relatively small, i.e., $\log p = o(n^{\frac{1}{2\alpha+1}})$, p has no effect on the convergence rate and the rate is purely driven by the "smoothness" parameter α. However, when p is large, i.e., $\log p \gg n^{\frac{1}{2\alpha+1}}$, p plays a significantly role in determining the minimax rate.

We should emphasize that the optimal choice of the tapering parameter $k \asymp n^{\frac{1}{2\alpha+1}}$ is different from the optimal choice for estimating the rows/columns as vectors under mean squared error loss. Straightforward calculation shows that in the latter case the best cutoff is $k \asymp n^{\frac{1}{2(\alpha+1)}}$ so that the tradeoff between the squared bias and the variance is optimal. With $k \asymp n^{\frac{1}{2\alpha+1}}$, the tapering estimator has smaller squared bias than the variance as a vector estimator of each row/column.

It is also interesting to compare our results with those given in Bickel and Levina (2008a). A banding estimator with bandwidth $k = \left(\frac{\log p}{n}\right)^{\frac{1}{2(\alpha+1)}}$ was proposed and the rate of convergence $\left(\frac{\log p}{n}\right)^{\frac{\alpha}{\alpha+1}}$ was proved. Take for example $\alpha = 1/2$ and $p = e^{\sqrt{n}}$. Their rate is $n^{-\frac{1}{6}}$, while the optimal rate in Theorem 1 is $n^{-\frac{1}{2}}$. Hence the banding estimator given in Bickel and Levina (2008a) is not rate optimal.

It is instructive to take a closer look at the motivation behind the construction of the banding estimator in Bickel and Levina (2008a). Denote $\hat{\Sigma} - \mathbb{E}\hat{\Sigma}$ by V and let $V = (v_{ij})$. An important step in the proof of Theorem 1 in Bickel and Levina (2008a) is to control the operator norm by the ℓ_1 norm as follows,

$$\begin{aligned}\mathbb{E}\left\|\hat{\Sigma} - \mathbb{E}\hat{\Sigma}\right\|^2 &\leq \mathbb{E}\left\|\hat{\Sigma} - \mathbb{E}\hat{\Sigma}\right\|_1^2 = \mathbb{E}\left(\max_{j=1,\dots,p}\sum_i |v_{ij}|\right)^2 \\ &\leq C\left(\frac{k}{\sqrt{n}}\sqrt{\log p}\right)^2 = C\frac{k^2\log p}{n}.\end{aligned}$$

Note that $\mathbb{E}\left[|v_{ij}|\, I\left\{|i-j| \leq k\right\}\right] \asymp 1/\sqrt{n}$, then $\mathbb{E}\sum_i |v_{ij}| \asymp k/\sqrt{n}$. It is then expected that $\mathbb{E}\left(\max_{j=1,\dots,p}\sum_i |v_{ij}|\right)^2 \leq C\left(\frac{k}{\sqrt{n}}\sqrt{\log p}\right)^2$ (see Bickel and Levina (2008a) for details) and so

$$\mathbb{E}\left\|\breve{\Sigma} - \Sigma\right\|_1^2 \leq C\frac{k^2\log p}{n} + Ck^{-2\alpha}$$

An optimal tradeoff of k is then $\left(\frac{\log p}{n}\right)^{\frac{1}{2(\alpha+1)}}$ which implies a rate of $\left(\frac{\log p}{n}\right)^{-\frac{\alpha}{\alpha+1}}$ in Theorem 1 in Bickel and Levina (2008a). This rate is slower than the optimal rate $n^{-\frac{2\alpha}{2\alpha+1}} + \frac{\log p}{n}$ in Theorem 1.

We have considered the parameter space $\mathcal{F}_\alpha$ defined in (3). Other similar parameter spaces can also be considered. For example in time series analysis it is often assumed the covariance $|\sigma_{ij}|$ decays at the rate $|i-j|^{-(\alpha+1)}$ for some $\alpha > 0$. Consider the collection of positive definite symmetric matrices satisfying the following conditions:

$$\begin{aligned}\mathcal{G}_\alpha = \mathcal{G}_\alpha(\varepsilon, M_1) &= \Big\{\Sigma : |\sigma_{ij}| \leq M_1 |i-j|^{-(\alpha+1)} \text{ for } i \neq j \\ &\qquad \text{and } 0 < \varepsilon \leq \lambda_{\min}(\Sigma) \leq \lambda_{\max}(\Sigma) \leq \varepsilon^{-1}\Big\}\end{aligned} \tag{25}$$

where $\lambda_{\max}(\Sigma)$ and $\lambda_{\min}(\Sigma)$ are respectively the maximum and minimum eigenvalues of the matrix Σ. Note that $\mathcal{G}_\alpha(\varepsilon, M_1)$ is a subset of $\mathcal{F}_\alpha(\varepsilon, M)$ as long as $M_1 \leq \alpha M$. Using virtually identical arguments one can show that

$$\inf_{\hat{\Sigma}}\sup_{\mathcal{G}_\alpha}\mathbb{E}\left\|\hat{\Sigma} - \Sigma\right\|^2 \asymp n^{-\frac{2\alpha}{2\alpha+1}} + \frac{\log p}{n}.$$

3.3.1 The Case of $p < 2n^{\frac{1}{2\alpha+1}}$

We have focused on the case $p \geq 2n^{\frac{1}{2\alpha+1}}$ in Sections 3.1 and 3.2. The case of $p < 2n^{\frac{1}{2\alpha+1}}$ can be handled in a similar way. The main difference is that in this case we no longer have a tapering estimator $\hat{\Sigma}_k$ with $k = n^{\frac{1}{2\alpha+1}}$ because $2k > p$. Instead the maximum likelihood estimator Σ^* can be used directly. It is easy to show in this case,

$$\sup_{\mathcal{F}_\alpha} \mathbb{E} \left\|\Sigma^* - \Sigma\right\|^2 \le C\frac{p}{n}. \tag{26}$$

The lower bound can also be obtained by a parameter space that is similar to $\mathcal{F}_{11}$ and the application of Assouad's Lemma. To be more specific, For an integer $1 \le m \le p/2$, define the $p \times p$ matrix $B_m = (b_{ij})_{p\times p}$ with

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

Define the collection of 2^k covariance matrices as

$$\mathcal{F}^* = \left\{\Sigma(\theta) : \Sigma(\theta) = I_p + \tau\frac{1}{\sqrt{np}}\sum_{m=1}^{k}\theta_m B(m,k), \quad \theta = (\theta_m) \in \{0,1\}^k\right\}. \tag{27}$$

Since $p < 2n^{\frac{1}{2\alpha+1}}$, then $\frac{1}{\sqrt{np}} < 2^{\alpha+1/2}p^{-(\alpha+1)}$. Again it is easy to check $\mathcal{F}^* \subset \mathcal{F}_\alpha(\varepsilon, M)$ when $0 < \tau < \min\left\{2^{-\alpha-1}M, (1-\varepsilon)/2\right\}$. The following lower bound then follows from the same argument as in Section 3.2.1

$$\inf_{\hat{\Sigma}} \sup_{\mathcal{F}^*} \mathbb{E} \left\|\hat{\Sigma} - \Sigma\right\|^2 \ge cp\left(\frac{1}{\sqrt{np}}\right)^2 \cdot \frac{p}{2} \cdot c_1 \ge c_2\frac{p}{n}. \tag{28}$$

Equations (26) and (28) together yield the minimax rate of convergence for the case $p \le 2n^{\frac{1}{2\alpha+1}}$,

$$\inf_{\hat{\Sigma}} \sup_{\mathcal{F}_\alpha} \mathbb{E} \left\|\hat{\Sigma} - \Sigma\right\|^2 \asymp \frac{p}{n}. \tag{29}$$

This together with equation (24) give the optimal rate of convergence

$$\inf_{\hat{\Sigma}} \sup_{\mathcal{F}_\alpha} \mathbb{E} \left\|\hat{\Sigma} - \Sigma\right\|^2 \asymp \min\left\{n^{-\frac{2\alpha}{2\alpha+1}} + \frac{\log p}{n}, \ \frac{p}{n}\right\}. \tag{30}$$

4 Rate Optimality under the Frobenius Norm

In addition to the operator norm, the Frobenius norm is another commonly used matrix norm for measuring the accuracy of an covariance matrix estimator. See, e.g., Lam and Fan (2007). In this section we consider the optimal rate of convergence for covariance matrix estimation under the Frobenius norm. The Frobenius norm of a matrix $A = (a_{ij})$ is defined as the ℓ_2 vector norm of all entries in the matrix,

$$\|A\|_F = \sqrt{\sum_{i,j} a_{ij}^2}.$$

This is equivalent to treating the matrix A as a vector of length p^2. It is easy to see that the operator norm is bounded by the Frobenius norm, i.e., $\|A\| \leq \|A\|_F$.

Again suppose we observe $\mathbf{X}_1, \ldots, \mathbf{X}_n \overset{iid}{\sim} N(\mu, \Sigma_{p\times p})$. The following theorem gives the minimax rate of convergence for estimating the covariance matrix Σ under the Frobenius norm based on the sample $\{\mathbf{X}_1, \ldots, \mathbf{X}_n\}$.

Theorem 4 *The minimax risk under the Frobenius norm satisfies*

$$\inf_{\hat{\Sigma}} \sup_{\mathcal{F}_\alpha} \mathbb{E}\frac{1}{p}\left\|\hat{\Sigma} - \Sigma\right\|_F^2 \asymp \min\left\{n^{-\frac{2\alpha+1}{2(\alpha+1)}},\ \frac{p}{n}\right\}. \tag{31}$$

This rate is identical to the optimal rate of convergence for estimating one row/column as a vector with the same constraint under squared error loss.

Remark 2 *It is easy to see that the optimal rate of convergence over the parameter space $\mathcal{G}_\alpha$ given in (25) also satisfies*

$$\inf_{\hat{\Sigma}} \sup_{\mathcal{G}_\alpha} \mathbb{E}\frac{1}{p}\left\|\hat{\Sigma} - \Sigma\right\|_F^2 \asymp \min\left\{n^{-\frac{2\alpha+1}{2(\alpha+1)}},\ \frac{p}{n}\right\}.$$

We shall establish below separately the minimax upper bound and minimax lower bound.

4.1 Upper Bound under the Frobenius Norm

The minimax upper bound is derived by again considering the tapering estimator (4). Under the Frobenius norm the risk function is separable. The risk of the tapering estimator can be bounded separately under the squared ℓ_2 loss for each row/column. This method has been commonly used in nonparametric function estimation using orthogonal basis expansions. Since

$$\begin{aligned} \mathbb{E}\tilde{\sigma}_{ij} &= \sigma_{ij} \\ \mathbb{V}ar(\tilde{\sigma}_{ij}) &= \mathbb{E}\left(\xi_{ij}\right)^2 = \frac{1}{n}\left(\sigma_{ii}\sigma_{jj} + \sigma_{ij}^2\right), \end{aligned}$$

for the tapering estimator (4), we have

$$\mathbb{E}\left(w_{ij}\tilde{\sigma}_{ij} - \sigma_{ij}\right)^2 = \left(1 - w_{ij}\right)^2 \sigma_{ij}^2 + w_{ij}^2 \frac{\sigma_{ii}\sigma_{jj} + \sigma_{ij}^2}{n}.$$

It can be seen easily that

$$\frac{1}{p}\mathbb{E}\left\|\breve{\Sigma} - \Sigma\right\|_F^2 \leq \frac{1}{p}\sum_{\{(i,j):k<|i-j|\}} \sigma_{ij}^2 + \frac{1}{p}\sum_{\{(i,j):|i-j|\leq 2k\}} w_{ij}^2 \frac{\sigma_{ii}\sigma_{jj} + \sigma_{ij}^2}{n} \equiv R_1 + R_2$$

The assumption $0 < \varepsilon \le \lambda_{\min}(\Sigma) \le \lambda_{\max}(\Sigma) \le 1/\varepsilon$ implies that $\varepsilon \le \sigma_{ii} \le 1/\varepsilon$ for all i. Since $|\sigma_{ij}|$ is uniformly bounded for all i and j, we immediately have $R_2 \le C\frac{k}{n}$.

Now we show that

$$\frac{1}{p}\sum_{\{(i,j):k<|i-j|\}} \sigma_{ij}^2 \le Ck^{-2\alpha-1}$$

which is clearly true if $|\sigma_{ij}| \le C_1 |i-j|^{-(\alpha+1)}$ for all $i \ne j$. However we do not have this condition. We shall show that Assumption (3) implies that there is a subsequence in each row satisfying that. Since

$$\sum_{\{(i,j):\lfloor k/2\rfloor<|i-j|\le k\}} |\sigma_{ij}| \le M\left(\frac{k}{2}\right)^{-\alpha}$$

which implies there is a $j_* \in \{j : \lfloor k/2\rfloor < |i-j| \le k\}$ such that $|\sigma_{ij_*}| \le M\left(\frac{k}{2}\right)^{-(\alpha+1)}$, then

$$\begin{aligned}\sum_{\{j:k<|i-j|\}} \sigma_{ij}^2 &\le \sum_{\{j:|i-j_*|<|i-j|\}} \sigma_{ij}^2 \le M\left(\frac{k}{2}\right)^{-(\alpha+1)} \sum_{\{j:|i-j_*|<|i-j|\}} |\sigma_{ij}| \\ &\le M\left(\frac{k}{2}\right)^{-(\alpha+1)} \sum_{\{j:\lfloor k/2\rfloor<|i-j|\}} |\sigma_{ij}| \le M\left(\frac{k}{2}+1\right)^{-(2\alpha+1)}.\end{aligned}$$

Thus

$$\mathbb{E}\frac{1}{p}\left\|\breve{\Sigma}-\Sigma\right\|_F^2 \le Ck^{-2\alpha-1} + C\frac{k}{n} \le C_2 n^{-\frac{2\alpha+1}{2(\alpha+1)}} \tag{32}$$

by choosing

$$k = n^{\frac{1}{2(\alpha+1)}} \tag{33}$$

if $n^{\frac{1}{2(\alpha+1)}} \le p$, which is different from the choice of k for the operator norm in equation (15). If $n^{\frac{1}{2(\alpha+1)}} > p$, we will choose $k = p$, then the bias part is 0 and consequently

$$\mathbb{E}\frac{1}{p}\left\|\breve{\Sigma}-\Sigma\right\|_F^2 \le \frac{p}{n}. \quad \blacksquare$$

Remark 3 *Under the Frobenius norm the optimal tapering parameter k is of the order $n^{\frac{1}{2(\alpha+1)}}$. The rate of convergence of the tapering estimator with $k \asymp n^{\frac{1}{2(\alpha+1)}}$ under the operator norm is*

$$\frac{\log p}{n} + n^{-\frac{\alpha}{\alpha+1}},$$

which is slower than $n^{-\frac{2\alpha}{2\alpha+1}} + \frac{\log p}{n}$ in equation (1). Similarly the optimal procedure under the operator norm is not rate optimal under the Frobenius norm.

4.2 Lower Bound under the Frobenius Norm

We now turn to the minimax lower bound. As in the case of estimation under the operator norm, we need to construct a finite parameter space $\mathcal{F}_2 \subset \mathcal{F}$ such that

$$\inf_{\hat{\Sigma}} \sup_{\mathcal{F}_2} \mathbb{E}\frac{1}{p}\left\|\hat{\Sigma}-\Sigma\right\|_F^2 \geq c\frac{k}{n}.$$

for some $c>0$ when $k=\min\left\{n^{\frac{1}{2(\alpha+1)}},\ p/2\right\}$.

We construct $\mathcal{F}_2$ as follows. Let $0<\tau<\min\{M,\ 1-\varepsilon\}$ be a constant. Define

$$\mathcal{F}_2=\left\{\Sigma(\theta):\Sigma(\theta)=I+\left(\theta_{ij}\tau n^{-\frac{1}{2}}I\left\{1\leq|i-j|\leq k\right\}\right)_{p\times p},\text{ for }\theta_{ij}=\theta_{ji}=0\text{ or }1\right\}$$

It is easy to verify that $\mathcal{F}_2\subset\mathcal{F}$. Note that $\theta\in\Theta=\{0,1\}^{kp-k(k+1)/2}$.

Applying Assouad's Lemma with d the Frobenius norm and $s=2$ to the parameter space $\mathcal{F}_2$, we have

$$\max_{\theta\in\mathcal{F}_2} 2^2 E_\theta\frac{1}{p}\left\|\hat{\Sigma}-\Sigma(\theta)\right\|_F^2 \geq \min_{H(\theta,\theta')\geq 1}\frac{\frac{1}{p}\left\|\Sigma(\theta)-\Sigma(\theta')\right\|_F^2}{H(\theta,\theta')}\frac{kp-k(k+1)/2}{2}\min_{H(\theta,\theta')=1}\left\|P_\theta\wedge P_{\theta'}\right\|.$$

Note that

$$\frac{\left\|\Sigma(\theta)-\Sigma(\theta')\right\|_F^2}{H(\theta,\theta')}=\frac{\left[\tau k^{-(\alpha+1)}\right]^2\sum\left|\theta_{ij}-\theta'_{ij}\right|^2}{H(\theta,\theta')}=\left[\tau k^{-(\alpha+1)}\right]^2$$

then

$$\min_{H(\theta,\theta')\geq 1}\frac{\frac{1}{p}\left\|\Sigma(\theta)-\Sigma(\theta')\right\|_F^2}{H(\theta,\theta')}=\frac{\tau^2}{p}n^{-1}.$$

It is easy to see that

$$\frac{kp-k(k+1)/2}{2}\asymp kp.$$

Lemma 8 *Let P_θ be the joint distribution of $\mathbf{X}_1,\ldots,\mathbf{X}_n\overset{iid}{\sim}N(0,\Sigma(\theta))$ with $\Sigma(\theta)\in\mathcal{F}_2$. Then for some constant $c_1>0$ we have*

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

We omit the proof of this lemma. It is very similar to and simpler than the proof of Lemma 6.

From Lemma 8 we have for some $c>0$

$$\min_{H(\theta,\theta')=1}\left\|P_\theta\wedge P_{\theta'}\right\|\geq c \tag{34}$$

thus

$$\max_{\theta\in\mathcal{F}_2} 2^2 E_\theta\frac{1}{p}\left\|\hat{\Sigma}-\Sigma(\theta)\right\|_F^2\geq c\min\left\{n^{-\frac{2\alpha+1}{2(\alpha+1)}},\ \frac{p}{n}\right\}$$

which implies the rate obtained in (32) is optimal. ∎

5 Estimation of the Inverse Covariance Matrix

The inverse of the covariance matrix Σ^{-1} is of significant interest in many statistical applications. The results and analysis given in Section 3 can be used to derive the optimal rate of convergence for estimating Σ^{-1} under the operator norm.

The following theorem gives the minimax rate of convergence for estimating Σ^{-1}.

Theorem 5 *The minimax risk of estimating the inverse covariance matrix Σ^{-1} over the parameter space $\mathcal{F}_\alpha$ defined in (3) satisfies*

$$\inf_{\hat{\Sigma}} \sup_{\mathcal{F}_\alpha} \mathbb{E} \left\| \hat{\Sigma}^{-1} - \Sigma^{-1} \right\|^2 \asymp \min \left\{ n^{-\frac{2\alpha}{2\alpha+1}} + \frac{\log p}{n}, \ \frac{p}{n} \right\}. \tag{35}$$

<u>*Proof of Theorem 5*</u>: We shall focus on the case $p \geq 2n^{\frac{1}{2\alpha+1}}$. The proof for the case of $p < 2n^{\frac{1}{2\alpha+1}}$ is similar. To establish the upper bound, note that

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

then

$$\left\| \hat{\Sigma}^{-1} - \Sigma^{-1} \right\|^2 = \left\| \hat{\Sigma}^{-1} \left(\Sigma - \hat{\Sigma} \right) \Sigma^{-1} \right\|^2 \leq \left\| \hat{\Sigma}^{-1} \right\|^2 \left\| \Sigma - \hat{\Sigma} \right\|^2 \left\| \Sigma^{-1} \right\|^2 .$$

It follows from Assumption (3) that $\left\| \Sigma^{-1} \right\|^2 \leq \varepsilon^{-2}$. It then suffices to show that there is a constant c such that $\mathbb{P}\left(\left\| \hat{\Sigma}^{-1} \right\| \geq c \right)$ decays faster than any polynomial of n, which follows easily from Lemmas 2 and 3.

The proof of the lower bound is almost identical to that of Theorem 1 except that here we need to show

$$\min_{H(\theta,\theta') \geq 1} \frac{\left\| \Sigma^{-1}(\theta) - \Sigma^{-1}(\theta') \right\|^2}{H(\theta, \theta')} \geq cka^2$$

instead of Lemma 5. For a positive definite matrix A, let $\lambda_{\min}(A)$ denote the minimum eigenvalue of A. Since

$$\Sigma^{-1}(\theta) - \Sigma^{-1}(\theta') = \Sigma^{-1}(\theta') \left(\Sigma(\theta) - \Sigma(\theta') \right) \Sigma^{-1}(\theta),$$

we have

$$\left\| \Sigma^{-1}(\theta) - \Sigma^{-1}(\theta') \right\| \geq \lambda_{\min}\left(\Sigma^{-1}(\theta) \right) \lambda_{\min}\left(\Sigma^{-1}(\theta') \right) \left\| \Sigma(\theta) - \Sigma(\theta') \right\| .$$

Note that

$$\lambda_{\min}\left(\Sigma^{-1}(\theta) \right) > \varepsilon, \ \lambda_{\min}\left(\Sigma^{-1}(\theta') \right) > \varepsilon,$$

then Lemma 5 implies

$$\min_{H(\theta,\theta')\geq 1} \frac{\left\|\Sigma^{-1}(\theta)-\Sigma^{-1}(\theta')\right\|^2}{H(\theta,\theta')} \geq \varepsilon^4 \min_{H(\theta,\theta')\geq 1} \frac{\left\|\Sigma(\theta)-\Sigma(\theta')\right\|^2}{H(\theta,\theta')} \geq cka^2$$

for some constant $c > 0$. ∎

6 Proofs of Auxiliary Lemmas

In this section we give proofs of auxiliary lemmas stated and used in Sections 3 - 5.

Proof of Lemma 1: It is easy to see

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

Proof of Lemma 2: Without loss of generality we assume that p is divisible by m. Set $\delta_l^{(m)} = M_l^{(m)} - \mathbb{E}M_l^{(m)}$ and $S^{(m)} = \sum_{l=2-m}^{p} M_l^{(m)}$. It follows from (6) that

$$\left\|S_l^{(m)} - \mathbb{E}S_l^{(m)}\right\| \leq \sum_{l=1}^{m} \left\| \sum_{-1\leq j\ <\ p/m} \delta_{jm+l}^{(m)} \right\|. \tag{36}$$

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

$$\left\|S_l^{(m)} - \mathbb{E}S_l^{(m)}\right\| \leq m \max_{1\leq l\leq m} \left\| \sum_{0\leq j\ <\ p/m} \delta_{jm+l}^{(m)} \right\| \leq m \max_{2-m\leq l\leq p} \left\|\delta_l^{(m)}\right\|. \tag{37}$$

Since $\delta_l^{(k)}$ are all sub-blocks of certain matrix $\delta_l^{(2k)}$ with $1 \leq l \leq p-2k+1$, Lemma 2 now follows immediately from Equations (37) and (6). ∎

Proof of Lemma 3: Fix $1 \leq l \leq p$ and $m \geq 1$. Let Q be the projection mapping $(a_1, \ldots, a_p)^T$ to $(a_l, \ldots, a_{l+m-1})^T$. Define

$$\Sigma_l^{(m)} = Q\mathbb{E}M_l^{(m)}Q^T,\ W = n\left(\Sigma_l^{(m)}\right)^{-1/2} QM_l^{(m)}Q^T \left(\Sigma_l^{(m)}\right)^{-1/2}.$$

Since $(x_{ij}, l \leq j < l+m)^T$ are i.i.d. $N(0, \Sigma_l^{(m)})$ vectors, W is an $m \times m$ matrix with the Wishart distribution $W_m(I, n)$. Moreover,

$$\left\|M_l^{(m)} - \mathbb{E}M_l^{(m)}\right\| \leq \|W/n - I_m\| \times \left\|\Sigma_l^{(m)}\right\|.$$

Since $\left\|\Sigma_l^{(m)}\right\| = \left\|\mathbb{E}M_l^{(m)}\right\|$, (16) follows from

$$\mathbb{P}\left\{\|W/n - I_m\| > (1+\sqrt{m/n}+t)^2 - 1\right\} \le \mathbb{P}\left\{|Z| > t\right\},$$

which is an immediate consequence of Theorem II.13 of Davidson and Szarek (2001). ∎

Proof of Lemma 5: Set $v = (1\{k \le i \le 2k\})$ and let

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

There are exactly $H(\theta, \theta')$ number of w_i such that $|w_i| = ka$ (just consider upper half of the matrix). This implies

$$\left\|\left[\Sigma(\theta) - \Sigma\left(\theta'\right)\right] v\right\|_2^2 \ge H(\theta,\theta') \cdot (ka)^2$$

and so $\|\Sigma(\theta) - \Sigma(\theta')\|^2 \ge H(\theta,\theta') \cdot (ka)^2/k \ge cka^2$. ∎

Proof of Lemma 6: When $H(\theta, \theta') = 1$, we will show

$$\begin{aligned}\|P_\theta - P_{\theta}\|_1^2 &\le K\left(P_{\theta'}|P_\theta\right) = n\left[\frac{1}{2}tr\left(\Sigma\left(\theta'\right)\Sigma^{-1}(\theta)\right) - \frac{1}{2}\log\det\left(\Sigma\left(\theta'\right)\Sigma^{-1}(\theta)\right) - \frac{p}{2}\right] \\ &\le n \cdot cka^2\end{aligned}$$

for some small $c > 0$. This immediately implies the L_1 distance between two measures is bounded away from 1, and then the lemma follows. Write

$$\Sigma\left(\theta'\right) = D_1 + \Sigma(\theta).$$

Then

$$\frac{1}{2}tr\left(\Sigma\left(\theta'\right)\Sigma^{-1}(\theta)\right) - \frac{p}{2} = tr\left(D_1\Sigma^{-1}(\theta)\right).$$

Let λ_i be the eigenvalues of $D_1\Sigma^{-1}(\theta)$, which are in $[-c_3ka, c_3ka]$ for some $c_3 > 0$. Then we have

$$\log\det\left(\Sigma\left(\theta'\right)\Sigma^{-1}(\theta)\right) = \log\det\left(I + D_1\Sigma^{-1}(\theta)\right) = tr\left(D_1\Sigma^{-1}(\theta)\right) + R_3$$

where

$$R_3 \le C\sum_{i=1}^{p}\lambda_i^2 \text{ for some } C > 0.$$

Since $\left\|\Sigma^{-1}(\theta)\right\|$ is finite, it follows from the fact that the Frobenius norm of a matrix remains the same after an orthogonal transformation that

$$\sum_{i=1}^{p}\lambda_i^2 = \left\|D_1\Sigma^{-1}(\theta)\right\|_F^2 \le C_1\|D_1\|_F^2 \le C_2ka^2. \quad ∎$$

References

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

[2] Bickel, P. J. and Levina, E. (2008a). Regularized estimation of large covariance matrices. *Ann. Statist,* **36**, 1, 199–227.

[3] Bickel, P. J. and Levina, E. (2008b). Covariance Regularization by Thresholding. *Ann. Statist.* To appear.

[4] Davidson, K. R. and Szarek, S. J. (2001). Local operator theory, random matrices and Banach spaces, *Handbook in Banach Spaces* Vol I, ed. W. B. Johnson, J. Lindenstrauss, Elsevier, 317–366.

[5] El Karoui, N. (2008). Operator norm consistent estimation of large dimensional sparse covariance matrices. *Ann. Statist.* To appear.

[6] Fan, J., Fan, Y., and Lv, J. (2006). High dimensional covariance matrix estimation using a factor model. *Journal of Econometrics*, to appear..

[7] Furrer, R. and Bengtsson, T. (2007). Estimation of high-dimensional prior and posterior covariance matrices in Kalman filter variants, *Journal of Multivariate Analysis*, 98 (2):227–255.

[8] Huang, J., Liu, N., Pourahmadi, M. and Liu, L. (2006), Covariance Matrix Selection and Estimation via Penalised Normal Likelihood, *Biometrika*, **93**(1), 85–98.

[9] Johnstone, I. M. (2001). On the distribution of the largest eigenvalue in principal components analysis, *Ann. Statist.*, **29**(2):295–327.

[10] Johnstone, I. M. and Lu, A. Y. (2004). Sparse principal components analysis, *J. Amer. Statist. Assoc.* Tentatively accepted.

[11] Lam, C. and Fan, J. (2007). Sparsistency and Rates of Convergence in Large Covariance Matrices Estimation. Princeton University. Technical report.

[12] Meinshausen, N. and Bühlmann, P. (2006). High dimensional graphs and variable selection with the Lasso. *Ann. Statist*, **34**, 1436–1462.

[13] Muirhead, R. J. (1987). Developments in eigenvalue estimation. *Advances in Multivariate Statistical*, (A. K. Gupta, Ed.), 277–288, Reidel, Boston.

[14] Rothman, A.J. , Bickel, P. J., Levina, E. and Zhu, J. (2008). Sparse Permutation Invariant Covariance Estimation. Electronic Journal of Statistics, 2:494-515.

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

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

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