Journal of Multivariate Analysis 116 (2013) 45–51

Contents lists available at SciVerse ScienceDirect

Journal of Multivariate Analysis

journal homepage: www.elsevier.com/locate/jmva

Minimax optimal estimation of general bandable covariance matrices

Lingzhou Xue, Hui Zou *

School of Statistics, University of Minnesota, Minneapolis, MN 55455, USA

ARTICLE INFO

Article history:
Received 17 August 2011
Available online 12 December 2012

AMS subject classifications:
primary 62H12
secondary 62F12
62G09

Keywords:
Adaptive minimax
Covariance matrix
Minimax optimal rates
Frobenius norm
Spectral norm
Tapering

ABSTRACT

Cai et al. (2010) [4] have studied the minimax optimal estimation of a collection of large bandable covariance matrices whose off-diagonal entries decay to zero at a polynomial rate. They have shown that the minimax optimal procedures are fundamentally different under Frobenius and spectral norms, regardless of the rate of polynomial decay. To gain more insight into this interesting problem, we study minimax estimation of large bandable covariance matrices over a parameter space characterized by a general positive decay function. We obtain explicit results to show how the decay function determines the minimax rates of convergence and the optimal procedures. From the general minimax analysis we find that for certain decay functions there is a tapering estimator that simultaneously attains the minimax optimal rates of convergence under the two norms. Moreover, we show that under the ultra-high dimension scenario it is possible to achieve adaptive minimax optimal estimation under the spectral norm. These new findings complement previous work.

1. Introduction

The problem of estimating a large covariance matrix has received a lot of attention in recent years. The classical sample covariance matrix estimator breaks down when the dimension greatly exceeds the sample size, as in contemporary high dimensional data. Various regularized covariance matrix estimators have been proposed to overcome the difficulty imposed by high dimensionality. Some popular proposals include Cholesky-based penalization [7,8,10], thresholding [3,5,9], banding [2,12] and tapering [6,4].

Let $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ be n independent and identically distributed observations from a p-variate distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma} = (\sigma_{ij})_{1\le i,j\le p}$. Let $\widehat{\boldsymbol{\Sigma}} = (\hat{\sigma}_{ij})_{1\le i,j\le p}$ denote a generic estimator of $\boldsymbol{\Sigma}$. Define $\|\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\|_F = \left(\sum_i \sum_j (\hat{\sigma}_{ij} - \sigma_{ij})^2\right)^{1/2}$ as the Frobenius norm of $\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}$. Let $\|\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\|_2$ denote the spectral norm of $\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}$, which is the largest singular value of $\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}$. In this work we only focus on large covariance matrices in which $p > n$ and often $p \gg n$. In addition we assume $\log(p) \ll n$ which is necessary for establishing consistency. We write $a_n \asymp b_n$ if there exist two finite positive constants c, C such that $cb_n \le a_n \le Cb_n$.

Cai et al. [4] was the first to study the minimax optimal estimation of $\boldsymbol{\Sigma}$ over a parameter space under the Frobenius norm and the spectral norm. It is assumed that the data follow a sub-Gaussian distribution in the sense that there is some constant $c > 0$ such that

$$\mathrm{pr}\{|\boldsymbol{v}^T(\boldsymbol{X}_1 - \boldsymbol{\mu})| > t\} \le e^{-ct^2/2} \quad \text{for all } t > 0 \text{ and } \|\boldsymbol{v}\| = 1.$$

* Corresponding author.
E-mail address: hzou@stat.umn.edu (H. Zou).

0047-259X/$ – see front matter © 2012 Elsevier Inc. All rights reserved.
doi:10.1016/j.jmva.2012.11.003

Note that the sub-Gaussian condition is only a technical relaxation from multivariate normality. As shown in [4] the minimax lower bounds are established for a class of multivariate normal distributions. Thus, one could even solely focus on the multivariate normal distributions without making the minimax estimation problem any easier.

Cai et al. [4] considered estimating $\boldsymbol{\Sigma}$ over the following parameter space:

$$\mathcal{F}_\alpha(M_0, M) = \{\boldsymbol{\Sigma} : \ |\sigma_{ij}| \le M|i-j|^{-\alpha-1}, \ \text{for } i \ne j; \ \lambda_{\max}(\boldsymbol{\Sigma}) \le M_0\}, \tag{1}$$

where $\lambda_{\max}(\boldsymbol{\Sigma})$ means the largest eigenvalue of $\boldsymbol{\Sigma}$ and α, M_0, M are positive constants. Similar parameter spaces were considered in [3]. The following minimax bounds were established [4]:

$$\inf_{\widehat{\boldsymbol{\Sigma}}} \sup_{\mathcal{F}_\alpha(M_0,M)} \mathbb{E}\|\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\|_2^2 \asymp n^{-2\alpha/(2\alpha+1)} + \log(p)/n, \tag{2}$$

$$\inf_{\widehat{\boldsymbol{\Sigma}}} \sup_{\mathcal{F}_\alpha(M_0,M)} p^{-1}\mathbb{E}\|\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\|_F^2 \asymp n^{-(2\alpha+1)/(2\alpha+2)}. \tag{3}$$

Furthermore, the minimax lower bounds can be attained by tapering [4]

$$\widehat{\boldsymbol{\Sigma}}_T = \left(\widehat{\boldsymbol{\Sigma}}_{ij}^{(s)} w_{ij}\right)_{1\le i,j\le p}, \tag{4}$$

where $w_{ij} = 2k^{-1}((k-|i-j|)_+ - (k/2-|i-j|)_+)$ and $\widehat{\boldsymbol{\Sigma}}^{(s)} = \frac{1}{n}\sum_{l=1}^n \boldsymbol{X}_l\boldsymbol{X}_l^T - \bar{\boldsymbol{X}}\bar{\boldsymbol{X}}^T$. When $k = k_S = n^{1/(2\alpha+1)}$, $\widehat{\boldsymbol{\Sigma}}_T$ attains the minimax risk bound under the spectral norm, while when $k = k_F = n^{1/(2\alpha+2)}$, $\widehat{\boldsymbol{\Sigma}}_T$ attains the minimax risk bound under the Frobenius norm. Since $k_S \gg k_F$ regardless of the value of α, [4] concluded that the optimal procedures for covariance matrix estimation are fundamentally different under spectral norm and Frobenius norm. In short, the minimax theory of Cai et al. [4] has two main points. First, an exactly banded matrix can uniformly approximate all covariance matrices in $\mathcal{F}_\alpha$. The larger α the better approximation. Second, the optimal procedure changes according to the choice of matrix norm.

To gain more insight into these two points, we consider the minimax estimation of $\boldsymbol{\Sigma}$ over a general parameter space

$$\mathcal{H}(M_0, M) = \{\boldsymbol{\Sigma} : |\sigma_{ij}| \le Mh(|i-j|), \ \text{for } i \ne j; \ \lambda_{\max}(\boldsymbol{\Sigma}) \le M_0\}, \tag{5}$$

where $h(t)$ is a general decay function. To avoid trivial cases, we require $h(t) > 0$ for all t. In addition, $h(t)$ has the following properties:

- $h(\cdot)$ is a strictly decreasing function in $[1, \infty)$;
- $h(\cdot)$ is integrable, i.e., $\int_1^\infty h(t)dt < \infty$.

When $h(t) = t^{-\alpha-1}$, the parameter space reduces to $\mathcal{F}_\alpha$.

We prove a general minimax theorem of estimating $\boldsymbol{\Sigma}$ over $\mathcal{H}_{M_0,M}$. The general theorem recovers the minimax theory of Cai et al. [4] when the decay function is polynomial. By applying the general minimax theorem to other types of decay functions, we discover some new interesting phenomena summarized as follows.

- *Simultaneous minimax estimation*. We provide two explicit parameter spaces over which the minimax optimal rates of convergence can be achieved by the same tapering estimator.
- *Adaptive minimax estimation*. We show that, under ultra-high dimensions, a universal tapering estimator can adaptively attain the minimax optimal rate of convergence under the spectral norm over some parameter spaces.

2. A general minimax theorem

For notation convenience, we use C and c throughout to denote generic constants in upper and lower bounds, respectively. We define the following quantities.

$$R_s(k) = k/n + \left(\int_{k/2}^\infty h(t)dt\right)^2, \qquad R_f(k) = k/n + \int_{k/2}^\infty h^2(t)dt; \tag{6}$$

$$k_n^s = n \cdot \min_k R_s(k), \qquad k_n^f = n \cdot \min_k R_f(k). \tag{7}$$

Theorem 1. *Assume that there exist two positive constants c_s and c_f such that*

C1. $\liminf_{n\to\infty} h(c_s k_n^s)(nk_n^s)^{1/2} = \gamma^* > 0$

C2. $\liminf_{n\to\infty} h(c_f k_n^f) n^{1/2} = \gamma^{**} > 0.$

The minimax risk of estimating $\boldsymbol{\Sigma}$ *over* $\mathcal{H}(M_0, M)$ *satisfies*

$$\inf_{\widehat{\boldsymbol{\Sigma}}} \sup_{\mathcal{H}(M_0,M)} \mathbb{E}\left\|\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\right\|_2^2 \asymp k_n^s/n + \log(p)/n, \tag{8}$$

$$\inf_{\widehat{\boldsymbol{\Sigma}}} \sup_{\mathcal{H}(M_0,M)} p^{-1}\mathbb{E}\left\|\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\right\|_F^2 \asymp k_n^f/n. \tag{9}$$

Remark 1. Note that c_s, c_f are free to be chosen. As a result, C1 and C2 are easily satisfied for well-behaved h functions. We go over some examples in the sequel.

Theorem 1 is proved by combining the minimax lower bounds in Lemma 1 and upper bounds in Lemma 2.

Lemma 1. *Under conditions* C1 *and* C2 *we have*

$$\inf_{\widehat{\boldsymbol{\Sigma}}} \sup_{\mathcal{H}(M_0,M)} \mathbb{E}\left\|\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\right\|_2^2 \geq ck_n^s/n + c\log(p)/n. \tag{10}$$

$$\inf_{\widehat{\boldsymbol{\Sigma}}} \sup_{\mathcal{H}(M_0,M)} p^{-1}\mathbb{E}\left\|\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\right\|_F^2 \geq ck_n^f/n. \tag{11}$$

Lemma 2. *Define* $k_s^* = \arg\min_k R_s(k)$ *and* $k_f^* = \arg\min_k R_f(k)$. *The tapering estimator of* $\boldsymbol{\Sigma}$ *with* $k = k_s^*$, *satisfies*

$$\sup_{\mathcal{H}(M_0,M)} \mathbb{E}\left\|\widehat{\boldsymbol{\Sigma}}_T - \boldsymbol{\Sigma}\right\|_2^2 \leq Ck_n^s/n + C\log(p)/n. \tag{12}$$

The tapering estimator of $\boldsymbol{\Sigma}$ *with* $k = k_f^*$ *satisfies*

$$\sup_{\mathcal{H}(M_0,M)} \mathbb{E}p^{-1}\left\|\widehat{\boldsymbol{\Sigma}}_T - \boldsymbol{\Sigma}\right\|_F^2 \leq Ck_n^f/n. \tag{13}$$

To demonstrate Theorem 1 let us revisit $\mathcal{F}_\alpha$ where $h(j) = j^{-\alpha-1}$ for some $\alpha > 0$. We have

$$R_s(k) = (k/2)^{-2\alpha}/\alpha^2 + k/n, \qquad R_f(k) = (k/2)^{-2\alpha-1}/(2\alpha+1) + k/n,$$

and then it follows that $k_n^s \asymp (n/\alpha)^{1/(2\alpha+1)}$, $k_n^f \asymp (n/2)^{1/(2\alpha+2)}$ and conditions C1 and C2 are satisfied with $c_s = c_f = 1$. Then by Theorem 1 we recover those minimax bounds in (2) and (3).

Remark 2. Lemma 2 gives a natural construction of tapering estimator to attain the minimax rates of convergence. In fact, there may be many other tapering estimators that can serve the same purpose. Define the sets $K_n(S)$ and $K_n(F)$ as follows:

$$K_n(S) = \{k : R_s(k) + \log(p)/n \asymp k_n^s/n + \log(p)/n\},$$
$$K_n(F) = \{k : R_s(k) \asymp k_n^f\}.$$

Then using any $k \in K_n(S)$, $\widehat{\boldsymbol{\Sigma}}_T(k)$ attains the minimax rate under the spectral norm. Likewise, using any $k \in K_n(F)$, $\widehat{\boldsymbol{\Sigma}}_T(k)$ attains the minimax rate under the Frobenius norm.

3. New parameter spaces and interesting phenomena

We have shown that Theorem 1 recovers the minimax results with a polynomial decay function. We now apply the general minimax results to some new parameter spaces and reveal some interesting new phenomena that are not seen from the minimax analysis in [4].

3.1. Simultaneous minimax optimal estimation

Following Remark 2, let us consider $\Lambda_n = K_n(s) \cap K_n(F)$. We immediately reach the following statement about simultaneous minimax optimal estimation:

If $\Lambda_n \neq \emptyset$ *for sufficiently large* n, *using any* $k \in \Lambda_n$, $\widehat{\boldsymbol{\Sigma}}_T(k)$ *simultaneously attains the minimax rates under both spectral and Frobenius norms.*

Although Cai et al. [4] claimed that the minimax optimal procedures are fundamentally different under spectral and Frobenius norms, their claim was proved for $h(t) = t^{-\alpha-1}$. However, we show that $\Lambda_n \neq \emptyset$ for some other decay functions, which implies that simultaneous minimax optimal estimation under two norms can be achieved by a tapering estimator.

Example 1. Consider

$$\mathcal{A}_\rho(M_0, M) = \{\boldsymbol{\Sigma} : \ |\sigma_{ij}| \le M\rho^{|i-j|}, \ \text{for } i \ne j; \ \lambda_{\max}(\boldsymbol{\Sigma}) \le M_0\}, \tag{14}$$

where $0 < \rho < 1$, $M_0 > 0$ and $M > 0$. In this case $h(t) = \rho^t$.

Remark 3. It is theoretically interesting to study $\mathcal{A}_\rho$ because it is the smallest parameter space defined in (5) that contains autoregressive covariance matrices which are widely used in applications to model spatial–temporal dependence. Asymptotically, $\mathcal{A}_\rho$ is a subspace of $\mathcal{F}_\alpha$ for any $\alpha > 0$. However, for a large but finite p, we can easily construct two correlation matrices $\boldsymbol{\Sigma}(1) \in \mathcal{A}_\rho$ and $\boldsymbol{\Sigma}(2) \in \mathcal{F}_\alpha$ such that $\boldsymbol{\Sigma}(1)$ is actually denser than $\boldsymbol{\Sigma}(2)$. For example, let $p = 500$ and design $\boldsymbol{\Sigma}(1)_{ij} = (0.9)^{|i-j|}$ and $\boldsymbol{\Sigma}(2)_{ij} = M|i-j|^{-1.1}I(i \ne j) + I(i = j)$. Let $M = 0.705$ which is the largest possible value for M that still keeps $\boldsymbol{\Sigma}(2)$ positive definite. Then $|\boldsymbol{\Sigma}(2)_{ij}| = 0.056$ when $|i-j| = 10$ but $|\boldsymbol{\Sigma}(1)_{ij}| = 0.349, 0.052$ when $|i-j| = 10, 28$.

By straightforward calculation we see that

$$R_s(k) = k/n + \rho^k/(\log(1/\rho))^2, \qquad R_f(k) = k/n + \rho^k/(2\log(1/\rho)).$$

Then we have

$$k_n^s = \log(n/\log(1/\rho))/\log(1/\rho), \qquad k_n^f = \log(n/2)/\log(1/\rho)$$

and conditions C1 and C2 hold with $c_s = 1/2$, $c_f = 1/2$. In addition, let $k(c) = c\log(n)/\log(1/\rho)$ for $c > 1$, we find

$$nR_s(k(c)) \asymp \log(n)/\log(1/\rho) \asymp k_n^s,$$
$$nR_f(k(c)) \asymp \log(n)/\log(1/\rho) \asymp k_n^f.$$

Thus, Theorem 1 yields the following corollary.

Corollary 1. *The minimax risk of estimating $\boldsymbol{\Sigma}$ over $\mathcal{A}_\rho(M_0, M)$ satisfies*

$$\inf_{\widehat{\boldsymbol{\Sigma}}} \sup_{\mathcal{A}_\rho(M_0,M)} \mathbb{E}\left\|\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\right\|_2^2 \asymp \log(p)/n,$$
$$\inf_{\widehat{\boldsymbol{\Sigma}}} \sup_{\mathcal{A}_\rho(M_0,M)} \mathbb{E}p^{-1}\left\|\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\right\|_F^2 \asymp \log(n)/n.$$

The tapering estimator with $k = c\log(n)/\log(1/\rho)$ simultaneously achieves the minimax optimal rates of convergence under the spectral and Frobenius norms.

Example 2. Let $h(t) = \exp(-\gamma t^{1/2})$ for some $\gamma > 0$. Consider

$$\mathcal{L}_\gamma(M_0, M) = \{\boldsymbol{\Sigma} : |\sigma_{ij}| \le M\exp(-\gamma|i-j|^{1/2}), \ \text{for } i \ne j; \ \lambda_{\max}(\boldsymbol{\Sigma}) \le M_0\}, \tag{15}$$

where $M_0 > 0$ and $M > 0$. Note that $h(t)$ is essentially (up to a constant) the density function of a squared $\Gamma(2, \gamma)$ random variable.

By direct calculation we have

$$R_s(k) = k/n + \gamma^{-4}(\gamma(2k)^{1/2} + 2)^2\exp(-\gamma(2k)^{1/2}),$$
$$R_f(k) = k/n + 2^{-1}\gamma^{-2}(\gamma(2k)^{1/2} + 1)\exp(-\gamma(2k)^{1/2}).$$

Then it is easy to show that

$$\log^2(n/\alpha^2)/(2\gamma^2) < k_n^s < \log^2(2n/\gamma^2)/\alpha^2,$$
$$\log^2(n/2)/(2\gamma^2) < k_n^f < \log^2(n/2)/\gamma^2.$$

Thus, we verify that conditions C1 and C2 hold with $c_s = 1$, $c_f = 1$. In addition, let $k(c) = c\log^2(n)/(2\gamma^2)$ for $c > 1$, we see that

$$nR_s(k(c)) \asymp \log^2(n) \asymp k_n^s,$$
$$nR_f(k(c)) \asymp \log^2(n) \asymp k_n^f.$$

Therefore, we reach the following corollary.

Corollary 2. *The minimax risk of estimating $\boldsymbol{\Sigma}$ over $\mathcal{L}_\gamma(M_0, M)$ satisfies*

$$\inf_{\widehat{\boldsymbol{\Sigma}}} \sup_{\mathcal{L}_\gamma(M_0,M)} \mathbb{E}\left\|\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\right\|_2^2 \asymp \log^2(n)/n + \log(p)/n,$$
$$\inf_{\widehat{\boldsymbol{\Sigma}}} \sup_{\mathcal{L}_\gamma(M_0,M)} \mathbb{E}p^{-1}\left\|\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\right\|_F^2 \asymp \log^2(n)/n.$$

The tapering estimator with $k = c\log^2(n)/(2\gamma^2)$, $c > 1$ simultaneously achieves the minimax optimal rates of convergence under the spectral and Frobenius norms.

3.2. Adaptive minimax optimal estimation under the spectral norm

Note that the optimal estimator in [4] critically depends on α. An important open question left in [4] is that whether there is an adaptive minimax optimal estimator of $\boldsymbol{\Sigma}$ without knowing α? We can ask the same question for parameter spaces $\mathcal{A}_\rho$ and $\mathcal{L}_\gamma$, because the optimal estimators in Corollaries 1 and 2 depend on the value of ρ, γ.

Here we provide a partially positive answer to the adaptive minimax question. We consider the ultra-high dimension scenario where we have

$$\log(p) \asymp n^\eta, \quad \text{for some constant } 0 < \eta < 1. \tag{16}$$

It follows that the minimax rate of convergence under the spectral norm is $\log(p)/n$ in $\mathcal{A}_\rho$ or $\mathcal{L}_\gamma$. Consider the tapering estimator with $k = \log(p)$. For $\mathcal{A}_\rho$ we

$$R_s(k) + \log(p)/n = 2\log(p)/n + \rho^{\log(p)}/(\log(1/\rho))^2.$$

Then under assumption (16), for $\mathcal{A}_\rho$ we have

$$R_s(k) + \log(p)/n \asymp 2\log(p)/n \quad \text{as } n \to \infty.$$

Similarly, for $\mathcal{L}_\alpha$ we have

$$\begin{aligned} R_s(k) + \log(p)/n &= 2\log(p)/n + \gamma^{-4}(\gamma(2\log(p))^{1/2} + 2)^2 \exp(-\gamma(2\log(p))^{1/2}) \\ &\asymp 2\log(p)/n \quad \text{under condition (16).} \end{aligned}$$

Therefore, we have the following corollary.

Corollary 3. *Under the ultra-high dimension condition (16), the tapering estimator with $k = \log(p)$ achieves the minimax optimal rate of convergence under the spectral norm in either $\mathcal{A}_\rho$ or $\mathcal{L}_\gamma$.*

4. Discussion

The general minimax theory in this work shows the role of decay function in determining the optimal rates and optimal procedures, which also indicates the difficulty in constructing an adaptive minimax procedure for estimating large covariance matrices. Corollary 3 shows that under ultra-high dimensions adaptive minimax estimation is possible. This is a partial positive answer because it still assumes that the parameter space is either $\mathcal{A}_\rho$ or $\mathcal{L}_\gamma$. It is desirable to construct an adaptive minimax procedure without assuming any specific knowledge about the parameter space. This is an interesting direction for future theoretical work.

Acknowledgment

This work was in part supported by a grant from the National Science Foundation, U.S.A. and a grant from the Office of Naval Research, U.S.A. The authors thank two referees for their helpful comments.

Appendix. Proofs of lemmas

Proof of Lemma 1. We use Fano's lemma and Assouad's lemma [1,13,11] to establish the minimax lower bounds. *Part I. The minimax lower bound for the spectral risk.*

Define $\mathcal{G}_0$ as follows

$$\mathcal{G}_0 = \left\{ \boldsymbol{\Sigma}_m = \beta \left(\boldsymbol{I}_{p\times p} + aI(i = j = m) \right) ; 1 \le m \le p \right\},$$

where $a = (\log(p)/16n)^{1/2}$ and β is a positive constant such that $\beta < \min(M, M_0)/2$. It is easy to verify that $\mathcal{G}_0 \subset \mathcal{H}(M_0, M)$ as $n \to \infty$. Let $\boldsymbol{\Sigma}_0^* = \beta I_{p\times p}$. Let $\mathcal{P}_0^*, \mathcal{P}_1, \ldots, \mathcal{P}_p$ be the normal distribution with mean zero and covariance matrix $\boldsymbol{\Sigma}_0^*, \boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\Sigma}_p$. For any $(i, j) : 1 \le i \ne j \le p$, we have

$$\|\boldsymbol{\Sigma}_i - \boldsymbol{\Sigma}_j\|_2^2 = \beta^2 a^2 = \beta^2 \cdot \log(p)/n. \tag{17}$$

The Kullback–Leibler divergence between $\mathcal{P}_j (1 \le j \le p)$ and $\mathcal{P}_0^*$ can be written as

$$\begin{aligned} K(\mathcal{P}_j || \mathcal{P}_0^*) &= 2^{-1} n[\mathrm{tr}(\boldsymbol{\Sigma}_j(\boldsymbol{\Sigma}_0^*)^{-1}) - \log\det(\boldsymbol{\Sigma}_j(\boldsymbol{\Sigma}_0^*)^{-1}) - p] \\ &= 2^{-1} n(a - \log(1 + a)) \le \log(p)/32. \end{aligned} \tag{18}$$

Combining (17) and (18), we apply Fano's lemma to obtain

$$\inf_{\widehat{\boldsymbol{\Sigma}}} \sup_{\mathcal{G}_0} \mathbb{E} \left\| \widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma} \right\|_2^2 \ge c \cdot \log(p)/n. \tag{19}$$

Let $k = c_s k_n^s/2$ and $d = (nk_n^s)^{-1/2}(4c_s)^{-1}$. Define another collection of covariance matrices as follows

$$\mathcal{G}_1 = \left\{ \boldsymbol{\Sigma}_m = \gamma \left(\boldsymbol{I}_{p\times p} + d \sum_{m=1}^{k} \theta_m \boldsymbol{B}(m,k) \right) ; \theta = (\theta_m) \in \{0,1\}^k \right\},$$

where $\boldsymbol{B}(m,k) = (b_{ij})_{1\le i,j\le p}$ with

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

We fix the positive constant γ such that $\gamma < \min\{M, M_0/2, 2M\gamma^*\}$. Then it is easy to check that $\mathcal{G}_1 \subset \mathcal{H}(M_0, M)$ as $n \to \infty$. Let $\mathcal{P}_{\boldsymbol{\Sigma}(\theta)} = \mathcal{P}_\theta$ be the normal distribution with mean zero and covariance matrix $\boldsymbol{\Sigma}(\theta)$. Use $H(\theta,\theta') = \sum_{1\le|i-j|\le k} |\theta_{ij} - \theta'_{ij}|$ to denote the Hamming distance between the binary vectors θ and θ'. Define $v = (v_j)_{1\le j\le p}$ with $v_j = I(k+1 \le j \le 2k)$. Then $(H(\theta,\theta'))^{-1}\|\{\boldsymbol{\Sigma}(\theta) - \boldsymbol{\Sigma}(\theta')\}v\|_2^2 = \gamma^2 k^2 d^2$, which implies that

$$\min_{H(\theta,\theta')\ge 1} \{(H(\theta,\theta'))^{-1}\|\boldsymbol{\Sigma}(\theta) - \boldsymbol{\Sigma}(\theta')\|_2^2\} \ge \gamma^2 k d^2. \tag{20}$$

Moreover, when $H(\theta,\theta') = 1$, we write $\gamma^{-1}\boldsymbol{\Sigma}(\theta) = \boldsymbol{\Sigma}^*(\theta)$. By Pinsker inequality [11] we have

$$\|\mathcal{P}_\theta - \mathcal{P}_{\theta'}\|_1^2 \le 2K(\mathcal{P}_\theta \| \mathcal{P}_{\theta'}) = 2K(\mathcal{P}_{\boldsymbol{\Sigma}^*(\theta)} \| \mathcal{P}_{\boldsymbol{\Sigma}^*(\theta')}). \tag{21}$$

Note that

$$\|\boldsymbol{\Sigma}^*(\theta) - \boldsymbol{I}_{p\times p}\|_2 \le (k_n^s/n)^{1/2}/4. \tag{22}$$

By definition of k_n^s, we have

$$k_n^s/n \le n^{1/2}/n + \left(\int_{j>n^{1/2}/2}^{\infty} h(t)dt \right)^2 \to 0,$$

which implies that

$$k_n^s/n \to 0. \tag{23}$$

Then combining (22) and (23) and using similar arguments for Lemma 6 in [4], we have that when $H(\theta,\theta') = 1$,

$$2K(\mathcal{P}_{\boldsymbol{\Sigma}^*(\theta)} || \mathcal{P}_{\boldsymbol{\Sigma}^*(\theta')}) \le 8nkd^2 = 1/4. \tag{24}$$

By Assouad's lemma and using (20), (21) and (24), we have

$$\inf_{\widehat{\boldsymbol{\Sigma}}} \max_{\boldsymbol{\Sigma}(\theta)\in\mathcal{G}_1} \mathbb{E}\left\|\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}(\theta)\right\|_2^2 \ge ck^2d^2 = ck_n^s/n. \tag{25}$$

Finally, combining (19) and (25) yields (11).

Part II. The minimax lower bound for the Frobenius risk.

We consider another collection of least favorable distributions defined as follows

$$\mathcal{H}_0' = \left\{ \boldsymbol{\Sigma}(\theta) = \beta \left(\boldsymbol{I}_{p\times p} + a \cdot \left[\theta_{ij} I_{\{1\le|i-j|\le k\}}\right]_{1\le i,j\le p} \right) : \theta_{ij} = \theta_{ji} = 0 \text{ or } 1 \right\},$$

with $k = c_f k_n^f$ and $a = n^{-1/2}/4$. Note that $\theta \in \{0,1\}^{kp-k(k+1)/2}$. Direct calculation shows that as long as we fix β such that $0 < \beta < \min\{M, M_0/(1+\log(1/q))\}$, then $\mathcal{H}_0'$ is a subclass of $\mathcal{H}_q(M_0, M)$ as $n \to \infty$. It is easy to see that

$$\begin{aligned}
&\min_{H(\theta,\theta')\ge 1} (H(\theta,\theta'))^{-1} p^{-1} \|\boldsymbol{\Sigma}(\theta) - \boldsymbol{\Sigma}(\theta')\|_F^2 \\
&= (\beta^2 a^2/p) \cdot (H(\theta,\theta'))^{-1} \sum_{1\le|i-j|\le k} (\theta_{ij} - \theta'_{ij})^2 \\
&= \beta^2 a^2/p = \beta^2/16 \cdot (np)^{-1}.
\end{aligned} \tag{26}$$

We write $\boldsymbol{\Sigma}(\theta) = \beta\boldsymbol{\Sigma}^*(\theta)$. By Pinsker inequality [11] we have

$$\|\mathcal{P}_\theta - \mathcal{P}_{\theta'}\|_1^2 \le 2K(\mathcal{P}_\theta \| \mathcal{P}_{\theta'}) = 2K(\mathcal{P}_{\boldsymbol{\Sigma}^*(\theta)} \| \mathcal{P}_{\boldsymbol{\Sigma}^*(\theta')}). \tag{27}$$

Note that

$$\|\boldsymbol{\Sigma}^*(\theta) - \boldsymbol{I}_{p\times p}\|_2 \le 2ka = 2^{-1} c_f k_n^f/n^{1/2}. \tag{28}$$

Write $h(c_f k_n^f)n^{1/2} = h(c_f k_n^f)k_n^f(n^{1/2}/k_n^f)$ and note that $h(c_f k_n^f)(c_f k_n^f/2) \le \int_{c_f k_n^f/2}^{c_f k_n^f} h(t)dt \to 0$. Thus, condition C2 implies that

$$n^{1/2}/k_n^f \to \infty. \tag{29}$$

Then combining (28) and (29) and using similar arguments for Lemma 6 in [4], we have that when $H(\theta, \theta') = 1$,

$$2K(\mathcal{P}_{\mathbf{\Sigma}^*(\theta)}||\mathcal{P}_{\mathbf{\Sigma}^*(\theta')}) \le 4na^2 = 1/4. \tag{30}$$

By Assouad's lemma and using (26), (27) and (30), we have

$$\inf_{\widehat{\mathbf{\Sigma}}} \sup_{\mathcal{H}_0'} \mathbb{E}p^{-1}\left\|\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}\right\|_F^2 \ge c(np)^{-1}k(p - (k+1)/2) = c \cdot k_n^f/n.$$

which automatically implies (11) since $\mathcal{H}_0'$ is a subspace of $\mathcal{H}(M_0, M)$. □

Proof of Lemma 2. The proof follows arguments in [4] with proper modifications. Without loss of generality we assume $\mu = 0$. Consider the tapering estimator with $k \ge 2j_0$. Define $\widetilde{\mathbf{\Sigma}} = (\tilde{\sigma}_{ij})_{1\le i,j\le p} = \sum_{l=1}^n \mathbf{X}_l\mathbf{X}_l^T/n$, $\widetilde{\mathbf{\Sigma}}_T = (\tilde{\sigma}_{ij}w_{ij})_{1\le i,j\le p}$, and $\mathbf{\Sigma}_T = (\sigma_{ij}w_{ij})_{1\le i,j\le p}$. Note that the remainder term $\widetilde{\mathbf{\Sigma}} - \widehat{\mathbf{\Sigma}} = \bar{\mathbf{X}}\bar{\mathbf{X}}^T$ has a negligible contribution to the risk after tapering (c.f. [4]; Remark 1). Hence we only need to focus on $\|\widetilde{\mathbf{\Sigma}}_T - \mathbf{\Sigma}\|_2^2$ and $\|\widetilde{\mathbf{\Sigma}}_T - \mathbf{\Sigma}\|_F^2$.

Note that $E\tilde{\sigma}_{ij} = \sigma_{ij}$ and $E\widetilde{\mathbf{\Sigma}}_T = \mathbf{\Sigma}_T$. Consider the triangle inequality:

$$\|\widetilde{\mathbf{\Sigma}}_T - \mathbf{\Sigma}\|_2^2 \le 2\|\widetilde{\mathbf{\Sigma}}_T - \mathbf{\Sigma}_T\|_2^2 + 2\|\mathbf{\Sigma}_T - \mathbf{\Sigma}\|_2^2.$$

By Lemmas 1–3 in [4] we have

$$\mathbb{E}\|\widetilde{\mathbf{\Sigma}}_T - \mathbf{\Sigma}_T\|_2^2 \le C((k + \log(p))/n). \tag{31}$$

By definition of $\mathcal{H}_q(M_0, M)$ and tapering weights, we have

$$\left\|\mathbf{\Sigma}_T - \mathbf{\Sigma}\right\|_2^2 \le \left\|\mathbf{\Sigma}_T - \mathbf{\Sigma}\right\|_1^2 \le C\left(\sum_{j>k/2}^{\infty} h(j)\right)^2 \le C\left(\int_{k/2}^{\infty} h(t)dt\right)^2. \tag{32}$$

Combining (31) and (32) and using (6) and (7), we obtain the upper bound in (12).

By direct mean-squared-error calculation we have

$$\mathbb{E}\|\widetilde{\mathbf{\Sigma}}_T - \mathbf{\Sigma}\|_F^2 = \sum_{(i,j)}\left(\sigma_{ij}^2(1 - w_{ij})^2 + w_{ij}^2\mathrm{var}(\tilde{\sigma}_{ij})\right).$$

Note that $\mathrm{var}(\tilde{\sigma}_{ij}) = n^{-1}(\sigma_{ij}^2 + \sigma_{ii}\sigma_{jj}) \le C/n$. Recall that the tapering weights satisfy $w_{ij} = 1$ for any $(i,j) : |i-j| < k/2$ and $w_{ij} = 0$ for any $(i,j) : |i-j| \ge k$. Then we have

$$\begin{aligned}
\mathbb{E}\|\widetilde{\mathbf{\Sigma}}_T - \mathbf{\Sigma}\|_F^2 &\le \sum_{(i,j):|i-j|>k/2} \sigma_{ij}^2 + \sum_{(i,j):|i-j|\le k} \mathrm{var}(\tilde{\sigma}_{ij}) \\
&\le C\sum_{j>k/2}^{\infty} ph^2(j) + Cpk/n \\
&\le Cp\left(\int_{k/2}^{\infty} h^2(t)dt + k/n\right).
\end{aligned}$$

Then by (6) and (7) we get the upper bound in (13). □

References

[1] P. Assouad, Deux remarques sur lestimation, C. R. Acad. Sci., Paris 296 (23) (1983) 1021–1024.
[2] P. Bickel, E. Levina, Covariance regularization by thresholding, Ann. Statist. 36 (2008) 2577–2604.
[3] P. Bickel, E. Levina, Regularized estimation of large covariance matrices, Ann. Statist. 36 (2008) 199–227.
[4] T. Cai, C. Zhang, H. Zhou, Optimal rates of convergence for covariance matrix estimation, Ann. Statist. 38 (2010) 2118–2144.
[5] N. El Karoui, Operator norm consistent estimation of large dimensional sparse covariance matrices, Ann. Statist. 36 (2008) 2717–2756.
[6] R. Furrer, T. Bengtsson, Estimation of high-dimensional prior and posterior covariance matrices in Kalman filter variants, J. Multivariate Anal. 98 (2007) 227–255.
[7] J. Huang, N. Liu, M. Pourahmadi, L. Liu, Covariance matrix selection and estimation via penalised normal likelihood, Biometrika 93 (2006) 85–98.
[8] C. Lam, J. Fan, Sparsistency and rates of convergence in large covariance matrix estimation, Ann. Statist. 37 (2009) 4254–4278.
[9] A. Rothman, E. Levina, J. Zhu, Generalized thresholding of large covariance matrices, J. Amer. Statist. Assoc. 104 (2009) 177–186.
[10] A. Rothman, E. Levina, J. Zhu, A new approach to Cholesky-based covariance regularization in high dimensions, Biometrika 97 (2010) 539–550.
[11] A. Tsybakov, Introduction to Nonparametric Estimation, Springer–Verlag, 2009.
[12] W. Wu, M. Pourahmadi, Banding sample covariance matrices of stationary processes, Statist. Sinica 19 (2009) 1755–1768.
[13] B. Yu, Assouad, fano, and le cam, in: D. Pollard, E. Torgersen, G. Yang (Eds.), Festschrift for Lucien Le Cam Research Papers in Probability and Statistics, Springer, Berlin, 1997, pp. 423–435.